

# 基于 Scrapy 的网络信息搜索工具

## 用户手册

2015 年 10 月 15 日

## 目录

1	引言 .....	3
1.1	编写目的 .....	3
1.2	项目背景 .....	3
1.3	定义.....	3
1.4	参考资料 .....	4
2	软件概述 .....	4
2.1	目标.....	4
2.2	功能.....	4
3	运行环境 .....	5
3.1	硬件.....	5
3.2	支持软件 .....	5
4	使用说明 .....	5
4.1	主界面.....	5
4.2	爬取功能 .....	7
4.3	配置功能 .....	8
4.4	查询功能 .....	9
5	附录 运行环境搭建说明 .....	12
5.1	安装 Python 2.7.....	12
5.2	安装 MongoDB .....	13
5.3	安装 Scrapy .....	14
5.4	安装 PyQt 4.....	14
5.5	安装 PyWin32.....	14

# 1 引言

## 1.1 编写目的

本用户手册的编写目的在于说明“基于 Scrapy 的网络信息搜索工具”的项目背景、运行环境、功能和使用方法，读者包括普通用户和开发人员。普通用户可以通过阅读本手册快速掌握“基于 Scrapy 的网络信息搜索工具”的使用方法，开发人员也可以通过阅读本手册了解到项目的开发情况，为软件维护和升级做准备。

## 1.2 项目背景

“基于 Scrapy 的网络信息搜索工具”项目是由南京航空航天大学计算机科学与技术学院张德平老师提出，由南京航空航天大学计算机科学与技术学院的李达（组长）、陈睿进、骆克云、吴钱胜、冯致远和邵玥六位同学完成开发。

## 1.3 定义

手册中使用的专业术语的定义如下：

- a. Python：一种面向对象、解释型计算机程序设计语言。
- b. Scrapy：Python 开发的一个快速，高层次的屏幕抓取和 web 抓取框架，用于抓取 web 站点并从页面中提取结构化的数据。Scrapy 用途广泛，可以用于数据挖掘、监测和自动化测试。
- c. MongoDB：一种非关系型数据库（NoSql），具备灵活的数据存储方式。
- d. PyQt 4：创建 GUI 应用程序的工具包。
- e. PyMongo：Python 操作数据库的工具包。

## 1.4 参考资料

- a. “基于 Scrapy 的网络信息搜索工具”项目开发计划，李达。
- b. “基于 Scrapy 的网络信息搜索工具”需求规格说明书，冯致远。
- c. “基于 Scrapy 的网络信息搜索工具”概要设计说明书，吴钱胜。
- d. “基于 Scrapy 的网络信息搜索工具”详细设计说明书，吴钱胜。
- e. “基于 Scrapy 的网络信息搜索工具”测试报告，邵玥。
- f. 使用 Scrapy 和 MongoDB 进行网络定向爬虫，骆克云。

## 2 软件概述

### 2.1 目标

本项目的目标是开发出一个基于 Scrapy 的网络信息搜索工具，它能够爬取指定网站的信息并存储到数据库中，用户可以查询爬取到的信息。

### 2.2 功能

“基于 Scrapy 的网络信息搜索工具”的具体功能如下：

- a. 爬取：爬取指定网站（即上海市政府采购中心中标公告汇总，网址是 [http://www.shzfcg.gov.cn:8090/new\\_web/cjxx/center\\_hz\\_zb.jsp](http://www.shzfcg.gov.cn:8090/new_web/cjxx/center_hz_zb.jsp)）上所有采购成功的公告中的项目名称、成交供应商、采购日期和成交金额，并将这些信息按采购项目逐条地全部显示在界面中；
- b. 配置：显示软件的相关配置信息，包括软件名称、爬取的网站网址、爬取功能模块名、参数设置、MongoDB 数据库参数设置；
- c. 查询：通过设置查询条件，包括待查询的采购项目名称、供应商、时间起始和采购金额范围，查询所有符合条件的项目并逐条地全部显示在界面中（每条信息包括项目名称、成交供应商、采购日期和成交金额）。

