

智能检索

实战

王家宝 李阳苗 壮

编著

内容实用

帮助读者实践对文本数据和图像数据的智能检索

技术新颖

涵盖人工智能领域最新的深度学习技术

案例翔实

描述详细，分析深入，提供完整的案例源代码

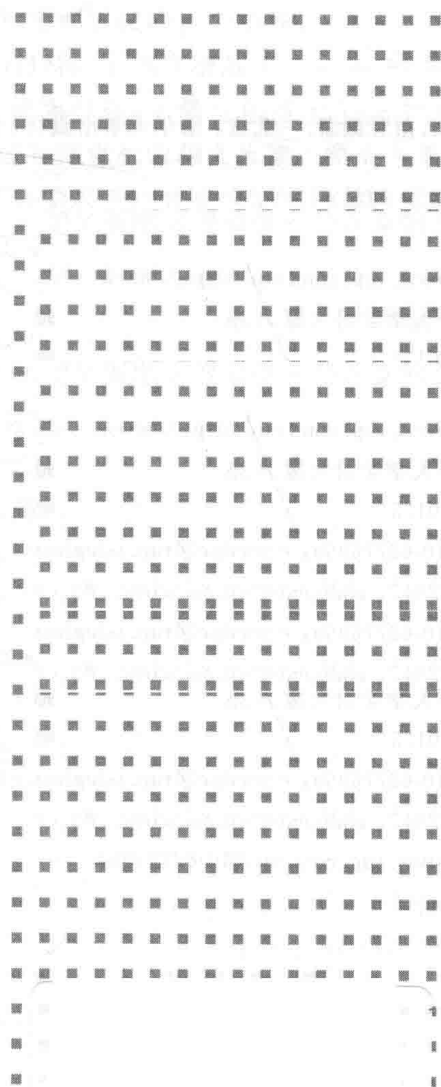


清华大学出版社

王家宝 李阳 苗壮

编著

智能检索实战



清华大学出版社
北京

内 容 简 介

智能检索是适应大数据和人工智能迅速发展的信息检索新方式。本书分为两个部分。第一部分以工程应用为目标,介绍了文本的本地检索和网络检索,基于全局特征、局部特征的图像检索,以及定制图像检索新特征和图像检索相关反馈;第二部分以技术研究为目标,介绍了利用深度学习特征提高检索精度,利用哈希特征提高检索速度,以及跨模态的深度哈希图文检索技术。全书从指导实践出发,附有所有实战源代码,为读者提供了联系实际、直接可用的检索系统和检索技术。

本书是信息检索及相关课程的教学参考书,适用于高等院校信息检索专业的大学生和研究生,也可供从事信息检索相关专业的研发人员参考。

本书封面贴有清华大学出版社防伪标签,无标签者不得销售。

版权所有,侵权必究。侵权举报电话:010-62782989 13701121933

图书在版编目(CIP)数据

智能检索实战/王家宝,李阳,苗壮编著. —北京:清华大学出版社,2018

ISBN 978-7-302-50518-1

I. ①智… II. ①王… ②李… ③苗… III. ①机器检索 IV. ①G254.929.9

中国版本图书馆 CIP 数据核字(2018)第 139433 号

责任编辑:袁勤勇

封面设计:常雪影

责任校对:时翠兰

责任印制:董 瑾

出版发行:清华大学出版社

网 址: <http://www.tup.com.cn>, <http://www.wqbook.com>

地 址:北京清华大学学研大厦 A 座 邮 编:100084

社 总 机:010-62770175 邮 购:010-62786544

投稿与读者服务:010-62776969, c-service@tup.tsinghua.edu.cn

质量反馈:010-62772015, zhiliang@tup.tsinghua.edu.cn

课件下载: <http://www.tup.com.cn>, 010-62795954

印 装 者:三河市铭诚印务有限公司

经 销:全国新华书店

开 本:185mm×260mm

印 张:14.25

字 数:328 千字

版 次:2018 年 10 月第 1 版

印 次:2018 年 10 月第 1 次印刷

定 价:39.00 元

产品编号:077678-01

前言

本书分为两个部分。第一部分以 Java 语言和 Web 应用为主体,实现文本和图像的搜索引擎,详细论述构建文本、图像搜索系统的实践过程;第二部分以 MATLAB 语言和深度学习技术为主,实现图像的深度特征、哈希特征检索,以及跨模态的图文检索技术,详细论述以研究为目标,设计和实现算法的各个步骤。

