第38卷 第13期
Vol.38 No.13

计 算 机 工 程
Computer Engineering

2012年7月
July 2012

·开发研究与设计技术· 文章编号：1000—3428(2012)13—0254—03 文献标识码：A 中图分类号：TP301.6

基于网络论坛的股市分析方法

吴 晶[1]，陈仪香[1]，刘道明[2]

(1. 华东师范大学上海市高可信计算重点实验室，上海 200062；2. 光大证券股份有限公司研究所，上海 200040)

摘 要：为更好地揣摩大众股民的心理及情感，提出一种基于网络论坛的股市分析方法。根据python实现相应的网络爬虫，利用该爬虫获取网络论坛中的所有帖子，对每日新帖子的数量进行统计分析，针对每个帖子中的文本内容设计分析工具，以进行情感分析，并将这些情感结果进行统计。实验结果表明，通过对比同一时期内的中国股市走势图，该方法能对其进行较为准确的分析。

关键词：网络爬虫；股市分析；情感分析；网络论坛

Internet-forum-based Stock Market Analysis Method

WU Jing[1], CHEN Yi-xiang[1], LIU Dao-ming[2]

(1. Shanghai Key Laboratory of Trustworthy Computing, East China Normal University, Shanghai 200062, China;
2. Research Institution, Everbright Securities Co., Ltd., Shanghai 200040, China)

【Abstract】In order to figure out the emotion of the public shareholders, this paper proposes an Internet-forum-based stock market analysis method. A python-based Web crawler is implemented. All posts on the Internet forum are downloaded using this crawler, and analyzes the number of new posts daily. The sentimental analysis tool is also designed to judge the posts' sentiment, and the final result is counted. Experimental results show that by comparing Chinese stock's movement in the same period, it can be found that these methods can be used in relatively accurate analysis.

【Key words】Internet crawler; stock market analysis; sentiment analysis; Internet forum

DOI: 10.3969/j.issn.1000-3428.2012.13.076

1 概述

随着互联网的快速发展，人们越来越多地依赖于网络。人们在网络上购物、聊天和消遣。而网络也逐渐成为广大网民展示内心情感的重要场所。同样的，网络也成为各类分析的重要数据来源。人们利用网络数据进行舆情分析、感情分析等。文献[1]利用人们在网络上的言论分析产品的特性和大众对该产品的情感倾向。而利用网络数据对股市进行一定的分析和预测也开始被一些学者所关注。文献[2]从Twitter上人们发表的各类状态提取情感并进行相应的统计，对股市进行了有效的预测，并已实际地应用于国外某量化基金。

近年来作为广大股民发表观点以及展示情感的场所，各类股吧论坛异常火爆。东方财富网下的股吧论坛(http://guba.eastmoney.com/)平均每天新发表的帖子达到30 000多条，而在线人数更是达到数百万。因此，该类股吧论坛或许可以成为获取相关股民数据的较好来源。本文提出一种分析方法对中国股市进行分析，以更好地揣摩大众股民的心理及情感。

2 分析方法

2.1 基于论坛发帖量的股市分析

在早期的股市中，流传着这样的一个指标，即证券营业部门口的自行车数量，该指标指出若发现营业部门口的自行车数目很少的时候，就可以买入股票了；当发现营业部门口的自行车数量暴增时，就可以卖出股票了。同样的，对于所拥有的论坛数据，希望可以利用每天论坛上新发表的帖子总数来替代营业部门口的自行车数目，将其做成一个指标，能够对于股市有一定的预测性。

同时，考虑到可能股票市场的下跌也会同样影响股民参与论坛的程度，那么论坛中新发表的帖子数量应该与股市成交量有较直接的同步关系，本文将对该结论进行验证。

2.2 基于论坛帖子主题情感的股市分析

若能够利用情感分析判别出大众股民针对股市的情感倾向，那么对于投资者做出相应的决策将会有很大的帮助。然而，由于中文的复杂特性，因此情感分析一直是中文处理的一个难点。但国内也有学者进行过研究，如文献[3-8]，但专门针对股市中的观点分析却少有学者涉及。在下面的实验中，将利用一个较为简单方法模型针对股市情感进行判别。而且，经过相应的测试，发现其有较高的正确率。

利用以下方法来判别论坛的帖子主题表明的立场，即看空还是看多。看多表明投资者认为大盘指数会上涨；看空则相反。具体的做法如下：

(1)构造一个含有较大数据的字典，该字典内包含了几乎所有的证券可能会用到的名词、动词、形容词以及相应的专有名词和各个行业常用词等。

(2)基于构造的字典以及一些常用词字典，对论坛中各个帖子主题进行分词。

(3)基于分好的词，对所有的词进行词频统计，针对于出现频率较大的词，挑选其中代表情感的词语，并对其人工进行标记，若认为其为正面，则标记为1，反之标记为−1。

基金项目：国家“973”计划基金资助项目(2011CB302802)；国家自然科学基金资助项目(61021004)；上海市自然科学基金资助项目(10ZR1410000)