## 3 运行环境

### 3.1 硬件

能够支持 Windows XP、Windows Vista、Windows 7 和 Windows 8 操作系统的家用 PC 台式电脑以及笔记本电脑。

### 3.2 支持软件

- a. 操作系统：Windows XP、Windows Vista、Windows 7 和 Windows 8
- b. 编译器：Python 2.7
- c. 开发框架：Scrapy 1.0.3
- d. GUI 开发工具：PyQt 4
- e. 数据库：MongoDB 3.0.6
- f. Python 操作数据库工具：PyMongo 3.0.3

## 4 使用说明

### 4.1 主界面

双击 Scraper.py，弹出主界面

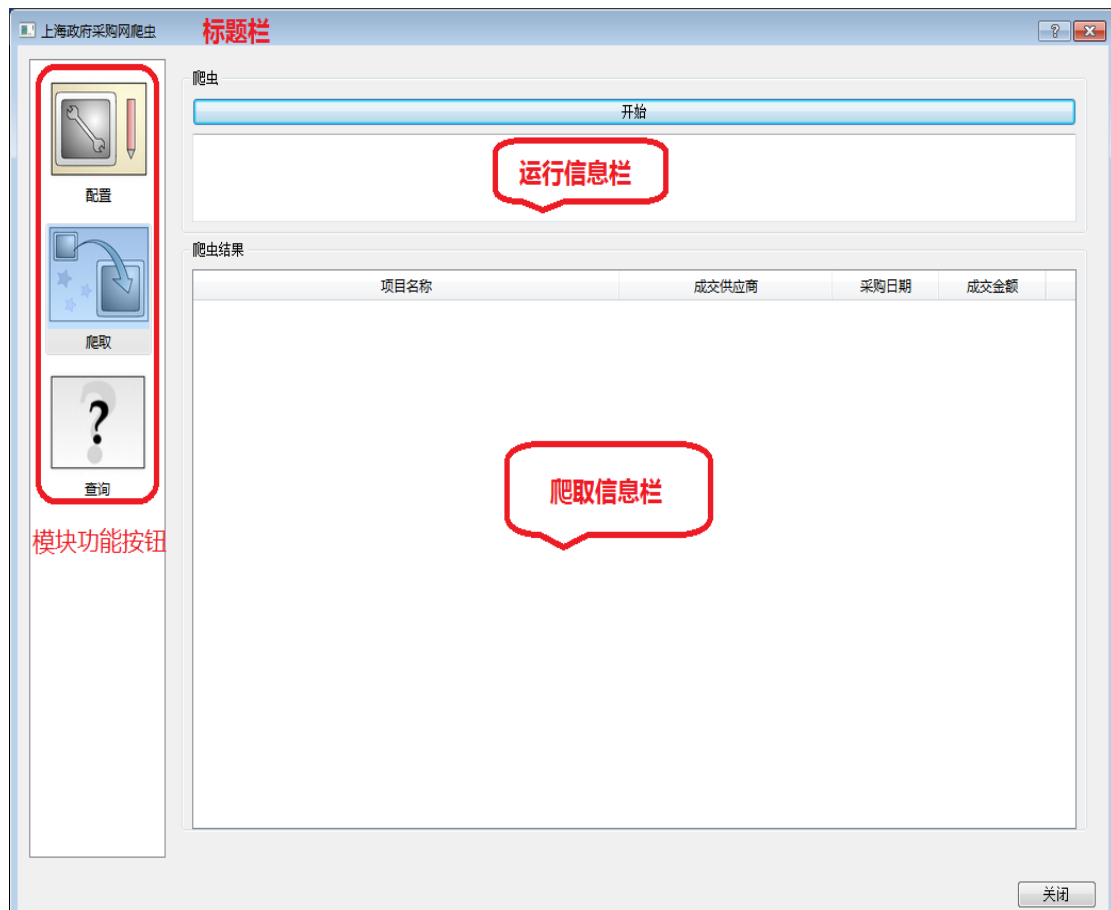


图 4.1 主界面

主界面由模块功能按钮、标题栏、运行信息栏和爬取信息栏组成，其功能如下：

- 模块功能按钮：点击模块功能按钮，进入相应的功能界面；
- 标题栏：显示主界面名称；
- 运行信息栏：显示 Scrapy 框架运行时的工作信息；
- 爬取信息栏：显示已经爬取到的信息。

