Inline Text Wrapping Picture

北京邮电大学

硕士研究生学位论文开题报告

学 号: 2015140089

姓 名: 蒋克层

学 院: 信息与通信工程学院

专业(领域): 电子与通信工程

研究方向: 宽带无线移动通信系统新理论及技术

导师姓名: 李立华

攻 读 学 位: 工程硕士

2016年12月29日

|  |  |  |  |
| --- | --- | --- | --- |
| 论文题目 | 社会热点事件辩论平台的设计与实现 | | |
| 选题来源 | 国家自然科学基金项目 | 论文类型 | 基础研究 |
| 开题日期 | 2016-12-15 | 开题地点 | 新科研楼723 |
| **一、立题依据（包括研究目的、意义、国内外研究现状和发展趋势，需结合科学研究发展趋势来论述科学意义；或结合国民经济和社会发展中迫切需要解决的关键科技问题来论述其应用前景。附主要参考文献目录）（不少于800字）**  在科技日新月异的今天,互联网信息技术的发展更加迅猛,网络对社会的生产和人们的工作、学习、生活和生存模式、思维模式产生着越来越大的影响。各种新技术的出现和应用不仅带来了网络产业的升级,也给旧的信息传播观念、传播模式带来了巨大冲击。用户成为信息制造的中心,拥有更多点对点传递信息的渠道[1]。  由于国内政府、企业等对网络舆情的关注度不断升温，使得关于互联网舆情领域的研究课题成为热点。微博、博客、论坛这样的互联网工具，是网络信息的载体，作为新闻或者信息的发布媒介，可以迅速地将某一“突发事件”发布在网上，随着参与“讨论”的网民的数量增加，有些会逐渐演变成为“热点话题”[2]。面对热点事件，社会各阶层的都有发表自己观点的需求，用户需要一个特别针对热点事件发表看法的平台。目前流行的互联网工具存在一定的局限性：微博是时下流行的热点事件互动平台，但是活跃在微博上的意见领袖，由于身份、社会地位等特殊性，在很多新闻热点事件及重大社会事件上，他们不仅主导着网络的话语权，左右着舆论，而且甚至影响整个事态的发展走向；博客平台对热点事件响应速度存在一定的滞后性，短时间的影响力较弱；论坛的用户数量规模以及用户身份的相似性使得热点事件的观点不能完整反映各个阶层的声音。  面对上诉问题，一个很好的解决方案是建立一个专门针对热点事件的辩论平台，降低各种身份的人参与的门槛，从而可以很好地反映出热点事件在不同的人群之中观点。观点的冲突在很多情况下会引发辩论，将不同立场下的言论聚集起来会增加对热点问题看法的广泛性。同时，群众的观点在民主的社会形态下扮演者不可获取的角色，例如政府在发布政策信息之前会一定程度上参考民众的意见，甚至法院在判决某个引起社会广泛关注的案件也会一定程度上考虑到判决产生的社会影响。因此，一个可以集中收集民意，反应民众针对时下热点事件观点态度的平台具有重要的社会意义与决策参考价值。  在现实世界中,可获取的大部信息是以文本形式存储在文本数据库中的,由来自各种数据源的大量文档组成,如新闻文档、研究论文、书籍、数字图书馆、电子邮件和页面。由于电子形式的文本信息飞速增涨,文本挖掘已经成为信息领域的研究热点。目前互联网中的信息大部分都以文字的形式呈现出来，所以，文本信息依然是非常重要的信息资源之一，这些文本信息通常包含在各个领域。  用户辩文内容的提交会产生文本消息，文本消息和实体信息是互联网文本数据的两种重要类型。社会化媒体上的文本消息，通常属短文本，并按其时间属性组织后形成文本消息流，包含着网民们的许多思想观念与倾向。然而，文本消息的不完整性、奇异性、海量性和动态性导致文本消息流的话题发现、倾向性分析和热点信息挖掘十分困难[3]。对热点事件的辩文内容进行进一步的文本挖掘分析，整理出用户创作内容质量高的辩文内容，进行推送与发布是自然而然的需求。通过文本内容分析，过滤掉包含敏感字段的辩文，收集出民众对热点事件的主要观点，汇总形成总结文档，具备重要的社会研究价值。  主要参考文献：  [1] 王丹.基于WEB2.0的信息服务研究[D].武汉.华中师范大学,2007  [2] 曲克伟.基于热点话题发现的BBS检索排序策略研究[D].北京.北京邮电大学,2012  [3] 黄九鸣.面向舆情分析和属性发现的网络文本挖掘技术研究[D].长沙.国防科学技术大学研究生院,2011  [4] 崔安颀.微博热点事件的公众情感分析研究[D].北京.清华大学,2013  [5] Paul Deck.Spring MVC[M].北京:人民邮电出版社,2015  [6] Josiah L.Carlson.Redis IN ACTION[M].北京:人民邮电出版社,2016  [7] Nick Dimiduk,Amandeep Khurana.Hbase IN ACTION[M].北京:人民邮电出版社,2016  [8] 马延辉等.Enterprise Application Development with HBase[M].北京:机械工业出版社,2015  [9] 单东林等.锋利的jQuery[M].北京:人民邮电出版社,2016 | | | |

