**使用注意事项：**

（1）本技术需配合SQL Server数据库使用，请参考附录安装SQL Server 2008。

（2）获取图片

在获取图片时，需要用户自己输入要抓取的根网址，这里的根网址需要保证其真实存在，如果不存在则会提示网页源代码获取失败，从而导致任务启动失败。

（3）搜索参数设置

在搜索参数设置的时候，需要慎重考虑其各个参数之间的制约关系，有以下几种情况需要说明。

如果搜索深度设置过大，那么就需要将抓取数据容量设置的大一些，否则任务会因为抓取容量不足而被迫停止抓取工作。

如果能够保证自己的网络连接正常，则选择不使用代理服务器。因为使用代理服务器，在任务执行过程中，如果代理服务器失效或者停止活动，则需要重新选择代理服务器，这中间会耗费大量时间来更换代理服务器。

在选择开始执行时间时，不易将执行时间设置的过长，因为设置的过长会使任务长期处于等待状态，建议只有在任务较少的情况下这样设置。

（4）关键字管理

在关键字管理中，需要注意的是在添加主关键字的时候，要同时添加一个与主关键字名字相同的子关键字，因为在图片进行分类的时候，首先比对的是子关键字，只有子关键字匹配之后才能确定其分类归属。

另外需要注意的是，删除主关键字时会将与其相关的所有子关键字一同删除。

（5）多任务

鉴于对任务耗费资源的考虑，同时执行的任务个数设定为5个，超过5个的任务只能处于等待状态，只有5个被执行任务中有结束的，处于等待的任务才能被执行。

（6）图片爬取工具支持网站：

|  |  |
| --- | --- |
| **编号** | **网址** |
| 1 | http://tupian.zol.com.cn |
| 2 | http://www.nipic.com |
| 3 | http://photo.sina.com.cn/ |
| 4 | http://www.veerchina.com/ |
| 5 | http://www.smugmug.com/ |
| 6 | http://www.worth1000.com/ |

自定义的爬取网站设置提取标记后也无法提取，属于正常现象，这是由于网站的特殊设计引起的。

### 附录1 安装SQL Server数据库及相关配置

1、安装SQL Server 2008

网络爬虫使用的数据库为SQL Server，使用前需要下载SQL Server 2008版本。

具体安装步骤可以直接在百度中搜索“sql server 2008 安装图解”关键字，参考其中的安装步骤进行，安装成功后，可以在应用程序中看到如下图所示。

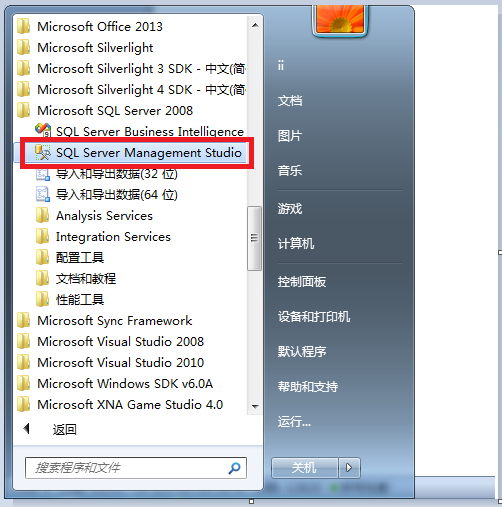


图1 安装成功

2、创建新登录名、数据库和数据表

启动连接服务器，首先以Windows身份验证进入服务器。



图2 以Windows身份验证登陆

创建新的登录名和密码，这个登录名和密码用作系统连接数据库时的用户名和密码，进入数据库后，选择【安全性】---> 【登录名】，右键选择【新建用户名】--->选择【SQL Server身份验证】，填写登录名和密码，其他选择默认即可。

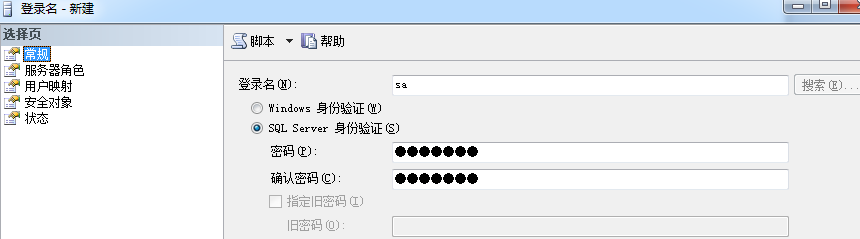


图3 新建登录名

查看【状态】选项，选择【授予】和【启用】，否则会导致使用该登录名登录失败，点击【文件】菜单项，选择断开连接，重新登录，选择SQL Server身份验证，输入新创建的登录名和密码，重新进入数据库。

