|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **东北大学软件学院** | | | | | | | | | |
|  | | | | | | | | | |
| **毕业设计（论文）开题报告** | | | | | | | | | |
|  | | | | | | | | | |
| **设计（论文）题目：** | | | **面向热点网络新闻的文本内容分类软件的设计与实现** | | | | | | |
| **姓 名：** | **康知修** | | | | **学号：** | | | **20144952** | |
| **专 业：** | **软件工程(国际班(英语))** | | | | **班级：** | | | **软英1401** | |
| **住 址：** | **3901 Parkview Ln, 7B. Irvine, CA.** | | | | | | | | |
| **电 话：** | **+1-9495016519** | | | | | | | | |
| **电子邮箱：** | **zhixiuk@uci.edu** | | | | | | | | |
| **实习单位：** | **University of California, Irvine** | | | | | | | | |
| **开题日期：** | **2018** | **年** | | **3** | **月** | | **23** | **日** |  |
| **指导教师评语及改进意见：** | | | | | | | | | |
| **指导教师签字：** | | | | | |  | | | |
| **审定教师签字：** | | | | | |  | | | |
| **日期：** | | | | | | **2018年3月23日** | | | |

# 1.课题提出

## 1.1课题背景、目的与意义

当今世界，其实我们的身边到处都是爬虫的产物，比如各种搜索引擎，他们之所以能为我们提取到这么多结果，其实就是他们爬取了很多信息，并展示给我们。再比如很多爬取各大电商网站进行比价的商业爬虫，极大的为我们的生活提供了便利并节约了金钱。同样，利用爬虫，可以帮助我们在漫天的新闻中爬取我们想要的新闻信息。

与此同时，在信息检索与数据挖掘的过程中，我们需要一种技术来挖掘文章中的关键词，来帮我们实现对爬取的大量数据之后的信息提取和概括，这时，TF-IDF这种通过衡量词频和相应词的权重的方法便可以帮我们解决这一问题，来实现上述要求。

然而，仅仅提取关键词并不能达到方便查阅各类新闻信息的要求，毕竟新闻数量繁杂，即使经过提取关键字，精简后的内容依然数量庞大。这时。我们需要一种分类算法帮我们将新闻进行分类，这时K最近邻算法、余弦相似度算法或SVM算法便可以帮我们解决这一问题。

## 1.2国内外现状

在搜索引擎中，对内容的抓取是消耗很大却又很重要的一部分，这就要求爬虫具有相当的效率同时又需要保持质量。近几年，国内外搜索引擎方面研究均较热，在上述背景下，爬虫技术和人工智能的结合便成为了研究的热点。

同时，在搜索引擎的日益流行中，信息过载问题日益严重，从大量爬取到的信息中找到符合用户要求的信息成为了非常迫切的需求，这时，文本分类就显得至关重要。在文本分类领域中，空间向量模型一直以来就在国内外研究中占有统治地位。其中，TF-IDF因其具有较高的准确率而一直受到相关研究人员和众多应用领域的青睐。TF-IDF通过判断一个词在特定文档中的频率和在文档中出现的范围，来区分此词在区别该文档内容属性方面的能力，这就解决了爬取网页内容后的文本表示的问题。

在这之后， 利用聚类算法对新闻进行分类就成了当务之急，在国内外的研究中，聚类方法大致包括划分聚类、层次聚类等，其中，划分聚类具有较低的时间复杂度和空间复杂度的有点，然而初始化的中心点对结果影响很大；而层次聚类具有质量高的有点，然而具有较高的复杂度。

通过本课题，可以将百度新闻信息的爬取、文本表示和分类实现耦合，从而解决新闻查询中信息过载的问题。

## 1.3研究（设计）内容

1. 通过python爬虫抓取百度新闻列表，提取百度新闻首页上的新闻信息并清洗新闻信息。
2. 通过TF-IDF算法计算每篇文章的向量值。这其中需要三步：

第一步，计算词频。

词频(TF) = 某个词在文章中的出现次数 / 文章的总词数

第二步，计算逆文档频率。

所谓逆文档频率(IDF), 用统计学语言表达，就是在词频的基础上，要对每个词分配一个权重。给予最常见的词如“的”，“是”最小的权重，给予较常见的词较小的权重，给予较少见的词较大的权重，因此，逆文档频率的大小与一个词的常见程度成反比。

