MapReduce 课程设计之

邮件自动分类

1 课程设计目标

本课程设计的目标是通过 MapReduce 编程来实现邮件的自动分类,通过本课程设计的学习,可以体会如何使用 MapReduce 完成一个综合性的数据挖掘任务,包括全流程的数据预处理、样本分类、样本预测等。

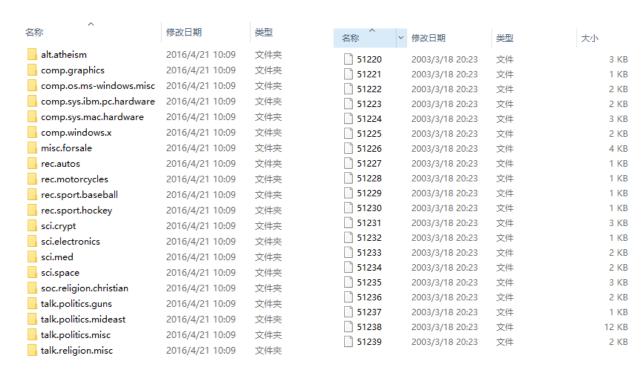


图 1 邮件分类。左边图是邮件类别,每个文件夹代表一个类别,右边图是每个类别下的邮件文本数据。

2 学习技能

通过本课程设计,可以熟悉和掌握以下 MapReduce 编程技能:

- (1) 在 Hadoop 中使用第三方的 Jar 包来辅助分析;
- (2) 掌握 MapReduce 算法设计:

- 1) 文本特征选择算法;
- 2) 文本特征权重计算算法;
- 3) 文本分类算法;

3 任务描述

在正常工作中我们会收到大量的邮件,不同的邮件包含不同的主题特征,本课程设计任务是通过 MapReduce 技术实现邮件的自动分类。邮件分类实质上和文本分类一样,具体包括如下的若干任 务,这些任务组合起来,就构成了一个完整的邮件分类流程。

任务 1 特征选择

本任务的主要工作是对原始的邮件文本中进行特征选择,选择出能够表征邮件主题的特征词,为后续的文本分类做准备。停词表一般用于去除文本分类无关的一些冗余词语,如数字、标点符号等。

输入输出

数据输入: 1.邮件训练样本全集(未分词); 2.停词表。

数据输出:邮件文本特征

示例

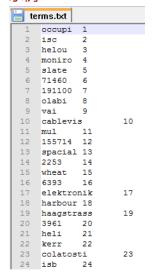


图 2 特征词

任务 2 特征向量权重计算

本任务主要是基于任务一输出的特征词向量,计算出每个邮件样本的特征词权重,特征词权重用来刻画特征词在描述此文本内容时所起的重要程度。

输入输出

输入: 1、任务 1 的输出; 2、邮件样本数据

输出:每个邮件类别中邮件样本的特征向量。

示例

任务 3 文本分类

当获取了每个邮件样本的特征向量后,剩下的就需要使用机器学习中分类算法来实现邮件的分类,具体采用哪种合适分类算法,同学们自定。(注:请至少使用两种分类算法进行比较分析)

任务 4 样本预测

根据任务三训练得到了分类模型,对邮件测试样本进行预测,输出预测结果,并统计预测的正确率,即预测正确的样本数与测试样本总数的比值,预测的正确率则反应了分类模型的质量。

参考资料

1. 文本分类概述 http://blog.csdn.net/chl033/article/details/4733647

4 提交材料

请各位同学提交如下材料。

- (1) 程序源代码,要求提供包含完整目录结构的 src 代码包,并且提供编译方法说明。
- (2) 程序可执行 jar 包以及 jar 包的执行方式。本题目的运行环境在 hadoop-2.7、jdk-1.7 或以上环境下,必须采用 MapReduce 编程模型。
- (3) 程序设计报告。报告内容包括程序设计的主要流程、程序采用的主要算法、进行的优化工作、优化取得的效果、程序的性能分析以及程序运行截图等。需要在课程设计报告中标明小组成员的分工情况。

实现指导

针对在课程设计的实现中可能遇到的问题,在这里整理了一份 FAQ,供参考。

关于任务 1

Q: 如何对原始邮件文本进行分词?

A: 可以使用 Lucene 的 Standard Analyzer 分词器。

Q:如何在 Hadoop 的 Mapper 或 Reducer 中使用第三方的 Jar 包?

A: 在打 Jar 包的时候,将第三方的依赖包也同步打入生成的 Jar 包。

- Eclipse:具体做法可以参考 Stack Overflow 上的问答。
- IntelliJ IDEA:在 IntelliJ IDEA 中,在创建 Artifacts 的时候,在 Output Layout 选项卡中,从 右侧的 Available Elements 区域中,选中所要引入的第三方 Jar 包,在右键菜单中选择 Extract Into Output Root。

关于任务 2

Q: 如何对邮件进行特征选择?

A: 可参考

http://blog.sina.com.cn/s/blog_4d1f33470100scz7.html

http://blog.sina.com.cn/s/blog_6622f5c30101datu.html

Q:如何计算文本特征词的权重?

A: 可考虑采取 TF-IDF 的计算方式,不过,你也可以选择其他更为精确的计算方式。

关于任务3

Q:如何进行文本分类?

A: 可考虑采用 KNN, 朴素贝叶斯或相应的改进算法, 也可以选择其他分类算法。

关于任务 4

Q:如何对预测结果进行统计分析?

A: 可考虑采用 Python 的 matplotlib 或 seaborn 绘图库进行可视化分析。