

毕业设计(论文)开题报告

（适用于工科类、理科类专业）

|  |  |  |  |
| --- | --- | --- | --- |
| 课题名称 | 一个面向特定领域的微博数据采集和分析系统 | | |
| 副 标 题 |  | | |
| 学院（系） | 软件学院 | | |
| 专 业 | 软件工程 | | |
| 学生姓名 | 陈薇伊 | 学 号 | 1252874 |
|  |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 2016 | 年 |  | 月 |  | 日 |

一、毕业设计（论文）课题背景（含文献综述）

随着网络的迅猛发展，各种网络社交平台快速发展并普及到人们的日常生活中。其中，微博作为一个基于用户关系信息分享、传播以及获取的平台，以其即时分享的时效性和随意性跻身成为最热门的社交网络平台之一。

2009年8月，新浪微博作为中国门户网站中的第一个微博网站进入人们的视野。据2015年微博发布的第三季度财报中显示，截止2015年9月30日，微博月活跃用户数(MAU)已经达到2.12亿人，较上年同期增长48%。基于大量的使用人数，新浪微博的状态信息更新频繁，信息转播速度快，且在微博平台中媒介用户占有率相对集中，因此，基于微博数据的研究成为了十分值得关注的研究方向。同时，由于抓取新浪微博数据涉及到登陆模式等相关问题，普通的网络爬虫的抓取模式很难运用到新浪微博的数据获取中[1]。所以，相对使用网络爬虫，使用新浪微博API接口能更加简洁地获取数据，但为了保证数据的规模，在使用新浪微博API获取的数据量不够的情况下，可以使用网络爬虫获取更多的数据。同时为了更方便快捷地使用新浪微博API和提高开发效率，系统将使用可读性强，可移植性高，使用广泛，易于操作文件的Python语言[5]进行开发。

通过新浪微博API获取到大量的数据后，需要对数据进行处理和储存。NoSQL数据库近年来在处理大数据上以高效率、高储存量、低成本的特性获得广泛的使用[3],其中的MongoDB是一个面向文档的数据库，通过面向文档的方法能够仅仅使用一条记录来表现复杂的层次关系[2]。同时，MongoDB不规定键值的类型和大小，还提供索引、聚合等独特的功能，这使得数据库的访问和数据的修改的速度极快[2] [3]。

将获得的数据存储之后，需要对所得的数据进行文本分析，即自然语言处理，它包括了分词，计算词频，信息检索，信息分类，关键词抽取等，NLTK（Natural Language Toolkit）便是一个基于Python的自然语言工具包，它包含了词性标识符、分类、分块、解析、语义解释等多模块进行自然语言处理[7]。由于仅通过单一的关键词对新浪微博进行数据抽取，会导致数据量不够或者部分相关数据缺失影响后续的分析，通过对数据进行文本分析，提取主题词，再以得到的主题词去进行相关数据的手机能扩大数据量。提取主题词首先需要分词，Jieba是基于Python的中文分词开源库，它支持精确模式、全模式和搜索引擎模式等三种分词模式，同时支持繁体分词和自定义词典，并以其扩展性好获得了广泛的使用。分好词后，可以通过主题建模进行主题词提取。通过TF-IDF可以找出文档集中便于搜索的主题词[4]，其主要思想是：如果某个词或短语在一篇文章中出现的频率TF高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类。同时，使用主题模型对已抓取的数据提取潜在的主题，并利用相关主题进行进一步的数据抽取可以使抽取的数据覆盖面更广。主题模型是对文字隐含主题进行建模的方法，其中LDA（Latent Dirichlet Allocation）是一个三级层次贝叶斯模型[6]，它可以用来识别语料库或者大规模文档集中隐藏的主题信息。其中，Gensim是一个基于Python的开源主题模型工具包，它可以有效地自动抽取予以主题，其中包括了LSA（Latent Semantic Analysis）,LDA和RP（Random Projections）等主题模型的算法。