作者简介：吴 晶(1987－)，男，硕士，主研方向：网络技术，可信计算；陈仪香，教授、博士生导师；刘道明，硕士

收稿日期：2011-11-01 **E-mail：**wujin05@yahoo.cn

(4)计算句子的情感。将每个句子分词，并按照各个标点分开，若每一个小句内出现标注了情感的词语，而该小句的情感为该词的情感。若出现多个该类词语，则该小句的情感为各个标记词语的情感之和除以标记词语的总个数，若该小句中出现各类否定词，则将小句的得分乘以−1，若有 2 个否定词，再乘以−1，依次递推。整个主题的情感为所有分句得分之和，若最终该得分为正，则认为该帖子为看多。反之，认为看空。即定义每个帖子的主题得分为：

$$score=\sum_i(-1)^{l_i}\left(\frac{\sum_{j<n_i}a_{ij}}{n_i}\right)$$

其中，i 为该主题中分句的个数；l_i 为分句中否定词的个数；n_i 为分句中标记的词语个数；a_{ij} 为标记的词语的得分。若一个主题的 $score$>0，就定义为看多，否则相反。

由于本文是从所有出现频率较高的词语寻找情感词，因此在一定程度上会使得寻找到的情感词比较贴近于论坛股民的用词方式，因此通过标注该类词，可以最大限度地使用较少的人工标注，判别较多的帖子。论坛的主题基本上均是单句，因此，使用上述的计分方式不会导致过多地问题。由于中文语言的博大精深与复杂的结构，因此这样的做法还是会导致一些错误识别，经过粗略的计算，基本可以识别出 70%的带有看多看空情感的帖子，在判别是看多情感还是看空情感上保持 85%以上的正确率。

3　方法实现

由于进行分析的数据是基于大量的网络数据，因此要获取这些数据，首先需要编写爬虫程序。

由于 python 简便的脚本编写方式，以及其具有较为成熟的网络抓取工具 urllib2 和网页解析工具 BeautifulSoup，因此选取 python 来编写爬虫。在程序中主要构建了 2 个类，ThreadUrl 及 DatamineThread。在 ThreadUrl 类中，通过输入网址，爬取相应的网页实体，而在 DatamineThread 中对 ThreadUrl 中获取的网页实体进行解析，以获取需要进一步抓取的网页网址，同时将需要的数据存入数据库中。其中，ThreadUrl 与 DatamineThread 类中的数据传递主要通过 2 个 Queue 类实体 Queue1、Queue2 实现，Queue1 中存储的是待爬网页的网址，Queue2 中存储的是 ThreadUrl 类获取的网页实体。在向 Queue1 类中，在输入相应的需要获取的网页网址时，需要利用数据库中存储的数据判断是否该网址对应的网页已被获取过，以免重复下载信息。爬虫结构如图 1 所示。

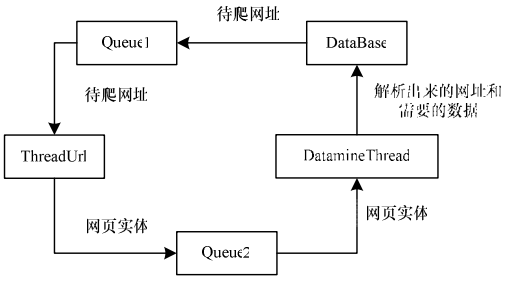

图 1　爬虫结构

在股吧论坛中让爬虫从股吧列表(http://guba.eastmoney.com/gblb.html)开始，由于在该列表中包含了所有子论坛的网址，因此能遍历完各个子论坛。而由于每个子论坛中均有下一页的链接，因此可以遍历该论坛。而由于仅需要获取帖子的名称和发表时间，因此仅需要遍历所有子论坛的帖子列表，而不用进入到各个帖子的主页，使得能够在较短的时间内收集所有的信息。

利用 python 可以较为方便地使用相应的接口调用分词工具。此外，python 现成的字典数据结构使得能够以较高效率完成各项计算。

利用该爬虫程序和相应的工具获取了相应的论坛数据中所有帖子的主题和发表时间，并且利用自行定义的情感计算方式，计算出各个帖子是看空、看多、没有情感。事实上，大概标记了 1 000 个较高频率的情感词。标记“涨”、“利好”、“超跌”、“红”、“抄底”等词为 1，而“跌”、“绿”、“垃圾”等词为−1。情感分析部分结果如图 2 所示。