第 1~3 章以文本搜索为目标,主要介绍基于 Lucene 的本地检索和基于 Solr 的网络检索。第 4~7 章以图像搜索为目标,主要介绍基于 Lire 的全局特征检索、局部特征检索,以及定制图像检索新特征和图像检索相关反馈。第 8、9 章以提高检索精度为目标,介绍利用预训练卷积特征和迁移学习卷积特征来完成图像检索的实践过程。第 10、11 章以提高检索速度为目标,介绍经典的局部敏感哈希技术和最新的深度哈希技术的实践过程。第 12 章以跨模态检索为目标,介绍跨模态的深度哈希图文检索实践过程。

本书是集体智慧的结晶,由王家宝、李阳、苗壮负责撰稿。另外,感谢李航、徐玉龙、张耿宁等同学为本书的撰写工作付出的辛勤工作,这些同学参与了本书的部分实验和素材整理工作,在此深表感谢。全书突出实践性,希望为读者提供联系实际、直接可用的检索系统或检索技术。全书内容经过多次讨论和修改才定稿,尽管每项内容都经过了细心检查,但难免存在缺点和遗漏,希望广大读者批评指正。

随书提供了所有实战源代码,读者可结合本书学习,应用、研究、开发面向特定需求的智能检索系统和技术。本书对于从事图文搜索研究与开发的技术人员有一定的指导和借鉴意义。

作 者

2018 年 1 月

目 录

- 第 1 章 搜索引擎初探..... 1
 - 1.1 Eclipse 开发环境 1
 - 1.1.1 Eclipse 简介 1
 - 1.1.2 JDK 与 JRE 1
 - 1.1.3 JDK 和 Eclipse 的安装 2
 - 1.1.4 用 Eclipse 开发 HelloWorld 项目 7
 - 1.2 Lucene 环境配置 11
 - 1.2.1 Lucene 简介 11
 - 1.2.2 Lucene 的安装 13
 - 1.3 Lucene 索引与检索示例 14
 - 1.3.1 索引和检索的概念 14
 - 1.3.2 一个简单的搜索应用程序 14
 - 1.4 小结..... 17
 - 参考文献 17
- 第 2 章 基于 Lucene 的本地检索 18
 - 2.1 Lucene 索引简介 18
 - 2.1.1 Lucene 索引 18
 - 2.1.2 Lucene 索引文件 20
 - 2.2 Lucene 的索引接口 21
 - 2.2.1 Directory 类 21
 - 2.2.2 Analyzer 类 21
 - 2.2.3 Document 类与 Field 类 21
 - 2.2.4 IndexWriter 类..... 22
 - 2.3 Lucene 的检索接口 23
 - 2.3.1 IndexSearcher 类 23
 - 2.3.2 Term 类 23
 - 2.3.3 Query 类 24

2.3.4	QueryParser 类	24
2.3.5	Sort 类和 Hits 类	24
2.4	中文分词	24
2.5	基于 Lucene 的本地检索实战	27
2.5.1	文本数据准备	27
2.5.2	Lucene 本地索引	27
2.5.3	Lucene 本地检索	31
2.6	索引可视化工具 Luke	32
2.7	小结	35
	参考文献	35
第 3 章	基于 Solr 的网络检索	36
3.1	Solr 简介	36
3.2	Solr 配置和使用	37
3.2.1	Tomcat 安装	37
3.2.2	安装并配置 Solr 至 Tomcat	38
3.2.3	新建并配置 core	42
3.2.4	配置和使用中文分词	43
3.3	基于 Solr 的网络检索实战	47
3.3.1	数据准备	47
3.3.2	Solr 网络索引	48
3.3.3	Solr 网络检索	51
3.4	小结	52
	参考文献	53
第 4 章	基于 Lire 全局特征的图像检索	54
4.1	Lire 简介	54
4.1.1	Lire 库导入	54
4.1.2	Lire 库分析	56
4.2	Lire 全局特征索引	57
4.2.1	Lire 全局特征索引方法	57
4.2.2	Lire 全局特征索引实现	57
4.3	Lire 全局特征检索	60
4.3.1	Lire 全局特征检索方法	60
4.3.2	Lire 全局特征检索实现	61
4.4	Caltech256 数据测试	63
4.4.1	测试数据和基本思路	63