## 4.2 爬取功能



点击模块功能按钮中的“爬取”按钮，显示的功能界面就是主界面。

点击“开始”按钮，爬虫程序开始从“上海市政府采购中心中标公告汇总”网站上爬取信息。信息爬取完毕后，运行信息栏将会显示本次爬取过程中爬虫程序的运行信息，爬取信息栏将会显示本次爬取结果——所有采购成功的公告中的项目名称、成交供应商、采购日期和成交金额。如果公告中没有成交供应商、采购日期或成交金额信息，则相应字段显示-1。

如果点击“关闭”按钮，则关闭界面退出程序。

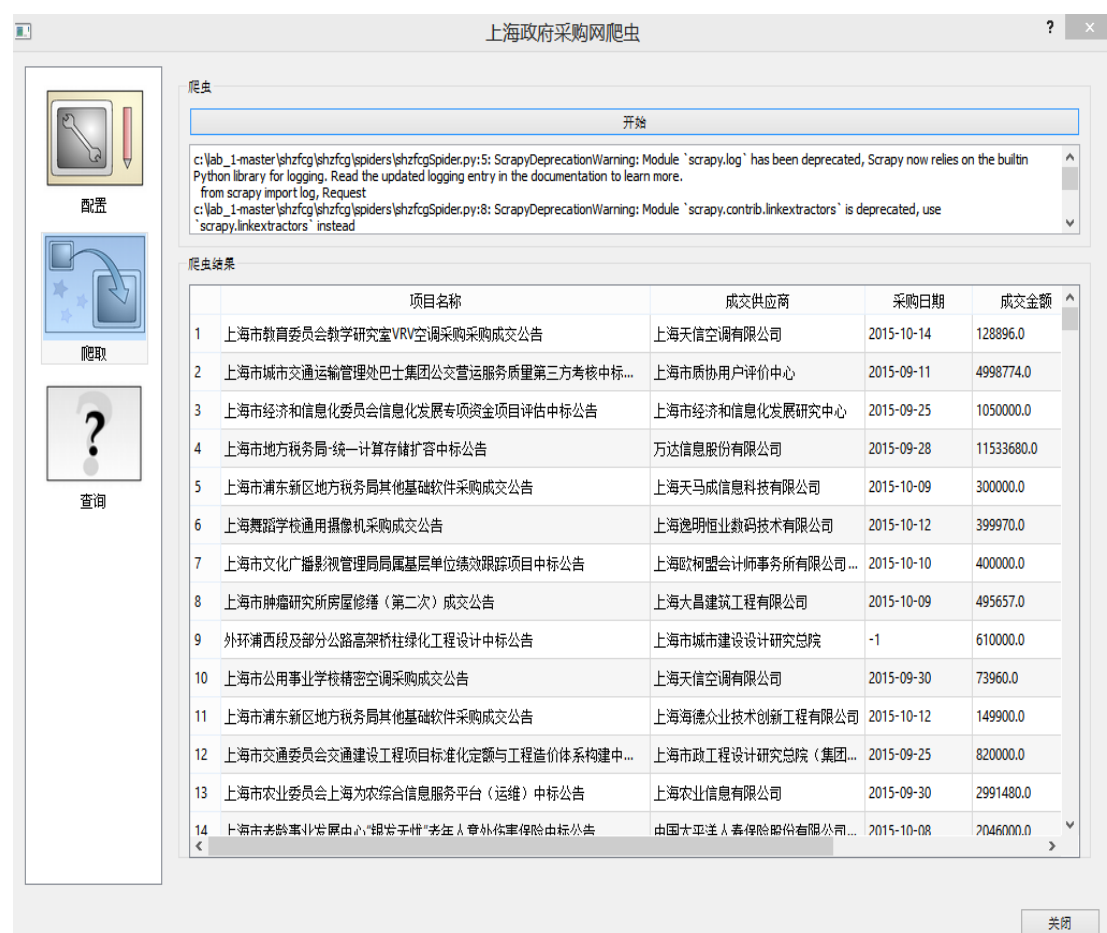


图 4.2 爬取界面

### 4.3 配置功能



点击模块功能按钮中的“配置”按钮，显示相应的配置界面。

