基于Web大数据挖掘的证券价格波动实时影响研究

杨 莎[1,2] 余 伟[1] 李石君[1] 曹晶晶[1] 刘 晶[1]
(武汉大学计算机学院 武汉 430079)[1] (汉口学院计算机科学与技术学院 武汉 430212)[2]

摘 要 随着Web大数据的发展,互联网中海量、快捷的信息为证券市场变化预测提供了丰富的数据支撑,如何利用大数据分析技术进行实时可靠的证券市场价格变化预测成为重要的科学问题。从证券市场价格变化的核心价值问题研究出发,分析了股票价值所反映的基本面要求,建立了影响股票价值内涵和价格表现的10项准确可度量的特征因素:经济周期、财政政策、利率变动、汇率变动、物价变动、通货膨胀、政治政策、行业变化、经营状况、上下游影响等。在此基础上,构造互联网中信息内容与各个特征因素的提取方法、变化关系和影响模型,提出了针对大盘、行业、个股的互联网信息指标来反映Web数据对其的支撑程度,最终实现了基于Web大数据的综合特征因素度量来预测证券市场的方法。实验表明,该方法具有良好的可行性,将带来明显的学术和商业价值。

关键词 数据挖掘,股票价格预测,Web大数据
中图法分类号 TP311 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.4.033

Research on Stock Price Real-time Fluctuation Influence Based on Web Big Data Mining

YANG Sha[1,2] YU Wei[1] LI Shi-jun[1] CAO Jing-jing[1] LIU Jing[1]
(School of Computer, Wuhan University, Wuhan 430079, China)[1]
(College of Computer Science and Technology, Hankou University, Wuhan 430212, China)[2]

Abstract With the growing of the Internet, the network of large data has become an important distribution center for the financial sector. These data give valuable opportunities and severe challenges to stock analysis and prediction. The development of Web2.0 allows investors to actively participate in all aspects of the creation of network of information, communication and accessing. This paper disclosed information on the Internet and generated complete and effective fundamental of information to get 10 actors that may impact the stock market, which are economic and cycle, fiscal policy, changes in interest rates and exchange rates, price changes, inflation, political policies, changes in the industry, operating conditions and the downstream effects. We proposed an algorithm for the stock market prediction based on these 10 factors. Experiments show that the method has good feasibility, and can bring significant academic and commercial values.

Keywords Data mining, Stock price forecast, Web big data

1 引言

信息对证券市场的影响是金融学研究的核心问题,也是广大股民进行投资决策的重要依据。长期以来,由于信息不对称造成的决策错误是投资失败的主要因素,因此如何有效地获得和分析信息并进行市场预测是产业界和学术界的研究热点。随着互联网技术的飞速发展,网络大数据中丰富的数据带来了证券市场从信息匮乏到信息过剩的转变,互联网中证券信息的快速增长成为了股票分析与预测的宝贵机遇和严峻挑战。一方面,各种新兴的网络社会媒体(例如微博、社区、论坛、博客等)成为了传递信息、沟通交流的主要平台,蕴含了海量的信息,为投资决策提供了丰富的数据支撑;另一方面,互联网发展带来的Web大数据的挑战和问题也困扰着对信息的有效分析和利用。随着人们逐渐依赖互联网信息对事务进行判断,网络也正日渐被投资者接受与使用。互联网不仅改变了投资者参与投资活动的方式,更是已经成为股市投资信息传播的主要渠道。越来越多的投资者利用网络搜索金融信息、与他人交流投资经验等。各类渠道的信息及时有效地在互联网上进行传播,使资本市场的信息传播方式产生了重大变化,有效解决了证券市场的信息不对称问题,而新的市场信息结构也对投资者的行为产生了巨大的影响,进而影响了对股票资产定价和金融资源的配置。

互联网中关于资本市场、上市企业和行业动态的信息,是影响股票价格的根本原因,也是股价基本分析法的主要研究

到稿日期:2014-04-02 返修日期:2014-07-18 本文受国家自然科学基金项目:面向过时信息自动发现的Web时态一致性研究(61272109),中央高校基本科研业务费专项资金项目:Web大数据环境下的数据时态一致性研究(2042014kf0057),湖北省人文社会科学基金项目:基于Web时间冲突性推理的智能信息过滤研究(14G461)资助。

杨 莎(1980-),博士生,主要研究方向为Web数据挖掘和互联网金融;余 伟(1987-),男,博士,讲师,主要研究方向为Web大数据质量评估、Web数据融合等,E-mail:yuwei@whu.edu.cn(通信作者);李石君(1964-),博士,教授,主要研究方向为Web社区分析、Web数据一致性等;曹晶晶(1982-),硕士生,主要研究方向为互联网金融;刘 晶(1981-),女,博士生,主要研究方向为Web数据挖掘。