|  |
| --- |
| **二、研究内容和目标（说明课题的具体研究内容，研究目标和效果，以及拟解决的关键科学问题。此部分为重点阐述内容）（不少于2500字）**  随着大数据相关技术的应用越来越广泛，互联网逐渐步入大数据时代，这使每个互联网产业参与者的行为都发生了改变。其中最大的改变莫过于普通民众的行为在大数据分析系统面前将是“可视化”。其中有效地整合民众在社交媒体上的言论和观点，并精准的对用户进行“画像”是高效利用大数据技术维护社会治安，评估舆论方向，进行精准营销等活动的前提。  本课题以广大微博用户在微博上留下的海量文本信息为基础，对用户关注，转发微博，发表微博，好友互动等行为进行分析，设计实现了微博用户的画像系统。本课题的研究内容主要包括以下两个部分：   1. 微博用户画像展示平台   展示平台初步设定的需求为，输入已知用户微博昵称或数字id，可显示用户基本信息，包括性别，年龄，归属地等，同时显示用户的画像信息包括兴趣爱好，关注内容等。另外本系统依据12维的社会态度特征向量对用户进行建模，描述了单一用户在社会态度方面的情绪倾向，并以此对该用户进行社会风险行为分级，系统可查询用户近几次的危险言论。与此同时可显示该用户的主要社交关系人，并显示社交网络中不同社交关系人的风险级别。   1. 微博数据的收集与分析   2.1 数据来源  本课题研究的的数据集主要来源于对微博数据的爬取，这部分由Java语言编写，可以同时爬取三部分内容：   1. 特定用户主页下的所有动态包括转发微博，原创微博，以及微博下与其他用户的评论互动。 2. 该用户的粉丝和关注者信息 3. 该用户自身的自述性信息，包括昵称，数字id，性别，年龄，归属地，自我简介   爬取数据的策略初步设定为：从某一用户出发，分别一1000为阈值，爬取该用户的粉丝和关注用户列表，并以每一个粉丝（或者关注用户）为新的结点继续查找新的粉丝和关注用户列表进行爬取。   * 1. 数据处理   在数据处理部分，主要进行三个步骤的处理，首先将数据进行分开建档，爬取到的数据，按照用户信息，粉丝，关注用户分别储存，其中用户个人的博文文本储存为一个长文本词袋，为后续的主题模型处理提供原始数据。  其次，对抓取后的数据进行预处理，预处理包括中文分词和停用词过滤两个部分。不同于英文数据中由空格符进行分词，中文的词与词之间没有明显的分隔符号，中文分词的目的再于将中文语句中的词与词分隔开来以供机器识别，中文分词也将直接影响到信息抽取、文本内容分析等的效果。同时因为用户的博文中为了使语义连贯，通常会使用一些“的”、“地”等停用词和“，”、“：”等标点符号，所以在进行分词处理后，需要将数据进行清洗。停用词过滤即为去除微博文本中所包含的对用户特征挖掘并无太大意义的词语或符号，这一步对于降低信息冗余度、提高兴趣分析准确度方面是非常必要的。  最后，利用算法实现用户特征信息模型的构建。由于微博文本信息具有文本短，表情多，不规范用语多，网络词汇多等特点，传统的文本挖掘技术如 VSM、，LSA，PLSA等已经不能有效地处理 Web 2.0 时代下社交网络的文本内容噪声多、用词不规范、词频较少等问题。主题模型可以更好地利用这些文本信息，于在线社交网站文本挖掘方面有着广泛的应用。  LDA(Latent Dirichlet Allocation)是一种文档生成模型。LDA 主题模型将文本分为文档层、主题层和单词层，文档和主题相关联主题和单词相关，联假设一篇文档是由主题集合中的各个主题按照一定的比例构成的，而每一个主题又是由单词表中的单词按照一定的比例混合而成的.通过机器学习的方法可以得到文档的主题。LDA 模型通过学习文档集中的单词挖掘出所有潜在的主题信息,并通过这些信息来挖掘该文档集以外的其他文档的主题分布，可直观看出网民对各种微博主题感兴趣的情况。  使用textmind工具可以很方便地获取特定用户的情绪向量，包括正向情感词，负向情感词，焦虑词，生气词，健康词，关爱词，脏话，惊叹词等情绪传达词的所占比率，根据这些情绪向量，连同微博的静态特征一起，通过多任务增量回归的方法，对用户的五大人格特征进行预测    1.社会热点事件线上辩论平台设计  随着 IT 技术的发展和企业对信息化需求的日益增加，在软件开发技术方面逐渐分为了两个阵营：Java 和.Net。因为 Java 在安全性、跨平台性以及分层次的架构方面有着不可比拟的优势，所以特别适合于大型企业级应用的开发，Java EE 就是 Java 平台中专门为企业应用制订的标准框架。  线上辩论平台初步需求为，用户可以自由选择进入热点事件辩文界面，在此界面用户可以浏览已经发表的辩文，可以对自己认可的辩文点赞，允许对辩文进行评论。用户提交的辩文信息，需要先经过敏感字段过滤的程序处理，处理完毕后，所有当前页面的用户均可以收到该辩文，实现辩文的广播。平台需要有管理个人信息入口，包括个人资料的修改与保存，头像信息的上传等。  针对热点事件线上辩论平台的设计需求，需要研究具体的实现技术。现有的B/S结构的互联网平台涉及到技术主要分为前端以及后端。后端主要包括数据库模块，架构模块，消息队列模块以及权限控制模块。