1 项目题目

基于 Nutch 爬虫框架的技术类博客数据抓取与分析

2 项目简介

爬取技术类博客，使用 Nutch 爬虫框架，爬取技术类博客的标题、标签、主要内容。将技术类博客进行汇总分类，如 java, python, Hadoop, spark 等。java 目录又分为算法，多线程，分布式等。最后编写前台界面，将技术类博客的分类结果展示出来，形成一个爬虫结果的比例分析。

2.1 项目优势

Nutch 是一种开源网络搜索引擎，初建时用 Java 编写，可以对 HTML 等文档格式的文件进行抓取、存储、解析等一系列的页面细节处理。Nutch 最初由 Doug Cutting（Lucene 和 HaDoop 的构建者）和 Mike Cafarella 合作构建。Nutch 系统是一个完整的搜索引擎，利用 Nutch 经过若干步骤的简单设置就可以建立自己内部网的搜索引擎，也可以针对 Internet 建立网络搜索引擎。在通用搜索引擎基础上构建主题搜索引擎的方法有三种：一是控制信息采集更新的网站范围，将索引和检索信息限制在特定的几个主题网站之内；二是在通用搜索引擎信息采集的基础上进行文本分类或过滤，提取主题信息进行检索和索引；三是实现主题 Crawler 来控制信息的采集，仅仅采集，索引主题相关的信息。

目前，国内基于 Nutch 构建主题搜索引擎的方法也是以上三种。一是实现主题 Crawler，主要有两种方法：一种是对网络爬虫进行算法改造使其成为主题爬虫；另一种是制定主题词库帮助网络爬虫发现主题资源。二是建立主题词库进行主题过滤，主要有两种方法：一是在爬虫模块与文本解析模块之间制定主题词库对网页信息进行主题过滤；另一种方法是改进分词技术。国外基于 Nutch 构建主题搜索引擎中可上线访问的主题搜索引擎较多，如 Ask About Oil 是关于石油产业的主题搜索引擎。

国内采用主题爬虫和构建主题词库的研究成果较为丰富，但是在控制信息采集范围方面，以往成果多是直接列出一到两个目标网站作为采集范围，并未进行系统的主题资源搜集。

在本项目中主要采取的实现策略是对爬虫抓取的数据进行主题过滤，即在建立特征词库

的基础上实现新的分词策略，对文本数据进行分词，进行词性标注，提取特征词汇，根据提取的特征词对技术类博客进行分类。分词是进行语义分析的前提。我们知道，在英文的行文中，单词之间是以空格作为自然分界符的，而中文只是字、句和段能通过明显的分界符来简单划界，唯独词没有一个形式上的分界符，虽然英文也同样存在短语的划分问题，不过在词这一层上，中文比之英文要复杂的多、困难的多。本文的优势是采用基于上下文无关文法的句法分析。本文采用的训练集是宾夕法尼亚大学的树库 treebank 作为语料库，对句子进行切分，得到了较好的结果。

