**四川省科技创新苗子工程**   
**资助项目申请书**

|  |  |  |  |
| --- | --- | --- | --- |
|  | 类型: | 项目 |  |
| 申 请 人: | 申乔木 |
| 在读(或所在孵化园)单位: | 四川大学计算机学院 |
| 邮政编码： | 610065 |
| 电    话： | 13608175215 |
| 传    真： |  |
| E\_mail： | 996547379@qq.com |
| 申请日期： | 2012-09-18 |

**四川省科技创新苗子工程项目管理办公室**

**二○一三年制**

**填 报 说 明**

1. 填写申请书前，请先查阅《四川省科技创新苗子工程实施暂行办法》和《四川省科技创新苗子工程指南》。申请书各项内容须实事求是，表达明确。外来语要同时用原文和中文表达。第一次出现的缩写词，须注明全称。
2. 申请书在A4纸上用5号宋体打印，上边距30mm，下边距20mm，左边距25mm，右边距20mm，统一标明页码，从网上将文件电子版提交给四川省科技创新苗子工程办公室，用于初选。经初选合格后由在读单位审核签署意见并加盖公章，报送四川省科技创新苗子工程办公室一式三份。
3. 如内容填写不下，请另加附页。如有其他反映申请人情况的材料可以用附件形式同时呈报。附件材料版面应与申请书相同。
4. 申请书格式可登陆www.tccxfw.com 下载。
5. 填好的申请书从网上提交。

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 项目名称 | 基于信息逻辑结构树的网络舆情可视化分析系统 | | | | 按项目指南分类编号 | | | 网络舆情分析技术 |
| 项  目  负  责  人 | 姓  名 | 申乔木 | 性别 | 男 | 年龄 | 24 | 身份证 | 140602198909299036 |
| 在读单位（或所在孵化园） | 四川大学计算机学院 | | | | 专业 | 计算机技术 | |
| 在读身份（或已取得的学历） | 硕士 | | | | | | |
| 通信地址 | 四川大学望江校区基础教学楼B318 | | | | 邮政编码 | 610065 | |
| 联系电话 | 13608175215 | | | | 电子邮箱 | 996547379@qq.com | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 项目组成员 | 姓 名 | 性别 | 年龄 | 在 读 学 校（或工作单位） 及 身 份 |
| 张永清 | 男 | 27 | 四川大学计算机学院博士研究生 |
| 李梦颖 | 女 | 24 | 四川大学计算机学院硕士研究生 |
| 高丽峰 | 女 | 24 | 四川大学计算机学院硕士研究生 |
| 符敏 | 女 | 20 | 四川大学计算机学院本科生 |
| 邵建荃 | 男 | 21 | 四川大学计算机学院本科生 |
| 任申元 | 女 | 20 | 四川大学计算机学院本科生 |
|  |  |  |  |

|  |
| --- |
| **1、项目主要内容、工作计划、资金使用计划概要：** |
| **1.1 项目主要内容：**  现有对网络舆情的研究与相关产品较为成熟，但多以文档或报表形式对其结果进行展示，使得用户陷入大量抽象的数据中，无法全面直观掌握数据特点，预测舆情发展趋势。  本项目将网络舆情分析检测与可视化技术相结合，研究和设计“基于信息逻辑结构树的网络舆情可视化分析系统”。该系统能用可视化工具对海量复杂的网络舆情数据进行处理，并将数据的特点、规律、发展趋势以图形视图的形式给予直观展示，从而帮助用户加快对舆情数据的理解，提高决策和处置水平。    **项目研究内容如下：**  **1、可视化体系建立：**   * 可视化数据格式与标准的研究：对大规模网络舆情数据进行采集，根据数据特点和可视化工具需要建立统一数据标准与格式； * 数据与工具关联研究：将统一格式的数据与可视化工具进行关联。   **2、可视化工具模型的建立与优化：**  对现有可视化工具及其使用案例进行分析，结合舆情数据特点，选择合适的可视化工具进行优化，达到最优的数据呈现效果。  **3、网络信息逻辑结构树的研究：**  以单个站点为单位，研究舆情信息在各个模块的分布情况，并最终以树结构的方式进行存储和表示；在上述基础上，整个互联网信息可以用森林的方式进行表示，从而实现对整个互联网信息的存储检索和展示。  **4、舆情信息全过程可视化方案研究：**   * 网络信息结构可视化方案：对网络信息的逻辑结构用树的形式进行展示，将多种树形工具相结合，在展示结果中达到全局数据和细节数据的平衡； * 宏观舆情热度变化展示方案：展示宏观热点事件热度变化情况 * 微观舆情生命周期展示方案：展示特定某个（或者某几个）事件从兴起到衰亡整个生命周期； * 平台影响力度可视化方案：展示不同类型平台（或同类型平台下不同网站）对信息传播的影响力度。   **5、基于信息逻辑结构树的网络舆情可视化分析系统设计：**  设计并实现完整的网络舆情可视化分析系统，满足用户快速获取重要舆情数据特征和预测趋势的需要，具有实际应用价值和市场推广价值。    **1.2 工作计划**  整个项目执行期为1年，从2012年11月15日～2013年11月15日，可以分为五个阶段，具体实施情况如下：  第一阶段：需求细化与数据获取 时间：2012年11月15日～2012年12月15日   * 细化系统的需求并完成系统的架构设计； * 通过数据获取模块获取需要的原始数据，分类保存到本地（包括URL信息表和舆情信息文件）；   第二阶段：数据预处理 时间：2012年12月16日～2013年2月15日   * 分析URL信息表，分析信息在互联网平台上的逻辑分布情况，并形成逻辑结构树； * 对舆情信息数据进行分析和统计，制定标准并统一将数据格式化处理； * 将数据入库或者保存到特定格式的文件   第三阶段：可视化模型建立 时间：2013年2月16日～2013年6月15日   * 根据数据库内容进行舆情信息分析报表的打印模块设计和实现 * 信息逻辑结构树可视化模型的建立 * 舆情信息可视化工具模型的建立   第四阶段：可视化方案设计与仿真 时间：2013年6月16日～2013年9月15日   * 通过可视化工具形成视图对实际数据进行展示 * 多视图表达方案设计 * 关联设计（包括视图间的关联和视图与报表之间的关联） * 交互设计 * 仿真平台下可视化工具的实现   第五阶段：系统集成与测试 时间：2013年9月16日～2012年11月15日   * 将各模块集成到系统中 * 对系统进行测试和改进 * 论文撰写和专利申请     项目工作计划图如下：      **图1 项目工作计划图**  **1.3资金使用计划：**   |  |  | | --- | --- | | **名称** | **金额（单位:元）** | | 论文版面费用 | 7500 | | 网络资料费用 | 2800 | | 书籍资料购置 | 2200 | | 软硬件购置 | 10000 | | 会议费用 | 3500 | | 文印费 | 1000 | | 实验材料费（样本采集差旅费等） | 3000 | | 合计 | 30000 |   ***2.1国内外使用现状***      该项目是一个网络舆情分析监控与可视化技术相结合的项目，其研究现状可以从网络舆情分析监控系统和可视化技术两个角度进行阐述：  **1、网络舆情分析与监控系统**      随着信息技术的发展，网络舆情分析逐步受到管理人员和学者的重视。      国内目前已有多家企业研发了在线网络舆情监控和分析平台，如军犬舆情监控系统[1]、本果互联网舆情处置系统[2]、北京拓尔思信息技术股份有限公司的TRS网络舆情监控系统[3]等，这些平台起步早，在数据提取的深度和广度上都有一定的优势，但这些系统多用中文信息处理与数据挖掘领域技术，自动获取、自动聚类、主题检测和专题聚焦实现对网络舆情的检测和分析，而分析结果多以文本的简报、分析报告、报表的形式呈现给用户，其表现形式单一，更新功能较少。与之相比较，云情报[4]网站则在数据展示方面使用了大量的可视化技术，将表单转化为形象的图形。国际上，美国的TDT（Topic Detection and Tracking）系统是最具代表性的网络舆情监控系统。TDT是美国的一个研究项目[5]，主要涉及5个研究内容，分别是针对新闻文档的连续文本分割，主题跟踪，主题发现，新事件发现和关联性发现[6]。话题识别与跟踪（TDT）领域的大部分研究都是借用信息检索的某些方法，只是通过调整某些参数来使这些方法更适合于处理话题（事件）。但是，话题识别与跟踪研究的某些特殊性，如面向话题、基于时间等，也决定了仅仅利用现有信息检索方法对进一步提升系统的性能是有限的[7]。      