架构模块采用教科书式的MVC框架Spring MVC，该框架结构设计合理，可扩展性强，针对底层存储技术以及顶层显示技术具有良好的兼容性。数据库模块作为平台的数据支撑，存储用户信息以及辩文的信息，对于结构化的信息，采用关系型数据库MySQL进行存储；关系离散的、数据量大的信息存储在非结构化数据库Hbase中，该数据库建立在集群环境之上，可扩展性好,数据规模大；对于访问频繁的数据，存放在Key-Value数据库Redis中，Redis中的数据均存储在内存中，响应速度快但存储数据量较小；音频、图片等文件存储在FastDFS中，FastDFS具备良好的文件存取属性。消息队列模块功能主要是加快平台的响应速度，优化用户体验，拟采用ActiveMQ技术实现。  辩论平台相比较于传统的拉取时B/S模式，需要增加的是广播的功能。当一个用户的辩文提交到平台并通过审核后，为了实现辩论动态化刷新效果，需要将辩文推送到当前所有处于本页面下的用户。传统的解决思路有Ajax轮询模式与long pull模式，Ajax轮询模式是前端界面在固定时间间隔下，定时向服务器请求数据，有更新则动态修改界面，无更新则不做处理；long pull模式是指阻塞性发送请求，一直等到回应才更新，然后发起下一次请求。这两种技术在技术均存在缺陷：Ajax轮询需要服务器有很快的处理速度和资源；long poll需要服务器具有处理高并发的能力。Websocket与STOMP协议结合可以很好地解决这个问题，Websocket只发送了一次http请求，当服务器有数据时再向浏览器推送数据，减少了带宽的使用以及服务器CPU使用率。STOMP即Simple (or Streaming) Text Orientated Messaging Protocol，简单(流)文本定向消息协议，它提供了一个可互操作的连接格式，允许STOMP客户端与任意STOMP消息代理（Broker）进行交互。  在用户使用平台时，服务器需要知道当前操作用户的身份与其所拥有的权限信息，传统的Session技术在处理多服务器数据同步上存在问题，Cookie技术将数据存放在浏览器端，安全性存在隐患。Apache Shiro与Redis结合可以很好地解决以上问题。Apache Shiro是一个强大的，易用的Java安全框架，常被被用作于认证，授权与加密，在前端与后台均可以方便实现对主体的角色与权限的认证。  面向用户的线上辩论平台服务于普通用户，对于产生的辩文等相关数据，需要设计后台数据管理系统进行管理。后台系统可以对用户的权限进行管理，对于恶意发布低俗、包含敏感字段的用户取消发布辩文资格等。同时，针对程序判定为包含敏感字段的辩文，需要后台系统提供人工复核的功能。对于某个热点事件产生的优质辩文，进行整理并发布。  2.辩文内容分析系统  文本挖掘就是以文本信息为对象，从中探索信息的模式、结构、模型等隐含的、具有潜在价值知识的过程。其所涉及到的技术主要包括文本表示、特征选择、文本分类、文本聚类、信息检索等。其中，文本表示是文本挖掘中必不可少的根本环节，特征选择、文本分类和文本聚类等技术则是其核心技术。  由于文本特征向量的高维性造成了文本信息处理困难，因此，文本表示成了文本挖掘首先要解决的问题。对于一种合理的文本表示方法应该满足以下两方面的要求：一方面在保留文本信息中关键特征的前提下，还要涵盖尽可能多的信息；另一方面特征也不宜过多，否则将会影响挖掘算法的效果。特征选择就是一种解决以上两方面问题的技术，能够完成对文本特征向量的约简。  对于本系统，主要运用文本挖掘的技术解决两个主要的问题，一个是敏感字段的过滤模块，对于不符合相关规定的辩文进行过滤处理；另外一个是用户创作内容质量分析，识别出可信度强，内容质量高的辩文。  热点事件的最主要的特点是短时间内的高并发数据量的生成，通常一个热点事件发生后的短时间内会产生超过70%的数据量，然后数据信息会处于递减的状态中。要求敏感字段过滤程序需要满足短时间内大批量处理任务的需求，考虑到包含敏感字段辩文广播后后果的严重性，需要考虑在牺牲部分精确率的条件下保证召回率。将包含敏感字段的程序发布给后台管理系统进行人工复审。  网站中的用户创作内容质量分析是本研究的另一个重点。内容质量优质是维系用户对网站长期依赖的重要支撑点。近一段时间，微博等互联网工具网站中充斥的竭言、不实信息，引起了各界的极大关注。新闻不断的报道，相关法规的出台，都说明了网络中的不实信息己经造成了十分恶劣的影响。虽然已经不断有相关的政策、法规的出台，对不实信息的发表和传播进行限制，但依然有恶意的用户，如水军、网络推手等在发布不实的信息，鼓动普通用户进行传播，造成了广泛的恶劣影响。文本挖掘是一些交叉的研究领域结合,它涉及到数据挖掘、信息检索、自然语言处理、机器学习等多个领域的内容,不同的研究者从各自的研究领域出发,对文本挖掘的含义有不同的理解,不同应用目的文本挖掘项目也各有其侧重点。因此，在辩文内容发布后针对内容质量进行评估，对于创作内容进行合理的标注，对于内容质量低下的可发人工复审，进行删除或者显示顺序推延等处理；对于创作质量高的可进行整理发布或者在当前热点事件下的辩文优先展示。 |

