**网络水军识别系统V1.0**

**用户手册**

**山西大学**

第一章 引言

## 1.1 编写目的

随着计算机技术、网络技术和通信技术的迅猛发展，互联网已成为海量数据信息的载体，广大网民可以通过其自由地浏览信息和发表个人观点，促使网络数据以一种爆炸式的模式增长，论坛已经成为新的信息传播媒介、话题的聚集地、网络舆情的发源地。由于话题讨论的多方参与，虚拟角色的隐蔽性等，使得在各种社交平台中出现了大量重复信息并且不断骚扰我们视线，随即也相应地出现了网络水军这一称号。发现并且去除水军，对于维护网络安全、网络信息的真实性具有重要的研究意义与研究价值。

## 1.2 项目背景

网络水军是指受雇于网络公关或营销公司，有组织、有计划地在集中炒作某个事件，以达成宣传、推销或诋毁效果的大量专业或非专业技术团体。随着移动互联网的广泛运用，论坛成为民众获取信息、发表言论的一种重要途径，而近些年来论坛上涌现出大量水军，肆意发表带有倾向性的言论，扰乱民众视听，混淆信息的真实性。水军引导了舆论走向，使民众无法发出真实声音，从而形成言论“一边倒”的形式。网络水军通过操纵软件机器人或傀儡账号在互联网中制造、传播虚假意见和垃圾信息等。现有网络环境催生了大量网络水军，其行为特点主要包括大规模团体行动，数量庞大，行为区别于正常用户等。网络水军影响了网络秩序、公民利益、社会稳定及国家政治。因此，如何发现并且去除水军，对于维护网络安全、网络信息的真实性具有重要的研究意义与研究价值。

本文提出基于文本内容与用户信息的水军识别方法研究算法，并通过对汽车之家论坛里的评论数据处理和分析，利用用户潜在行为数据得出用户行为特征，采用LDA主题模型得出用户评论的主题特征，采用Word2Vec词向量模型得出用户评论的文本内容特征，采用K-Means聚类得出水军文本类别，结合用户行为特征，最终实现了对网络水军的识别。

## 1.3 相关技术

**1.3.1 任务定义**

本系统针对汽车之家论坛评论的七大任务，即：数据搜集，数据预处理，LDA模型主题表示，Word2Vec模型文本表示，K-Means聚类，用户行为分析，用户属性判定。

（1）数据收集

本文的数据来源于从汽车之家论坛爬取得到数据。数据的类型主要有文本信息：从汽车之家论坛爬取的各种车型的评论数据；用户信息：在论坛所爬取的文章的作者信息，包括关注数、粉丝数、里程值、注册地点、注册时间。数据的规模为汽车之家论坛3000条文本信息，并且获取了其中1000人的用户信息。另外，收集了停用词表，高频词。

（2）数据预处理

针对汽车评论数据，预处理操作包括分词、去停用词两部分。

1. 本系统采用jieba分词工具对爬取得汽车评论数据进行分词；
2. 将分词后的结果与收集到的停用词表作对比，去除每条评论数据中的停用词，高频词。

（3）LDA模型主题表示

使用预处理后的数据训练得到LDA模型，再利用训练后的LDA模型对新的文本进行计算主题分布概率（主题特征）。

（4）Word2Vec模型文本表示

利用Word2Vec词向量模型对文本进行向量化（文本特征）。

（5）K-Means聚类

对主题分布概率和词向量的融合向量进行K—Means聚类（两类）。

（6）用户行为分析

利用爬取的用户信息计算用户行为特征：好友粉丝比，里程时间比。

其中，用户行为特征定义如下：

1. 好友粉丝比（FA）

反映用户关注数和粉丝数的比例，其粉丝数和关注数的比例越低，其可信度越低.设用户i关注数位ANi，粉丝数为FNi，则好友粉丝比定义如下：

 （1）

设定阈值为0.5，比值低于0.5时，认为其信誉度较低；当比值远大于0.5时，则可基本认定其不是水军；当比值接近于0.5时，与其它信息一起进行综合评判。

1. 里程时间比（TM）

里程时间比反映注册用户对注册车辆的使用频率，比值越低，可信度也就越高.设用户里程值为MLi（公里），注册时间为RTi（天），则历程时间比如下：

 （2）

设定阈值为0.05，小于该值时，认为其可信度较高。

备注：按数据统计一个月里程值至少为600公里为正常值，因此设30/600为阈值，作为评判标准。

（7）用户属性判定

根据用户评论文本数据的聚类结果和用户行为特征值，计算p，q值，再根据图1-1的用户属性判定图判断用户属性。

 （3）

 （4）

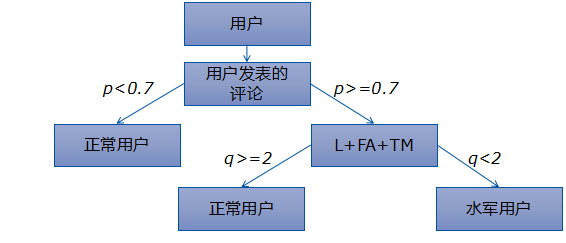


图1-1 用户属性判定

* + 1. **评价标准**

**评价指标**

对于该方法的评价，我们采用准确率P,召回率R，F值作为评测标准，计算公式如下：

准确率（P）  （3）

召回率 （R）：  （4）

F值  （5）

第二章 系统概述

## 2.1 系统总体功能

本系统总体上分为四个部分，包括输入文件、导入文件、预处理和分析用户。

用户通过系统总界面来输入所需分析的数据文件路径，在系统中文件以txt文本形式进行存储与操作，对导入文件进行分析处理，系统总体功能结构如2-1图所示。

网络水军识别系统

分析用户

预处理

导入文件

输入文件

图2-1 总体功能结构图

(1) 输入文件路径：输入待处理数据文件的路径。

(2) 导入文件：点击导入文件按钮导入输入文件路径下的文件。

(3) 预处理：点击预处理按钮，对所有用户的评论数据进行预处理，统计用户个数。

(4) 分析用户：将每个用户的所有评论数据分别通过LDA和Word2Vec向量化，计算每个数据点与通过K-Means聚类稳定下来的两类类中心的距离，给每个评论数据分类，根据用户主页网址爬取用户信息并计算用户行为特征值，判断用户属性，显示正常用户个数，水军用户个数以及所有用户的评判结果。