此外，网络舆情分析也正成为高校和研究院所的关注热点，比如南京大学谷尼舆情研究实验室与谷尼国际软件(北京)有限公司（曾研发Goonie互联网舆情监控系统）合作，对网络舆情分析领域展开研究，通过发挥双方优势、整合双方资源，让网络传播研究成果能为市场服务，而研究基地亦将以市场需求为导向，推出可定制的个性化服务[8]；复旦大学舆情研究实验室专注于网络舆情的生成规律及传播机制的研究；中国人民大学舆论研究所[9]、新传媒网、新传媒产业联盟联合建立新传媒网络舆情技术实验室，致力于网络舆情前沿技术与管理方法研究与推广等。此外，很多高校设立了专门实验室对舆情分析进行跟踪研究，但目前很少有提出可视化与舆情分析相结合的研究，尤其是在信息分布逻辑结构可视化方面，尚未进行相关研究。      综上所述，目前的舆情分析系统所采用的技术多侧重于数据挖掘方面，而呈现结果则以分析报告、简报或数据报表为主，普遍存在可视化程度低，对于海量舆情分析结果阅读和理解不直观的问题。  **2、信息可视化技术**      近年来，随着计算机和因特网在各行各业的广泛应用，各类数据也越积越多，对数据处理分析的能力要求也越来越迫切。可视化和可视分析作为新兴的研究领域，旨在将单一的数据信息转换为视觉上的图像或图形呈现给用户，其优点是直观、明了，通过添加相应的用户交互，帮助用户根据数据需求从不同的角度对数据进行分析。该项目在可视化方面的研究关键是网络舆情逻辑结构的呈现，以及从宏观和微观两方面对热点事件的跟踪与可视化表达。      目前，有许多国内外学者和科研人员针对不同的社交网络进行了一定的分析和展示，文献[10]就提出了针对社交网络舆情的预测和网民对社会政策的态度反应。文献[11]提到将GIS信息与可视化技术结合来建立更加友好全面的网络舆情分析系统的思想。文献[12]提到了对Web空间和现实社会空间中的新闻事件演变进行可视化的相关技术。文献[13]提出了一种对树图的表达方式，即采用了多个同心圆环，不同的圆环表示不同级别的节点，以此来对树图进行展示，文献[14]采用了结点链接图的方式来表现树。文献[15]提出了themeriver的概念，表示某个信息随着时间的变化过程。文献[16]提出了一种对文本信息的可视分析系统，结合了多种可视化方案和数据挖掘技术，有一定的前瞻意义。文献[17]提出了一种可视化的方法分析人们对2008年美国大选这一事件的态度，文中采用形状和颜色映射来表示民众的态度，并添加一定的交互手段对民众态度进行分析。文献[18]采用散点图的形式对人们对于某事件发表的评论进行了可视化展示。许多互联网公司也对信息可视化进行了深入的研究，并以平台的方式进行发布，图2[19]为一个文字云（Word Cloud）的示例，这种方法可通过单词的大小、颜色等特征来表示文本中词汇的频率等属性值。图3[20]是北京大学可视化实验室在微博信息可视化方面的应用案例，这种方法能通过相应图形表达传播途径、热度变化等信息。图4[21]为流图（Theme River）的典型案例，其中横轴表示时间，纵轴表示所关注的某个属性值，如新闻访问量等，不同的事件在图中用颜色予以区分，流的宽度则表示该事件在某个时间点上访问量大小，由此形成的流图能同时反映多个事件在同一属性值（如访问量等）的变化情况。图5[22]为新闻树图，它是树图（Tree Map）的一种，可用于表示当前的热点新闻，其中每个颜色区域表示一个地区（在树结构中表示非叶子结点），同一颜色中更小的矩形区域表示该地区发生的新闻事件（即为叶子结点），叶子结点的大小表示新闻的热度。      综上所述，在当今信息、网络广泛使用的今天，可视化无论是在应对与日俱增的海量数据展示分析上，还是在对某一具体信息的微观表达上都具备相当的优势。采用可视化的手段呈现数据或数据分析的结果直观，易于理解。        综合上述对舆情监控系统、可视化与可视分析国内外研究现状的分析表明，对于网络舆情分析的研究多侧重于中文信息处理与数据挖掘领域，在可视化领域研究较少，且尚未提出对舆情信息逻辑结构进行可视化分析研究的案例。为此，基于信息逻辑结构的网络舆情可视化分析系统的研究与实现是非常必要的。  **参考文献：**  [1] 如军犬舆情监控系统 http://www.54yuqing.com/yuqing\_3.html  [2] 本果互联网舆监控系统http://www.ibenguo.cn/products/wlyqjk/tyb.htm [3] TRS：http://www.bjhd.gov.cn/zt/hdcxpt/cxcp/dzxx/201003/t20100305\_182565.htm [4] 云情报. http://www.kcis.cn/ [5] Topic Detection and Tracking (TDT) Evaluation Workshop. The 2002 topic detection and tracking task definition and evaluation plan [EB/OL].   ftp://jaguar.ncsl.nist.gov/tdt/tdt2002/,2006 [6]  Charles L Wayne.  Multilingual Topic Detection and Tracking: Successful Research Enabled by Corpora and Evaluation. Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC),2000 [7] 李保利、俞士汶.话题识别与跟踪研究. 国家自然科学基金项目 [8] 谷尼国际：http://www.cnii.com.cn/20080623/ca508624.htm [9] 人民大学舆情研究所：http://baike.baidu.com/view/5364596.htm [10] Robert Kleinfeld, Louay Bassbouss, Iosif Alvertis, George Gionis. Empowering Civic Participation in the Policy Making Process through Social Media. When the City Meets the Citizen, Vol.12, No.3, 2012, pp. 20-24. [11] 傅文棋. 基于GIS 的突发事件网络舆情可视化研究 [B].  地理空间信息 2012 [12] 刘晓娟,陈嘉勇,刘世希. 文本可视化在新闻事件演变中的应用.图书情报工作 2010.9 [13] Liang Gou, Xiaolong (Luke) Zhang: TreeNetViz: Revealing Patterns of Networks over Tree Structures. IEEE Trans. Vis. Comput. Graph. 17(12): 2449-2458.2011 [14] S. Jürgensmann, H.-J. Schulz. A visual survey of tree visualization. InfoVis’10: Poster Compendium of the IEEE Conference on Information Visualization. 2010. [15] S. Havre, B. Hetzler, and L. Nowell. ThemeRiver: Visualizing theme changes over time. In IEEE Symposium on Information Vizualization, 2000 [16] Furu Wei , Shixia Liu, Yangqiu Song. Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, 2010. [17] Franz Wanner, Christian Rohrdantz, Florian Mansmann, Daniela Oelke, Daniel A. Keim. Visual Sentiment Analysis of RSS News Feeds Featuring the US Presidential Election in 2008. Workshop on Visual Interfaces to the Social and the Semantic Web [18] Siamak Faridani,Ephrat Bitton,Kimiko Ryokai,Ken Goldberg.Opinion Space:A Scalable Tool for Browsing Online Comments. CHI 2010 [19] 文字云实例 http://www-958.ibm.com/software/data/cognos/manyeyes/ [20] 微博可视化分析系统 http://vis.pku.edu.cn/weibova/weiboevents/ [21] 流图实例 http://www-958.ibm.com/software/data/cognos/manyeyes/ [22] 新闻树图实例 http://newsmap.jp/  **2.2 项目意义**      本项目提出了舆情信息逻辑结构树的概念，用以描述信息在站点（或者整个互联网）上的分布情况；选用最优的可视化工具并加以改进，运用可视化技术将舆情分析的结果直观形象的呈现给用户，克服数据表现形式抽象、单一的缺陷。将可视化技术融入到网络舆情分析与监控中，帮助用户快速建立关于网络舆情数据及其规律的心理模型，更好更快地对舆情趋势做出判断和决策。项目意义如下：  **1、提出了网络舆情信息逻辑结构树的概念：**网络舆情信息逻辑结构是以每个站点为基本单位，对网络舆情信息在该网站中的分布结构进行分析，通过去噪、聚类等方式最终获得树形结构的完整表达形式。其中，非叶子结点用来表示网站中对信息的分类（模块），叶子结点表示舆情信息。在信息更新方面，通常只要对叶子结点或者接近叶子结点的部分进行操作即可。这种树结构即为上文提到的信息逻辑结构树，对于整个互联网则可以通过森林的形式进行展示；同时，对树型进行变化可以将关注的事件信息所在的叶子结点变为树根，而发布该事件的站点为叶子结点。  **2、探索可视化技术在舆情数据分析展示中的应用：**项目弥补当前舆情分析系统可视化程度低，数据展示形式单一的缺点，使舆情信息的分析结果呈现不再仅仅是单一的简报或报表；从宏观和微观两个角度对信息的生命周期进行可视化描述，采用关联技术和交互技术，使数据之间的规律更加明显；多角度比较，在对舆情数据进行比较描述的同时，也对平台的影响力进行评估和比较；  **3、探索在相应数据集下的工具模型：**专用的可视化工具模型，在现有技术上进行扩展和改进，使得呈现效果更加优化、表达信息更加全面；  **4、建立具有一定通用性的可视化实现体系：**针对网络舆情信息可视化设计专用的数据格式，使数据到图形的转换能够分层实现，从而是可视化工具具有更强的通用性和扩展性，这对以后研究可视化在具体领域的应用有着重要的作用；  **5、实现“基于信息逻辑结构树的网络舆情可视化分析系统”：**针对上文提到的目标和内容进行系统的设计和实现，使其具有较强实用性、安全性，同时弥补现有系统的缺陷，能够满足市场需求，具有一定的市场竞争力。    ***2.3 市场需求***      巨大的网民数和网站数都使当今互联网蕴藏着巨大的信息爆发力和广泛的影响力，从而使得网络舆情的分析与研究备受关注。如2012年北京暴雨事件，其传播途径如下：媒体报道→网民关注→网民质疑→官方表态→网民热议→媒体持续报道(来自“鹰眼舆情观察室”发布的《2012年7月社会热点》)。由此可见，网络舆情在事件信息获取与快速传播中起到了至关重要的作用。      中国互联网络信息中心（CNNIC）在京发布《第30次中国互联网络发展状况统计报告》显示，截至2012年6月底，中国网民数量达5.38亿，互联网普及率为39.9%。相较于其他传播方式，网络具有极强的隐蔽性和复杂性，这使得网络舆情具有信息的多元性、形成突发性、传播随意性、演化放大性等特点。而通过互联网及移动互联网得到的原始数据多为文本数据，其数据量大、数据类型繁多，且多为半结构化甚至非结构化数据，这些特点使得舆情数据的分析成为近年来理论和应用研究的热点领域之一，如何及时把握当前舆论导向，这也是当前学者研究的重点方向之一。      目前，越来越多的企业、研究所对舆论分析系统进行研究。现已有多家公司推出了自己的舆情分析系统，如乐思、军犬等等，这些系统信息采集面广泛、处理速度高、感情倾向性分析准确，但是普遍存在如下问题：可视化程度低、数据呈现形式单一，通常是以报表或者简单的统计图来呈现，使用户仍然需要面对大量复杂的信息，从而无法及时把握当前舆论状态做出快速判断和相应的处理。      网络舆情可视分析系统的实现，将使大量复杂的舆情数据不再仅仅通过报表的形式反馈给用户，而是通过直观的可视化工具来展示，同时配以友好的交互方式，使得用户在展示方式上能根据自己的需要进行调整，可以更好、更快、更准确的理解数据，做出决策。      根据上述背景，按照最低标准估算，本项目初步估计将推广到舆情监控及分析部门，具有巨大的市场潜力。  ***2.4 项目创新点和特色***  本项目是一个以实际需求为依托，多种技术相结合的研发项目，从学术研究、技术实现以及实际应用三点来考虑，其特色及创新点如下：  **1、将可视化技术与网络舆情分析相结合：**该项目弥补了应用研究领域网络舆情可视化的研究空白，使用直观的可视化技术对复杂海量的可视化数据进行分析和展示，对提高科学决策和处置水平都将具有重大的现实意义。  **2、提出了信息逻辑结构树的概念：**信息逻辑结构树是该项目的核心内容之一，项目拟用树状结构（或者是森林）来表示舆情信息在互联网上的分布情况，为舆情信息的存储和检索提供了较为便利的途径。此外，通过对信息逻辑结构树的可视化展示可以比较某一站点不同模块上信息的访问情况，对网站的合理规划具有一定的参考价值。  **3、建立规范标准的可视化体系：**可视化体系的建立使得可视化不再是简单的数据一次性转换成图形的过程，而是用分层的方式实现从底层数据到可视化成果发布的过程。使系统具有较强的通用性和可扩展性。  **4、针对网络舆情数据对可视化工具的模型进行改进：**根据数据特征选取可视化工具并就其算法进行改进。  **5、多视角比较：**在展示宏观数据的同时还注重对数据的比较，包括：不同事件生命周期的比较，不同类型平台对信息传播影响力度的比较，同一类型平台不同站点对信息传播力度的比较等等； |

|  |
| --- |
| **3、项目内容和方案** |
| **（１）项目目标、主要内容、技术途径、开发平台、语言工具** |
| ***3.1.1******项目目标***  本项目在研究网络舆情数据分析和相应可视化技术的基础上，以国内几个典型平台为例，如：新浪新闻（新闻类），天涯论坛（论坛类）以及新浪博客（博客类），进行数据的获取和研究。该系统能够对网络舆情数据进行自动获取、处理和分类，并通过可视化技术，从宏观和微观两个角度对舆情信息进行分析并对其数据特征进行可视化呈现，帮助用户在短时间内对当前的热点信息及其相关属性数据形成心理模型。  系统拥有友好的交互性，用户可通过选择数据集、时间段、具体网站、关注信息等，在数据展示过程中融入自己的操作，实现数据的“按需展示”。此外，该系统具有打印报表、权限管理、工作日志记录等功能，可按时间点对可视化数据进行备份。  ***3.1.2 主要内容***  该项目以系统的实现为主要核心工作，并对系统中涉及的到的技术、构架等进行深入研究。  该系统的硬件架构如下：  **图6 硬件架构图**  系统的外部由两部分组成，一部分是Internet网络资源，另一部分是用户的客户机。内网部分由三部分组成：网页采集服务器、数据库服务器、WEB服务器。其中数据库服务器为系统提供数据库相关服务，网页采集服务器负责对网页资源进行采集，WEB服务器负责向客户机提供服务。  该系统的逻辑工作流程如下：  **图7系统工作流程图**  **1、数据源选择**  选择几类典型网络平台进行研究：新闻类：如新浪新闻、搜狐新闻、腾讯新闻等;论坛类：如天涯论坛、百度贴吧等;博客类：如新浪博客、网易博客等。以上平台主要有三个特点：  1）知名度高，覆盖面广：能够广泛地获取舆情信息，对于互联网信息的舆论倾向有一定的指导性。  2）格式规范：这些案例格式规范，方便数据的提取  3）提供接口：这些平台提供了比较丰富的接口，方便信息的提取和分析  **2、数据提取与预处理**  通过爬虫技术以及部分平台提供的数据接口，将获取相关XML数据、HTML数据、图片等数据。通过一定的数据分析，在预处理环节将其转换成关键字，并将其时间信息和热度信息等保存在数据库中。同时，将提取该网站的URL信息，形成URL信息表，通过一定的算法和分析，最终得到信息的逻辑结构树。  **3、可视化方案设计**  首先，可视化体系建立，分层实现数据到图形的转换，并将各层模块化；      然后，对可视化工具进行研究和优化，使其从展示大量宏观数据和展示重要微观数据中得到平衡，并支持数据的动态更新。在系统中我们将考虑采用流图和树图作为主要的可视化工具进行研究。  树图主要用来对信息逻辑结构树进行呈现，在选取树图模型过程中，不仅要在宏观大量舆情数据和微观舆情信息细节中找到平衡点，还要在叶子结点（表现信息）和非叶子结点（结构信息）中得到平衡。      流图主要作用如下：   * 比较不同类别（如：新闻类，博客类，论坛类）网络平台对事件传播力度和影响力度； * 比较同一类别不同站点（如：新闻类的网易新闻，搜狐新闻）对事件传播力度和影响力度； * 比较不同主题随着时间热度的变化； * 显示单个事件的整个生命周期；   此外，还将根据实际实施情况以及数据的特征选择其他可视化工具，如文字云（关键信息比较）和平行坐标（高维数据可视化）等等。