进入服务器后，右键【数据库】--->选择【新建数据库】，输入数据库的名字，在图片系统中数据库名为CrawlWorm-Picture，如图4所示。

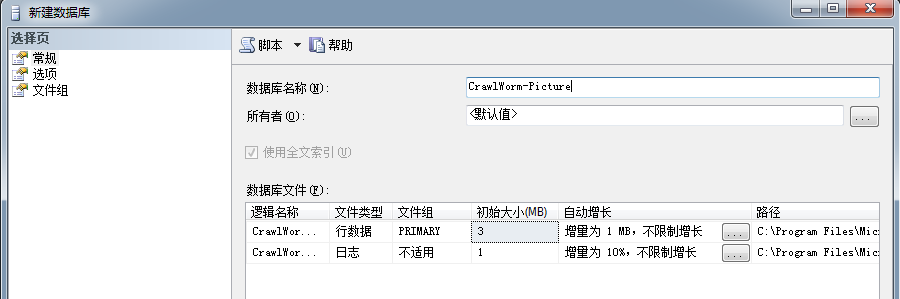


图 4 新建数据库

新建数据库后，右键【数据库】选择【刷新】，点击左边的小加号可以看到新建的数据库，右键新建的数据库，选择【编写数据库脚本为】--->【CREATE到】--->【新查询编辑器窗口】，可以看到右边出现的一个带脚本的窗口。清空里面的脚本，然后通过脚本创建数据库，创建的过程类似，只需要清空脚本之后在脚本窗口里输入创建相应表的脚本即可（具体的脚本请查看附录2）。

利用脚本创建各表之后，选择【数据库】下的【表】，右键【表】选择【刷新】，可以看到创建的表，如图5所示。

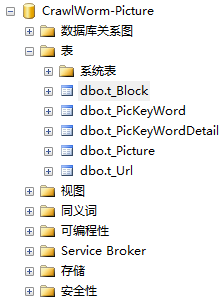


图5 已创建表

3、配置数据源

创建数据库之后，还不能被系统使用，需要配置数据源。点击【开始】→选择【控制面板】→【管理工具】→【数据源】。选择【系统DSN】选项卡，添加“SQL Server”，然后按照图6、图7、图8所示的信息填写，点击【下一步】，直到最后完成。完成之后需要测试数据源，测试成功后即完成配置数据源，如果失败，请重新配置。

