基于 Hadoop MapReduce 并行计算框架的邮件自动分类

曹佳涵　白家杨　刘笑今
2019st04 小组

摘　要

本实验的目标是通过 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。我们使用 Hadoop MapReduce 并行计算框架对原始邮件文本进行特征选择、特征向量权重计算、文本分类和样本预测任务，从而搭建起一个完整的文本分类处理流程。我们采用了不同的特征提取方法和分类算法并进行比较和分析。此外，我们使用 Spark 同样完成了实验。

关键词： 邮件分类，并行计算，MapReduce，Spark

1　引入

此模板是基于 LaTeX 的标准文类 article 设计，也即意味着你可以把 article 文类的选项传递给本模板，比如 `a4paper`，`10pt` 等等（推荐使用 `11pt`）。本模板支持 PDFLaTeX 和 XeLaTeX[1] 两种编译方式。

数学字体的效果如下：

$$(a+3b)^n = \sum_{k=0}^{n} C_n^k a^{n-k}(3b)^k \tag{1}$$

1.1　全局选项

我在这个模板中定义了一个语言选项 `lang`，可以选择英文模式 `lang=en`（默认）或者中文模式 `lang=cn`。当选择中文模式时，图表的标题引导词以及参考文献，定理引导词等信息会变成中文。你可以通过下面两种方式来选择语言模式：

```
\documentclass[lang=cn]{elegantpaper} % or
\documentclass{cn}{elegantpaper}
```

[1] 中文字体均使用 ctex 包设置。

1.2 自定义命令

在此模板中，并没有修改任何默认的命令或者环境，所以，你可以在此模板使用原来的命令和环境。另外，我自定义了 3 个命令：

1. \email：创建邮箱地址的链接；
2. \figref：用法和 \ref 类似，但是会在插图的标题前添加 < **图 n**> ；
3. \tabref：用法和 \ref 类似，但是会在表格的标题前添加 < **表 n**>；
4. \keywords：为摘要环境添加关键词。

1.3 列表环境

你可以使用列表环境（itemize、enumerate、description），示例如下：

```
\begin{itemize}
   \item Routing and resource discovery;
   \item Resilient and scalable networks;
   \item Distributed storage and search.
\end{itemize}
```

- Routing and resource discovery;
- Resilient and scalable networks;
- Distributed storage and search.

1.4 插图

插图的命令和以前一样，也是使用 figure 环境。**图 1** 显示了插图的效果。你可以把你的图放到当前工作目录的如下子目录下 (./image/, ./img/, ./figure/, ./fig/)。

```
\begin{figure}[htbp]
  \centering
  \includegraphics[width=0.6\textwidth]{scatter.pdf}
  \caption{Scatter Plot Example \label{fig:scatter}}
\end{figure}
```

1.5 表格

我强烈建议你使用 booktabs 宏包，这个宏包有三个命令 \toprule、\midrule 和 \bottomrule 能方便你制作三线表。**表 1** 是一个示例：

```
\begin{table}[htbp]
  \small
  \centering
  \caption{Auto MPG and Price \label{tab:reg}}
    \begin{tabular}{lcc}
    \toprule
                    &       (1)          &         (2)       \\
    \midrule
```

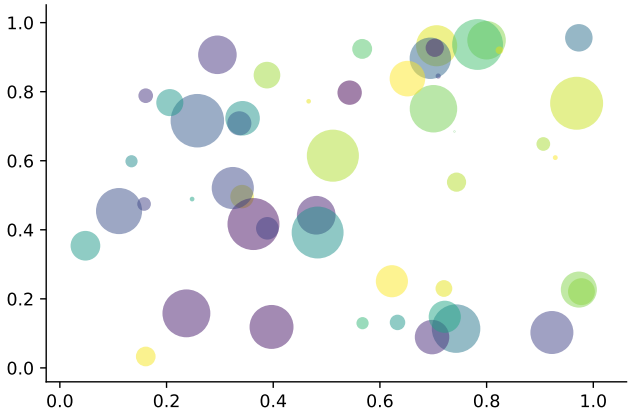

图 1: Scatter Plot Example

```
    mpg             &     -238.90***      &       -49.51      \\
                    &      (53.08)        &       (86.16)     \\
    weight          &                     &       1.75***     \\
                    &                     &       (0.641)     \\
    constant        &      11,253***      &        1,946      \\
                    &      (1,171)        &       (3,597)     \\
    obs             &         74          &          74       \\
    $R^2$           &       0.220         &       0.293       \\
    \bottomrule
    \multicolumn{3}{l}{\scriptsize Standard errors in parentheses} \\
    \multicolumn{3}{l}{\scriptsize *** p<0.01, ** p<0.05, * p<0.1} \\
    \end{tabular}%
\end{table}%
```

表 1: Auto MPG and Price

	(1)	(2)
mpg	-238.90***	-49.51
	(53.08)	(86.16)
weight		1.75***
		(0.641)
constant	11,253***	1,946
	(1,171)	(3,597)
obs	74	74
R^2	0.220	0.293

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

1.6 参考文献

此模板使用了 BibTEX 来生成参考文献，默认使用的文献样式（bib style）是 GB/T 7714-2015[2]。参考文献示例：？使用了中国一个大型的 P2P 平台（人人贷）的数据来检验男性投资者和女性投资者在投资表现上是否有显著差异。

你可以在谷歌学术，Mendeley，Endnote 中获得文献条目（bib item），然后把它们添加到 wpref.bib 中。在文中引用的时候，引用它们的键值（bib key）即可。注意需要在编译的过程中添加 BibTEX 编译。如果你想在参考文献中添加未引用的文献（部分或者全部），可以使用

```
\nocite{EINAV2010, Havrylchyk2018} % add the two reference.
\nocite{*}   % add all the reference in the bib file.
```

如果你想修改参考文献的样式（比如改为 aer），你可以在导言区将下面代码注释掉。

```
\usepackage[authoryear]{gbt7714}
```

并且文档末尾添加

```
\bibliographystyle{aer}
```

2 数据分析

共有 20 个类别，20000 个文件.....

3 MapReduce 处理流程

3.1 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。

对于输入的未分词的邮件训练样本全集和停词表，我们需要输出全局邮件文本特征，并对它们进行相应的编号，此外，对于训练数据集的目录，将目录名（即文本类别）转换为相应的类别序号。

对于该步骤，我们采用了两种计算方法，分别为非并行化和并行化计算方法。

3.1.1 非并行化计算方法

非并行化计算方法主要思路是顺序执行，首先读取 20_newsgroup 文件夹下的子文件夹，将子文件夹名（即类别名）转化为类编号并进行存储。

[2]通过调用 gbt7714 宏包

3.1.2 并行化计算方法

4 示例

在这部分，我提供一个示例文档：

```
\documentclass[lang=cn]{elegantpaper}

% title information
\title{A Working Paper Example}
\author{ddswhu}
\institute{Elegant\LaTeX{} Group}
\version{1.00}
\date{\today}

\begin{document}

\maketitle

\begin{abstract}
Your abstract goes here.
\keywords{keyword1, keyword2}
\end{abstract}

\section{Introduction}
The content of introduction section.

\section{Conclusion}
The content of conclusion section.

% include the noncited reference
\nocite{ref1, ref2}
\bibliographystyle{aer}
\bibliography{wpref}
\end{document}
```