1J5-OS-13b-4

Extraction of Factors and Related Stocks of Individual Stocks Using Multiple Textual Data

Takuya Shikichi*1 Jinyang Fan*1 Kiyoshi Izumi*1*2 Kenta Yamada*1*3

*1 School of Engineering, The University of Tokyo
*2 CREST, Japan Science and Technology Agency
*3 PRESTO, Japan Science and Technology Agency

In this study, we proposed a new method for extracting factors and related stocks which affect individual stocks. We combined two text-mining methods which are the CPR method for news articles and the TF-IDF method for summary of financial statements. We showed how individual stocks are connected through factors in each term.

1. Introduction

1.1 Background

Recently, many individual investors have come to participate in the equity investment. However, it is very difficult, especially for those individual investors, to make investment decisions instantly because there are numerous factors that affect the stock market. Therefore, the need for technology to help investors has been increasing. And furthermore, the textual information that is directly or indirectly relevant to equity investment is explosively increasing on the web in recent years. In consequence, there is a significant increase in the research on discovering the relationship between textual information and market movements by using text mining technique [1]. It is expected that the numerical information such as economic index and technical indicators of market can be extracted quickly and automatically from the textual information.

1.2 Related work

Several studies have been conducted to analyze the movements of the financial market by using textual data. Zhang et al. [2], evaluated the newspaper articles, and successfully showed the correlation between the evaluation value and the stock price volatility of individual stocks. However, they didn't give detailed influencing factors of that. Izumi et al. [3], used Monthly Report of Recent Economic and Financial Developments and the CPR method to analyze the factors of long-term movements in the financial markets. Due to lack of information that related to specific companies' activities, the factors which have been extracted are always the factors that have influence on the whole market. Therefore, it is difficult to interpret their relationship with individual stocks. In this study, by complementally using multiple textual data, we are trying to extract regularity that even individual investors can interpret from the market.

1.3 Objective

In this study, by using multiple textual data, we extract the influencing factors of individual stocks and correlated stocks. While taking advantage of the CPR method which is able to analyze the factors of market movements, we also use another textual data which contain information about company activities. By doing this, it is considered that factors that affect individual stocks can be extracted and we can also obtain a factor-based stock relation network.

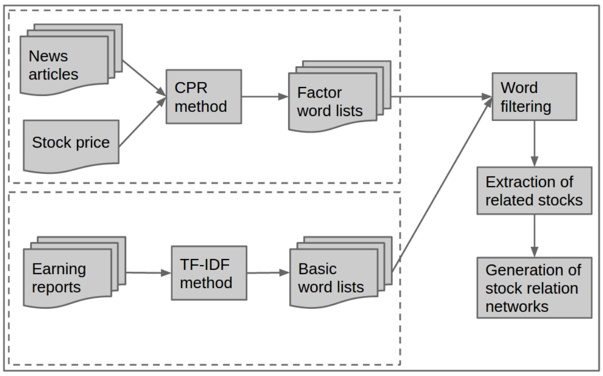

Figure 1: Analysis method overview

2. Analysis Method

2.1 Overview

Figure 1 provides an overview of the analysis method in this study. On the basis of newspaper articles and stock price, we use the CPR method to extract the most important words that affect individual stocks. Then, for each stock, we can obtain a factor word list which consists of the influencing words and the value that shows how much it affects this stock. On the other side, TF-IDF method based on earning reports of each company is utilized to help us extract the basic word list for each stock. When these two lists have been successfully extracted, we conduct the word filtering by combining these two lists. Finally, the common words are used to generate the stock relation networks.

2.2 Generation of factor word list

The CPR method consists of three steps that include Co-occurrence Analysis, Principal Component Analysis and Regression Analysis. First, in the Co-occurrence Analysis, morphological analysis was performed to divide a sentence in each article into words. Among the words we obtained, only nouns, verbs, adjectives are extracted. We regard each pair of adjacent words as a combination, And the words in the combination will be counted if there is at least one word in the combination appears in the Nikkei thesaurus [4] which is a comprehensive dictionary in financial area. At that time, only the words that their times of appearance over a stated threshold value will be used to generate an appearance pattern matrix. Then, Principal Component Analysis is performed for the matrix to reduce the dimensionality of original data. Finally, we use these principal components as independent variables and use relative rate of change of stock price as dependent variable to perform Multiple Regression Analysis. Equation (1)(2) show the definition of relative rate of change of stock price. Since the news article data we use is data from Nikkei Newspaper morning edition, after being issued on the morning of the day, the market is considered to be reflected accordingly from the openning price of that day. Therefore, the rate of change of stock price $r_{i,t}$ is defined by $O_{i,t}$ which means opening price of day t, and $C_{i,t-1}$ which stands for the closing price of day $t-1$.