内容。股价能在一定程度上反映企业的市场价值，因此企业前景和利润变化直接影响股票价格的变化。表1中列举了近期的几个例子。

表1 网络信息影响市场变化的例子

网络信息	信息类型	影响结果
一季度钢铁全行业处于严重亏损状态	市场现状	钢铁板块-负面影响
香港瑞安地产急抛内地物业跑路，沈阳项目烂尾	市场现状	房地产板块-负面影响
时隔近18个月IPO审核重启4家公司30日上会	政策公告	基本面-积极影响
全面布局移动互联网二六三去年营收增长近九成	企业业绩	002467个股-积极影响
与雷士渠道整合深化德豪润达LED产业开始发力	企业市场	002005个股-积极影响

这些信息直接或间接地影响证券市场的板块或个股，而互联网已经成为股市投资信息传播的主要渠道，各类新闻事件、内幕信息、传闻消息等及时广泛地在互联网中传播，具有重要的参考价值和分析意义。因此从互联网信息中挖掘隐含的、新颖的、影响股票市场的知识和信息，帮助市场参与者优化投资决策、规避风险，具有创新性和现实意义。本文将深入分析证券市场的各类信息类型和信息对板块及个股的影响程度，最终形成基于网络信息对证券价格变化的预测方法。

2 国内外研究现状

Web大数据具有丰富的新闻资讯、用户交流和网络名人，基于互联网的证券价格变化趋势研究已经成为许多研究者关注的问题。

目前，证券投资分析法可分为基本分析法和技术分析法两类[1]。基本分析法主要依据经济学、投资学、金融学等基本原理推导出结论，不能适应快速变化的市场。而技术分析法则在各种指标的选择与图表的分析上过多地依靠经验判断，股票市场相关信息的复杂性决定了这种方法的可靠性在很大程度上得不到保证[2]。基于时间序列分析的股票预测研究方法是一种常见的技术。根据技术手段的不同，时间序列法可分为模型法、指标法和数据图法3类[3]。一般的时间序列很难处理高难度的非线性问题，而股票市场具有典型的非线性特点[4]，这些问题导致时间序列方法很难在股票预测方面取得理想结果。

股票市场经济功能的发挥依赖于股票价格与相关信息的相互影响[5]，比如各类互联网信息和股评家发表的观点、分享的信息构成了具有价值的股票知识库。如何从浩如烟海的互联网数据中识别和抽取专家的有效知识并进行股票价格变化预测是非常重要的。而过去的股票预测方法存在的共同问题是难以有效获取股票相关的信息。通过建立互联网股票信息知识库，结合网络数据挖掘与特征提取方法可以有效地对某行业股票波动进行预测。

目前，也有研究者就基于互联网中的证券行业新闻信息来进行市场的预测展开了相关的研究。梁循等人在SVM的预测模型中引入了互联网金融信息的褒贬值，通过实验揭示出金融市场波动率与互联网上新闻的相关性，提出了有效的股市预测方法[6]。但是这些方法仅研究了倾向性的比重情况与金融市场公司的股价波动率的关系，对于信息本身的内容没有展开深入的研究。

大数据带来的是信息的极大丰富，有效提取大数据中的信息并转换成有用的知识主要在于数据特征的提取。数据特征提取可以看作是用映射(或变换)的方法将原始特征进行线性结合从而得到能更好地描述数据集的新特征。经典的特征提取方法有主成分分析法(Principle Component Analysis-PCA)和线性判别分析法(Linear Discriminant Analysis-LDA)。在知识提取方面，张云璐等人研究了在数字图书馆资源中提取用户想要的知识和服务[7]；王海涛等人提出了基于领域本体的文本知识自动获取方法，其通过引入领域本体，实现了半结构化文本知识的完全自动获取[8]；钟秀琴等人通过构造可共享、可重用、可扩展的几何学本体，用RDF/OWL的形式进行描述，形成了一套较完整的几何学知识获取和知识表示体系[9]；张清华等人提出一种基于最大粒的规则获取算法，该算法根据条件属性域形成的分层递阶的划分空间，自顶向下逐渐地提取最大粒对应的规则，提高了粗糙集的泛化能力，实现了基于规则的知识提取[10]。

本文首先将这些方法与证券技术分析指标相结合，建立股票知识库，结合宏观的经济环境，对行业兴衰、企业经营、盈利现状和前景进行分析，对上市公司的股票做出预测。

3 价格变化与影响关系分析

同所有商品一样，股票在证券市场的价格也不是一成不变的，通常我们会通过技术分析和基本面分析两种方式来判断股市价格的未来走向。技术分析是通过各种图表、技术指标来研究股票市场行为，以判断股票市场价格变化趋势的方法。而基本面分析则是对股市波动成因的分析，是对长线投资的定性分析。对基本面的分析其实是研究上市企业或证券市场的真实理论价值，也是对影响股票价格走势的综合因素的分析，主要包括3个方面：宏观类影响因素、中观类影响因素和微观类影响因素。这3类因素是股票市场基本面分析的主要组成部分。股票预期价格就是反映这些因素条件下证券市场的理论价值体现，然而因为受到多种因素的影响而频繁变化，所以要对影响股票价格的各个因素进行综合分析。本文针对这3类因素，根据实际的市场情况，建立了10个证券市场信息因素指标，包括：经济周期、财政政策、利率变动、汇率变动、物价变动、通货膨胀、政治政策、行业变化、经营状况、上下游影响等。