4.4.2 测试实现	63
4.5 基于 Lire 全局特征的图像检索实战	66
4.5.1 主体框架构建	66
4.5.2 外部依赖包导入	67
4.5.3 搜索引擎界面实现	69
4.5.4 搜索引擎后台实现	71
4.5.5 搜索引擎配置和部署	77
4.5.6 搜索引擎操作和效果	78
4.6 小结	79
参考文献	80

第 5 章 基于 Lire 局部特征的图像检索

5.1 词袋模型简介	81
5.2 Lire 局部特征索引	83
5.2.1 Lire 局部特征索引方法	83
5.2.2 Lire 局部特征索引实现	84
5.3 Lire 局部特征检索	85
5.3.1 Lire 局部特征检索方法	85
5.3.2 Lire 局部特征检索实现	86
5.4 Lire 中 SIFT 特征的改进	87
5.5 Caltech256 数据测试	92
5.6 基于 Lire 局部特征的图像检索实战	93
5.6.1 主体框架构建	93
5.6.2 搜索引擎界面实现	94
5.6.3 搜索引擎后台实现	94
5.6.4 搜索引擎配置和部署	98
5.6.5 搜索引擎操作和效果	98
5.7 小结	99
参考文献	99

第 6 章 面向 Lire 定制图像检索新特征

6.1 Lire 特征类的结构	101
6.2 Lire 颜色布局特征	102
6.3 添加新的图像特征	105
6.4 矩特征的索引和检索	110
6.5 小结	112
参考文献	112

第7章 面向 Lire 定制图像检索的相关反馈	114
7.1 基于 SVM 的相关反馈原理	114
7.2 相关反馈实战	115
7.2.1 主体框架构建	115
7.2.2 外部依赖包导入	115
7.2.3 搜索引擎界面实现	116
7.2.4 搜索引擎后台实现	117
7.2.5 搜索引擎配置和部署	118
7.2.6 搜索引擎操作和效果	119
7.3 关键代码解析	120
7.3.1 生成索引阶段	120
7.3.2 查找检索阶段	121
7.3.3 反馈检索阶段	123
7.4 小结	127
参考文献	127
第8章 基于预训练卷积特征的图像检索	128
8.1 卷积神经网络技术	128
8.1.1 卷积	129
8.1.2 池化	130
8.1.3 ReLU	130
8.1.4 全连接	130
8.2 卷积神经网络模型简介	131
8.2.1 AlexNet 网络模型	131
8.2.2 VGGNet 网络模型	132
8.2.3 ResNet 网络模型	133
8.3 基于预训练卷积特征的图像检索实战	135
8.3.1 环境配置	135
8.3.2 数据准备	137
8.3.3 预训练网络特征提取	137
8.3.4 预训练网络检索评测	140
8.3.5 预训练网络检索效果	144
8.4 小结	146
参考文献	146