我们将通过一定的交互功能，实现用户对数据显示方式的控制，使显示结果更容易被用户接受。  **4、数据报表**  通过入库数据可以生成数据报表，数据报表将详细展示所有数据的分析结果。  **5、界面设计**  将可视化结果发布在界面上，通过合理的布局和交互接口，为用户提供友好的交互体验。  ***3.1.3技术途径***  本项目开发的技术途径在于结合可视化技术，将大量格式复杂、内容冗余的原始数据转换为形象直观的视图反馈给用户，帮助用户快速形成心理模型。整个项目包含内容比较丰富，从最底层的数据获取预处理，可视化体系建立和工具改进，再到最后的可视化效果展示。其主要的技术途径如下：    **图8 技术途径**    **1、原始数据的获取与预处理：**   * 采用基于Scrapy框架的web抓取技术获取原始数据，此时将获得XML和HTML类型文件；同时获得的还有URL表； * 对于URL表进行进一步分析，针对特定的站点（如新浪新闻）形成该平台的结构树，将所关注的舆情信息存放在树结点，并将该结构保存于文件； * 对于HTML文件，系统将采用开源的HtmlCleaner将其处理得到XML类型文件；再用DOM4J工具对全部的XML文件进行解析，将半结构化的XML数据转化成结构化的DOM树结构； * 通过DOM树结构提取文本信息，热点信息，时间信息等等关键属性；对于文本信息采用ICTCLAS4J工具进行中文分词，方便于进一步的挖掘； * 设计数据库并将数据入库；   **2、可视化体系的建立：**   * 对事件进行分类并对其相关属性进行量化和统计； * 对可视化工具的输入值进行范围限定，并将量化数据规约到该范围内； * 针对不同工具建立“数值-颜色”表，以及“数值-长度”表等关联关系； * 为可视化工具制定输入数据的接口标准；   **3、可视化工具改进：**   * 树图：有多种表现方式，如结点链接图、放射状图和同心圆图等等。为了在表示清楚树结构的同时更好地体现叶子结点数据信息，系统采用同心圆图和放射状图结合的方式； * 流图：将文字云嵌入到流图中，使流图表示更多信息；通过一定的交互可以选择显示重点，采用高亮放大的方式突出某一个流表示的信息； * 文字云：在文字云的布局过程中，采用文字坐标半锁定的方式，使同一个文字的位置在不同的视图中仅会在一定的范围内浮动，以达到不同文字云比较的目的；   **4、信息可视化实现：**   * 事件逻辑分布表示：采用树图进行表示：叶子结点表示舆情事件信息，非叶结点表示网站对信息的分类，根结点表示某一个网站；所有的网站及事件信息可以用一个森林进行表示。 * 单个生命周期表示：采用折线图的方式表示：横轴表示时间，纵轴表示事件当前点所受的关注度。 * 多个事件生命周期展示：采用流图进行表示：由多个并排的流状图进行表示，横轴表示时间，每个流在纵轴方向上的宽度为该流表示事件在当前时间点的受关注度； * 热点事件展示：采用文字云进行表示：文字云表示当前受关注度排名一定范围内的事件名称，大小表示某一定时间内的受关注度； * 不同类型平台（或者不同网站）对事件传播影响力度比较：采用流图进行表示：一个流图对应一个事件，多个并排的流分别表示不同平台（或网站），其横轴表示时间，纵轴上某一流的宽度表示该事件平台（或网站）上的受关注度； * 多视图：使用多线程针对多个窗口视图进行绘制，使系统可以同时完成多个可视化任务。同时建立全局索引机制，使用数据的索引进行可视化，数据的变化会使全部视图刷新。 * 报表与可视化视图关联：在报表中进行索引设置，通过对可视化视图的交互操作，可以打开报表并自动检索到报表中的相应位置。     ***3.1.4 开发语言及平台***     |  |  | | --- | --- | | 开发平台 | Windows XP/Win 7 | | 开发工具 | Eclipse | | 数据库管理系统 | MySql | | 开发语言 | Java | |
| **（２）项目进展计划和预期成果** |
