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

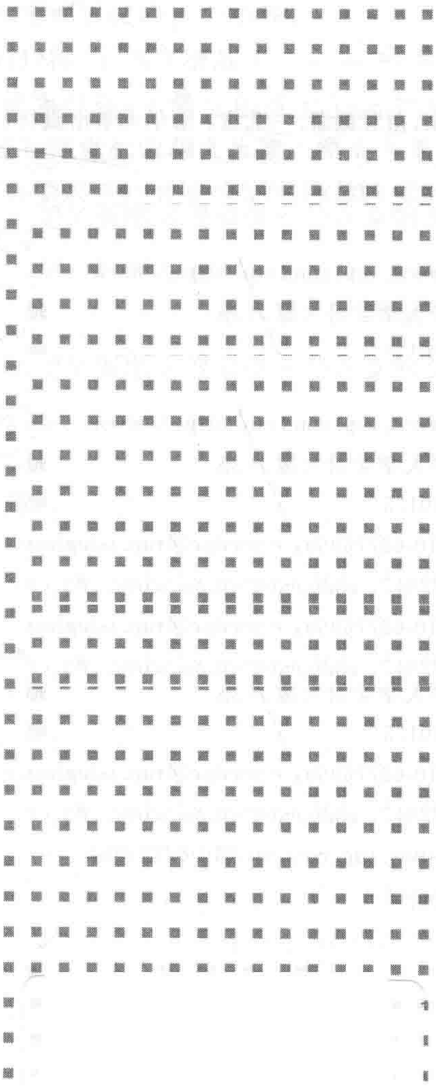
案例翔实

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。

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

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。

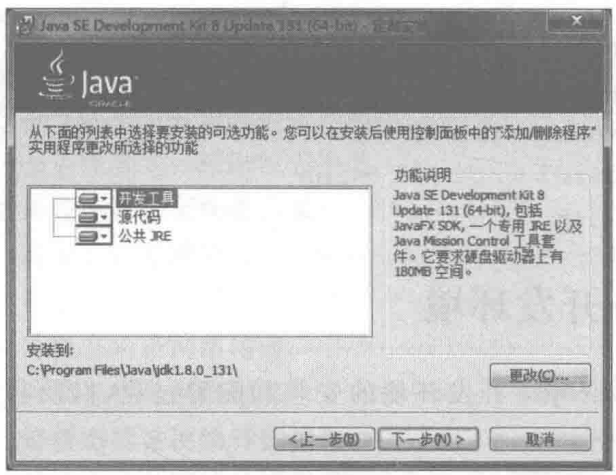


图 1-1 安装 JDK 时选择安装路径界面

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 程序在运行时会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。