## 2.2 数据格式

1）输入文件路径格式：磁盘名：/文件夹名/…/文件名

2)文件格式：名称；网址；原文

其中名称为汽车之家用户名称；网址为用户主页网址，其格式为：//i.autohome.com.cn/id/

home.html；原文为用户评论内容。导入的文本文件如图2-2所示。

3）预处理

1. 评论数据以txt文本文件存储，一行为一条评论数据，如图2-3所示。
2. 分词去糙，去停用词，去高频词，结果如图2-4所示。

4）分析用户

（1）LDA主题特征表示将评论数据表示为主题分布概率向量，一行为一条评论的主题特征表示，一共100维，表示结果如图2-5所示。

（2）Word2Vec词向量模型将评论数据表示成词向量，再将各个词向量加权平均得出该评论数据的文本特征表示，一共100维，表示结果如图2-6所示。

（3）将融合文本特征向量和话题特征向量后的评论数据进行K-Means聚类，聚为两类，结果如图2-7所示。

（4）爬取用户信息

根据用户的主页网址爬取用户的行为数据，行为数据包括关注数、粉丝数、里程值、注册地点、注册时间，计算用户行为特征值。爬取结果如图2-8所示。

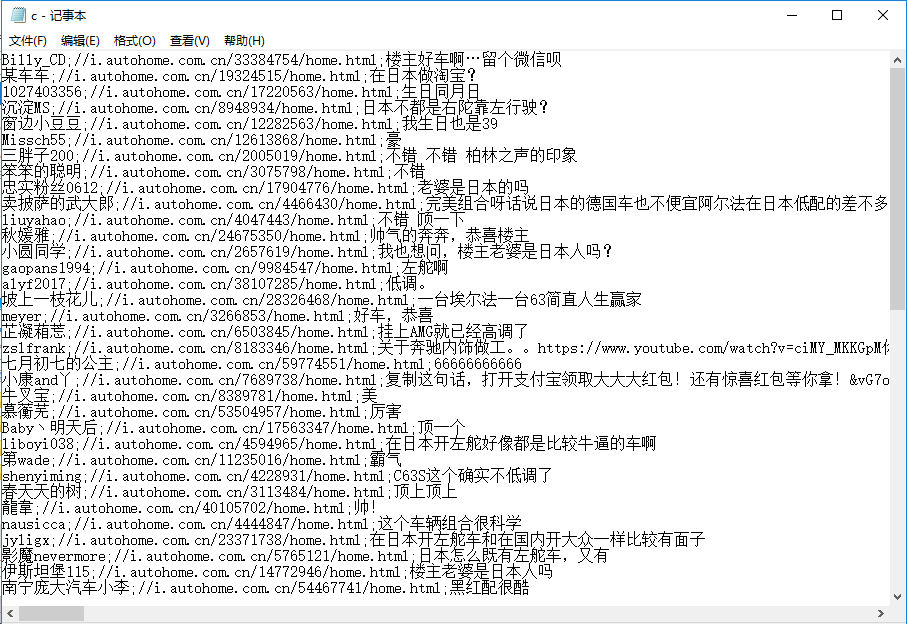


图2-2 导入的文本文件

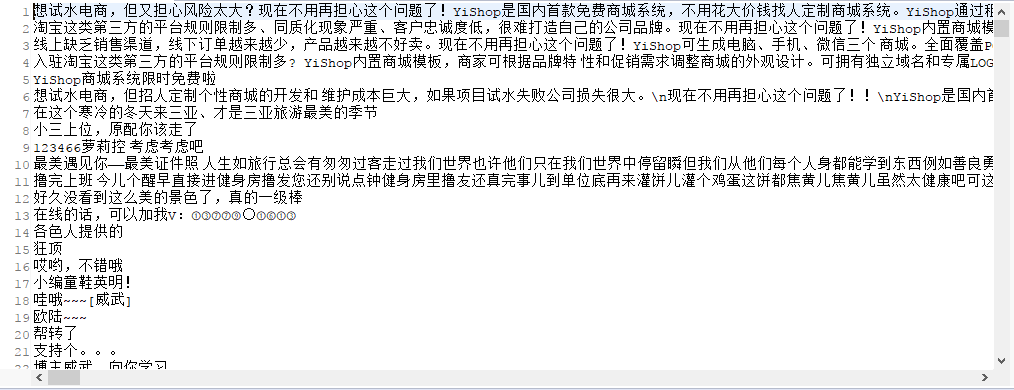


图2-3 评论数据

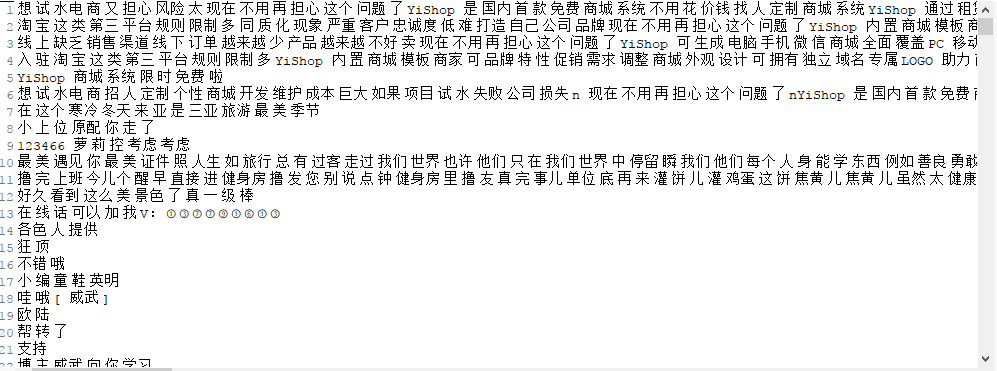


图2-4 分词去糙结果

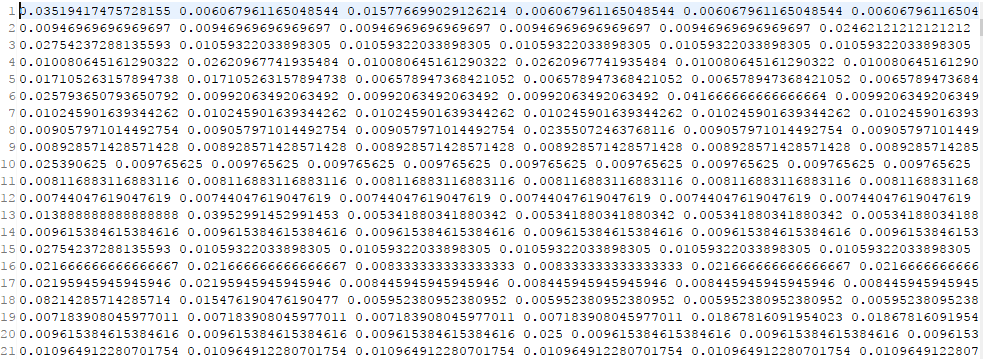


图2-5 LDA主题特征表示结果

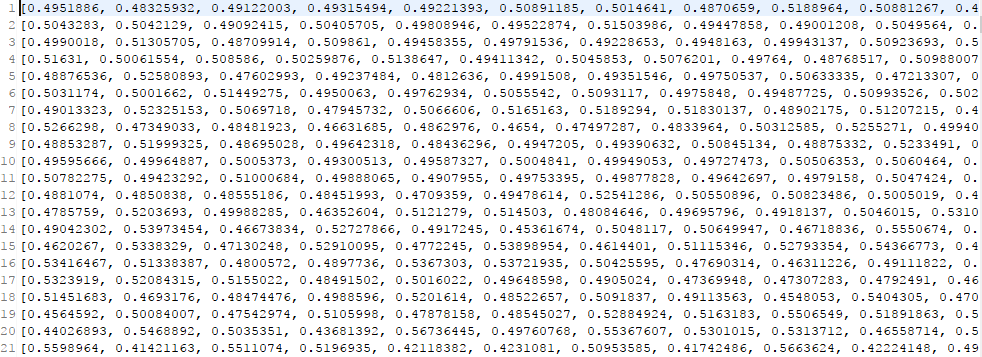


图2-6 Word2Vec文本向量表示结果

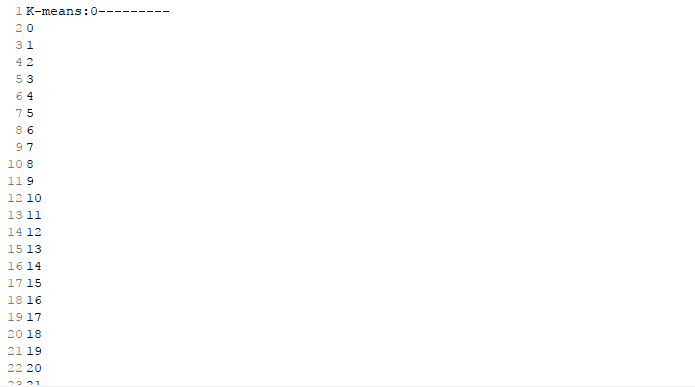


图2-7 K-Means聚类结果