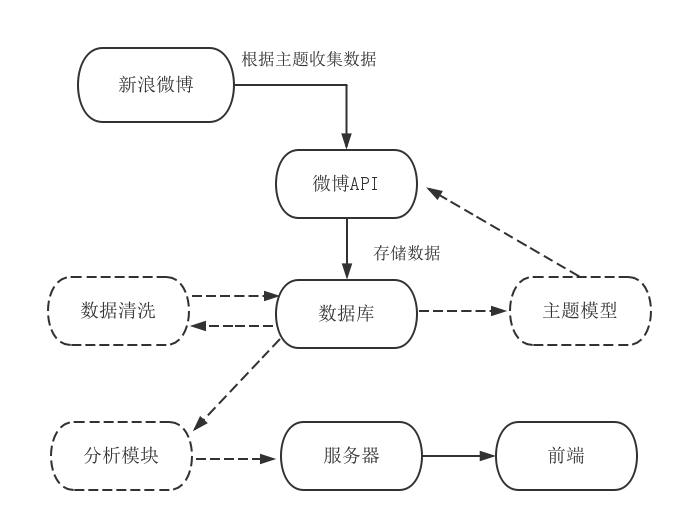
将数据处理好之后，最后要进行数据分析。由于从微博采集到的数据是微博用户分享日常生活的各个方面的评论和社会舆论，因此从这些数据中可以进行情绪分析[8][10]。通过有效的分析，可以确定某一个状态或评论的情绪是正面，负面或者是中性的[9]。

本课题便是在这个背景下提出，旨在设计并实现一个有效的面向特定领域微博数据收集分析系统。

二、毕业设计（论文）方案介绍（主要内容）

本课题最终会建立一个微博数据收集分析系统，根据主题从新浪微博抓取数据，对数据进行预处理并将数据存入MongoDB数据库中，并对数据进行简单的分析。

系统的大致工作流如下：



1．数据收集

使用Python编写程序调用新浪微博API，根据主题词定时对新浪微博上的最新数据进行抓取，并将数据存储到MongoDB数据库中。

2．数据清洗

将通过新浪微博API获取到的JSON格式的数据进行重新审查和校验，并将数据中的重复信息，无用信息删除，纠正存在的错误，使数据保持完整和一致性。

3．主题模型

获得一定量的有效数据后，使用主题模型对已有的数据进行主题提取。根据LDA的思想，对每一条状态，从主题分布中抽取一个主题，然后从上述被抽到的主题所对应的单词分布中抽取一个单词，重复上述过程直至遍历状态中的每一个单词。

4.数据分析

基于收集到的数据，提取有用的信息，并对其进行情绪分析，判断其情绪为正面，负面或者是中性的。

三、毕业设计（论文）的主要参考文献

[1]廉捷, 周欣, 曹伟,等. 新浪微博数据挖掘方案[J]. 清华大学学报:自然科学版, 2011(10):1300-1305.

[2] Chodorow, Kristina. MongoDB: the definitive guide[M]. " O'Reilly Media, Inc.", 2013.

[3]Han J, Haihong E, Le G, et al. Survey on NoSQL database[C]// Pervasive Computing and Applications (ICPCA), 2011 6th International Conference on. IEEE, 2011:363-366.

[4] Ramos, Juan. "Using tf-idf to determine word relevance in document queries." Proceedings of the first instructional conference on machine learning. 2003.

[5] Lutz, Mark. Learning python[M]. " O'Reilly Media, Inc.", 2013.

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

[7] Bird, Steven, Ewan Klein, and Edward Loper. Natural language processing with Python[M]. " O'Reilly Media, Inc.", 2009.

[8] Pang B, Lee L. Opinion Mining and Sentiment Analysis[J]. Foundations and Trends in Information Retrieval，2008， 2(1–2):1-135.

[9] Pak, Alexander, and Patrick Paroubek. Twitter as a Corpus for Sentiment Analysis and Opinion Mining[C]// Conference on International Language Resources & Evaluation. 2010: 1320-1326

[10] Tan, Songbo, and J. Zhang. An empirical study of sentiment analysis for chinese documents[J]. *Expert Systems with Applications* 34.4(2008):2622-2629.

四、审核意见

|  |
| --- |
| 指导教师审核意见：（针对选题的价值及可行性作出具体评价）  指导教师签名  年 月 日 |
| 专业审核意见：  负责人签名  年 月 日 |