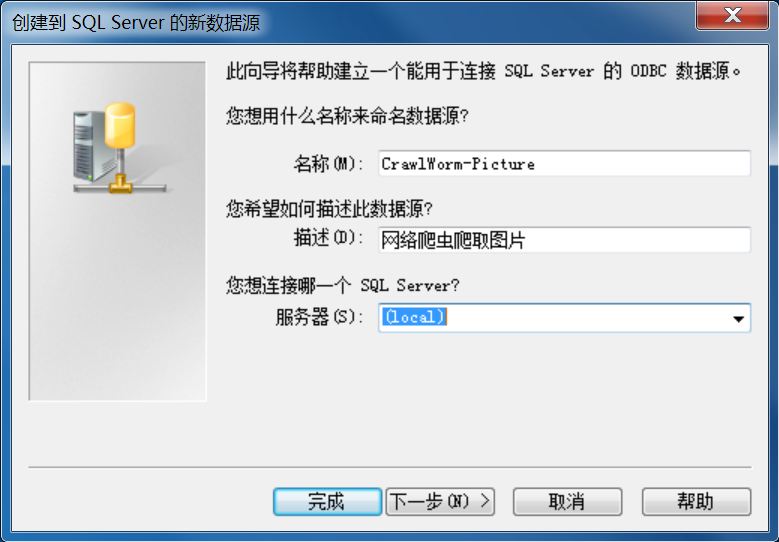


图6

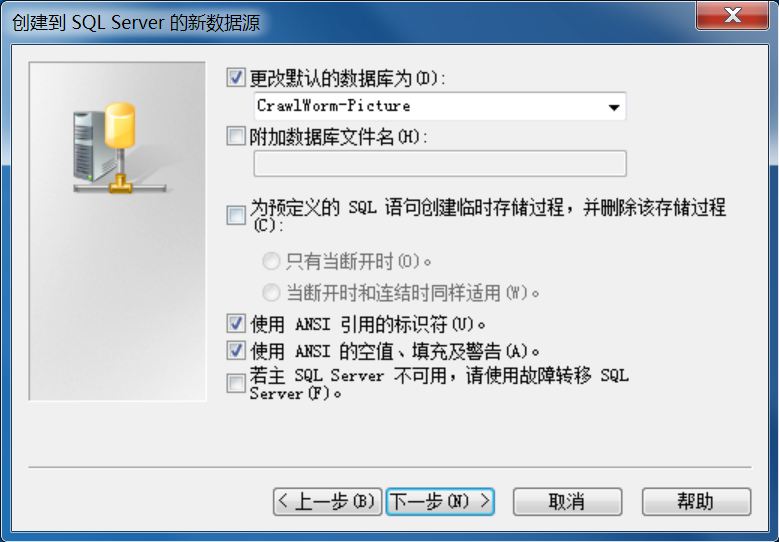


图7



图8

4、创建配置文件

在运行程序之前，需要检查数据库连接的配置文件是否存在，其中的配置内容是否正确。如果不正确会提示数据库连接失败，并且程序无法启动或者启动后无法继续其他操作。如果想要修改数据库连接，可以通过修改配置文件中的内容实现。

配置文件内容样例如下：

[SQLConnection-SQLServer]

Provider="SQLOLEDB"

Server="192.168.0.119"

Database="CrawlWorm-Picture"

uid="sa"

pwd="User123"

其中需要修改的参数有三个，分别是Server、Database、uid和pwd，其他的保持不变。其中Server为所连接数据库所在的服务器IP，Database为所创建的数据库的名字，这里数据库名必须为CrawlWorm-Picture，uid为登陆SQL Server数据库的用户名，pwd为登陆密码。如果运行程序后提示连接数据库失败，首先检查这些参数设置是否正确。

### 附录2 创建CrawlWorm-Picture数据库各表相关脚本

（1）创建t\_Url表的脚本

CREATE TABLE [dbo].[t\_Url](

[ID] [int] IDENTITY(1,1) NOT NULL,

[Url] [nvarchar](max) NULL,

[RootUrl] [nvarchar](max) NULL,

[ParentUrl] [nvarchar](max) NULL,

[Depth] [int] NULL,

[AddTime] [datetime] NULL,

CONSTRAINT [PK\_t\_Url] PRIMARY KEY CLUSTERED

(

[ID] ASC

)WITH (PAD\_INDEX = OFF, STATISTICS\_NORECOMPUTE = OFF, IGNORE\_DUP\_KEY = OFF, ALLOW\_ROW\_LOCKS = ON, ALLOW\_PAGE\_LOCKS = ON) ON [PRIMARY]

) ON [PRIMARY]

GO

EXEC sys.sp\_addextendedproperty @name=N'MS\_Description', @value=N'网址ID' , @level0type=N'SCHEMA',@level0name=N'dbo', @level1type=N'TABLE',@level1name=N't\_Url', @level2type=N'COLUMN',@level2name=N'ID'

GO

EXEC sys.sp\_addextendedproperty @name=N'MS\_Description', @value=N'网址' , @level0type=N'SCHEMA',@level0name=N'dbo', @level1type=N'TABLE',@level1name=N't\_Url', @level2type=N'COLUMN',@level2name=N'Url'

GO

EXEC sys.sp\_addextendedproperty @name=N'MS\_Description', @value=N'首页网址' , @level0type=N'SCHEMA',@level0name=N'dbo', @level1type=N'TABLE',@level1name=N't\_Url', @level2type=N'COLUMN',@level2name=N'RootUrl'

GO

EXEC sys.sp\_addextendedproperty @name=N'MS\_Description', @value=N'父网址' , @level0type=N'SCHEMA',@level0name=N'dbo', @level1type=N'TABLE',@level1name=N't\_Url', @level2type=N'COLUMN',@level2name=N'ParentUrl'

GO

EXEC sys.sp\_addextendedproperty @name=N'MS\_Description', @value=N'搜索深度' , @level0type=N'SCHEMA',@level0name=N'dbo', @level1type=N'TABLE',@level1name=N't\_Url', @level2type=N'COLUMN',@level2name=N'Depth'

GO

EXEC sys.sp\_addextendedproperty @name=N'MS\_Description', @value=N'添加时间' , @level0type=N'SCHEMA',@level0name=N'dbo', @level1type=N'TABLE',@level1name=N't\_Url', @level2type=N'COLUMN',@level2name=N'AddTime'

GO

ALTER TABLE [dbo].[t\_Url] ADD CONSTRAINT [DF\_t\_Url\_AddTime] DEFAULT (getdate()) FOR [AddTime]