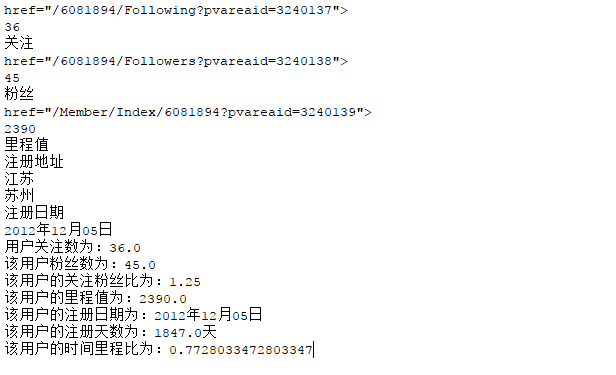


图2-8 爬取用户信息结果

（5）用户属性判定

计算用户的p，q值，判定用户属性，判定结果如图2-9所示。



图2-9 用户判定结果

所有数据文件均以txt文本格式存储，采用GBK编码。

## 2.3 系统主界面

2.3.1系统主页

本系统总体上分为四个部分，包括输入文件、导入文件、预处理和分析用户。

双击入口程序Main.java，进入到系统主界面，如图2-10所示。



图2-10 系统主页

2.3.2输入文件路径

输入文件路径提供将文件路径导入工程的功能。

在输入框中输入待处理数据文件的路径，如图2-11所示。



图2-11 输入文件路径

2.3.3导入文件

导入文件提供将文件导入工程，并提示导入文件路径的功能，以此来保证接下来的操作的正确性。

输入待处理数据文件的路径后，点击“导入文件”按钮，待执行完毕会弹出显示导入成功和导入文件路径的提示界面，如图2-12所示。

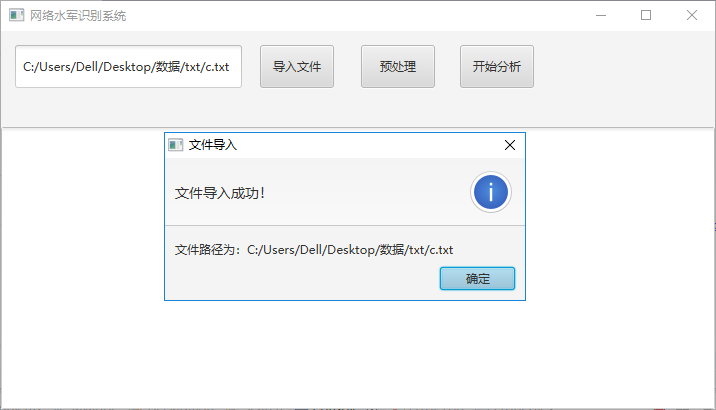


图2-12 导入文件

2.3.4预处理

数据预处理提供对文本数据的切割、统计用户个数、去杂、分词、去停用词以及去高频词六个功能。

用户点击“预处理”按钮，系统便开始执行上述六个功能，待执行完毕便会弹出提示预处理完成和发现用户个数的界面，如图2-13所示。

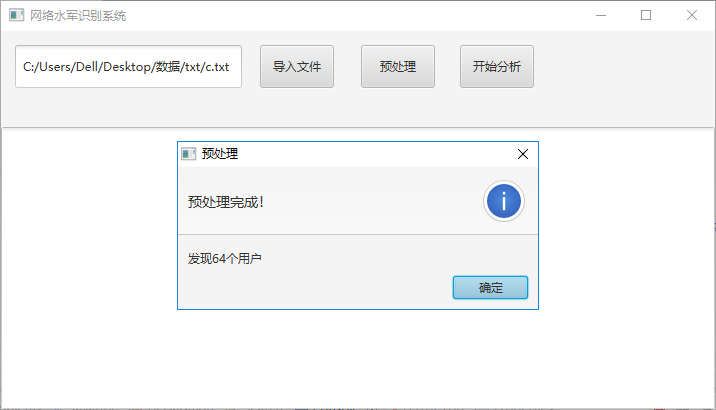


图2-13 预处理

2.3.5分析用户

分析用户提供对用户评论数据进行LDA主题特征表示、Word2Vec词向量文本特征表示、文本相似度计算、爬取用户信息、及算用户行为特征值、判定用户属性以及显示判定结果七个功能。

点击“开始分析”按钮，系统便开始执行上述七个功能，待执行完毕系统会在文本框区域显示结果，结果包括正常用户个数，水军用户个数以及各个用户的分析结果，结果如图2-14所示。