普准后把握三大热点个股	1
铁股铮铮：市场再酝酿830式的下挫！！！！！！！！！	-1
股市T+0利好你？我？券商？	1
加息日期已定，	-1
惊现第二只西藏矿业！最有投资价值的极度超跌超底新能源股！	1
中信国安三网融合＋锂电池新能源＋子公司楚天网络将上市	1
大盘今日回踩确认，明日翻红！	1
如果你的股票突然跌了一半	-1
给点时间，让市场慢慢造底	0
赶紧抄底，上车机会来了	1
急跌后投资者能否抄底？	0
电力设备：电网招标接踵而至 行业销售进入旺季	1
大盘进入底部构造过程	1
量价分析之经验之谈	0
今天走势不算差正常调整，明天延续反弹（26号走势分析）	1
重点解析长三角概念股	0
私募说了连涨3个板板不值你就重点关注20元见资产重组	1
亚海种业一季度利润大幅增长每股收益0.63元海通证券买入评级	1

图 2　情感分析部分结果

由图 2 可知，通过该方法是可以判别正确的情感。但也会有一些问题，如图 2 中的第 3 行是一个反问句，而却将其标为正面，但总体上通过该方法可以得到一个较好的结果。

最后，对各项数据按周进行统计。同时，通过大智慧下载相应时间的上证指数、个股股价和成交量。

4　结果分析

4.1　基于论坛发帖数量的股市分析结果

统计了 2010 年 7 月-2011 年 6 月每周论坛上新发表的帖子总数，并将其与上证指数在相应日期内的收盘指数做成时间序列，发帖量与上证指数(1)如图 3 所示，其中，横坐标 201027 表示 2010 年第 27 周，201037 表示 2010 年第 37 周，依此类推。

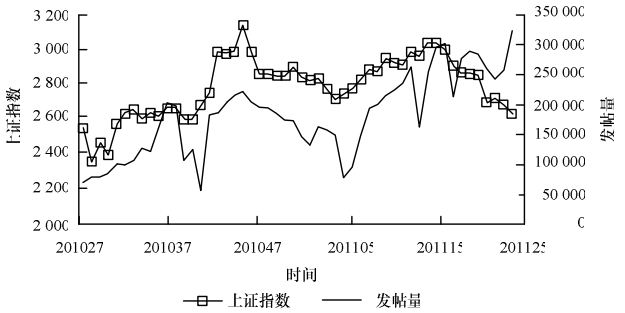

图 3　发帖量与上证指数(1)

由图 3 可知，在论坛新发表的帖子数量与上证指数的关系基本满足传统的营业部自行车数量指标效果，即由于上证指数的上升引起了论坛帖子数量的上升；而当帖子数量急剧上升到一定高度时，指数开始下降；指数上升的起点在帖子数量在低位的时候，因此，通过该指标，可以设计出如下的操作策略，即当发帖量在低位的时候，进行买入操作；当发帖量在高位的时候，进行卖出操作。

针对图 3 进行更进一步的分析，将其以增加过去 4 周 2 倍标准差的布林线为上下限，得到发帖量与上证指数(2)如

图 4 所示。

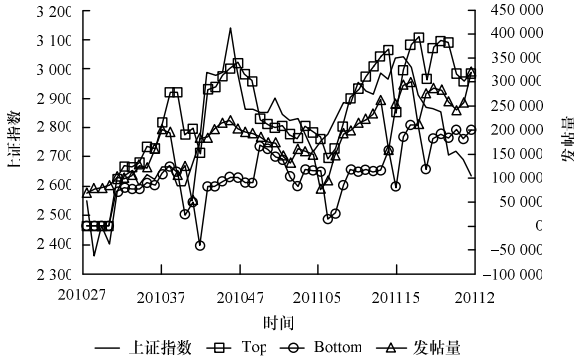

图 4 发帖量与上证指数(2)

根据图 4 设计如下策略：当发帖量到达下限时，买入；当发帖量到达上限时，卖出。根据该策略，在 2010 年 7 月-2011 年 7 月的 10 次策略触发时，可以判断正确 7 次大盘后续的涨跌，但也会错过 10 年 10 月的那一轮牛市行情，因为在牛市开始后不久，该指标就发出卖出信号，同样在 11 年的下跌行情中也存在误判。