第 9 章 基于迁移学习卷积特征的图像检索	147
9.1 迁移学习技术	147
9.2 迁移学习方法简介	148
9.2.1 迁移学习的定义与分类	148
9.2.2 深度迁移学习	148
9.2.3 卷积神经网络的迁移	149
9.2.4 迁移学习抑制过拟合	150
9.3 基于迁移学习卷积特征的图像检索实战	151
9.3.1 迁移学习网络设计	152
9.3.2 数据准备	152
9.3.3 迁移学习网络构建	155
9.3.4 迁移学习网络训练	157
9.3.5 迁移学习网络检索评测	159
9.3.6 迁移学习网络检索效果	159
9.4 小结	162
参考文献	162
第 10 章 基于局部敏感哈希的图像检索	163
10.1 局部敏感哈希技术	163
10.1.1 哈希简介	163
10.1.2 近似最近邻搜索问题	164
10.2 局部敏感哈希方法简介	165
10.2.1 LSH 算法	166
10.2.2 E2LSH 算法	167
10.3 基于局部敏感哈希的图像检索实战	168
10.3.1 局部敏感哈希软件包	168
10.3.2 局部敏感哈希函数功能介绍	168
10.3.3 局部敏感哈希测试数据集	170
10.3.4 局部敏感哈希索引建立	170
10.3.5 局部敏感哈希索引分析	173
10.3.6 局部敏感哈希检索效果	178
10.4 小结	181
参考文献	181
第 11 章 基于深度哈希的图像检索	182
11.1 深度哈希技术	182

11.2	深度哈希方法简介	183
11.3	基于深度哈希的图像检索实战	185
11.3.1	深度哈希网络设计	185
11.3.2	深度哈希网络构建	187
11.3.3	深度哈希网络训练	189
11.3.4	深度哈希网络检索评测	190
11.3.5	深度哈希网络检索效果	192
11.4	小结	193
	参考文献	193
第 12 章	跨模态的深度哈希图文检索	195
12.1	跨模态检索技术	195
12.2	跨模态检索方法简介	196
12.2.1	基于典型相关性分析的跨模态检索	196
12.2.2	基于深度学习的跨模态检索	197
12.3	跨模态的深度哈希图文检索实战	199
12.3.1	跨模态哈希网络设计	199
12.3.2	数据准备	200
12.3.3	跨模态哈希网络构建	201
12.3.4	跨模态哈希网络训练	205
12.3.5	跨模态哈希网络特征提取	207
12.3.6	跨模态哈希网络检索评测	209
12.4	小结	211
	参考文献	212
附录 A	信息检索评价指标	213
A.1	召回率与准确率	213
A.2	F1 分数指标	213
A.3	mAP 指标	214
A.4	CMC 曲线	215

第 1 章

搜索引擎初探

本章主要介绍 Eclipse 开发环境、Lucene 开发包等相关软件及其安装配置,目的是使初学者能够使用 Eclipse 开发环境编写调用 Lucene 索引和检索基本功能的代码。本章的重点是体验和理解 Eclipse 开发环境、Lucene 的索引和检索过程,难点是 Java 环境的安装和配置。

开发环境: Windows 7 操作系统

软件及开发包:

- jdk-8u131-windows-x64.exe
- eclipse-jee-mars-1-win32-x86_64.zip
- lucene-4.2.1.tgz

1.1 Eclipse 开发环境

本节主要介绍 Eclipse 开发环境的安装和配置过程,以及利用 Eclipse 开发一个 HelloWorld 项目。

1.1.1 Eclipse 简介

Eclipse 是一个由个人和组织协作组成的、商业友好的开源软件社区。Eclipse 项目关注于构建一个开放的开发平台,该平台涵盖可扩展框架、工具和运行环境,以及部署和管理软件^[1]。2001 年 11 月,IBM 创建了 Eclipse 项目,并得到多个软件供应商联盟的支持,包括 Borland、MERANT、Rational Software、Red Hat、SuSE、TogetherSoft 和 Webgains 等。到 2003 年年底,该联盟的成员已经超过 80 名。2004 年 1 月,Eclipse 基金会作为一个独立的非营利公司而成立,充当 Eclipse 社区的管理人员,帮助开发开放源码社区、互补产品和服务的生态系统。

Eclipse 作为著名的跨平台集成开发环境(Integrated Development Environment, IDE),其本身只是一个框架和一组服务,通过插件、组件构建强大的开发环境。Eclipse 最初主要是用来做 Java 语言开发。随着其插件服务的不断发展,通过安装不同的插件可以支持不同的计算机语言,如 C/C++、Python、PHP 等。