图2-14 分析结果

第三章 运行环境

## 3.1 软件环境

eclipse：eclipse3.7 64bit；

操作系统：Windows 7 、Windows 8 、Windows 10

## 3.2 硬件环境

CPU：4逻辑核心、8G RAM、500G HDD 及以上。

第四章 软件部署

本系统软件包括：eclipse主程序、windowbuilder相关插件2个组成部分。

## 4.1 数据部署

首先需在汽车之家汽车论坛爬取帖子数据，将数据整理格式，以txt文本格式存储，编码方式为GBK，具体格式如图4-1所示。



图4-1 数据格式截图

其次，需收集停用词表，以保证后续分词去糙的正确性，停用词表如图4-2所示。

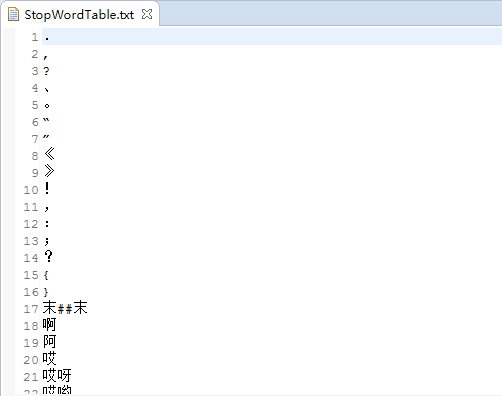


图4-2停用词表

最后，以从汽车之家爬取的2000条帖子数据作为训练集，人工进行标注，以txt形式存储，采用GBK编码方式，具体格式如图4-3所示。

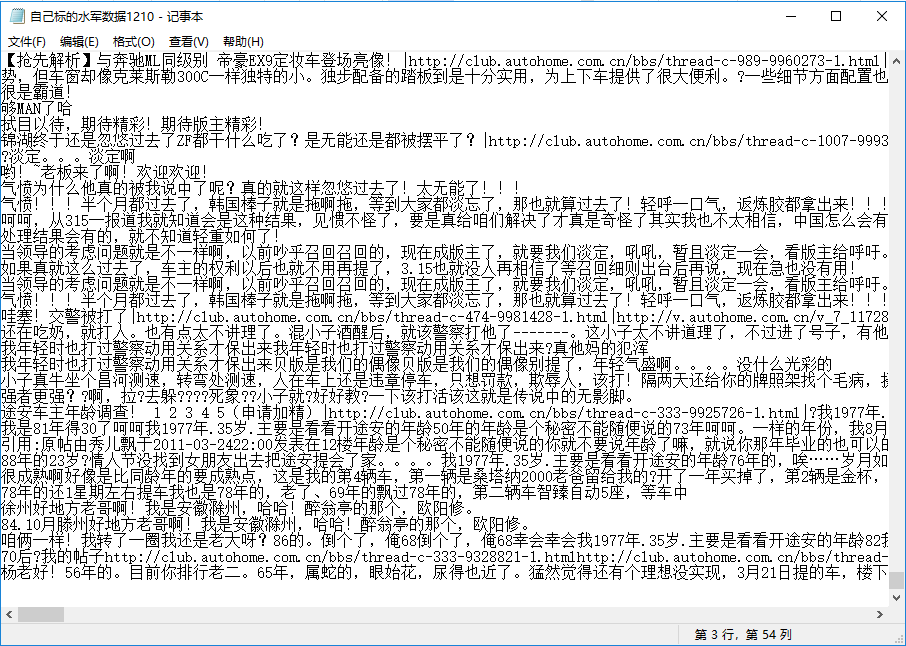