股票的基本分析就是通过丰富全面的信息进行综合研判，从宏观的经济环境开始研究，逐步开始中观的行业兴衰分析，再进行微观的企业经营、盈利现状和前景进行分析，从而对各上市公司的股票做出接近事实的评价，进而尽可能地预测未来的变化。

4 互联网信息与证券市场变化影响模型

本节采用多元统计中的因素分析(factor analysis)模型，根据股票价格变化的影响参数，结合互联网中实时发布的信息，将股票的日收益率作为被解释变量，并认为其由许多信息因素(经济周期、财政政策、利率变动、汇率变动、物价变动、通货膨胀、政治政策、行业变化、经营状况、上下游影响)共同作用决定。这些信息因素对公司价值起到直接的决定作用，本节研究包含一些因素(factors)的可逐日连续观测变量(proxies)，试图通过检验可连续观测变量与日收益率间的关系，提出关于因素与收益率间关系的科学猜想。以互联网中的各类信息作为数据来源，构建了一个新的“互联网信息指标”(ISI)来作为成交量指标的补充和检验。ISI指标具有以下特征：

①ISI 指标与证券市场间的相关性，据此判断互联网中获得的信息内容是否与证券市场变化显著相关；

②ISI 指标、成交量指标共同存在的模型对证券市场变化的影响程度，并据此研究基于指标数据对证券市场变化的影响变化趋势分析。

构建该指标的主要经济学目的在于找到一个能够增强对证券市场变化解释能力的指标，该指标的经济含义是作为对“证券相关”互联网信息的一个加总；并考虑该指标能否直接反映开源信息的质量，以及对“噪音”的处理能力。

根据分析角度的不同，ISI 指标从 3 个层次进行研究，分别为大盘综合互联网信息指标 AISI、行业互联网信息指标 IISI 和个股互联网信息指标 CISI。

定义互联网中可以获得的因素，包括经济周期、财政政策、利率变动、汇率变动、物价变动、通货膨胀、政治政策、行业变化、经营状况、上下游影响，并构造成可逐日连续观测变量向量：

$$V=\{v_1,v_2,v_3,v_4,v_5,v_6,v_7,v_8,v_9,v_{10}\} \tag{1}$$

其中，v_i 表示第 i 个影响因素。

针对当前时刻 T，如果在近期的一个连续观测时间段 $[t_s,T]$ 内，互联网中共出现了 k_i 个关于 v_i 的影响因素消息信息，其中第 j 个消息信息 $N_j(j=1,\cdots,k_i)$ 对大盘市场的综合影响度为 f_j，f_j 定义为对大盘影响的指数变化值，如果为负数则表示对大盘市场有负面影响，如果为正数则表示对大盘市场有正面影响，该值是通过历史的机器学习和经验判断针对各个类型的信息事件进行判断而获得的。该消息 j 的发生时间为 t_j，随着时间变化的影响变化权重函数为 $w(N_j,\Delta t)$，其中 $\Delta t=T-t_j$，表示距发生时间的时间间隔，描述了随着时间的变化该消息对当前时刻的市场情况的影响力度。根据我们的研究经验，该变化函数服从正态分布。

大盘综合互联网信息指标 AISI 模型如下：

$$AISI=\sum_{i=1}^{10}\sum_{j=1}^{k_i}f_i\times w(N_j,\Delta t) \tag{2}$$

行业的互联网信息指标与上类似，同样受到各个消息信息的影响，在当前时刻针对行业 M，如果在近期的一个连续观测时间段 $[t_s,T]$ 内，互联网中共出现了 $\bar{k}_i$ 个关于行业 M 的 v_i 的影响因素消息信息，其中第 j 个消息信息 $\bar{N}_j(j=1,\cdots,k_i)$ 对大盘市场的综合影响度为 $\bar{f}_j$，$\bar{f}_j$ 定义为对该行业 M 的指数变化值，如果为负数则表示对行业 M 有负面影响，如果为正数则表示对行业 M 有正面影响，该值通过历史的机器学习和经验判断针对各个类型的信息事件进行判断而获得。消息 j 的发生时间为 t_j，随着时间变化的影响变化权重函数为 $w(N_j,\Delta t)$，其中 $\Delta t=T-t_j$，表示距发生时间的时间间隔，描述了随着时间的变化该消息对当前时刻的市场情况影响力度。根据我们的研究经验，该变化函数服从正态分布。