1.1.2 JDK 与 JRE

JDK(Java Development Kit)是 Java 开发工具包,是用于开发和测试 Java 程序的工具。JRE(Java Runtime Environment)是 Java 运行环境,通常会集成在 JDK 中。简单地

说：JDK 用于开发程序，JRE 用于运行程序。

1.1.3 JDK 和 Eclipse 的安装

本小节主要介绍 JDK 和 Eclipse 的安装和配置过程。

1. 安装 JDK

下面以安装 jdk-8u131-windows-x64.exe^① 为例进行介绍。

(1) 安装 JDK 开发包

双击运行安装程序，单击“下一步”按钮直到安装结束，最后单击“关闭”按钮即可。安装耗时约几分钟，内容包含 JDK 和 JRE 两部分。安装过程的主要界面如图 1-1 和图 1-2 所示，其中图 1-1 用于选择 JDK 安装路径，图 1-2 用于选择 JRE 安装路径。

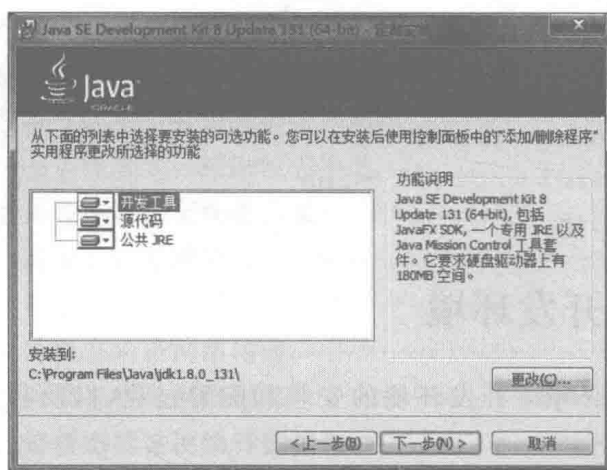


图 1-1 安装 JDK 时选择安装路径界面



图 1-2 安装 JRE 时选择安装路径界面

① <http://www.oracle.com/technetwork/java/javase/downloads/>。

(2) 配置系统环境变量

Java 程序在运行时调用相关软件包和资源,故需要配置环境变量 Path 和 ClassPath,以告知 Java 程序获取相关软件包和资源的路径。

右击“我的电脑”图标,在弹出的菜单中单击“属性”菜单项,出现如图 1-3 所示的控制面板关于系统信息的界面。

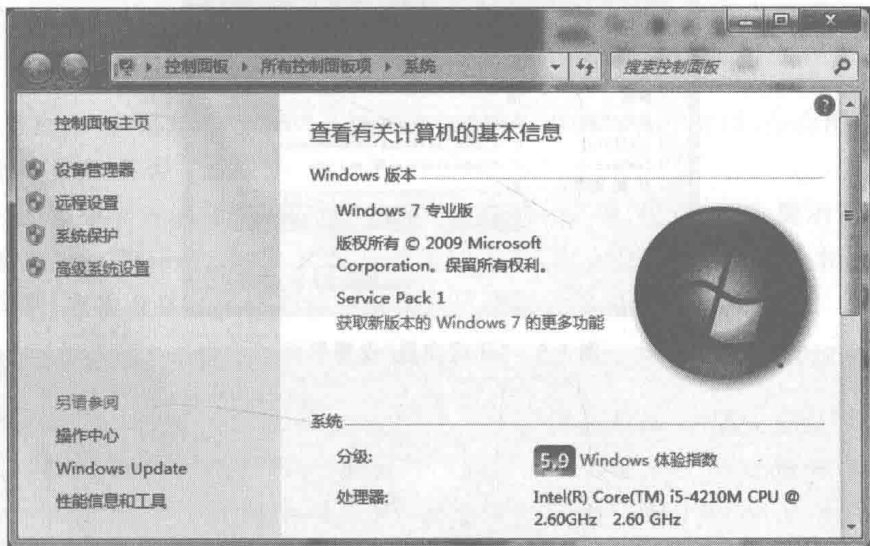


图 1-3 控制面板关于系统信息的界面

单击左侧的“高级系统设置”选项,弹出“系统属性”对话框并自动进入如图 1-4 所示的“高级”选项卡。

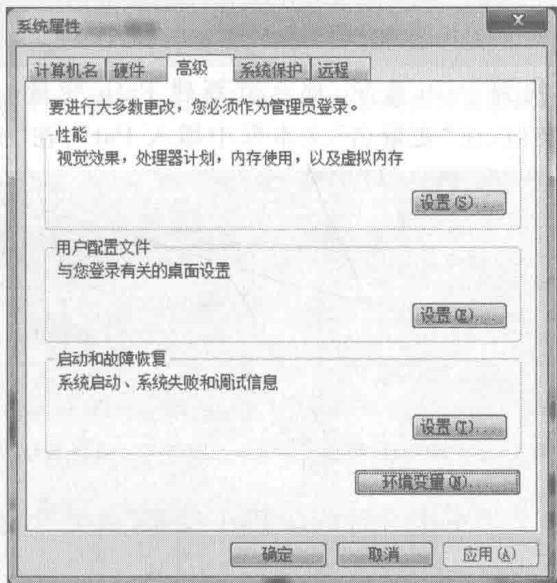


图 1-4 系统属性中的“高级”选项卡界面