图4-3 训练数据格式

## 4.2 软件包部署

用户可通过打开eclipse，在File菜单下点击import来导入程序，如图4-4，4-5所示。

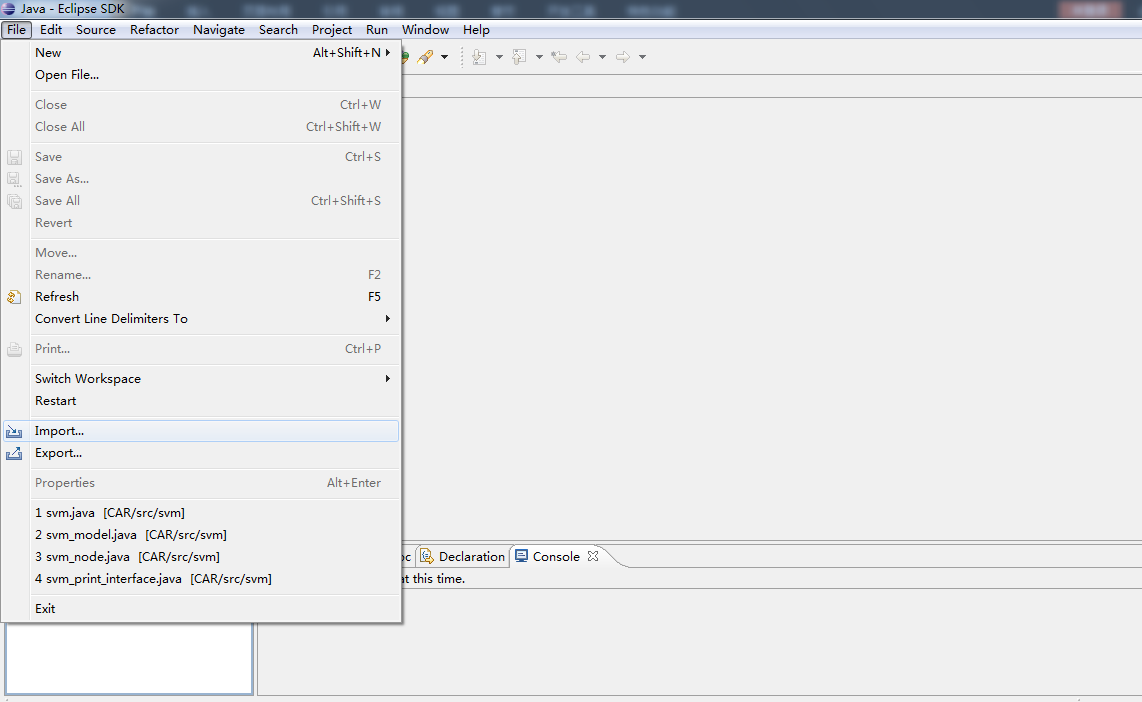


图4-4 eclipse导入程序

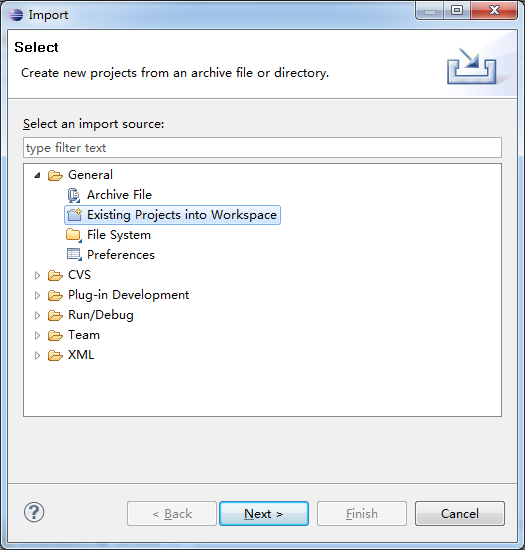


图4-5 eclipse导入程序

点击next按钮，弹出如图4-6所示界面

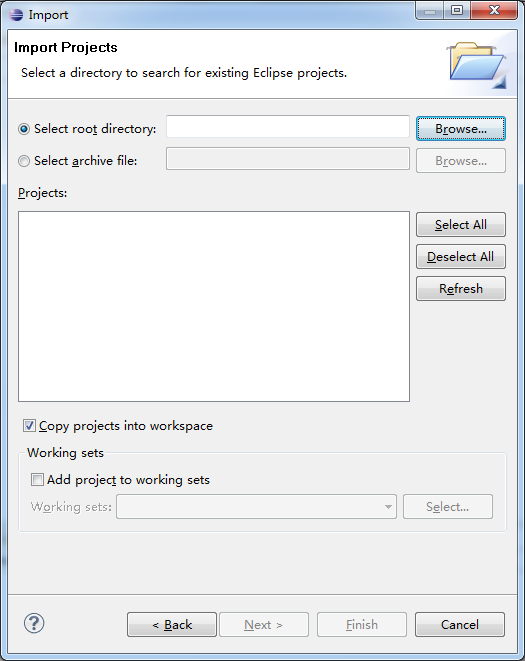


图4-6 选择导入工程路径

点击browse按钮,选择导入的工程如图4-7所示。

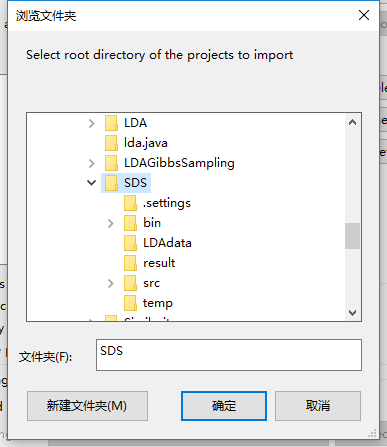


图4-7 选择所要导入的程序

然后点击确定按钮，再点击finish，便将工程导入到eclipse工作空间，如图4-8所示

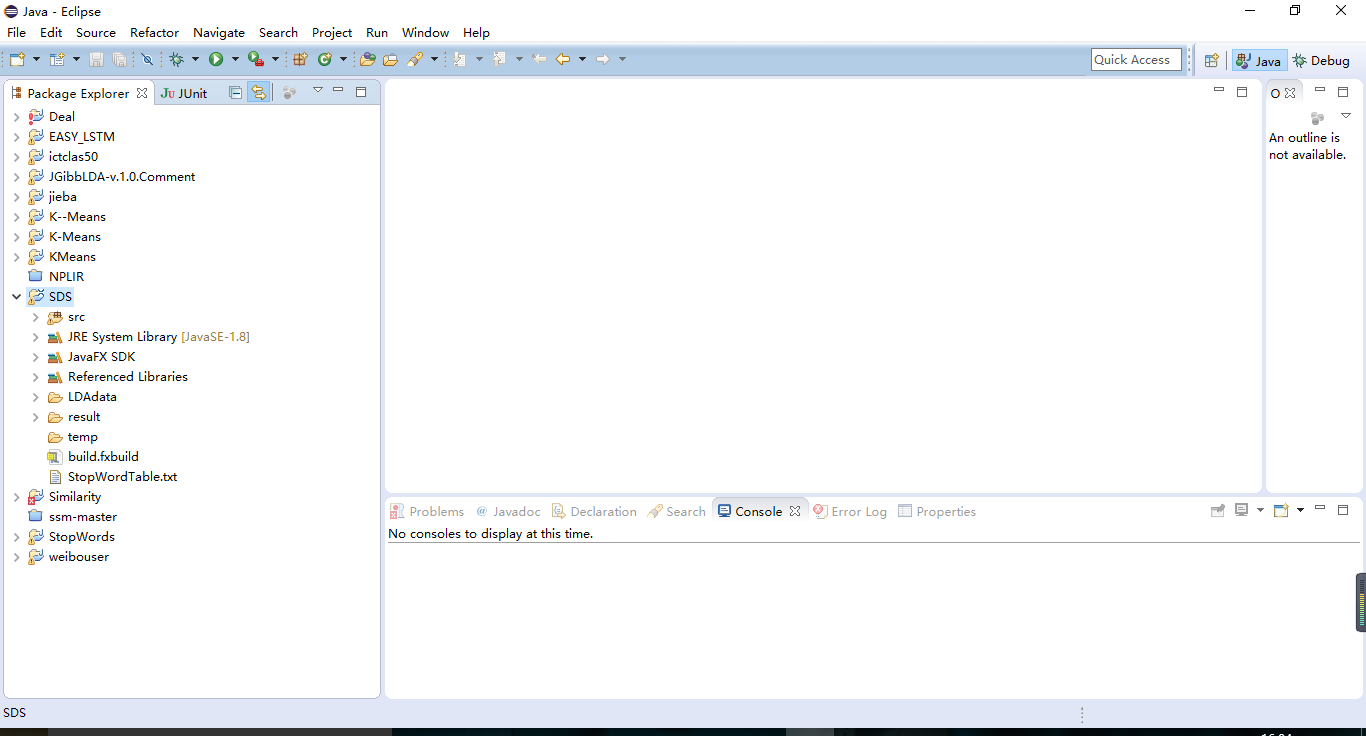
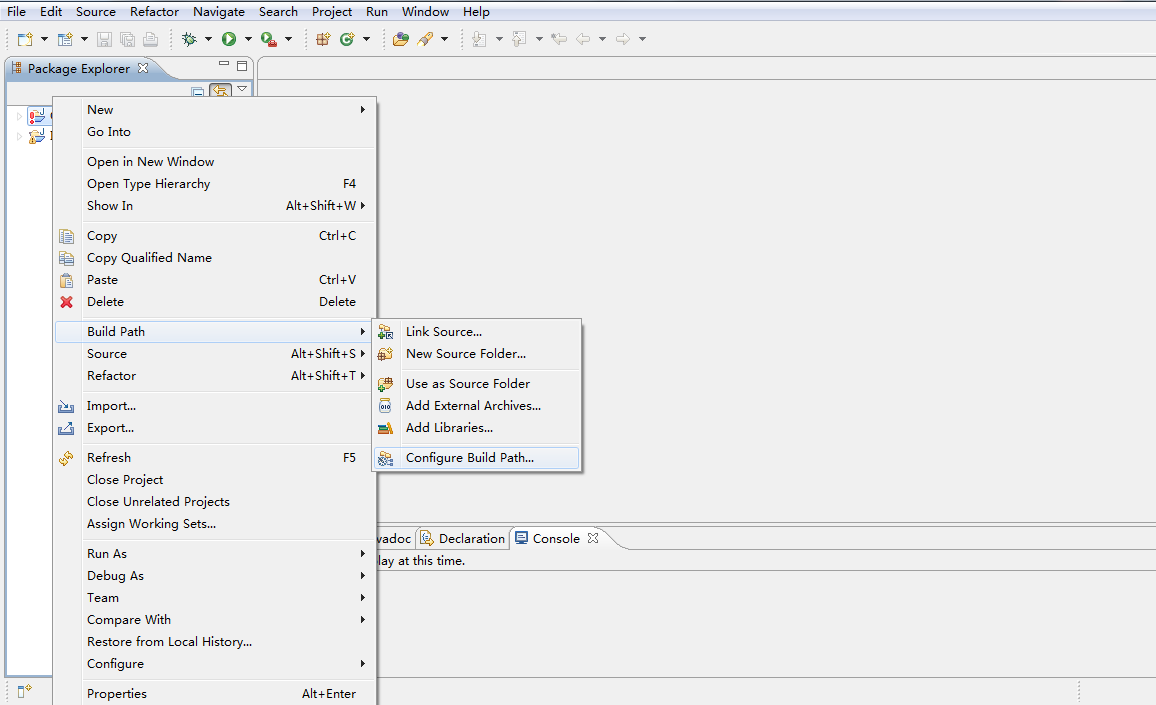


