毕业设计（论文）任务书

毕业设计（论文）题目：

面向热点网络新闻的文本内容聚类方法的研究和实现

基本内容：

基于 python 语言设计实现网络爬虫，爬取网络热点文本新闻内容，对新闻文本执行数据清洗，训练被爬取内容的表示方法，基于流行的分类或聚类方法实现对爬取内容的分类或聚类，最后以软件的形式予以展示。翻译一篇与毕设内容相关的外文资料，译文汉字字数不少于 4000 字。

毕业设计（论文）专题部分：

题目：

基本内容：

学生接受毕业设计（论文）题目日期

第 1 周

指导教师签字：

2018 年 3 月 9 日

Research and Implementation of Text Content Clustering Method for Breaking News

Abstract

At present, the Internet has become the main platform for people to obtain information and exchange information. In the Internet, the breaking news rankings are the window to explore the most valuable news. By studying breaking news, users can better understand the news topics they care about. At the same time, website builders can also learn about the Internet through analysis results. This could improve the quality of the site.

This paper first analyzes the development status of crawler technology and the research status of text clustering. The commonly used crawler techniques are described at the same time. Methods for obtaining text feature values and methods commonly used in clustering are also has been discussed. During the implementation of the system, the web crawler is improved so that it can properly handle the garbled data when crawling data, thus ensure the quality of the collected breaking news information. For the collected breaking news data, it is formed into a document list and Chinese words are cut for each document to form a format that can be processed in the next step. After this, the TF-IDF algorithm is used to characterize the data after the word segmentation and the matrix data is generated. In this process, some high-frequency phrases with little value are discarded. Finally, the value of the matrix data in the K-means algorithm is tested by selecting the number in a specific range as the clustering number of the K-means algorithm. The point with least error is selected as the k-value of the cluster. Based on the best clustering k-values, the system could learn the clustering rules in the matrix data and store the predicted categories and the original texts.

This paper summarizes data mining, cluster analysis and introduces the structure of each module of the entire system. At the same time, it has conducted in-depth research on modules such as web crawlers, Chinese word segmentation, feature extraction, weight calculation and cluster analysis. Finally, by analyzing the clustering results generated by different k values of this system, the accuracy of the selected k-value structure and the accuracy and stability of the clustering results are verified.

Key words: Web Crawling, Chinese words segmentation, TF-IDF algorithm, K-means clustering algorithm

面向热点网络新闻的文本内容聚类方法的研究和实现

摘 要

目前，互联网成了人们获取信息，交流信息的主要平台。在互联网中，热点新闻排行榜是发掘最具价值新闻的窗口，通过研究第三方对外提供的热点新闻排行榜比如百度风云排行榜，能够使用户更好更快的了解自己关心的新闻话题，同时，网站建设者也能通过分析结果了解到互联网用户的喜好，从而提高网站质量。

本文首先分析了国内外的爬虫技术的发展现状和文本聚类的研究现状，引出了研究热点网络新闻的必要性。同时阐述了目前常用的爬虫技术，并探讨了获取文本特征值的方法和在聚类中常用的方法。在系统的实现过程中，通过对通用的网络爬虫进行改进，使之可以正确处理爬取数据时的乱码问题，保证了收集到的热点新闻信息的质量。对于收集到的热点新闻数据，将其形成文档列表并对每个文档进行中文切词,形成可供下一步进行处理的格式。这之后，通过 TF-IDF 算法对切词后的数据进行特征化，并生成矩阵数据，在这一过程中，剔除了一些在文档中经常出现的没有信息量的高频词组。最后，通过在特定范围内选取数值依次作为 K-means 算法的聚类数，测试矩阵数据在 K-means 算法中的误差，选择误差改善最好的点作为聚类的 k 值。基于最佳的聚类 k 值学习矩阵数据中的类群规律，并将预测的类别与原始文本进行储存。

本文对数据挖掘和聚类分析进行了概括和总结，介绍了整个系统各个模块的结构。同时，对网络爬虫、中文分词、特征提取、权重计算和聚类分析等模块进行了深入的研究。最后通过对这一系统的不同 k 值所产生聚类结果的分析，验证了选取 k 值结构的准确性和聚类结果的准确性与稳定性。

关键词：网络爬虫，中文分词，TF-IDF 算法，K-means 聚类算法