（2）创建t\_Picture表的脚本

CREATE TABLE [dbo].[t\_Picture](

[ID] [int] IDENTITY(1,1) NOT NULL,

[Picture] [nvarchar](max) NULL,

[RootUrl] [nvarchar](max) NULL,

[ParentUrl] [nvarchar](max) NULL,

[PictureContent] [text] NULL,

[AddTime] [datetime] NULL,

CONSTRAINT [PK\_t\_Picture] PRIMARY KEY CLUSTERED

(

[ID] ASC

)WITH (PAD\_INDEX = OFF, STATISTICS\_NORECOMPUTE = OFF, IGNORE\_DUP\_KEY = OFF, ALLOW\_ROW\_LOCKS = ON, ALLOW\_PAGE\_LOCKS = ON) ON [PRIMARY]

) ON [PRIMARY] TEXTIMAGE\_ON [PRIMARY]

GO

EXEC sys.sp\_addextendedproperty @name=N'MS\_Description', @value=N'图片ID' , @level0type=N'SCHEMA',@level0name=N'dbo', @level1type=N'TABLE',@level1name=N't\_Picture', @level2type=N'COLUMN',@level2name=N'ID'

GO

EXEC sys.sp\_addextendedproperty @name=N'MS\_Description', @value=N'图片名称' , @level0type=N'SCHEMA',@level0name=N'dbo', @level1type=N'TABLE',@level1name=N't\_Picture', @level2type=N'COLUMN',@level2name=N'Picture'

GO

EXEC sys.sp\_addextendedproperty @name=N'MS\_Description', @value=N'首页网址' , @level0type=N'SCHEMA',@level0name=N'dbo', @level1type=N'TABLE',@level1name=N't\_Picture', @level2type=N'COLUMN',@level2name=N'RootUrl'

GO

EXEC sys.sp\_addextendedproperty @name=N'MS\_Description', @value=N'父网址' , @level0type=N'SCHEMA',@level0name=N'dbo', @level1type=N'TABLE',@level1name=N't\_Picture', @level2type=N'COLUMN',@level2name=N'ParentUrl'

GO

EXEC sys.sp\_addextendedproperty @name=N'MS\_Description', @value=N'图片内容' , @level0type=N'SCHEMA',@level0name=N'dbo', @level1type=N'TABLE',@level1name=N't\_Picture', @level2type=N'COLUMN',@level2name=N'PictureContent'

GO

EXEC sys.sp\_addextendedproperty @name=N'MS\_Description', @value=N'添加时间' , @level0type=N'SCHEMA',@level0name=N'dbo', @level1type=N'TABLE',@level1name=N't\_Picture', @level2type=N'COLUMN',@level2name=N'AddTime'

GO

ALTER TABLE [dbo].[t\_Picture] ADD CONSTRAINT [DF\_t\_Picture\_AddTime] DEFAULT (getdate()) FOR [AddTime]

GO

（3）创建t\_PicKeyWordDetail表的脚本

CREATE TABLE [dbo].[t\_PicKeyWordDetail](

[ID] [int] IDENTITY(1,1) NOT NULL,

[KeyWordDetail] [nvarchar](50) NULL,

[KeyWordID] [int] NULL,

[AddTime] [datetime] NULL

) ON [PRIMARY]

GO

ALTER TABLE [dbo].[t\_PicKeyWordDetail] ADD CONSTRAINT [DF\_t\_PicKeyWordDetail\_AddTime] DEFAULT (getdate()) FOR [AddTime]

（4）创建t\_PicKeyWord表的脚本

CREATE TABLE [dbo].[t\_PicKeyWord](

[ID] [int] IDENTITY(1,1) NOT NULL,

[KeyWord] [nvarchar](50) NOT NULL,

[IsEnglish] [bit] NULL,

[AddTime] [datetime] NULL

) ON [PRIMARY]

GO

ALTER TABLE [dbo].[t\_PicKeyWord] ADD CONSTRAINT [DF\_t\_PicKeyWord\_AddTime] DEFAULT (getdate()) FOR [AddTime]

（5）创建t\_Block表的脚本

CREATE TABLE [dbo].[t\_Block](

[ID] [int] IDENTITY(1,1) NOT NULL,

[HostName] [nvarchar](50) NULL,

[StartBlock] [nvarchar](max) NULL,

[EndBlock] [nvarchar](50) NULL,

[AddTime] [datetime] NULL

) ON [PRIMARY]

GO

ALTER TABLE [dbo].[t\_Block] ADD CONSTRAINT [DF\_t\_Block\_AddTime] DEFAULT (getdate()) FOR [AddTime]

GO