统计去年 7 月-今年 6 月每周沪市成交量，并与相应的每周发帖数量作对比，得到发帖量与成交量如图 5 所示。

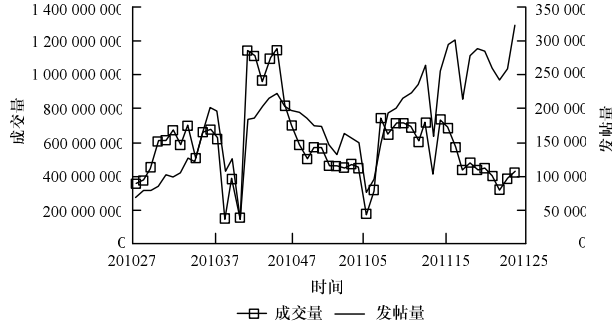

图 5 发帖量与成交量

由图 5 可知，当某一周内成交量放大时，论坛中新发表的帖子数也同步地急剧上升，而当成交量萎缩时，论坛上新发表的帖子数也相应地减少了，其具有相同的拐点时间。

同时，选取了东方电子(000682)股票，对该个股子论坛 2010 年来新发表得帖子数目按周统计，与该股相应日期的股价和成交量进行对比，对数据进行了相应的平滑处理，即对初始数据按 4 周进行移动平均计算得平滑值，得到东方电子发帖量与股价如图 6 所示，和东方电子发帖量与成交量如图 7 所示。由图 6、图 7 可知，在东方电子(000682)股票分论坛上发表的帖子数与该股票相应的股价与成交量具有一定的相关性，但由于个股论坛中的数据量还是较小，因此其中的效果比大盘中分析的效果差。

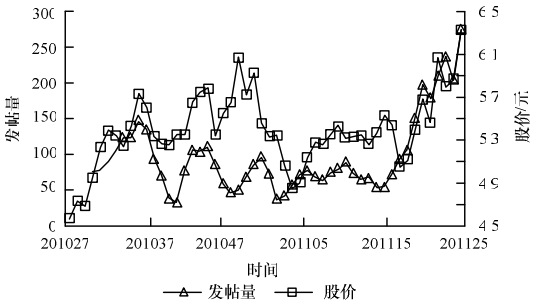

图 6 东方电子发帖量与股价

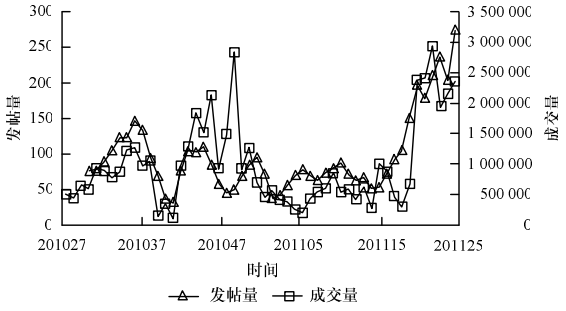

图 7 东方电子发帖量与成交量

4.2 基于论坛帖子主题情感的股市分析结果

基于第 2 节中对情感的分析方法按周统计各周类看多和看空的帖子数，定义多空比为看多帖子数目除以看空帖子数目，将其与相应的日期的上证指数进行对比，得到多空比与上证指数(1)如图 8 所示。

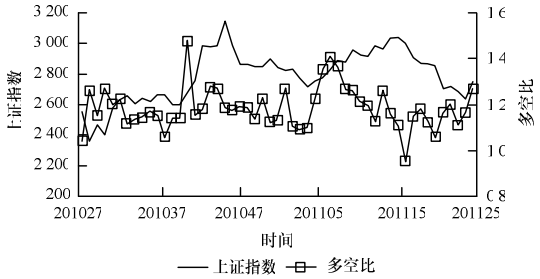

图 8 多空比与上证指数(1)

由图 8 可知，当定义的多空比数值急剧上升到一定高度时，预示着一轮较好行情的来临，而当多空比数值急剧下降到底点时，预示着一轮较差行情。

同样的，增加过去 4 周 2 倍标准差的布林线为上下限，得到多空比与上证指数(2)如图 9 所示。

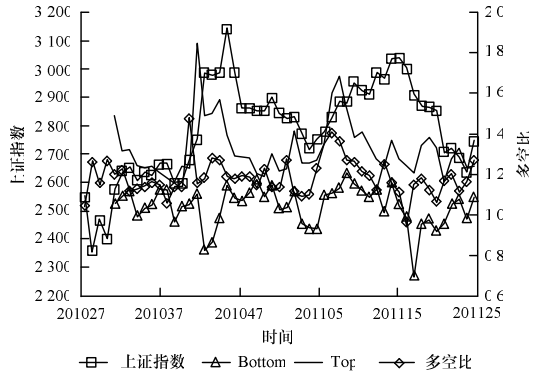

图 9 多空比与上证指数(2)

那么可以制定策略：当多空比穿过上限线时，买入；当多空比穿过下限线时，卖出。根据该策略，在 2010 年 7 月-2011 年 7 月的 15 次策略触发的时候，可以判断正确 12 次大盘后续的涨跌情况。

5 结束语

本文提出一种基于网络论坛的股市分析方法。利用论坛上获取的数据对股市进行分析。实验结果表明该方法的有效性。今后的研究方向为：基于论坛数据获取相应表示愤怒的帖子，统计其总数与大盘指数的关系；按照同样的框架以及中文处理方法，完成相应的爬虫下载财经新闻，并判别新闻关注热点以及个股相关性。

(下转第 259 页)