独立的行业 M 互联网信息指标 $\hat{IISI}$ 模型如下：

$$\hat{IISI}=\sum_{i=1}^{10}\sum_{j=1}^{k_i}f_i\times w(N_j,\Delta t) \tag{3}$$

然而事实上行业不是独立存在的，因此要考虑行业之间的影响关系，行业之间由于上下游产业链的关系而相互影响各自的市场状况，图 1 反映了“布”行业的上下游关系。

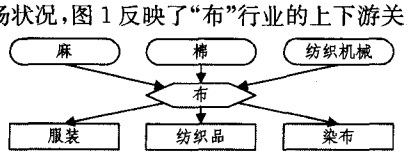

图 1 “布”行业的上下游分布

上游的变化会影响行业的建设成本和生产能力，下游的变化会引起行业的市场需求和供给利润，设行业 M 的上游行业有 a 个，分别为 $\overset{\frown}{M}=\{M_1,\cdots,M_a\}$，每个上游行业 M_i 对该行业的影响因子为 A_i，行业 M 的下游行业有 b 个，分别为 $\underset{\smile}{M}=\{M_1,\cdots,M_b\}$，每个下游行业 M_j 对该行业的影响因子为 B_j。因此，考虑到上下游影响关系的行业 M 互联网信息指标 $IISI$ 模型如下：

$$IISI=\hat{IISI}_M+\sum_{i=1}^{a}A_i\,\hat{IISI}_i+\sum_{j=1}^{b}B_j\,\hat{IISI}_j \tag{4}$$

个股的互联网信息指标与上类似，同样受到各个消息信息的影响，在当前时刻针对个股 S，如果在近期的一个连续观测时间段 $[t_s,T]$ 内，互联网中共出现了 $\bar{k}_i$ 个关于个股 S 的 v_i 的影响因素消息信息，其中第 j 个消息信息 $\bar{N}_j(j=1,\cdots,k_i)$ 对大盘市场的综合影响度为 $\bar{f}_j$，$\bar{f}_j$ 定义为对该个股 S 的指数变化值，如果为负数则表示对个股 S 有负面影响，如果为正数则表示对个股 S 有正面影响，该值通过历史的机器学习和经验判断针对各个类型的信息事件进行判断而获得。该消息 j 的发生时间为 t_j，随着时间变化的影响变化权重函数为 $w(N_j,\Delta t)$，其中 $\Delta t=T-t_j$，表示距发生时间的时间间隔，描述了随着时间的变化该消息对当前时刻的市场情况影响力度。根据我们的研究经验，该变化函数从正态分布。

因此独立的个股 S 互联网信息指标 $\hat{CICS}$ 模型如下：

$$\hat{CICS}=\sum_{i=1}^{10}\sum_{j=1}^{k_i}f_i\times w(N_j,\Delta t) \tag{5}$$

然而事实上个股不是独立存在的，一方面，个股受它所在的大盘环境影响；另一方面，个股受它所在的行业状况影响。另外个股还受到其他公司的影响。

例如个股 S 的主营范围决定了它所在的行业，并且根据他的主营范围比重受到了行业的不同程度的影响。图 2 和表 2 反映了个股 S 的主营范围和行业的关系。

布	21714.81	−14.17%	59.14%	729.66
工程机械国际贸易	4238.95	99.40%	11.54%	474.10
服装	3798.75	−50.44%	10.35%	392.71
纱	3356.64	25.20%	9.14%	97.41
纺织机械等设备进口	1668.85	190.24%	4.54%	122.97
家纺产品国际贸易	706.86	28.10%	1.92%	17.46
陶瓷国际贸易	631.74	20.34%	1.72%	74.76
其它(补充)	603.73	—	1.64%	—
纺织业务	29581.40	−13.09%	80.56%	1534.57
其他业务	6535.19	21.35%	17.80%	374.50
其它(补充)	603.73	—	1.64%	—
境外销售	30706.33	−8.06%	83.62%	—
境内销售	5410.27	−10.19%	14.73%	—
其它(补充)	603.73	—	1.64%	—

图 2 某上市公司的主营构成(大智慧软件)

表 2 某上市公司主营相关行业所在比重构成

主营分类	比重
国内贸易-纺织业务-布	59.14%
国内贸易-纺织业务-服装	10.35%
国内贸易-纺织业务-纱	9.14%
国内贸易-机械设备-纺织机械	4.54%
国际贸易-纺织业务-家纺产品	1.92%
国际贸易-机械设备-纺织机械	11.54%
国际贸易-艺术品-陶瓷	1.72%

另一方面，企业和企业之间具有各类控股或被控股的关系，使得他们的市场变化相互影响。图 3 反映了个股的控股公司和被控股公司的关系。

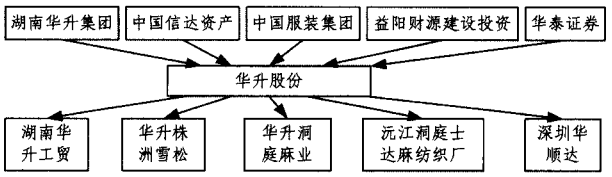

图 3　某上市公司控股公司和被控股公司关系

如果企业 C 的控股公司一共有 t 个，构成了企业 C 的控股公司集，用 K 表示：

$$K=\{\langle C_1,w_1\rangle,\langle C_2,w_2\rangle,\cdots,\langle C_t,w_t\rangle\}$$