|  |
| --- |
| **三、研究方案设计及可行性分析（包括：研究方法，技术路线，理论分析、计算、实验方法和步骤及其可行性等）（不少于800字）**  本论文拟设计并实现社会热点事件的线上辩论平台以及辩文内容分析系统，并针对线上辩论平台实现其相应的后台管理系统，同时对辩文进行敏感字段过滤处理以及辩文内容质量分析。具体的，本论文将沿袭一下的技术路线完成热点事件的下上辩论平台以及分析系统：  1. 对相关领域的技术进行充分调研  充分而有效的调研可以保障在工程设计上不会出现技术路线的偏差，成熟的技术架构在满足设计需求的同时也可以为开发提供可靠的保障。  首先，在确定系统的实现的前景时，充分考虑了时下用户对社会热点事件迫切表达观点的需求。设计平台的主要功能的时候，收集并归纳了用户的反馈。对当前互联网常用架构进行了分析与总结，并与导师和领域的专家进行了充分的交流和讨论，同时与项目组成员进行阶段性的总结报告。充分了解与学习实验室之前进行过的工程项目，确定具体的实现技术。  其次，在工程实现的具体过程中，扎实的理论基础具有很重要的作用，对于涉及到的相关技术，提前购买了相关的书籍，进行充分的学习。工程性的项目在实现的过程中会在所难免的出现这样那样的错误，充分利用搜索引擎结合已经了解的理论基础，可以迅速定位到程序的出错点。  最后，在工程有了一定的进展时，需要对已经实现的工程项目进行测试，包括正常的功能性测试以及人为异常输入，对数据进行合理化的备份处理，才能将程序部署到云平台上，完成论文的编写。  2. 利用书籍、博客以及技术文档等进行理论分析  本论文工程项目的设计与实现需要对Java EE架构下涉及到的技术非常的熟悉，对于后端常用的数据库技术有很多实践类型的书本可以参考。部分技术博客立足工程实践的具体需求，具备很强的可再现性，对于理论的集体实践有一定的指导价值。相关技术的开发文档是最直接的资料，技术涉及到相关的使用指南均可以在开发文档中找到。  3．利用常用的测试方法等对项目进行测试评估  在工程项目的实现过程中，为了测试代码正确性以及健壮性，在每部分代码完成后，需要编写单元测试。在单个模块单元的功能实现以后，内部测试是发现代码漏洞的有效手段，完成常规的用户测试时。在不同的阶段，分别进行单元测试、集成测试、系统测试以及验收测试。  4. 依托项目进行相关交流  项目在实现过程中需要不断地与合作方进行交流汇报，完成阶段性的项目进度总结报告。对于技术上的瓶颈，可以在交流中寻求合作方的技术指导。团队的合作以及不同想法间的相互交流十分重要，有助于迅速发现问题，提高解决问题的效率。 |

