基于网络爬虫的股票信息预警系统的设计与实现

摘要:通过网络爬虫抓取股票数据,研究股票信息预警系统的设计与实现,通过系统为用户提供股票多条件的预警与公告推送功能。

关键词:网络爬虫；股票预警；WEB挖掘；

1绪论

伴随着互联网等相关技术的发展股票交易方式由早期的手工操作、电话委托、柜台交易直至今天的在线股票交易平台。同时用户所能获取的股票相关信息无论是在数量上还是纬度上都呈爆炸增长的趋势。用户所能获得的信息从原来的匮乏时代一下子来到了信息过载时代。因此用户需要一种手段可以对这些信息进行过滤，只获取自己关注的信息。

本文的主要目标就是通过使用爬虫技术对于股票这个垂直领域中的数据、信息进行采集、针对采集的网页进行解析、挖掘、存储。在通过用户的预设条件进行实时的信息预警与公告推送。

2系统分析

面对大量的实时股票信息， 只有通过手段进行有效的过滤，才能满足用户的需求。通过使用网络爬虫技术与数据库相结合的方式，来实现股票信息预警系统。网络爬虫是一种按照一定的规则自动抓取互联网信息的程序或脚本，可以采集分散在网络上的信息，以便更有效的获取所需的信息。从而大大提高用户对于股票数据、信息获取的效率。

本文采用C# .Net基类提供的工具自己构建一个简易的网络爬虫来实现股票数据、信息的采集。配合Windows Service的预警推送与Windows Forms 客户端的设置，实现从采集到设置到推送的股票信息预警系统。爬虫部分首先通过HttpWebResponse获取网页。之后通过Xpath 提取网页中的信息。最后对抓取到的信息进行处理存入数据库中。最后通过比对数据库中的预设条件与数据，通过短信、邮件的形式将满足条件的信息推送给用户。

3系统设计

3.1 工作原理

系统由客户端获取用户设置的股票预警条件与公告推送要求。将这些信息处理后存入数据库中。然后利用网络爬虫对指定的数据源进行实时抓取。由于部分股票指数需要历史数据的参与，因此爬虫将对所有股票进行信息、数据抓取。本文的爬虫抓取列表为上证A股。对抓取的数据进行处理写入数据库。最后预警Service 比较动数据库中的最新数据与用户设定的预警。如果符合预警条件，进行推送，如果不符合条件，则继续等待下次比对。

4具体实现

本系统采用模块化开发，将系统分为了股票信息获取模块、股票预警模块、数据浏览模块等部分。由于篇幅的限制，主要介绍股票信息获取模块。

4.1股票信息获取

本系统获取的股票数据包括股票快照数据、日交易数据、历史交易数据、财务数据以及股票发布公告。为了较全面的开发爬虫可抓取的网页，数据的主要来源分别有新浪公用数据接口，上海证券交易所官网（动态加载），东方财富网（静态页面）。

对于从数据源获取的数据需要进行格式化处理。由于考虑到股票信息字段越多则用户可定制的预警条件越丰富，本系统不舍弃数据源的信息。比如快照数据包括：股票名称、

行情时间、最新价、昨收盘、今开盘、涨跌额、最低、最高、涨跌幅、成交量、成交额、竞买价、竟卖价、委比、买一一买五、卖一一卖五。通过JsonToObject的方式将接口数据转化为预设的快照对象并最后存入数据库中。对于其他数据源本系统也进行类似的处理过程区别在于对于静态页面采用Xpath进行对象化处理，而动态加载页面则通过分析网站的Js请求，获取对应数据。

具体的实现方法以新浪股票公用数据源获取数据为例：

* 1. 从数据库中获取股票列表，并对列表中的股票代码进行遍历
  2. 拼接URL字符串。新浪的共用接口调用地址为<Http://hq.sinajs.cn/list=shXXXXXX>，其中XXXXXX为具体股票代码。
  3. 通过爬虫发送get请求，获取当前的股票信息。
  4. 对获取的数据进行格式化处理，并转化为对象。
  5. 数据保存至数据库。

4.2股票预警

实现股票信息的预警，需要两个要素。其一是用户设置的预警条件，其二是股票相关的实时信息。通过实时的对这两个要素进行比较，就可以实现对用户设置条件的预警。由于用户的设置需要持久化保存，同时存在大量需要比较的数据因此将这两个要素都存储在数据库中。

在运行时预警系统首先获取用户设置的预警信息，取得需要优先抓取的数据。然后执行SQL语句，根据当返回值（0或1）为1时按用户选择的方式推送预警内容。当预警被出发后，修改该条用户预警设置的状态为close，避免重复对用户进行预警。对于未能触发的预警则将其保持其状态为open以待下一次程序执行。

5总结

本文采用C#的 .Net基类提供的工具自己构建一个简易的网络爬虫，通过详细的系统设计分析，完成了用户设置信息、获取股票数据、信息，股票信息预警，公告推送，短信发送，邮件发送等模块的开发，实现股票信息预警系统从而提高了用户对于股票信息获取的效率。

参考文献

[1]胡婧, 叶建木. 基于微博信息的股票交易预测研究[J]. 财政监督, 2017(5):108-111.

[2]陈亮华. 基于网络爬虫的基金信息抽取与分析平台[D]. 华南理工大学, 2010.

[3]温玲. 基于爬虫技术的股价分析系统的设计与实现[D]. 北京大学, 2014.

[4]Griffiths LAN, Programing C#5.0 2012

[5]张世明等.C#编程语言基础和应用 中国铁道出版.2011

[6]Kumar N, Singh M. Framework for Distributed Semantic Web Crawler[C]// International Conference on Computational Intelligence and Communication Networks. IEEE, 2016:1403-1407.