单击下方的“环境变量”按钮,进入如图 1-5 所示的“环境变量”设置界面。



图 1-5 “环境变量”设置界面

关于环境变量的配置,具体过程如下:

首先,在“系统变量”列表框下单击“新建”按钮弹出“新建系统变量”对话框。在“变量名”文本框中填入 JAVA_HOME,在“变量值”文本框中填入 JDK 安装路径,如图 1-6 所示。

单击“确定”按钮即可完成 JAVA_HOME 环境变量的添加。

然后,在“系统变量”列表框中找到 Path 变量并双击弹出如图 1-7 所示的“编辑系统变量”对话框,其中“变量名”文本框中的字符串 Path 不用修改,在“变量值”文本框的最后添加“;%JAVA_HOME%\bin;%JAVA_HOME%\jre\bin”。切记要用“;”将它与前一个值分开,如“\tools;%JAVA_HOME%\bin;%JAVA_HOME%\jre\bin”。如果在“系统变量”列表框中没有找到 Path 变量,那么就新建 Path 变量;和上面新建系统变量 JAVA_HOME 的方法类似,在“变量名”文本框中填入 Path,在“变量值”文本框中填入“%JAVA_HOME%\bin;%JAVA_HOME%\jre\bin”。

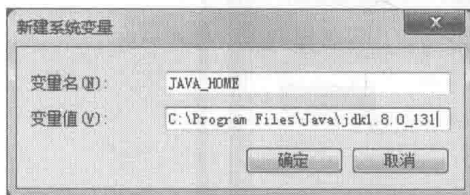


图 1-6 新建系统变量 JAVA_HOME 界面

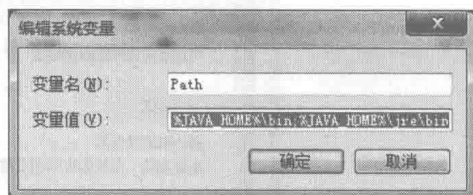


图 1-7 编辑系统变量 Path 的界面

其次,在“系统变量”列表框中找到 ClassPath 变量,双击弹出“编辑系统变量”对话框,其中“变量名”文本框中的字符串 ClassPath 不用修改,在“变量值”文本框的最后添加“;%JAVA_HOME%\lib\dt.jar;%JAVA_HOME%\lib\tools.jar;”。切记要用“;”将它与前一个值分开。如果没有找到 ClassPath 变量就新建变量 ClassPath,在“变量值”文本框中添加“%JAVA_HOME%\lib\dt.jar;%JAVA_HOME%\lib\tools.jar;”,如