图4-8 成功导入程序

在工程上点击右键，点击build path选项，再点击configure build path，如图4-9所示。 图4-9工程build path界面



点击Libraries菜单，弹出如图4-10所示界面。

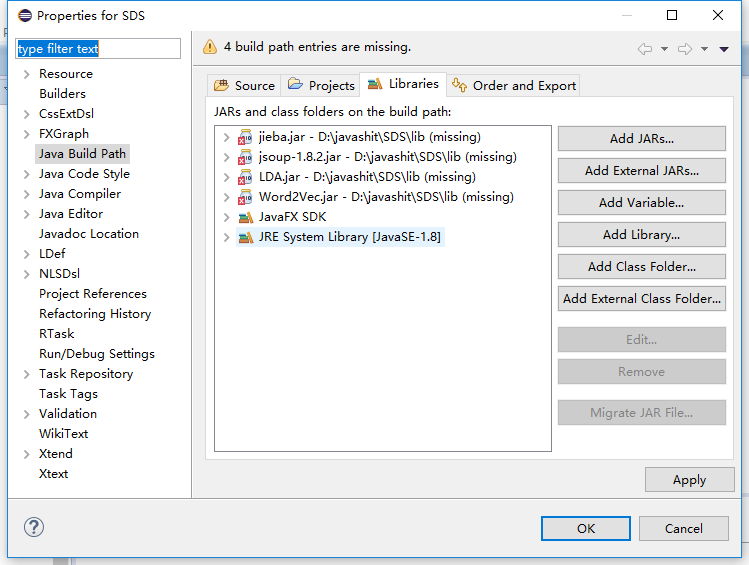


图4-10 工程建立Libraries界面

选中有错误的两个包，点击Edit,在所导入工程下找到lib文件夹，将这四个包外部包导入，如图4-11所示。

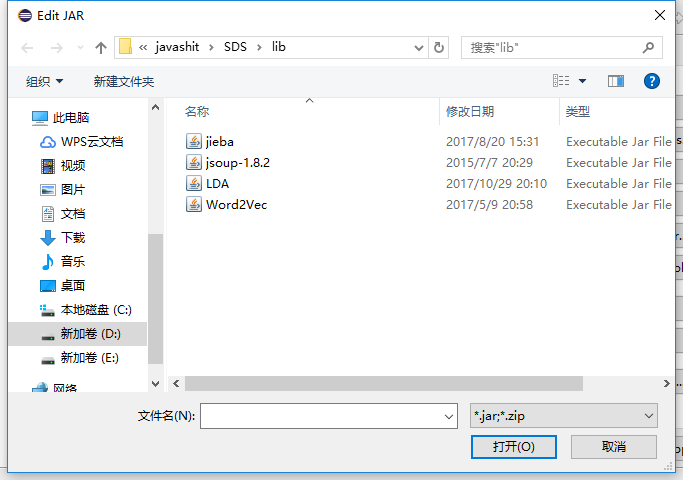


图4-11 导入外部包界面

成功导入后，之前有错误的地方就正确了，这是点击ok按钮就行了。如图4-12所示。

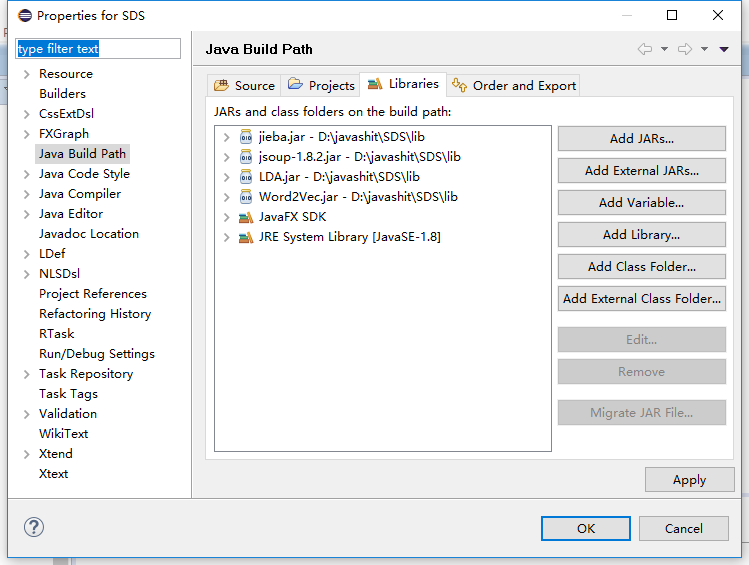


图4-12成功导入外部包界面

用户在导入的工程中找到入口程序main.java，便可运行了，如图4-13所示。

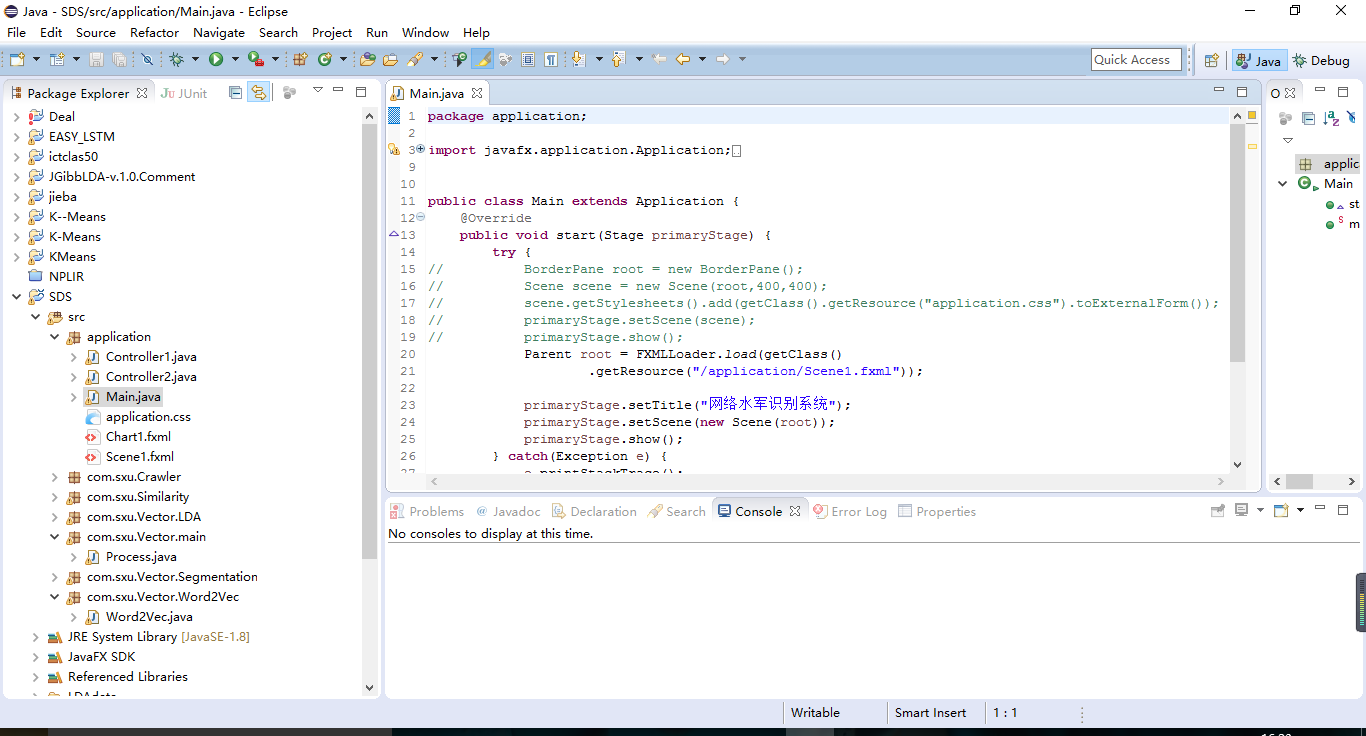


图4-13找到入口程序界面

## 4.3 工具软件安装

软件运行环境包括eclipse主程序以及Windowsbuilder相关插件。

下载地址

Eclipse：http://www.eclipse.org/downloads/

Windowsbuilder:<http://www.eclipse.org/windowbuilder/d>[ownload.php](http://www.eclipse.org/windowbuilder/download.php)

Eclipse安装安装eclipse之前确保安装了JDK，然后下载eclipse3.7 32位，解压后找到eclipse的应用程序便可直接运行不需安装，如图4-14所示。