其中，w_i 表示企业 C 在公司 C_i 中占有个股份比例，其中 $0\leqslant w_i\leqslant 1$。

企业 C 的股东一共有 $m+n$ 个，其中 m 个企业，n 个个人，构成了 C 的股东集，用 G 表示：

$$G=\{\langle C_1,w_1'\rangle,\langle C_2,w_2'\rangle,\cdots\langle C_m,w_m'\rangle,\langle P_1,w_1''\rangle,\cdots\langle p_1,w_n''\rangle\}$$

其中，w_i' 表示企业 C_i 在公司 C 中占有额股份比例，w_1'' 表示个人 P_i 在公司 C 中占有的股份比例，$0\leqslant w_i'\leqslant 1$，$0\leqslant w_i''\leqslant 1$，$\sum_{i=1}^{m}w_1'+\sum_{i=1}^{n}w_1''=1$。

定义 1(控股)　对于企业 C_i 和企业 C_j，如果存在 $c_j\in K_{c_i}$，则表示企业 C_i 对 C_j 控股，C_i 在 C_j 中占有的股份为 $w_{ij}=K_{C_i}\cdot C_j\cdot w$，控股的算法表示为：

$$\Theta C_i\xrightarrow{w_{ij}}C_j$$

定义 2(循环持股)　如果企业 C_i 和企业 C_j 相互之间存在控关系，即存在 $\Theta C_i\xrightarrow{w_{ij}}C_j$，也存在 $\Theta C_j\xrightarrow{w_{ji}}C_i$，则表示企业 C_i 和企业 C_j 循环持股，表示为：

$$\Theta C_i\xrightarrow{\langle w_{ij},w_{ji}\rangle}C_j$$

定义 3(间接控股)　如果企业 C_i 和企业 C_j 之间，存在 $C_1,C_2,\cdots,C_n$，满足 $\Theta C_i\xrightarrow{w_{i1}}C_1,\Theta C_1\xrightarrow{w_{12}}C_2,\cdots,\Theta C_n\xrightarrow{w_{nj}}C_j$，则表示企业 C_i 对企业 C_j 间接控股，表示为：

$$\Theta C_i\overset{n}{\Rightarrow}C_j$$

其中，n 表示企业 C_i 对企业 C_j 间接持股的间接度，$n>0$，C_1，$C_2,\cdots,C_n$ 表示传递节点。

间接控股具有传递性，即如果 $\Theta C_i\overset{n}{\Rightarrow}C_j,\Theta C_j\overset{m}{\Rightarrow}C_k$，且 $i\neq k$，则 $\Theta C_i\overset{n+m}{\Rightarrow}C_k$。

定义 4(交叉持股)　如果存在企业 C_i、企业 C_j，既满足 $\Theta C_i\overset{n}{\Rightarrow}C_j$，也满足 $\Theta C_j\overset{m}{\Rightarrow}C_i$，则表示企业 C_i 和企业 C_j 交叉持股，表示为：

$$\Theta C_i\overset{n,m}{\Rightarrow}C_j$$

交叉持股满足交换律和结合律。

如果企业主营构成包括 n 类，构成了企业的主营构成集，用 Y 表示：

$$Y=\{\langle G_1,w_1\rangle,\langle G_2,w_2\rangle\cdots\langle G_n,w_n\rangle\}$$

其中，G_i 表示第 i 项主营构成，w_n 表示第 i 项主营构成在整个公司所占的比重，满足：

$$\sum w_i=1$$

且任意企业的主营构成中，每一项 G_i 都互不相交。

控股公司的变化会影响该公司的战略规划和发展格局，被控股公司的变化会影响该公司的财务状况和业务能力。设个股 S 的控股公司有 a 个，分别为 $\hat{S}=\{S_1,\cdots,S_a\}$，每个控股公司 S_i 对该个股的影响因子为 A_i，个股 S 的下游行业有 b 个，分别为 $\underset{\sim}{S}=\{S_1,\cdots,S_b\}$，每个被控股公司 S_j 对该行业的影响因子为 B_j。

同时个股 S 的主营结构对应了 c 个行业，第 i 个行业对该个股 S 的影响度为 C_i，因此，考虑到股权结构影响关系和个股的行业状况的个股 S 互联网信息指标 CISI 模型如下：

$$CISI=\hat{CISI}_S=\sum_{i=1}^{a}A_i\hat{CISI}_i+\sum_{j=1}^{b}B_j\hat{CISI}_j+\sum_{k=1}^{c}C_i\hat{CISI}_k \tag{6}$$

5　互联网中证券影响因素信息获取

为了有效地验证和实现信息的提取和市场的预测，本文选择了 10 个样本行业和 100 个样本个股，关于样本行业的选择问题有如下 2 点说明：

①为了反映相互的影响，所选择的 10 个样本行业(板块)部分具有上下游关系；

②所选择的 10 个样本行业是近期较为重点和关注的行业，市场波动较大，消息面比较频繁。

所选择的 10 个样本行业的情况如表 3 所列。

表 3　本文所研究的 10 个行业

序号	行业板块	总市值(亿)	个股数量
1	大数据	323.13	16
2	智能穿戴	150.47	15
3	节能环保	1047.90	65
4	特斯拉	1120.62	24
5	国企改革	6219.92	97
6	新能源	3399.32	105
7	燃料电池	1203.99	13
8	锂电池	1644.36	43
9	物联网	1178.93	42
10	中字头	16855.37	35