|  |
| --- |
| **四、本研究课题可能的创新之处（不少于500字）**  面对热点事件，社会各阶层的都有发表自己心声的需求，用户需要一个特别针对热点事件发表看法的平台。用户对于社会热点事件的表达看法的需求越来越高，而目前流行的互联网工具例如微博、博客以及论坛等均存在一定的局限性。本课题的创新点应该在于设计并实现了针对热点事件的辩论平台和针对辩文内容的敏感字段过滤机文本质量分析，具体包括：  1.社会热点事件的辩论平台  （1）.功能：社会热点事件的话题辩论。现有的互联网工具内容涉及到日常生活的方方面面，比较宽泛，本平台进行了内容话题内容的界定，将话题讨论的范围缩小在社会热点问题上，内容更加集中。  （2）.实现技术：页面广播。现有的互联网工具页面的更新手段采用是Ajax的定时轮训方式，对于辩文信息的广播，采用的WebSocket protocol协议，建立全双工通道，进行通信。  2.社会热点事件的辩论分析系统  （1）.高速敏感字段过滤。在处理辩文的敏感字段时，需要实时进行敏感字段的过滤，短时间内识别出设计敏感信息的辩文，进行过滤处理。要求敏感字段过滤程序需要满足短时间内大批量处理任务的需求，考虑到包含敏感字段辩文广播后后果的严重性，需要考虑在牺牲部分精确率的条件下保证召回率。  （2）. 用户创作内容质量分析。现有的互联网工具在用户创作内容的质量分析存在部分欠缺，在辩文内容发布后针对内容质量进行评估，对于创作内容进行合理的标注，对于内容质量低下的可发人工复审，进行删除或者显示顺序推延等处理；对于创作质量高的可进行整理发布或者在当前热点事件下的辩文优先展示。 |
| **五、研究基础与工作条件（1.与本项目相关的研究工作积累基础 2.包括已具备的实验条件，尚缺少的实验条件和拟解决途径）（不少于500字）**  1. 技术支撑  经历了互联网将近20年的高速发展，现有的用于实现Java Web的技术以逐渐趋于成熟，企业级轻量化网站架构的设计与实现很多已经发展成熟。当今的互联网在人类生活中扮演了十分重要的角色。特别是随着 Web2.0 技术的发展，微博、论坛、博客和电子商务交易平台等应用大量涌现，互联网已成了人类社会活动和传播信息的重要平台。文本是互联网上数据的主要形式，对其深入挖掘有重大的应用价值和学术意义。对于各种新型互联网应用上的文本数据，现有的文本挖掘技术也处于飞速发展阶段。文本挖掘是一些交叉的研究领域结合,它涉及到数据挖掘、信息检索、自然语言处理、机器学习等多个领域的内容,伴随着各个领域的发展也为文本挖掘提供了技术条件。  2. 合作交流以及开源技术  项目的实现在一定程度上可以依靠合作方的技术积累，合作方对于基础开发框架的搭建的技术支持为开发提供的条件。目前，开源社区开放了很多的开源技术，充分利用现有的开源代码进行相关的整合与修改会大大减少代码量。  3. 软硬件支持  实验室为项目的开发提供了良好的软硬件条件，包括新购置的多台服务器，提供了用于分布式存储系统的小型集群，同时也包括 Eclipse，Xshell，Xftp，Navicat，Redis Studio等软件。  4. 具备的实验条件  实验室目前的研究方向涵盖移动通信领域的基础理论，关键技术的探索和实现，移动通信系统的标准化，基于个体和群体特征的网络配置与优化等。研究领域涵盖物理层、网络层、服务环境、系统最优化和演示系统的建设。作为通信技术的国内主要研究力量之一，实验室的成果丰富。综上，实验室提供的优秀科研条件将为论文课题的进行提供有力的支撑。 |