$$r_{i,t} = \frac{O_{i,t} - C_{i,t-1}}{C_{i,t-1}} \tag{1}$$

$$r'_{i,t} = r_{i,t} - R_t \tag{2}$$

On the basis of factor loadings and the regression coefficients of the principal components that we acquired by the CPR method, we can define absolute influencing degree $E_{i,k}$ on a specific stock i for word k as following equation.

$$E_{i,k} = \sum_{j=1}^{n} |a_{i,j}\mu_{k,j}| \tag{3}$$

$a_{i,j}$ refers to the regression coefficient of the j-th principal component during the regression analysis of word k. $\mu_{k,j}$ indicates the factor loadings for the word k has on the j-th principal component. Furthermore, considering the mean μ_i and the standard deviation s_i of the absolute influencing degree $E_{i,k}$ for stock i, we define normalized influencing degree $E'_{i,k}$ as equation (4).

$$E'_{i,k} = \frac{10(E_{i,k} - \mu_i)}{s_i} + 50 \tag{4}$$

For each individual stock, we generate a word list with the influencing degree $E'_{i,k}$ for each period. This is called factor word list.

2.3 Generation of basic word list

TF-IDF method is a commonly used method for weighting words in the field of information retrieval. The weight of keyword i in document d is defined by equation (5). $tf_{i,d}$ represents the frequency of occurrence of the keyword i in document d. N is the total number of documents. df_i indicates the number of documents including keyword i.

$$w_{i,d} = tf_{i,d} \times \log \frac{N}{df_i} \tag{5}$$

The tf-idf value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in all documents, which helps to adjust for the fact that some words appear more frequently in general. Therefore, we conduct tf-idf method to the data of earning reports of individual stocks. Here, we only extract nouns after morphological analysis and then calculate tf-idf value for each word. Finally, we can generate a basic word list with weight value for each stock.

2.4 Word filtering and extraction of related stocks

The next phase is called word filtering which will use the factor word list obtained by the CPR method and basic word list obtained by TF-IDF method. If the word included in one list is contained in the string in the other list, or if there is a partial match for the string in both lists that has at least three characters, the words in the two lists are determined to be synonyms and the word in the factor word list will be hold. From the final word list after word filtering, we can extract correlation between stocks through the highly influencing words. According to the factor-stock relations we obtained, we generate an undirected graph to represent the network. Factors and stocks are shown as node, and edge indicates a correlation between them. In this study, we call this network stock relation network. Besides, during the generation of the network, the edge with very low influencing degree and nodes with very low or high degree is excluded.

3. Experiment

3.1 Data and parameters

The stocks are chosen from TOPIX100 which is an index that consists of 100 most liquid and highly market capitalized stocks in the Tokyo Stock Exchange 1st Section. However, we only use following 82 stocks due to lack of earning reports or stock split in 2012.

Table 1: Individual stocks being used

国際石開帝石	大和ハウス	武田薬品	積水ハウス
アサヒ GHD	キリン HD	味の素	セブン＆アイ HD
東レ	旭化成	信越化学	三菱ケミカル
花王	武田薬品	アステラス製薬	エーザイ
第一三共	オリエンタルランド	富士フィルム	資生堂
JX	ブリヂストン	旭硝子	新日鐵住金
JFE	住友金属鉱山	住友電気工業	SMC
クボタ	ダイキン工業	日立製作所	東芝
三菱電機	日本電産	富士通	パナソニック
ソニー	京セラ	村田製作所	日産自動車
いすゞ自動車	トヨタ自動車	ホンダ	スズキ
富士重工業	ニコン	キャノン	リコー
大日本印刷	任天堂	伊藤忠商事	丸紅
三井物産	東京エレクトロン	住友商事	三菱商事
三菱 UFJ	りそな HD	三井住友トラスト	三井住友 FG
横浜銀行	みずほ FG	オリックス	大和証券 G
野村 HD	MS ＆ AD	第一生命	東京海上
T ＆ D HD	三井不動産	三菱地所	住友不動産
JR 東日本	JR 西日本	ヤマト HD	日本電信電話
中部電力	関西電力	東京ガス	大阪ガス
セコム	ファーストリ		

The whole data set covers the period from April 4, 2012 to December 28, 2012. Data of each 30 days is used for learning and the data of the next day is used for forecast test. In this way, we separate the data into 217 periods. The news data we use is Nikkei Newspaper morning edition and earning report data is provided by [5]. They have extracted textual data from the PDF files that have been posted on the web page of each company. In the setting of the CPR, We set the threshold of appearance pattern matrix to 2 and set the upper limit of the number of principal components to 15. In addition, for the validation of the prediction accuracy of the CPR method, we use the average percentage of correct answers of each individual stock in each period.