关于样本个股的选择问题有如下 3 点说明：

①之所以选择主板中盘面较大的个股，主要是两方面权衡的结果。主板中业绩较好的个股盘面较大，相对能防止受人为操作的剧烈波动。

②根据证券交易所报告中 100 家具有代表性的主板公司组成，选样空间是在上海和深圳证券交易所主板上市交易且满足下列条件的股票：上市时间超过一年(流通市值排名在样本数 10%范围内的不受此限)；非 ST、非 ST * 股票；公司最近一年无重大违规、财务报告无重大问题；公司最近一年经营无异常、无重大亏损；考察期内股价无异常波动。此方法是计算入围个股在考察期的平均流通市值及平均成交金额所占市场比重，将上述指标按 2:1 权重进行加权计算，再将结果从高到低排序，选取排名前 100 名的股票构成指数成份股。成份股每年 1、7 月进行样本股定期调整。本文采用的是 2013 年 1 月调整后的样本股。

③这些个股在互联网上都有较为全面的信息披露，也与本文选择的 10 个样本行业有着相互关系。

因此所选择的 100 个样本个股中部分如表 4 所列。

表 4　本文所研究的 100 个股(部分代表)

代码	名称	最新(元)	总量(手)	金额(万)
002335	科华恒盛	14.87	7079	1047
002368	太极股份	28.90	22571	6453
000977	浪潮信息	48.06	11699	5634
002657	中科金财	27.00	5432	1448
300302	同有科技	27.83	1849	512
600804	鹏博士	14.06	133331	18765
300245	天玑科技	25.20	19133	4868
300231	银信科技	18.78	9048	1709
300229	拓尔思	20.42	16338	3350
300290	荣科科技	14.00	2286	320
002432	九安医疗	20.50	87534	17563
002681	奋达科技	28.79	155885	45257
002273	水晶光电	17.54	63623	11118
300061	康耐特	12.86	10762	1375
002241	歌尔声学	27.80	78893	21555
600584	长电科技	7.48	107317	8041
002655	共达电声	10.23	18457	1887
002512	达华智能	13.12	21247	2791
002230	科大讯飞	24.99	49480	12327
600636	三爱富	13.53	343302	45374
300072	三聚环保	17.83	53767	9545
002335	科华恒盛	14.87	7079	1047
002665	首航节能	25.55	22928	5886
002648	卫星石化	26.12	18879	4896
000723	美锦能源	5.17	13149	678
002499	科林环保	21.16	6133	1292
300338	开元仪器	21.14	15510	3260
300140	启源装备	14.93	8361	1246
600874	创业环保	7.65	32508	2487

通过互联网爬虫每天实时抓取互联网上出现的各类证券相关信息，主要的信息来源有 3 类。

新闻：主要是从各大新闻网站中获取最新的相关信息；

微博：从新浪微博中获取最新的相关事件信息；

论坛：从各大论坛尤其是股票论坛中获取最新的相关事件信息。

除了各自的来源不同外，本文将它们作为统一的消息来源进行分析，并对其进行建模。信息模型将信息的基本属性进行了定义，对于信息 N，定义为

$$N=\{tt,stt,at,ws,ct,tm,tp,ul\}$$

其中，tt 表示标题，stt 表示副标题，at 表示作者，ws 表示来源站点，ct 表示正文，tm 表示发表时间，tp 表示该信息在该来源站点的分类，ul 表示该信息的链接 URL。tp 是一个 n 元组，$tp=\{t_1,t_2,\cdots,t_n\}$，$n$ 表示该站点的分类最大级数。

然后对信息进行内容分析，提取信息中关于证券相关的信息内容，例如信息“中国人民银行决定，从 2011 年 1 月 20 日起，上调存款类金融机构人民币存款准备金率 0.5 个百分点”属于影响因素“财政政策”中“存款准备金调整”的分类，行为为“上调”，程度为“0.5%”，发生时间为“2011 年 1 月 20 日”，根据分类知识，该新闻对时刻“2011 年 1 月 21 日”大盘的影响度为 −3，延时因子为 1。

针对行业板块，例如信息“购新能源车最高补贴 12 万”，属于“政治政策”中“产业扶持”的分类，行为为“新能源购车补贴”，程度为“120000”，发生时间为“2014 年 4 月 29 日”，根据分类知识，该新闻对时刻“2014 年 4 月 30 日”的“新能源汽车”行业的影响度为 2，延时因子为 0.2。

针对个股，例如信息“酒鬼酒被指塑化剂超标 260% 引发白酒行业塑化剂门”，属于“经营状况”中“质量问题”的分类，行为为“塑化剂超标”，程度为“260%”，发生时间为“2012 年 11 月 19 日”，根据分类知识，该新闻对时刻“2012 年 11 月 20 日”的“酒鬼酒 000799”的影响度为 −10，延时因子为 1。