图 1-8 所示。

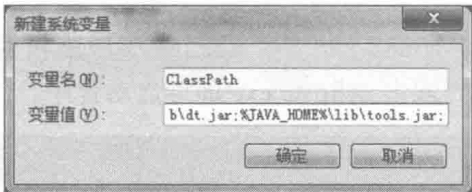


图 1-8 新建系统变量 ClassPath 的界面

最后，依次单击“环境变量”和“系统属性”对话框中的“确定”按钮，保存配置。

(3) 检验 JDK 是否安装成功

在“开始”菜单下方的搜索框内输入 cmd 并按 Enter 键，出现“命令提示符”界面。在其中依次输入 java 和 javac，如果 JDK 环境配置成功，那么会出现如图 1-9 和图 1-10 所示的界面。否则，应检查系统环境变量中各变量值的设置是否正确。



图 1-9 输入 java 命令验证 JDK 环境配置是否成功

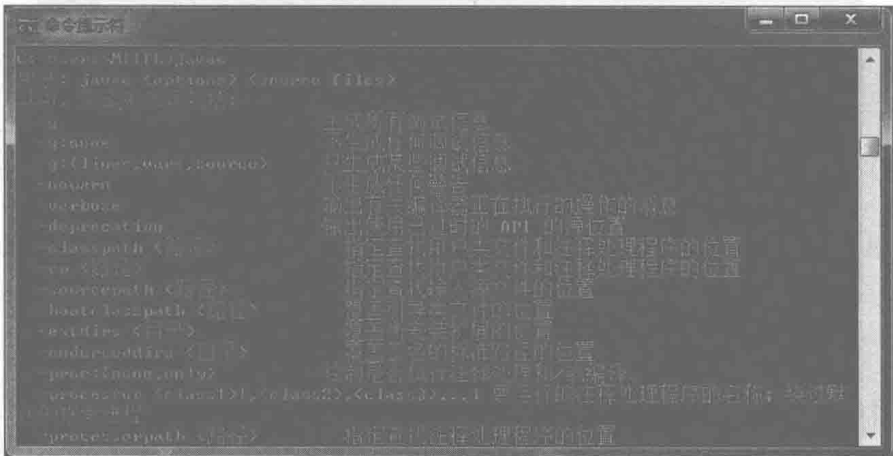


图 1-10 输入 javac 命令验证 JDK 环境配置是否成功

2. 安装 Eclipse

下载 eclipse-jee-mars-1-win32-x86_64.zip^①,解压压缩包后出现 eclipse 文件夹。进入文件夹后,如图 1-11 所示,双击 eclipse.exe 即可启动程序。



图 1-11 Eclipse 软件文件目录

如果启动出现 Failed to create java virtual machine 错误提示,则需要用写字板打开 eclipse.ini,如图 1-12 所示。将其中的数值 256M 改成 512M,表示运行内存增加一倍,然后再双击 eclipse.exe 启动软件。

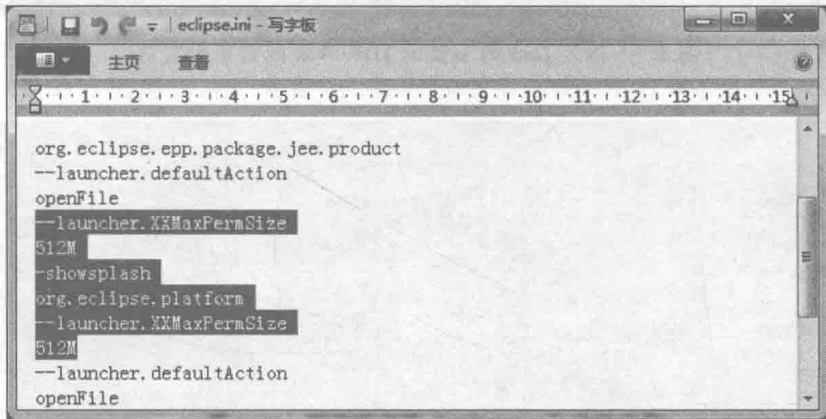


图 1-12 用写字板打开 eclipse.ini 文件

正常启动后,首次运行会出现设置工作空间选项,如图 1-13 所示。可勾选左下方的

^① <https://www.eclipse.org/downloads/packages/eclipse-ide-java-ee-developers/mars1>。

