The Conception of Stock Price Volatility Analysis System Based on News Events

Liu Xin[1,a],Sheng Mingcai[2,b],Huang Xi[3,c] and Su Ganya[*]

[1]*Statistics institute, Chengdu university of information technology, Chengdu, Sichuan, China*
[2]*Statistics institute, Chengdu university of information technology, Chengdu, Sichuan, China*
[3]*Statistics institute, Chengdu university of information technology, Chengdu, Sichuan, China*
[a]*lx001201@163.com,* [b]*925394476@qq.com,* [c]*1129858063@qq.com*
**Corresponding author*

Abstract.
News events are one of the factors influencing stock price fluctuations. The analysis system proposed in this paper quantifies the influence of historical news events into "limitation period" and "stock price fluctuation range", predicts the possible impact of real-time news events on stock price, and gives the limitation period and stock price fluctuation range. Keyword factors are used to link historical news events with real-time news events.
Keywords: *news events, stock price volatility, factor of corpus, text segmentation*

基于新闻事件的股价波动分析系统构想

刘鑫[1,a]，沈明财[2,b]，黄茜[3,c]，苏甘雅[*]

[1]*成都信息工程大学统计学院，成都，四川，中国*
[2]*成都信息工程大学统计学院，成都，四川，中国*
[3]*成都信息工程大学统计学院，成都，四川，中国*
[a] *lx001201@163.com,* [b] *925394476@qq.com,* [c] *1129858063@qq.com*
[*]*通讯作者*

中文摘要：
新闻事件是股价波动的影响因素之一。本文构想的分析系统将历史新闻事件的影响量化为“时效期”和“股价波动幅度”两部分，预测实时新闻事件对股价可能产生的影响,给出时效期和股价波动范围。历史新闻事件和实时新闻事件之间用关键词因子进行连接。
关键词：*新闻事件；股价波动；因子库；文本分词。*

1. 数据处理

1.1 数据有效性

数据是否有效，会直接影响数据库的质量。目前研究事件风险的瓶颈在于难以保证数据的有效性。对于股市事件，确定其是否对股价有影响是关键。本文将这种影响分为两部分来刻画——时效性与波动幅度。一般来说，一则新闻事件的发生，其影响有短有长，时效性是衡量新闻事件对股价影响持续长短的最常用指标。波动幅度是衡量新闻事件对股价变动影响程度的指标，超过正常波动幅度，视为有效新闻事件。

具体表现为，利用前后两次新闻事件时间间隔中的每日收盘价建立 Var 模型，进行脉冲响应分析，确定前一新闻事件的时效性。新的事件的时效性是通过后续匹配到的新闻事件的时效期长度的平均值。最后将有效期内的收盘价减去开盘价除以时效天数，得到一个均值波动幅度，超过这个波动幅度，才能视为有效波动幅度，相对应的新闻事件才是有效新闻事件。

1.2 历史新闻事件库

本文将新闻事件分为三类：公司类、行业类、宏观类，建立对应历史新闻事件子库。公司类历史新闻子库的标志是标题或者内容中明确有公司名称或者股票代码；行业类历史新闻子库的标志是标题或者内容中明确

有行业名称或者概念板块；其余没有明显特征的新闻事件归于宏观类历史新闻子库。每则新闻事件文本都保存了其提取的对应的关键词因子，即在历史新闻库和因子库之间建立连接关系。

1.3 因子库

新闻事件的表现形式是中文文本，而“汉语中具有完整语义信息的最小单元是词，且汉语自然语言的句法、语义和语境分析，静态、动态语义网构建，以及搜索引擎到倒排索引建立等技术的处理和分析对象都是词。因此，中文自然语言处理的第一步就是将由汉字连续组成的字串切分为词的序列，即中文分词。”[1] 目前分词系统主要采用的技术是词典分词和机器学习分词。[2]

本文采用的是词典分词，分词词典就是由在整个事件文本中所出现的可能会引起股价波动的关键词构成的因子库。从历史新闻事件中能否准确地切分出有效关键词将会影响到后续关键词因子的提取，以及后续匹配过程和匹配过程所产生的结果。因此，因子库是整个系统的关键，是匹配过程的标签值，充当实时新闻事件与历史新闻事件之间的桥梁。