3.2 Result and discussion

Average prediction accuracy of all 82 stocks in 217 periods results in 59.2%.

Then, we focus on a specific time period and extract the factors that affect the individual stocks and related stocks. On June 8, 2012, former Prime Minister Noda announced “there is a need to restart the Oi Nuclear Power Station.”. Such news is likely to have a significant impact on the market, therefore we focus on the period 93 that from May 21, 2012 to June 29, 2012. Also, we show an example of the experiment result by focusing on TOSHIBA which is one of nuclear-ralated stocks. Table 2 shows the words that have high influencing degree in the factor word list of TOSHIBA in the period 93. Table 3 shows the words that have high weight value in the basic word list of TOSHIBA. Subsequently, the result of word filtering is shown in Table 4.

Table 2: Impact word list(top 20) of TOSHIBA in the period 93

ＴＯＰＩＸ	マンション	ニュー	フェイス
富士電機	いじめ	サーバー	ブック
ポスト	南シナ海	和歌山	従業員
分かれる	資本流出	ラジオ番組	ヘルスケア
京都大学	金融危機	避ける	医療保険

Table 3: Basic word list(top 20) of TOSHIBA in the period 93

東芝	電子デバイス	プロダクツ	電器
東芝テック	ウェスチングハウス	医用	ストレージ
セミコンダクター	ランティス・ギア	LSI	IBM
白物家電	半導体	パソコン	照明
フラッシュメモリ	家庭	ケ月	フラッシュ

Table 4: Example of result of word filtering of TOSHIBA in the period 93

マンション	ニュー	フェイス	サーバー
ヘルスケア	マイナス	ニュース	電力消費量
DRAM	買収防衛	生産拠点	対外証券投資
パソコン	太陽光発電	証券投資	インテック
参事官	ベアリング	水力発電	昭和電工

As can be seen from the factor word list of TOSHIBA in Table 2, the words that affect the entire market such as “TOPIX” and “金融危機 (financial crisis)” are extracted. According to the basic word list in Table 3, the words that seem to be related to TOSHIBA’s business activities are extracted. Result of word filtering in Table 4 shows that the words related to electricity have been successfully extracted such as “電力消費量 (power consumption)”, “太陽光発電 (Solar power)“, “水力発電 (hydro power)”. Compared with the factor word list in Table 2, general words like “TOPIX” decreased effectively. As a consequence, proportion of the words that exactly affect TOSHIBA increased.

Table 5 shows the result of extraction of stocks related to “原子力発電 (nuclear power)” in the period 93. In the table, only the stocks with influencing degree over 60 are defined as related. For comparison, Table 5 also lists the same result in the period 170 (September 6, 2012 ∼ October 19, 2012) on the right side.

Table 5: Related stocks of “原子力発電 (nuclear power)” in different periods

Period 93		Period 170	
Stock	Influencing degree	Stock	Influencing degree
中部電力	67.1	三菱電機	71.3
関西電力	65.2	住友金属鉱山	70.6
大阪ガス	64.8	丸紅	64.3
東京ガス	61.4	東芝	61.2
		第一生命保険	61.2
		日立製作所	60.1

For the period 93 in Table 5, electric power companies such as “中部電力 (CEPCO)” and “関西電力 (KEPCO)” and companies related to infrastructures like gas such as “大阪ガス (OG)” and “東京ガス (TG)” have been extracted as related stocks. In the period 170, nuclear reactor manufacturers such as “TOSHIBA” and “日立製作所 (HITACHI)” and nuclear power plant manufacturers such as “三菱電機 (ME)” are extracted. Such results illustrate that related stocks can be successfully extracted. Furthermore, different period shows different related stocks.

To represent the stock relation network, we choose 7 stocks which are “国際石油開発帝石 (INPEX)”, “積水ハウス (Sekisui House)”, “ブリヂストン (Bridgestone)”, “東芝 (TOSHIBA)”, “トヨタ自動車 (TMC)”, “ホンダ (Honda)”, “関西電力 (KEPCO)”. The result of generating the stock relation network is shown in Figure 2.

Then, it is necessary to verify if the extracted relationship is legitimate. We conduct a questionnaire survey about the stock relation networks we have extracted. Subjects are selected from participants of an academic meeting of financial IT technology. The survey is designed to ask the general awareness of the relationship that shown in the stock relation networks. The awareness is divided into 4 levels including “well understood”, “understood”, “somewhat understood”, “poorly understood”.