Use this as the default and do not ask again 选项,这样再次启动程序时会自动加载该工作空间。

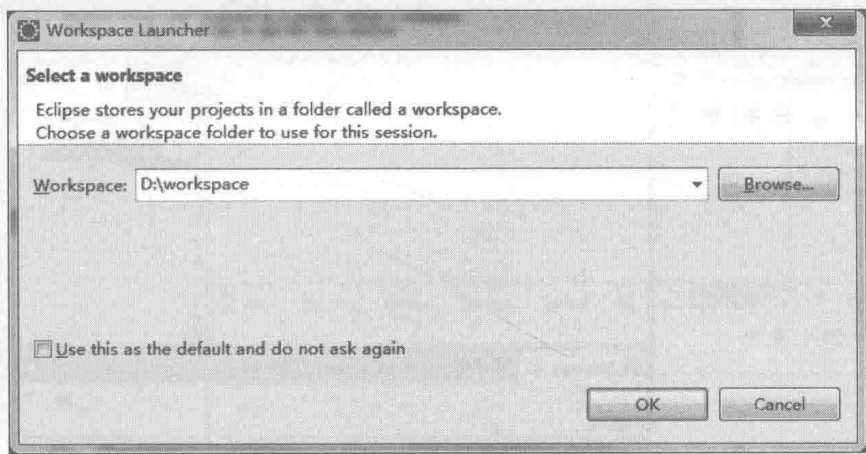


图 1-13 Eclipse 软件启动时设置工作目录

单击 OK 按钮后进入 IDE 界面,如图 1-14 所示。

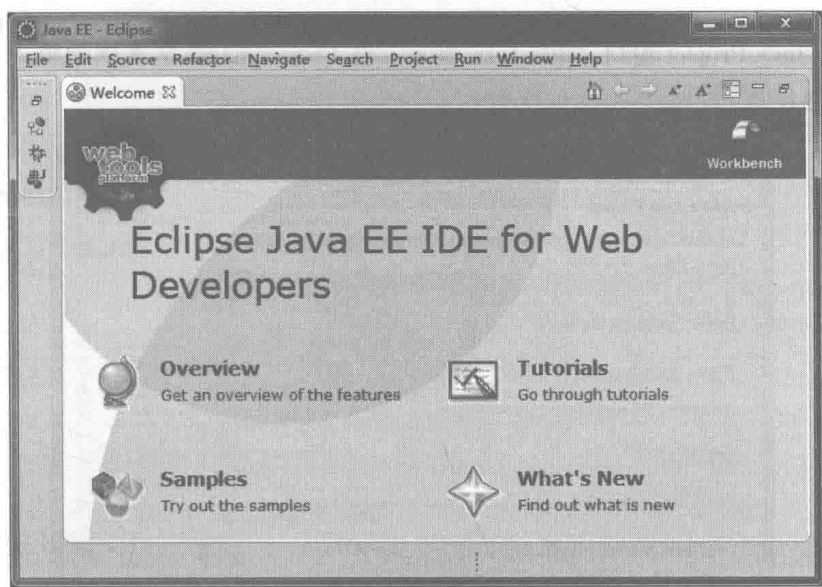


图 1-14 Eclipse IDE 界面

单击页面右上角的 Workbench 图标,出现如图 1-15 所示的 Eclipse 开发界面。

1.1.4 用 Eclipse 开发 HelloWorld 项目

本小节主要介绍利用 Eclipse 新建项目、类并运行项目的过程。

1. 新建 Java 项目

选择 File 菜单中 New 子菜单下的 Project...,选择 Java Project 选项,单击 Next 按