分词的速度与效果都取决于词典的结构[3]。初始因子库由股市基本术语大全和常用语大全直接导入。需要说明的是，在股票市场上，各主体的用词已经形成了一套完整的体系，体系内的用语习惯大体相同，没有太大的差异，因而因子库是已经包含了绝大部分关键词。汉语的词汇系统总是在不断发展和变化，“未登录词”的数量在理论上是无限的[4]，因此初始的因子库还需要不断训练和强化，处于一个动态更新的过程。

1.4 指标数据库

1.4.1 股价变动幅度库

每日的开盘价P_1与收盘价P_2的平均值作为波动均值，设该值为$Y=\frac{P_1+P_2}{2}$，将最高价P_{max}与波动均值之差作为最高价的波动距离，设该值为$M_1=P_{max}-Y$，$\frac{M_1}{Y}$为最高价的相对波动幅度；同样地，设最低价P_{min}与波动均值之差作为最低价的波动距离，设该值为$M_2=P_{min}-Y$，$\frac{M_2}{Y}$为对最低价的相对波动幅度。

假设某个事件的影响时效为 n 日，第 i 日的开盘价为P_{1i}，收盘价为P_{2i}，若最高价为P_{maxi}，若最低价为P_{mini}，则最高价波动幅度为$M_{1i}=\frac{P_{maxi}-Y_i}{Yi}$，最低价的波动幅度为$M_{2i}=\frac{P_{mini}-Y_i}{Yi}$，第 i 天日的波动幅度为$M_i=\frac{P_{maxi}-P_{mini}}{Yi}$。

在脉冲响应分析过程中，将脉冲响应程度作为新闻事件对每日股价变动的影响力。假设第 i 天的影响力为F_i，将影响力进行百分化，则第 i 日的权重为$Wi=\frac{Fi}{\sum_{i=1}^{n}Fi}$。那么 n 日内股价的相对波动范围的下限为：

$$MIN=\sum_{i=1}^{n}(w_i*M_{2i})$$
$$=\sum_{i=1}^{n}[\frac{Fi}{\sum_{i=1}^{n}Fi}*\frac{P_{min}-(\frac{P_{1i}+P_{2i}}{2})}{\frac{P_{1i}+P_{2i}}{2}}],\quad(1)$$

上限为：

$$MAX=\sum_{i=1}^{n}(w_i*M_{1i})$$
$$=\sum_{i=1}^{n}[\frac{Fi}{\sum_{i=1}^{n}Fi}*\frac{P_{max}-(\frac{P_{1i}+P_{2i}}{2})}{\frac{P_{1i}+P_{2i}}{2}}],\quad(2)$$

有效期内股价的波动幅度为：$\{MIN,MAX\}$。

1.4.2 宏观指数数据库

大盘指数和行业指数，均采用股价变动幅度库的建立思路，只是将股价换成相应的指数，就可以得到指数在新闻事件有效期内的相对波动幅度。将计算出的各个新闻有效期内的波动幅度存放在指标数据库中，并与历史新闻库建立对应的联系。

2. 匹配

2.1 常规匹配过程

一则新闻事件被爬取到以后，首先进入因子库通过最大匹配算法（MM）进行切分。对拆分完之后的关键词因子进行识别，若关键词因子中存在公司类关键词（公司名称、股票代码），将其作为优先关键词，返回该公司类新闻事件子库与其他关键词进行匹配，返回路径存在一对多的关系。匹配出的多个相似度记为向量组$\boldsymbol{\alpha_g}=$（α_{g1}，α_{g2}，$\alpha_{g3}\ldots\ldots\alpha_{gx}$）；若没有存在公司类关键词，则进入行业类与宏观类关键词识别。若出现行业名称、概念板块等关键词，将其作为优先关键词，返回提取出关键词的历史行业类新闻事件子库与其他关键词进行匹配，返回路径同样存在一对多的关系。行业类的记为$\boldsymbol{\alpha_h}=$（α_{h1}，α_{h2}，$\alpha_{h3}\ldots\ldots\alpha_{hx}$）；若未存在行业类关键词，则返回提取出关键词的历史宏观类新闻事件子库与其他关键词进行匹配，记为$\boldsymbol{\alpha_d}=$（α_{d1}，α_{d2}，$\alpha_{d3}\ldots\ldots\alpha_{dx}$）。

将相似度作为权重，进一步的返回历史事件所对应的股价波动幅度。每一个α_i（指一般匹配相似度）都对应一个$[Max,Min]$。

特别地，在公司类匹配过程中，有本公司有效消息，则匹配本公司的历史新闻事件、相应的股价及波动幅度数据。无本公司的历史数据，则用同行业其他公司数据估计替代。为了结果的准确，选择替代的公司在资产规模、销售规模、生产规模等方面尽可能的与本公司一致。