**学位论文工作计划**

|  |  |  |
| --- | --- | --- |
| 时间 | 研究内容 | 预期效果 |
| 2015年12月至2016年2月 | 学习调研阶段。调研并了解Java WEB开发常用的技术，阅读《Spring MVC》、《Hbase企业应用开发实战》、《锋利的jQuery》、《Redis实战》等工具书籍。同时查看技术博客以及文本挖掘相关论文。 | 对项目中涉及到的数据库知识、框架知识以及前端知识有深入的了解，掌握调试方法，了解使用机制。 |
| 2016年2月至2016年5月 | 方案设计阶段。根据第一阶段的调研结果，通过理论分析指导，确定项目明确使用的技术方案，设计出项目的整体框架。 | 根据制定的研究内容，明确方案，验证框架切实可行。 |
| 2016年5月至2016年7月 | 程序开发阶段。按照设计需求，逐步实现程序功能，解决开发过程中出现的问题。修正框架中不合理的地方。 | 程序开发顺利，全部功能得以实现，系统运行流畅。 |
| 2016年7月至2016年10月 | 验证完善阶段。进行程序的验证测试，完善平台的中细节。开始论文的撰写工作。 | 对系统进行进一步的优化，并将成果转化为学术论文。 |
| 2016年10月至2016年12月 | 论文撰写阶段。整理课题进行期间的研究资料，撰写硕士学位论文。 | 完成硕士学位论文的文档性工作。 |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 评  定  小  组  成  员 | 姓 名 | 职 称 | 单位名称 | 职务 |
| 王莹 | 教授 | 北京邮电大学 | 组长 |
| 李立华 | 副教授 | 北京邮电大学 | 成员 |
| 田辉 | 教授 | 北京邮电大学 | 成员 |
| 翟文军 | 副教授 | 北京星河亮点技术股份有限公司 | 成员 |
|  |  |  |  |
|  |  |  |  |
| 导师意见： | | | | |
| 同意开题 | | | | |
| 导师（签名）：  日期： 年 月 日 | | | | |
| 开题报告小组意见： | | | | |
| 组长（签名）：  日期： 年 月 日 | | | | |
| 学院意见（签章）： | | | | |
| 负责人：  日期： 年 月 日 | | | | |