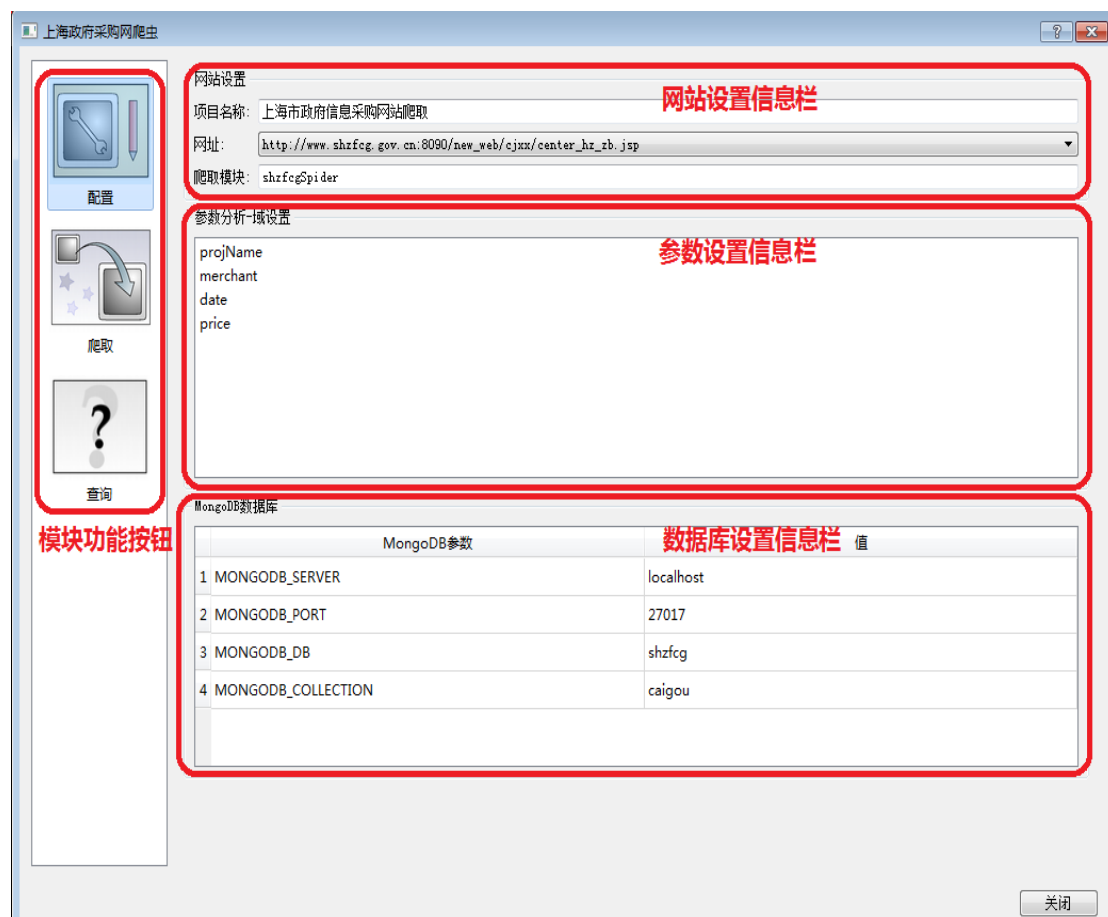


图 4.3 配置界面

配置界面包括模块功能按钮、网站设置信息栏、参数设置信息栏、数据库设置信息栏，其功能如下：

- 模块功能按钮：点击模块功能按钮，进入相应的功能界面；
- 网站设置信息栏：显示项目名称、爬取网站的网址以及爬取模块名称。

如果需要更换爬取网站网址，可以从“网址”的下拉菜单中选择其他网址，默认网址是

[http://www.shzfcg.gov.cn:8090/new\\_web/cjxx/center\\_hz\\_zb.jsp](http://www.shzfcg.gov.cn:8090/new_web/cjxx/center_hz_zb.jsp) (上海市政府采购中心中标公告汇总)；