2.2 额外匹配过程

额外匹配过程主要是针对不能直接匹配出权重的情况。由于在股市消息中，会出现一些特殊情况，即其消息文本中没有公司名称或者股票代码，而是相关持有人，相关概念，其上下游产业等内容。这种情况则需要用到关联匹配的方式。本文所指的关联匹配是与优先关键词相互关联的词语，这种关联的关系，特别指与该只股票、板块、概念相关的词语。如股东、实际控制人、与该公司相关的活动等等，这些词语往往不会直接出现优先关键词的任何一个字。

匹配过程及逻辑关系见图 1。

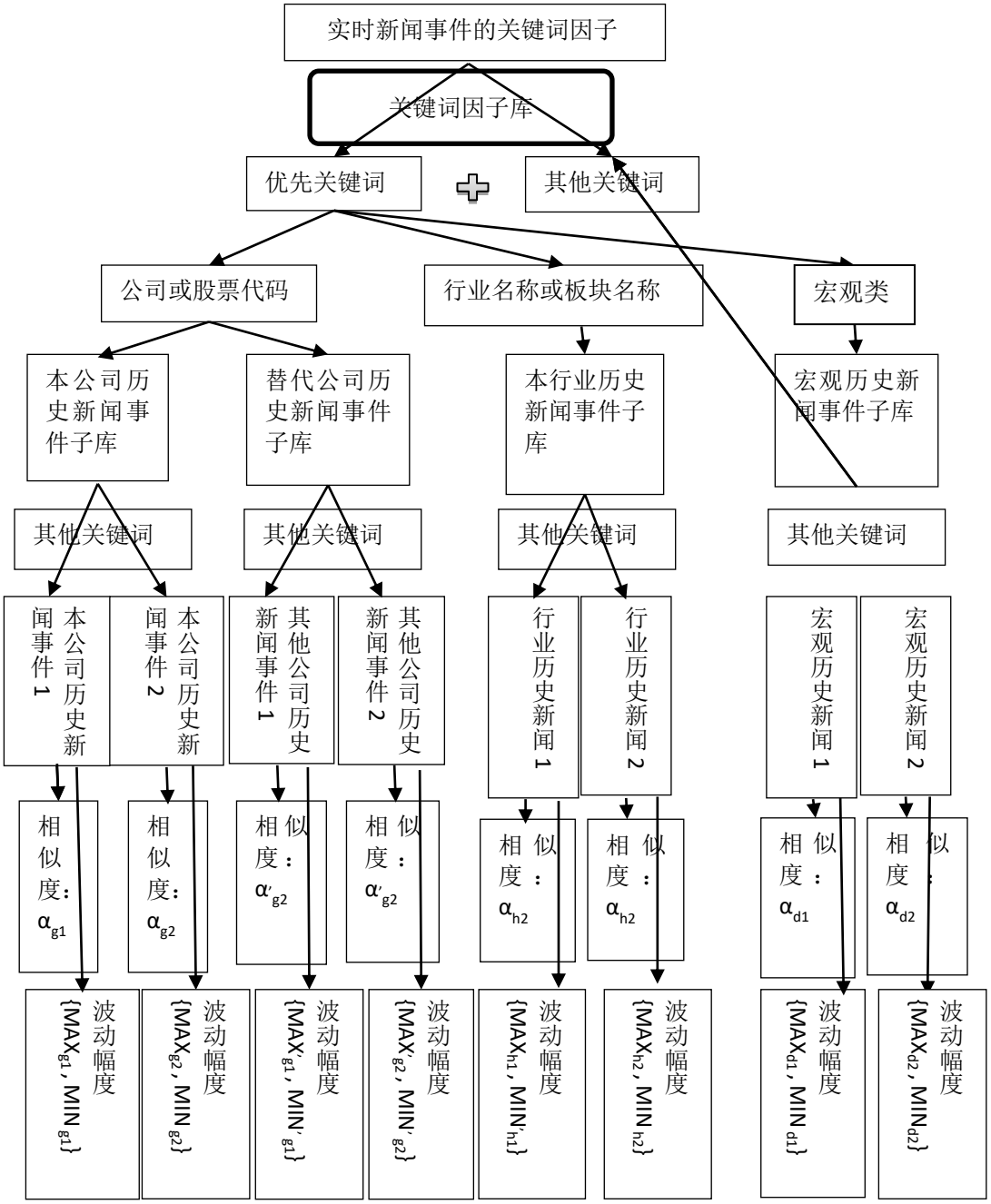

图 1 匹配过程图

3. 预测股价波动幅度

由于匹配结果全是有效匹配，本文把投资者对匹配相似度的认可程度分为高、中、低三档。高认可度是指可以接受所有匹配相似度，无论匹配度有多低；中认可度是指接受匹配相似度大于 50%，即匹配度小于 50%的匹配新闻事件不予考虑；低认可度是指接受匹配相似度大于 85%的新闻事件，小于 85%的事件不予考虑。排除两种极端情形（当日开盘价即为当日涨停价且同时是当日收盘价，以及当日开盘价即为当日跌停价同时也是当日收盘价）后，返回到历史新闻事件库时，设分别有 X 则历史新闻事件与之对应。对相应相似度做一个数据处理，使得所有匹配相似度之和为 1，相对相似度用在向量组$\vec{\alpha}$中所占权重来表示，第x事件的相对相似度表示为：

$$\alpha_x^* = \frac{\alpha_x}{\sum_{x=1}^{X}\alpha_x}\ ,\ \sum_{x=1}^{X}\alpha_x^* = 1 \qquad (3)$$

每一个匹配到的历史新闻事件都有与之对应股价的相对波动范围$\{MAX, MIN\}$，设M_{max} 为当日该新闻事件发生时相关股票对应的最大涨幅，M_{min}为当日相关股票向下的最大幅度。新的新闻事件进入系统，对应算出当日围绕均值最高涨幅、最高跌幅如下：

$$M_{max} = \sum_{x=1}^{X}\alpha_x^* \times MAX_x \qquad (4)$$

$$M_{min} = \sum_{x=1}^{X}\alpha_x^* \times MIN_x \qquad (5)$$

MAX_x表示第x事件在其事件有效期内所对应的加权涨幅；MIN_x表示第x事件在其事件有效期内所对应的加权跌幅；该新事件的时效性为匹配出的历史事件的有效期的平均值。

行业类、宏观类的计算步骤与上述相同，不同在于最后计算出来是指数的波动幅度。可进一步通过 CAPM 模型[5]($R_i - r_f = \beta_i(R_M - r_f)$)的 β 推导出具体某只股票的波动幅度。β 可通过回归估计。

$$M_{max} - r_f = \beta_m(\sum_{x=1}^{X}\alpha_x^* \times MAX_x - r_f) \quad (6)$$