2.2 系统整体设计

本次项目主要是抓取数据、数据处理、数据分析和挖掘、展示分析和挖掘的成果。根据这些需求确定出了系统的整体设计图和整体功能结构图。

本次项目主要是搭建 Nutch 爬虫框架、抓取数据、数据处理、数据分析和挖掘、展示分析和挖掘的成果。根据这些需求确定处理系统的整体设计图和整体功能结构图。

系统的整体设计图如图 2.1：

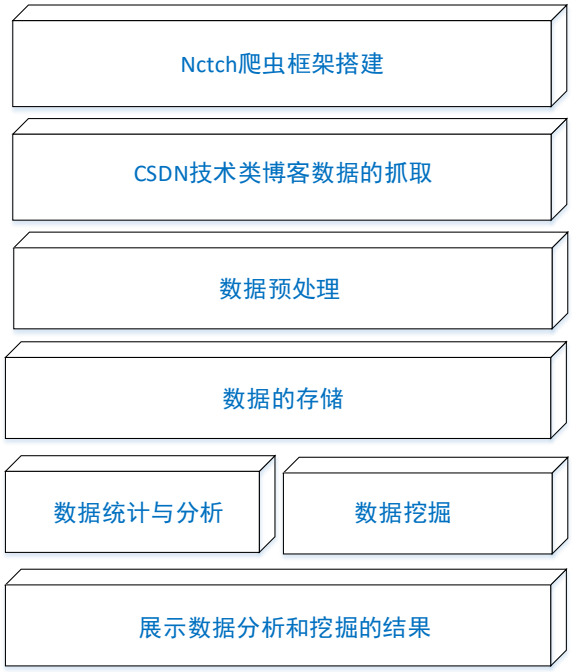

图 2.1　整体设计图

2.3 系统模块划分

本项目分为四个模块：

1.抓取数据

用 Nutch 框架爬取数据，将 CSDN 技术类博客作为主页，设置爬去深度、并发进程数量、每层抓取的最大页面数、存放目录等。将爬取的技术类博客内容存放至指定文件。

2.数据处理和存储

把获取到的数据通过除杂和统一格式等处理成能直接使用的格式并存到文本文件。抓取的数据主要供 Tomcat 显示，通过搭建的 Apache 服务器，通过网页访问，利用 Nutch 框架的集成 readseg 功能，将抓取的数据转换为可读的文本格式。

3.数据处理

对抓取的数据进行分词、词性标注、关键词提取，选取具有技术类特征的名词后进行分类。

4.数据处理结果展示

将数据处理结果以图表形式展示。

2.4 系统开发步骤

在上面划分模块的基础上，按着各个模块的逻辑关系和每个模块内部的依赖关系，确定的开发步骤如下：

1. 配置 Nutch 运行基础环境；
2. 配置和应用 Nutch，安装 Nutch，配置爬虫相关参数，并将爬去的内容存入指定文件夹。
3. 将爬去的内容通过 Tomact 验证是否成功，并将爬去内容格式转化为文本文件。
4. 分词、数据挖掘。
5. 前端界面开发。

3 所采用的技术

3.1 Nutch：

Nutch 是一个开源 Java 实现的搜索引擎。它提供了我们运行自己的搜索引擎所需的全部工具。包括全文搜索和 Web 爬虫。

使用 nutch 的原因：

(1) 透明度：Nutch 是开放源代码的，因此任何人都可以查看他的排序算法是如何工作的。

(2) 扩展性：Nutch 是非常灵活的：他可以被很好的客户订制并集成到你的应用程序中，使用 Nutch 的插件机制， Nutch 可以作为一个搜索不同信息载体的搜索平台。

在 Nutch 中，Crawler 操作的实现是通过一系列子操作的实现来完成的。这些子操作 Nutch 都提供了子命令行可以单独进行调用。下面就是这些子操作的功能描述以及命令行，命令行在括号中。

1. 创建一个新的 WebDb (admin db -create).
2. 将抓取起始 URLs 写入 WebDB 中 (inject).
3. 根据 WebDB 生成 fetchlist 并写入相应的 segment(generate).
4. 根据 fetchlist 中的 URL 抓取网页 (fetch).
5. 根据抓取网页更新 WebDb (updatedb).
6. 循环进行 3－5 步直至预先设定的抓取深度。
7. 根据 WebDB 得到的网页评分和 links 更新 segments (updatesegs).
8. 对所抓取的网页进行索引(index).
9. 在索引中丢弃有重复内容的网页和重复的 URLs (dedup).
10. 将 segments 中的索引进行合并生成用于检索的最终 index(merge).

3.2 Stanford Postagger

Stanford Postagger，由斯坦福大学自然语言处理小组开发的词性标注工具，最新发布于 2006 年，该词性标注工具主要分为标注、测试和训练三个模式。本文所用到其主要功能为对文本中的单个词汇进行词性判断并在每个单词后进行词性标注，例如形容词（/JJ）、副词

（/RB）、动词（/VB）等词性，以方便开发人员对相关词汇进行筛选，减小数据处理时间，提高工作效率。该开源词性标注词典对英文单词的标注准确率达到 97%。除此之外，该词性标注工具还可对中文、德文以及法文等进行词性标注。