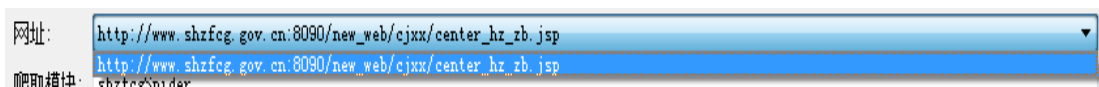

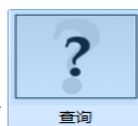


图 4.4 更改网址

- c. 参数设置信息栏：显示爬取信息对应的参数名；
- d. 数据库设置信息栏：显示数据库的相关配置信息。

如果点击  按钮，则关闭界面退出程序。

## 4.4 查询功能



点击模块功能按钮中的“查询”按钮，显示相应的查询界面。



图 4.5 查询界面

查询界面包括模块功能按钮、查询条件设置栏、查询结果信息栏，其功能如下：

- a. 模块功能按钮：点击模块功能按钮，进入相应的功能界面；
- b. 查询条件设置栏：设置查询条件，包括项目名称、供应商、时间起始范围和金额范围，具体操作如下：

1) 设置项目名称。可以在“项目名称”后的文本框

项目名称:  中输入待查询内容，也可以不输入任何内容，查询结果为项目名称包含输入字段的项目；

2) 设置供应商。可以在“供应商”后的文本框

供应商:  中输入待查询内容，也可以不输入任何内容，查询结果为供应商名称包含输入字段的项目；

3) 设置时间起始范围。在图 4.6 所示的日历中选择起始日期(默认为 2015 年 1 月 1 日)，在图 4.7 所示的日历中选择截止日期(默认为操作系统显示的日期)；




图 4.6 选择开始日期

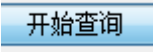


图 4.7 选择截止日期

4) 设置金额范围。在

金额下限: 0 元 和  
金额上限: 10000 万元 中可以手动输入  
数字, 也可以通过点击  来改变数字大小。默认的金额下限为 0 元,  
上限为 10000 万元。设置的金额下限必须在 0~1000000 元之间, 金额  
上限必须在 0~1000000 万元之间。