因此对互联网中的信息进行自动抽取最新的信息，然后将其分析整理后得到统一的数据模型，信息 F 可以被描述为：

$$F=\{Title,Context,Addtime,Factor,Type,Behavior,Degree,Appendtime,IF\} \tag{6}$$

各元素分别表示信息标题、信息描述、发表时间、所属影响因素、所属分类、具体行为、行为程度、发生时间，IF 表示该信息对大盘、行业、个股等所产生的影响关系，可以描述为：

$$IF=\{if_1,\cdots,if_n\}$$

其中，n 表示该新闻会产生的影响的目标个数，对第 i 个目标的影响为

$$if_i=\{Otype,Object,Impact,Delay\}$$

其中，$Otype$ 表示类型，为固定枚举型，1 表示大盘，2 表示行业，3 表示个股。Object 表示影响的具体目标，如果为个股，则描述股票代码；如果为行业，则描述自定义的行业编码；如果为大盘，则记录为 0。

6　基于互联网信息对证券市场变化预测实验

本文根据近期的时间，进行了数据采集、分析和预测实现，并根据结果与后来发生的情况进行实际校验。通过采集 2014 年 3 月 1 日－2014 年 3 月 15 日半个月内的所有发生的相关信息事件，对其去重后一共有 12314 条相关的信息事件，平均每天约 821 条，每天的数据分布如表 5 所列。

表 5　所有发生的相关信息事件数量分布(3 月 1 日－3 月 15 日)

日期	大盘	行业	个股
* 2014 年 3 月 1 日	29	38	934
* 2014 年 3 月 2 日	10	107	674
2014 年 3 月 3 日	33	89	924
2014 年 3 月 4 日	45	57	625
2014 年 3 月 5 日	46	139	732
2014 年 3 月 6 日	47	42	1050
2014 年 3 月 7 日	56	80	625
* 2014 年 3 月 8 日	18	77	1043
* 2014 年 3 月 9 日	53	73	307
2014 年 3 月 10 日	31	59	706
2014 年 3 月 11 日	34	49	491
2014 年 3 月 12 日	21	33	984
2014 年 3 月 13 日	45	135	525
2014 年 3 月 14 日	40	65	633
* 2014 年 3 月 15 日	38	68	406

其中带 * 号的表示是周末，因为周末证券市场不交易，因此我们认为周末的消息对周一的市场影响时延为 0，即表示及时的影响。

为了降低消息长时间影响所造成的计算复杂度，本文假设消息的时间周期为 5 天，即消息发生 5 天后对市场不再造成影响，根据大盘综合指数：

$$AISI=\sum_{i=1}^{10}\sum_{j=1}^{k_i}f_i\times w(N_j,\Delta t) \tag{7}$$

每个交易日的大盘综合指数由近 5 天内的消息事件综合决定，根据时间间隔按函数 $w(N_j,\Delta t)$ 产生因子变化。

分布图如图 4 所示。

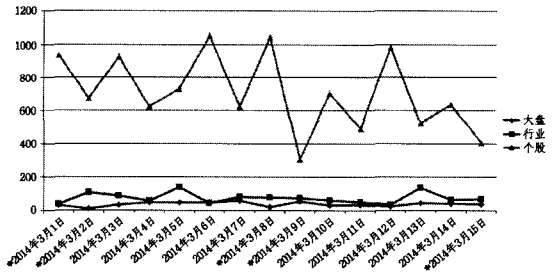

图 4　3 月 1 日－3 月 15 日所有发生的相关信息事件数量分布

根据每个信息的数据情况，对这 15 天内的 10 个交易日的 AISI 指数按模型进行计算。表 6 列出根据 AISI 预测的每个交易日的价值变化指数。

表 6　每日 AISI 值(3 月 1 日－3 月 15 日)

日期	预测变化值
2014 年 3 月 3 日	－8
2014 年 3 月 4 日	2
2014 年 3 月 5 日	－10
2014 年 3 月 6 日	－2
2014 年 3 月 7 日	8
2014 年 3 月 10 日	－40
2014 年 3 月 11 日	1
2014 年 3 月 12 日	－10
2014 年 3 月 13 日	20
2014 年 3 月 14 日	－18

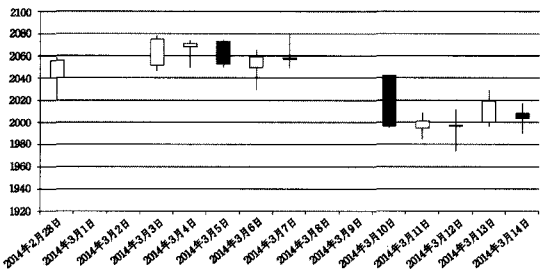

图 5　11 个交易日的实际交易 K 线图(2 月 28 日－3 月 15 日)