Overall, 11 financial experts and researchers responded to the questionnaire. From the results, 43 pairs of relationship are obtained. Relationships with low awareness which are replied as ‘somewhat understood” and “poorly understood” are shown in Figure 3. For instance, on May 30, 2012 which in the period 93, “電力・都市ガス大手が一斉値上げ 7月、現行制度で全社最高値 (Big electric power stocks and city gas stocks raise simultaneously. Reaches highest price in the present system in July.)” was announced [6]. It turns out that relationships that noticed by financial practitioners such as “都市ガス (city gas)” and “関西電力 (KEPCO)” have

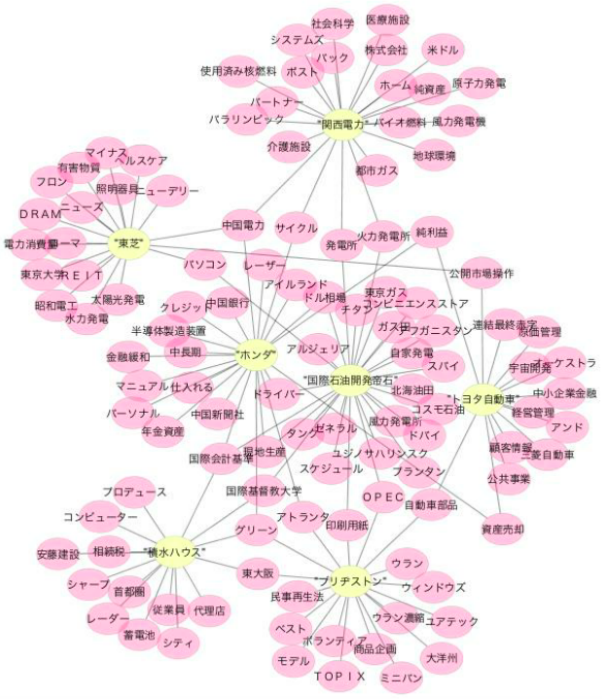

Figure 2: Stock relation network in the period 93

been successfully extracted.

4. Summary and outlook

By using multiple textual data, it is possible to eliminate the words that affect the entire market, and to extract the words that affect the individual stocks. In addition, resonable relationships of stocks and keywords, meaningful to financial practitioners and individual investors can be extracted from the generated network. What ’s more, it turns out that some surprising and less aware relationship can be extracted.

Considered further work will focus on two aspects as follows:

- Further improvements in filtering algorithm can be achieved by generating a specialized synonym dictionary.
- Develop investment support system that can make response to the key words in news that affect the individual stocks, and display the stocks and factors associated therewith.

References

[1] 和泉潔, 松井藤五郎.“金融テキストマイニングの紹介”. コーパスとテキストマイニング. 石田基広, 金明哲 編. 共立出版, 2012, p. 15-24, ISBN 978-4-320-11033-5

[2] 張へい, 松原茂樹. “ 株価データに基づく新聞記事の評価. ” 人工知能学会全国大会論文集 JSAI08 (2008): 81-81.

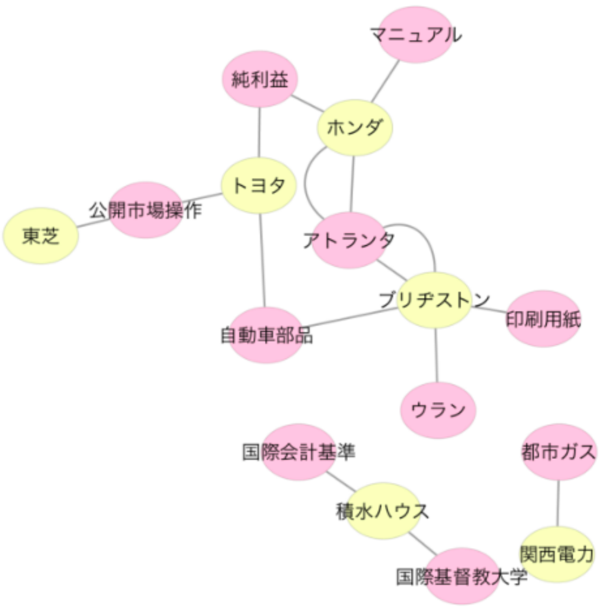

Figure 3: Relationships with low awareness in the period 93

[3] 和泉潔, 後藤卓, 松井藤五郎. “ テキスト情報による金融市場変動の要因分析. ”人工知能学会論文誌 25, no. 3 (2010): 383-387.

[4] 日本経済新聞デジタルメディア. 日経シソーラス http://telecom21.nikkei.co.jp/help/contract/price/23/help_KIJI_thes.html

[5] 坂地泰紀, 酒井浩之, and 増山繁. “ 決算短信 PDF からの因果関係抽出に基づく過去事象間の関連表示システム. ”人工知能学会全国大会論文集 28 (2014): 1-4.

[6] 日本経済新聞オンライン版「電力・都市ガス大手が一斉値上げ　７月、現行制度で全社最高値」http://www.nikkei.com/article/DGXNASDD300HH_Q2A530C1TJ1000/