c. 查询结果信息栏: 显示符合查询条件的所有项目信息。

查询条件设置完毕后, 点击  按钮, 程序将会在已经爬取到的项目  
信息中查询符合条件的项目, 并将结果显示在查询结果信息栏中。

点击  按钮, 程序将会清除查询结果信息栏中的所有内容。

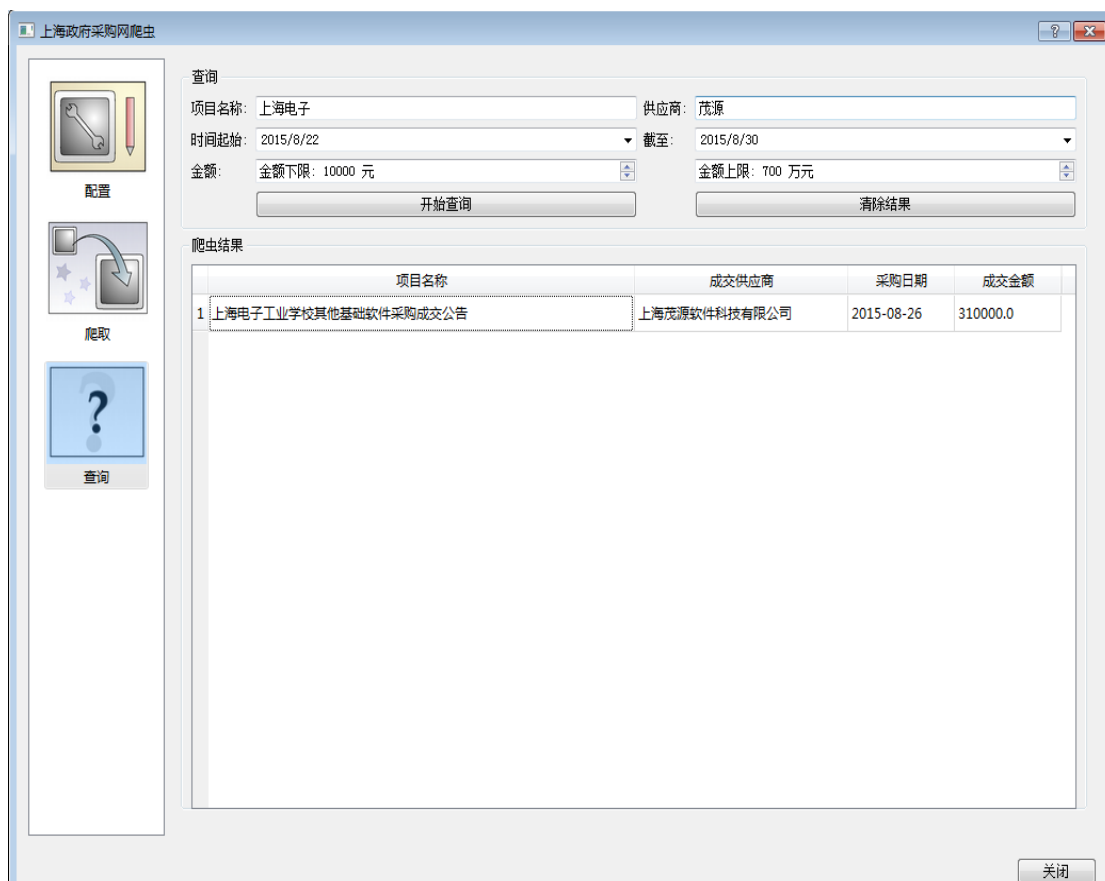
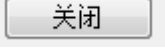


图 4.8 查询示例

如果点击  按钮，则关闭界面退出程序。

## 5 附录 运行环境搭建说明

运行环境可以搭建在 Windows XP、Windows Vista、Windows 7 和 Windows 8 操作系统上。

### 5.1 安装 Python 2.7

从 Python 官方网站上下载 Python 2.7.10.msi，运行 Python 2.7.10.msi 安装好 Python 2.7 后将 Python 的根目录和 Scripts 目录添加到环境变量中。

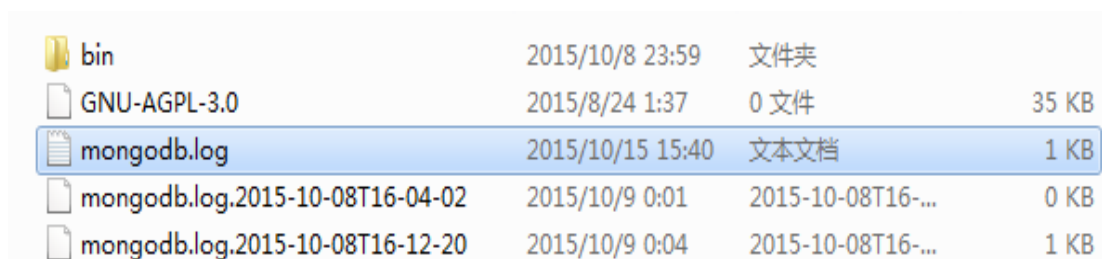
## 5.2 安装 MongoDB

下载并运行 MongoDB 的安装程序 `mongodb-win32-i386-3.0.6-signed.msi`，假设安装目录是“C:\Program Files (x86)\MongoDB”，则在该安装目录下创建 db 文件夹，在“C:\Program Files (x86)\MongoDB\Server\3.0”目录下创建一个 `mongodb.log` 文件。



db	2015/10/9 10:47	文件夹
Server	2015/10/8 23:59	文件夹

图 5.1 创建 db 文件夹



bin	2015/10/8 23:59	文件夹	
GNU-AGPL-3.0	2015/8/24 1:37	0 文件	35 KB
mongodb.log	2015/10/15 15:40	文本文档	1 KB
mongodb.log.2015-10-08T16-04-02	2015/10/9 0:01	2015-10-08T16-...	0 KB
mongodb.log.2015-10-08T16-12-20	2015/10/9 0:04	2015-10-08T16-...	1 KB