而计算逆文档频率时，需要一个语料库，来模拟语言的使用环境。

因此，逆文档频率(IDF) = log(语料库的文档总数 / 包含该词的文档数 + 1)

第三步，计算TF-IDF

TF-IDF = 词频(TF) \* 逆文档频率(IDF)

1. 通过分类算法找出相似文章并分类。 由于通过TF-IDF，已经计算出了每篇文章的向量值，所以计算两篇文章的相似程度，就变成了计算两个向量的相似程度。而我们知道，通过计算两个向量夹角的大小，便可以计算两个向量的相似程度，夹角越小，说明两个向量越相似。

而两个向量之间夹角的公式为;

使用这个公式，我们可以得到两篇文章的词频向量的余弦，进而比较两篇文章的相似程度。比较所有文章之间的相似程度，便可以完成分类。

# 2．设计方案论证（可行性研究）

首先，通过分析页面HTML，分析在百度新闻中要抓取的列表的标题、href, 经过分析后，可以通过正则表达式来实现对新闻内容的清洗。

其次，通过python的scikit-learn包，通过TF-IDF算法可以计算出每篇文章的向量值，继而计算余弦相似度。在这之后，可以实现对不同文章的分类。

# 3．开发环境及系统实现

操作系统: Mac OS X

开发工具: JetBrains PyCharm

# 4．毕业设计进度计划

**表4.1 毕业设计进度计划表**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **序号** | **起止日期** | **任务** | **提交的阶段成果** | **备注** |
| 1 | 3月5日—3月18日（1-2周） | 调研 | 调研报告 |  |
| 2 | 3月19日—3月25日（3-3周） | 查阅文献资料 | 论文综述 |  |
| 3 | 3月26日—4月8日（4-5周） | 系统分析 | 建立系统业务模型，功能模型，数据模型 |  |
| 4 | 4月9日—4月29日（6-8周） | 系统设计 | 功能设计，数据库设计，模块设计 |  |
| 5 | 5月30日—5月13日（9-11周） | 系统实现与调试 | 经过单元测试的源程序 |  |
| 6 | 5月14日—5月20日（12-12周） | 整理论文 | 论文全文 |  |
| 7 | 5月21日—5月27日（13-13周） | 修改、审核论文 | 论文全文 |  |
| 8 | 5月28日—6月4日（14-14周） | 准备答辩 | 各种答辩材料 |  |

# 5．参考文献

1. Christian S. Perone. Machine Learning: Cosine Similarity for Vector Space Models (Part III) [EB/OL].http://blog.christianperone.com/2013/09/machine-learning-cosine-similarity-for-vector-space-models-part-iii/.

2. Juan Ramos. Using tf-idf to determine word relevance in document queries [D]，Rutgers University，2003

3. Ho C W, Robert W P L, Kam F W. Interpreting TF-IDF term weights as making relevance decisions [J], ACM Transactions on Information Systems, 2008, Article No. 13.

4. Sandeep T, Jignesh M. Estimating the selectivity of tf-idf based cosine similarity predicates [J], ACM SIGMOD Record, 2007, Pages 7-12.

5. DL Lee, H Chuang. Document ranking and the vector-space model [J], IEEE Software, 1997, Volume: 14 Issue: 2.

6. Akiko Aizawa. An information-theoretic perspective of tf–idf measures [J], Information Processing & Management, 2003, Pages 5-9.

7. Lailil M, Baharum B. Document Clustering Using Concept Space and Cosine Similarity Measurement [J], Information Processing & Management, 2003, Pages 5-9.

8. Joachims Thorsten. Making large-scale SVM learning practical [J], ACM Transactions on Information Systems, 1998, Pages 14-17.

9. NM Nasrabadi. Pattern recognition and machine learning [J], Journal of electronic imaging, 2007, Pages 3-9.

10. T Joachinms. Text categorization with support vector machines: Learning with many relevant features [J], European conference on machine learning, 1998, Pages 6-10.