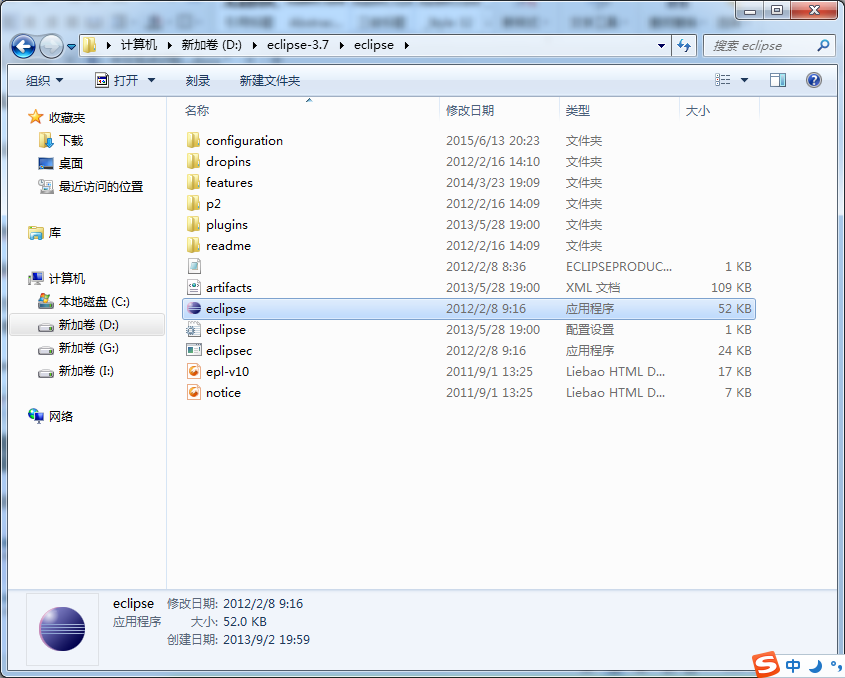


图4-14 eclipse应用程序

双击eclipse应用程序，打开eclipse的工作空间，如图4-15所示。

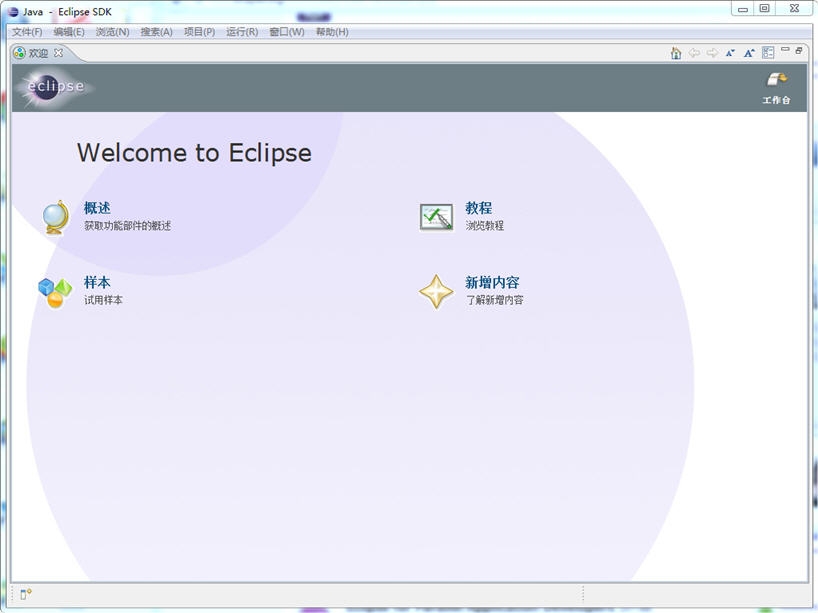


图4-15 eclipse工作空间

Windowsbuilder插件安装：

首先在官网上下载eclipse对应的windowbuilder版本，如图4-16所示。

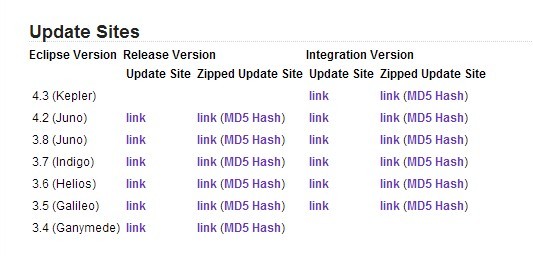


图4-16Windowbuilder对应版本下载

打开eclipse，选择Help菜单中的Install New Software，如图4-17所示。

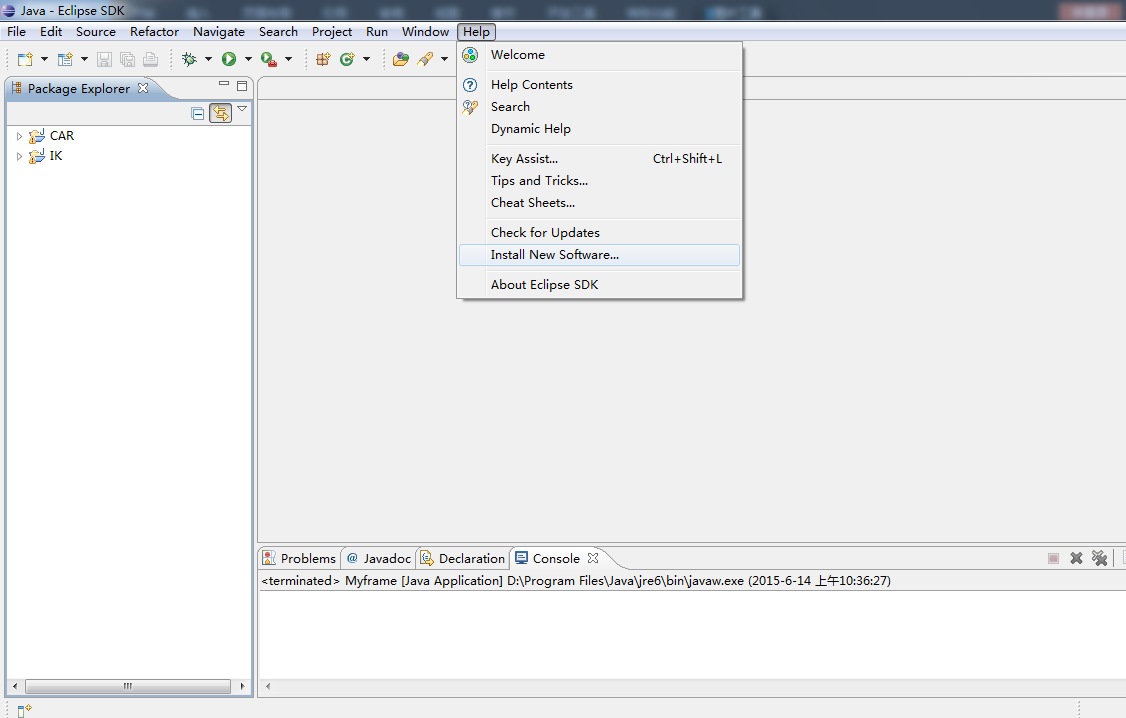


图 4-17 eclipse中Help菜单

点击Add按钮，如图4-18所示。

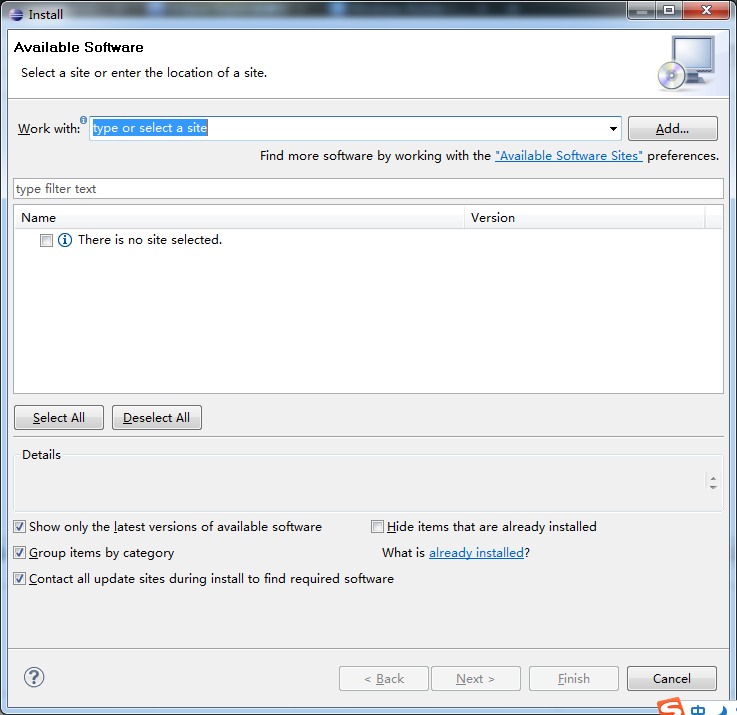
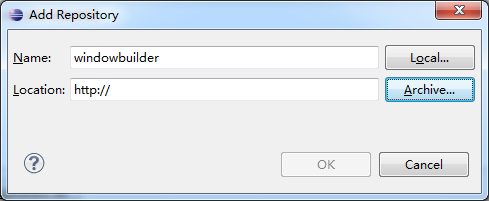