图 5.2 创建 mongodb.log

在目录“C:\Program Files (x86)\MongoDB\Server\3.0\bin”下的命令程序执行“`mongod -dbpath “C:\Program Files (x86)\MongoDB\db” --logpath “C:\Program Files (x86)\MongoDB\Server\3.0\mongodb.log” --install --serviceName “MongoDB”`”命令，设置数据路径并安装 MongoDB 服务。

在目录“C:\Program Files (x86)\MongoDB\Server\3.0\bin”下的命令程序执行“`net start MongoDB`”，启动 MongoDB 服务。

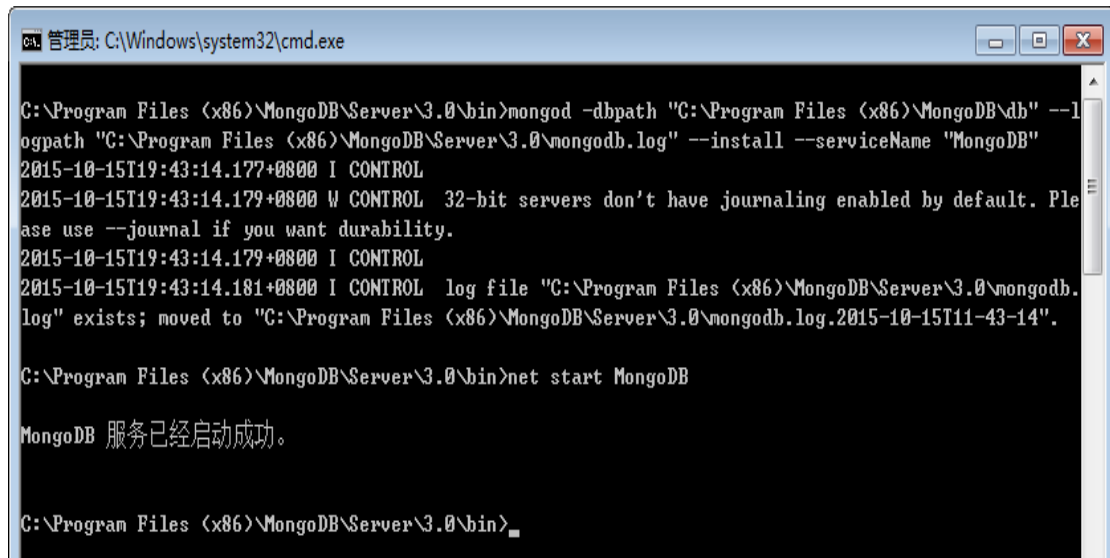


图 5.3 设置数据路径并安装 MongoDB 服务

在命令程序中执行“pip install pymongo”命令，安装 PyMongo。

### 5.3 安装 Scrapy

在命令程序中执行“pip install scrapy”命令。

### 5.4 安装 PyQt 4

下载 PyQt 4 的安装包 PyQt-4.11.4-gpl-Py2.7-Qt4.8.7-x32.exe 和 sip-4.16.9.zip，将 sip-4.16.9.zip 解压到 python 安装目录下的 Lib\site-packages 目录下。打开 Visual Studio 2008 Command Prompt，定位到 sip-4.16.9.zip 的解压目录，先执行“python configure.py”命令，再执行“nmake”命令，最后执行“nmake install”命令。运行 PyQt-4.11.4-gpl-Py2.7-Qt4.8.7-x32.exe 安装程序成功后，PyQt 4 安装完毕。

### 5.5 安装 PyWin32

以管理员身份在命令程序中运行“pip install pywin32”命令。

完成了 5.1 至 5.5 的所有步骤后“基于 Scrapy 的网络信息搜索工具”所需的运行环境搭建完成。