$$M_{min} - r_f = \beta_m(\sum_{x=1}^{X}\alpha_x^* \times MIN_x - r_f) \quad (7)$$

β_m表示第 m 只股票价格变动与行业指数变动的敏感性。

结语

本文基于惯性原则,运用爬虫技术，将权威性的各大网站和软件上的实时新闻事件以及对应时间段内的股价进行爬取建立本地数据库。利用文本匹配技术，将各类事件作为系统中的初步关键词进行提取，同时采取 Var 模型的脉冲响应分析确定事件变化的时间范围。以关键词建立因子库，实时事件提取关键词与因子库进行匹配，预测与之相应的新闻事件和股价波动幅度。

致谢

本文为国家级 2019 年大学生创新创业训练计划项目《基于 AI 量化交易的科创板投资策略的探索》(S201910621021)的阶段性成果之一。

REFERENCES

[1] Zhou Jun. Method of Chinese words rough segmentation based on improving maximum match algorithm [A]. Computer Engineering and Applications,2014,50(2):124-128

[2] Fang Tingting. Empirical study on the fusion of dictionary segmentation and model segmentation in Chinese word segmentation [D]. Guangxi: Guangxi normal university,2019

[3] Wang Ruilei. An improved forward maximum matching algorithm for Chinese word segmentation [J]. Computer application and software,2011,28(3):195-197

（6）

[4] Qu Jianju. Sense prediction of Chinese unknown words based on knowledge base [J]. Journal of Chinese information processing, 2018,32(1):34-42

[5] Sun Jinhua. An empirical study on manufacturing sector coefficient based on CAPM -- a case study of manufacturing industry in Jiangsu province. Wuhan university of technology (Social science edition), 2016,29(6):1185-1189, DOI:10.3963/j.issn.1 671-6477.2016.06.0027