图4-18 eclipse安装插件界面

填写name，点击Archive按钮，如图4-19界面。

图4-19填写安装插件name界面



选择下载的windowbuilder.zip，点击打开按钮，如图4-20所示。

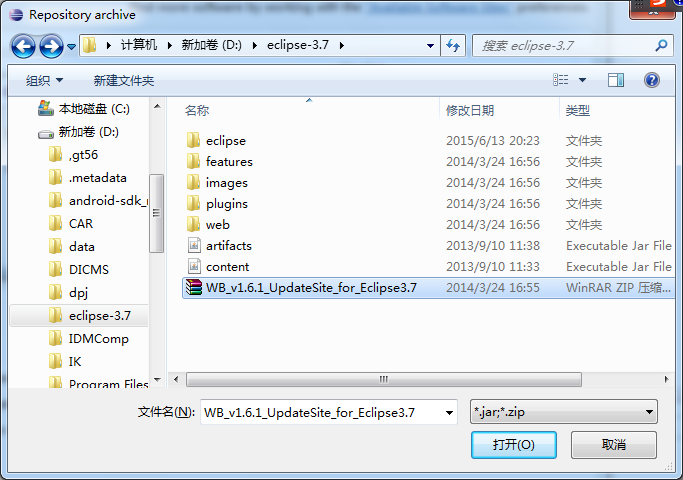


图4-20 选择下载windowbuilder压缩包

选择Select All，然后点击Next，如图4-21所示

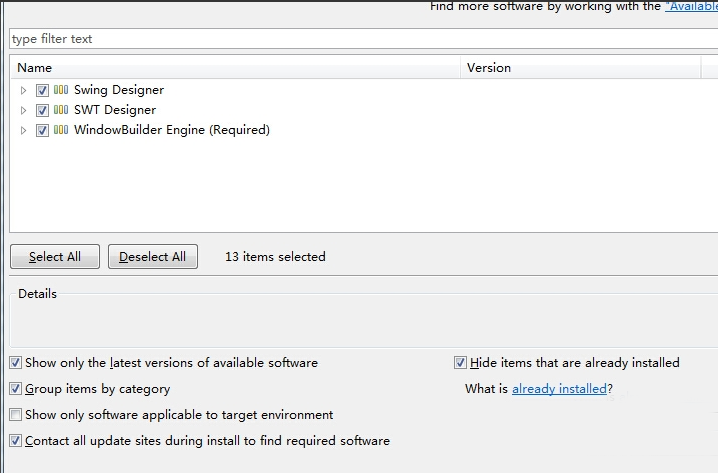


图4-21 安装windowbuilder所有组件

一直根据提示点击下一步就可以了，最后重启Eclipse。之后点击New——other，就能看到WindowBuilder了，如图4-22所示。

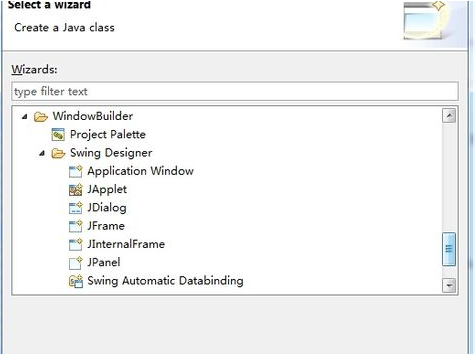


图4-22 windowbuilder安装成功界面

## 4.3版权声明

该软件是山西大学独立开发的软件，并拥有全部版权，任何非授权个人和单位不得用做商业目的。