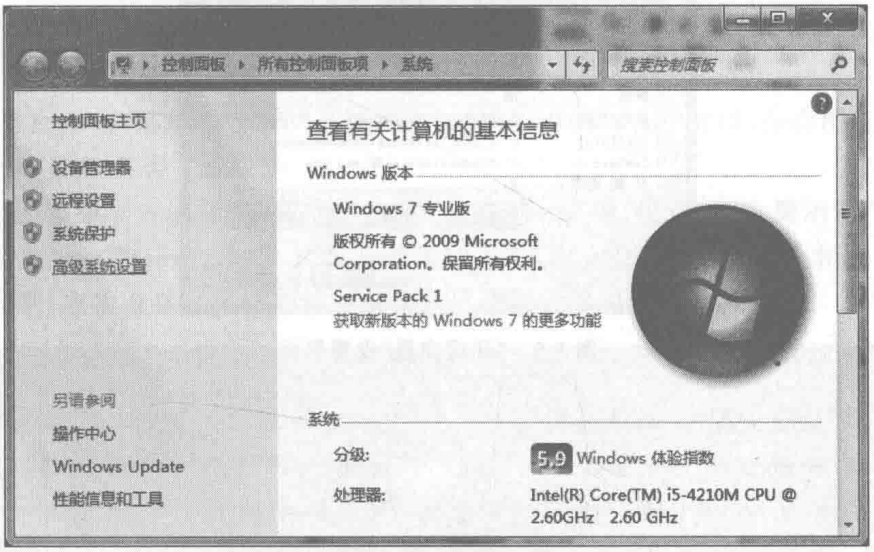


图 1-3　控制面板关于系统信息的界面

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。

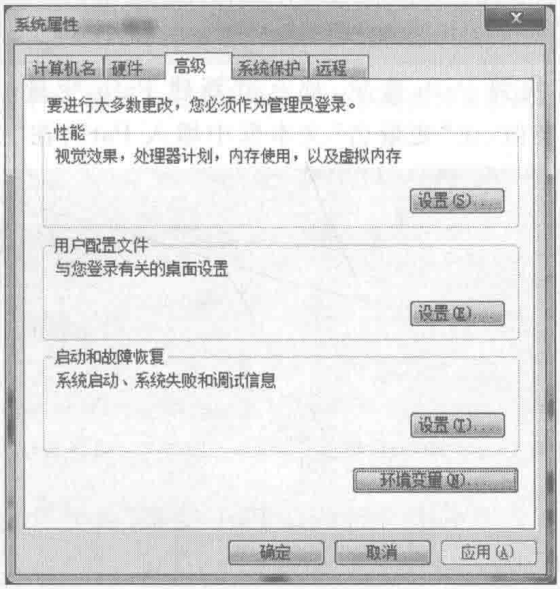


图 1-4　系统属性中的“高级”选项卡界面

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”。

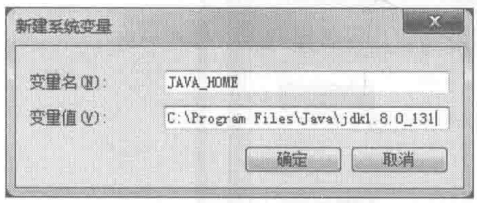


图 1-6 新建系统变量 JAVA_HOME 界面

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。

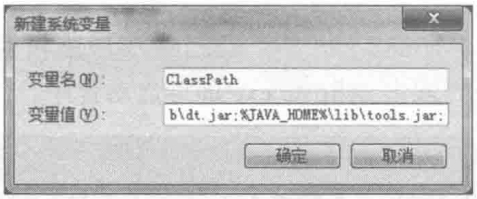


图 1-8　新建系统变量 ClassPath 的界面

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。