| 3.2.1 项目进展计划  详细进展计划   |  |  |  | | --- | --- | --- | | 阶段 | | 时间区间 | | 第一阶段: 需求细化与数据获取 | 1、明确系统的需求分析 2、完成系统的架构设计 | 2012年11月15日—2012年12月31日 | | | 1、获取源数据 2、处理数据信息，并分类保存到本地 | 2013年01月01日—2013年01月31日 | | | 第二阶段: 数据预处理 | 解析url信息表，形成逻辑结构树 | 2013年02月01日—2013年02月15日 | | 对数据信息进行分析和统计，并归一化 | 2013年02月16日—2013年03月15日 | | 将数据入库或者保存到特定格式的文件 | 2013年03月16日—2013年03月23日 | | 第三阶段: 可视化模型建立 | 设计和实现舆情分析报表的打印模块 | 2013年03月24日—2013年03月31日 | | 建立信息逻辑结构树可视化模型 | 2013年04月01日—2013年04月10日 | | 建立舆情信息可视化工具模型 | 2013年04月11日—2013年04月30日 | | 第四阶段: 可视化方案设计与仿真 | 1、以可视化视图展示实际数据 2、多视图表达方案设计 3、关联设计 4、交互设计 | 2013年05月01日—2013年07月31日 | | | | | 仿真平台下可视化工具实现 | 2013年08月01日—2013年08月15日 | | 第五阶段: 系统集成与测试 | 系统集成 | 2013年08月16日—2013年08月31日 | | 系统测试和改进 | 2013年09月01日—2013年09月30日 | | 论文撰写和专利申请 | 2013年10月01日—2013年11月15日 |     ***3.2.2 预期成果***      本项目研究的预期成果：   1. 实现网络舆情可视分析系统 2. 发表高质量学术论文2-4篇； 3. 申请软件著作权或发明专利一项； 4. 培养研究生3-4人； 5. 与应用单位，力争实现部分成果的转化和应用推广。 |

|  |
| --- |
| **4、项目组分工与合作方案** |
| **4.1** **项目分工**  为了使项目的开发具有延续性，团队成员由包括了博士研究生、硕士研究生和本科生共计6人组成。人员比例如下图：    **图9 人员比例**  项目主体是系统开发，因此从软件开发的角度将团队分为算法设计、模块开发和系统测试三个方面。    **图10项目分工**  为使项目整体进展顺利，并达到预期成果，详细分组如下：  **项目经理（申乔木）：**  对项目进行总体规划，包括人员安排、进度与风险控制、协调各部门间合作以及制定项目经费使用计划等。  **核心算法组（申乔木，张永清，高丽峰）：**  该小组研究解决问题的核心算法：数据分析算法，可视化工具模型的建立算法，可视化图形绘制算法等等。在研究的同时与其他成员沟通协作，制定可行性好、通用性高的方法，并在理论上进行充分论证，在保证可行性后进行核心代码实现。  **系统实现组（申乔木，李梦颖，邵建荃）：**  对解决方案进行编程实现。该小组在协助核心算法组完成算法同时完成外围程序设计，包括完成UI(用户界面)设计、辅助程序、相关接口、稳定性保障等方面程序，此小组为系统最基础部分，工作量最大，因此人数安排需要较多系统测试：  **系统测试组：（邵建荃，符敏）：**  完成单元测试及集成测试。在各个单元模块完成设计后，测试小组均需采用不同方法对功能模块进行测试，确保模块能正常运行，在所有模块集成后该小组需进行集成测试，对整体进行测试，从实际上验证方案可行。    ***4.2 合作方案***   1. 需求分析阶段，所有小组共同学习，全面了解项目主要任务； 2. 系统建模阶段，程序开发组及测试组部分成员加入核心算法组，协助解决困难，尽早制定解决方案； 3. 开发阶段测试组及核心算法组均增加相应成员到程序开发组，保证程序设计全部能按计划完成。 4. 在测试阶段，所有开发人员均担当测试员，首先对自己开发程序进行测试，保证基本正常运行后交由测试组，减少因沟通带来的时间延迟。       总体上来说，所有组中大部分成员均需灵活调度，保证项目总体进展顺利，并取得预期成果。 |

|  |  |
| --- | --- |
| **5、在读学校或工作单位审核意见** | |
|  | |
| 负责人（签章） 在读学校或工作单位(盖章) | 年  月  日 |
| **6、专家组评审意** | |
|  | |
| 组长（签章） | 年  月  日 |
| **7、项目管理办公室意见** | |
|  | |
| 主任（签章） | 年  月  日 |