表 7　3 月 1 日－3 月 15 日的市场行情与比较情况

日期	盘价	高价	最低	收盘	预测变化	实际最大值	实际最小值	误差
2014 年 2 月 28 日	2040	2058	2020	2056				
2014 年 3 月 3 日	2052	2078	2047	2075	－8	3	－28	0
2014 年 3 月 4 日	2068	2074	2050	2071	2	－1	－25	3
2014 年 3 月 5 日	2073	2074	2050	2053	－10	3	－21	0
2014 年 3 月 6 日	2050	2065	2030	2059	－2	12	－23	0
2014 年 3 月 7 日	2058	2079	2050	2057	8	20	－9	0
2014 年 3 月 10 日	2042	2042	1995	1996	－40	－15	－62	0
2014 年 3 月 11 日	1994	2008	1985	2001	1	12	－11	0
2014 年 3 月 12 日	1996	2011	1974	1997	－10	10	－27	0
2014 年 3 月 13 日	2000	2029	1996	2019	20	32	－1	0
2014 年 3 月 14 日	2008	2017	1990	2004	－1	－2	－29	1

图 5 描述了 2 月 28 日－3 月 15 日之间 11 个交易日的实际交易数据(为了反映 2 月 28 日到 3 月 1 日之间的变化情况，所以加入 2 月 28 日的市场行情)。

根据实际变化与预测变化进行比较，结果如表 7 所列。

比较结果如图 6 所示。

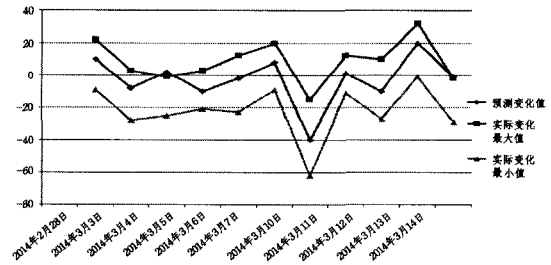

图 6　3 月 1 日－3 月 15 日的市场行情与比较情况分布

由结果可以看出，预测的变化结果的趋势基本与实际情况符合，交叉情况属于预测失败的异常值。关于行业的 IISI 与个股的 CISI 原理和方法基本相同。

结束语　过去几十年中，互联网的产生与飞速发展带来的不仅仅是技术的革命、产业的升级、生产力的提升，也显著地改变了个体的信息获取与决策方式以及群体的行为。

本文通过对互联网上的信息事件进行挖掘分析，对影响证券市场变化的经济周期、财政政策、利率变动、汇率变动、物价变动、通货膨胀、政治政策、行业变化、经营状况、上下游影响这 10 个因素分别进行了建模，使得在一段时间内的信息事件对这 10 个因素的影响程度是准确可度量的。最终根据模型综合分析得出市场变化的总体结论。该方法是基于互联网中信息本身的影响程度来对证券市场的基本面进行定量分析的，具有客观性和独立性，反映了证券市场价格变化的本质，与传统的定性分析具有本质的区别。

然而证券市场是一个复杂、庞大的开放式系统，主观性与客观性是一个矛盾对立的辩证双方。本文虽然利用主观性和客观性的方法分别进行了研究，但是我们知道矛盾对立的双方在一定程度上可以相互作用、相互转换，现实市场中也会间歇性出现恐慌性群体影响，因此如何有效地把握主观性和客观性分析的一个平衡关系是一个难题，本文也没有对此进行深入的研究，这将是今后研究的一个重点。

参 考 文 献

[1] 吴云勇，范树杰. 证券投资分析方法研究[J]. 中国市场，2012，(27)：76-77

[2] 许泱. 基于神经网络的股票市场预测研究[D]. 武汉：华中科技大学，2008

[3] 李嵩松. 基于隐马尔可夫模型和计算智能的股票价格时间序列预测[D]. 哈尔滨：哈尔滨工业大学，2011

[4] 臧玉卫，张慎峰，吴育华，等. 中国股票市场的非线性分析[J]. 天津大学学报：社会科学版，2005，7(6)：417-420

[5] 刘长虎，陶建格，崔衍秋，等. 股票价格指数的投资功能[J]. 市场论坛，2004(6)：71-72

[6] 王超，李楠，李欣丽，等. 倾向性分析用于金融市场波动率的研究[J]. 中文信息学报，2009，23(1)：95-99

[7] Yu Wei，Zhang Yun-lu，Gan Lin. Automatic Evaluation for Engineering Pathway Premier Award Winners [J]. Applied Mathemtics & Information Sciences，2014，8(5)：381-394

[8] 王海涛，曹存根，高颖. 基于领域本体的半结构化文本知识自动获取方法的设计和实现[J]. 计算机学报，2005，28(12)：2010-2018

[9] 钟秀琴，符红光，佘莉，等. 基于本体的几何学知识获取及知识表示[J]. 计算机学报，2010，33(1)：167-173

[10] 张清华，王国胤，刘显全. 基于最大粒的规则获取算法[J]. 模式识别与人工智能，2012，25(3)：388-396