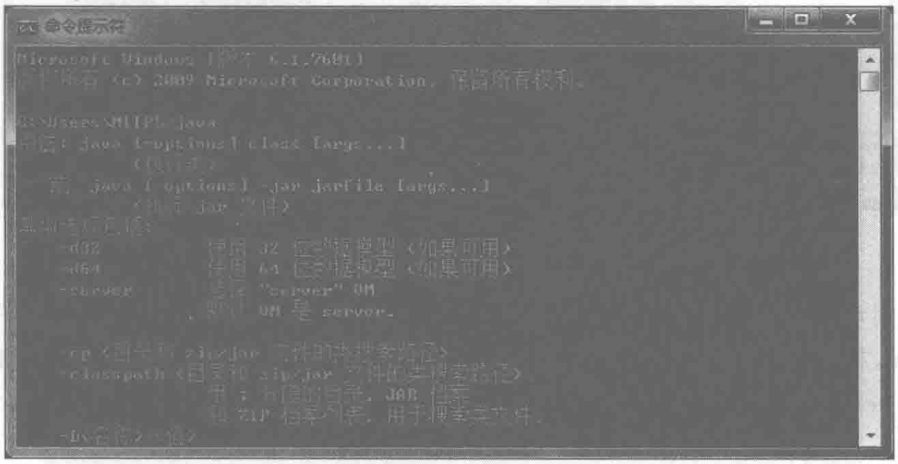


图 1-9　输入 java 命令验证 JDK 环境配置是否成功

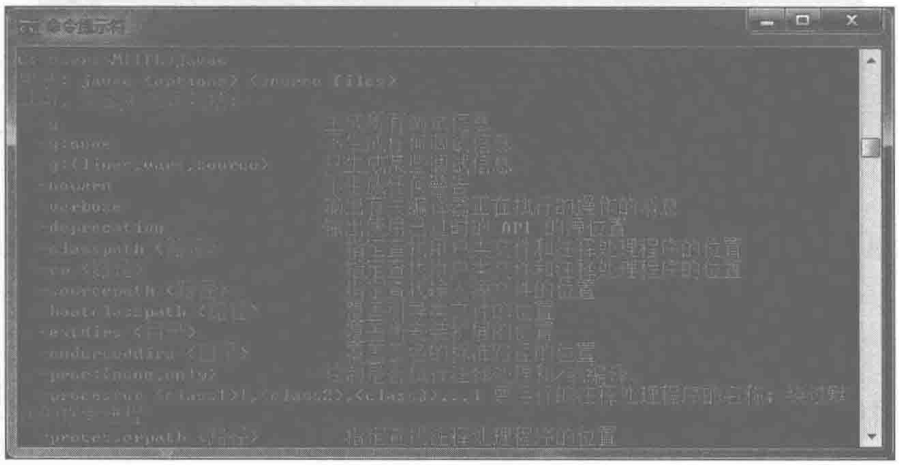


图 1-10　输入 javac 命令验证 JDK 环境配置是否成功

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。

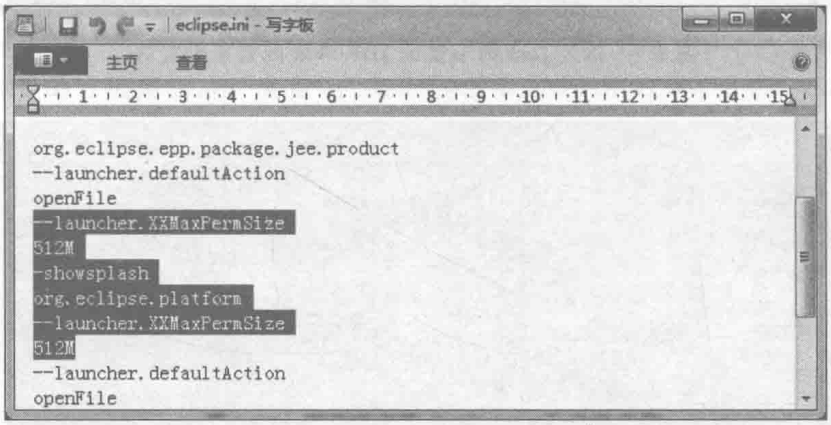


图 1-12 用写字板打开 eclipse.ini 文件

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。

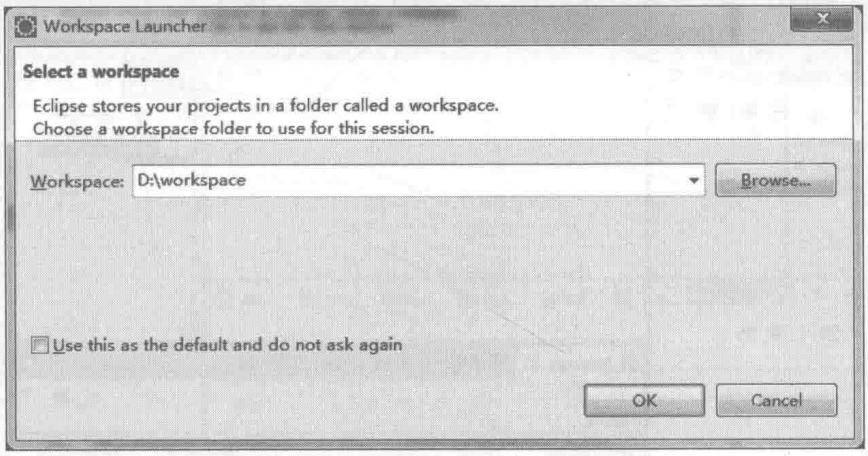


图 1-13　Eclipse 软件启动时设置工作目录

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