

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。

名称	修改日期	类型	名称	修改日期	类型	大小
alt.atheism	2016/4/21 10:09	文件夹	51220	2003/3/18 20:23	文件	3 KB
comp.graphics	2016/4/21 10:09	文件夹	51221	2003/3/18 20:23	文件	1 KB
comp.os.ms-windows.misc	2016/4/21 10:09	文件夹	51222	2003/3/18 20:23	文件	2 KB
comp.sys.ibm.pc.hardware	2016/4/21 10:09	文件夹	51223	2003/3/18 20:23	文件	2 KB
comp.sys.mac.hardware	2016/4/21 10:09	文件夹	51224	2003/3/18 20:23	文件	3 KB
comp.windows.x	2016/4/21 10:09	文件夹	51225	2003/3/18 20:23	文件	2 KB
misc.forsale	2016/4/21 10:09	文件夹	51226	2003/3/18 20:23	文件	4 KB
rec.autos	2016/4/21 10:09	文件夹	51227	2003/3/18 20:23	文件	1 KB
rec.motorcycles	2016/4/21 10:09	文件夹	51228	2003/3/18 20:23	文件	1 KB
rec.sport.baseball	2016/4/21 10:09	文件夹	51229	2003/3/18 20:23	文件	1 KB
rec.sport.hockey	2016/4/21 10:09	文件夹	51230	2003/3/18 20:23	文件	1 KB
sci.crypt	2016/4/21 10:09	文件夹	51231	2003/3/18 20:23	文件	3 KB
sci.electronics	2016/4/21 10:09	文件夹	51232	2003/3/18 20:23	文件	1 KB
sci.med	2016/4/21 10:09	文件夹	51233	2003/3/18 20:23	文件	2 KB
sci.space	2016/4/21 10:09	文件夹	51234	2003/3/18 20:23	文件	2 KB
soc.religion.christian	2016/4/21 10:09	文件夹	51235	2003/3/18 20:23	文件	3 KB
talk.politics.guns	2016/4/21 10:09	文件夹	51236	2003/3/18 20:23	文件	2 KB
talk.politics.mideast	2016/4/21 10:09	文件夹	51237	2003/3/18 20:23	文件	1 KB
talk.politics.misc	2016/4/21 10:09	文件夹	51238	2003/3/18 20:23	文件	12 KB
talk.religion.misc	2016/4/21 10:09	文件夹	51239	2003/3/18 20:23	文件	2 KB

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(数据来自经典的 20 Newsgroups 数据集，)。

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：
 - a) 文本特征选择算法；

- b) 文本特征权重计算算法；
- c) 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 and 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，常见的文本特征选择方法包括卡方检验、信息增益等。

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



1	occupi	1
2	isc	2
3	helou	3
4	moniro	4
5	slate	5
6	71460	6
7	191100	7
8	olabi	8
9	vai	9
10	cablevis	10
11	mul	11
12	155714	12
13	spacial	13
14	2253	14
15	wheat	15
16	6393	16
17	elektronik	17
18	harbour	18
19	haagstrass	19
20	3961	20
21	heli	21
22	kerr	22
23	colatosti	23
24	isb	24

图 2 特征词

任务 2 特征向量权重计算

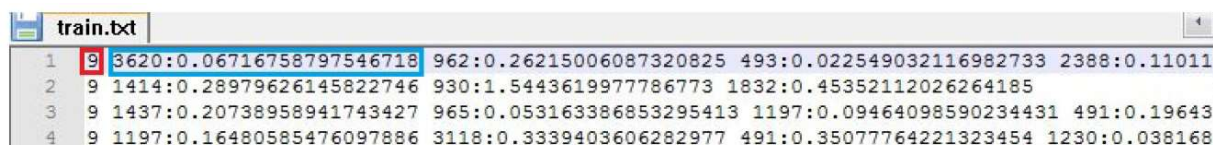
本任务主要是基于任务一输出的特征词向量，计算出每个邮件样本的特征词权重（如TF、TF-IDF），特征词权重用来刻画特征词在描述此文本内容时所起的重要程度。

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



1	9	3620:0.06716758797546718	962:0.26215006087320825	493:0.022549032116982733	2388:0.11011
2	9	1414:0.28979626145822746	930:1.5443619977786773	1832:0.45352112026264185	
3	9	1437:0.20738958941743427	965:0.053163386853295413	1197:0.09464098590234431	491:0.19643
4	9	1197:0.16480585476097886	3118:0.3339403606282977	491:0.35077764221323454	1230:0.038168

图 2 样本特征向量

上述示例每一行代表一个邮件样本，其中第一列表示类别编号，其余每一列是词及其权重，使用冒号分隔，如“3620:0.06716758797546718”表示编号为3620的词，对应的特征值为0.06716758797546718。如果特征有N个，那么每条记录就对应着一个N维向量。

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。

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。

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。

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，你也选择其他更为精确的计算方式。