4I1-OS-16-3

Indirect Factors to the Stock Price Prediction via Google Trends

Wang Peng[1], Kiyoshi Izumi[12], Shinobu Yoshimura[1]

[*1] Department of Systems Innovations,School of Engineering,the Univ. of Tokyo

[*2] CREST, JST

Stock price can be effected by many factors. This paper aims to put forward a framework of detecting and evaluating the indirect factors to specific stock. Our framework is constructed to obtain the economic behaviour of retail investors by appearance frequency of a word in the Google search result. (It is named as indirect factor because in some case it is difficult to explain the relationship between the high frequency words and a specific stock.) Then we prove the existence of correlation between the trade data and the search volume of keywords. As a sample set in this paper, daily stock price series of five corporations in Japanese stock market are selected from 2010 to 2012. The experimental result shows good performance that the trend of correlated keywords are possible to predict the trend of stock price before a few days indirectly.

1. Introduction

Recently many studies have shown that the search volume in the Internet may cause some economic behaviour, which has been proven by numerous researchers. ZHI DA et al.[1] indicate clearly that the weekly search volume did have performance in the mid-long term prediction among groups of experiments. According to their research, the search queries show the attention related market from the retail investors, which may lead to an economic activity in a future time. Foucault et al.[2] found that the retail investor’s activities regularly increase the volatility of individual stocks through an experiment taken in France.Our research paid close attention to similar objects as well, however, which is in the Japanese stock market.

Thomas Dimp and Stephan Jank[3] have proven that the volatility of the search amount of the word “dow” has a strong relationship with the volatility of DJI(Dow Jones index). Note that they only focuses on the search volume of a single word, which is extended to that of multiple words in this paper. This kind of method even can be used to predict the real world. Hyunyoung Choi et al.[4] have found the relationship between the search amount and economic indicators, for example, the unemployment rate. However, contemporaneous forecasting is the main topic of their research.

There exists a problem that many key terms are prescient and one-sided. In some researches, these keywords are even obtained by hands. In this paper, we tentatively put forward a novel and automatic framework to capture the related terms/words of a specific stock. Then the correlation between this words search volume and the trade volume or price of an individual stock is proved by regression.

This paper contributes to the multiple and manifold terms based prediction via GT data. We focus on Japanese stock market here. Future more, this framework can be applied into various of market, hence this framework is easily extensible and improvable as a baseline method.

Contact:Wang Peng, Department of Systems Innovations,School of Engineering,the Univ. of Tokyo, 7-3-1 Hongo Bunkyo-ku 113-8656 Tokyo, Japan, E-mail: wangpengtoo@gmail.com

2. Framework and method

2.1 Generalization process

There are four principal steps in this framework. As a basic assumption, google search can provide relevant content of the searched terms as far as possible. Instead of using the text contents of the homepage, we prefer to obtain text from a search engine, which provides more comprehensive information. Another benefit is clawed data coming from many different and objective informants. Subsequently all of the contents are segmented into independent terms. Original related word set is contained in the noun segments while the words of other part of speech are not picked out from the corpus. The final keyword set that will be used to download the GT data is extracted from the original noun set.

Through the aforementioned process, all necessary data have been captured. The second procedure of the framework is correlation detection. After that, regression is applied to analyze the relationship between the GT of keywords and the everyday trade data of an individual stock. Even we can analyze the direct and inverse proportionality of the independent variables and dependent variables. If there exists a correlation (actually the experiments prove that there are some correlations indeed), it is necessary to test the reason giving rise to this relevance. In consequence, Granger causality test this dependency relationship.

The brief procedures in this framework are shown as follows :

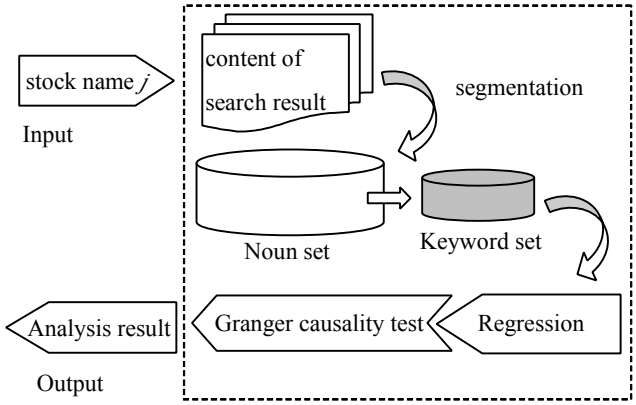


Fig.1 basic procedures of correlation detecting model

This is a preliminary method to reduce artificial selection that may be pretty time consuming. The input of our framework is a stock name, while the out put is relevance estimation.

As a very important way to obtain the search volume, we use GT (Google Trends), from which we can download the weekly even daily search data of a specific word. There are already many researchers use these data to predict the future. For instance, Kira Radinsky et al. predict top terms that will prominently appear in the future news via GT data.[5]

2.2 Word selection

Word selection is an approach proposed reducing the redundancy and selecting the most probable relevant words. The same as the basic hypotheses of TF-IDF, it is supposed that the frequence of terms are associated with the relevance to the target word. Nevertheless we do not concern about the term importance in semantic implication, but the importance of search volume series. So TF-IDF is not employed here.

Clawing text data begins with searching the stock name in Google. As the relevant contents are obtained from the search engine, the original noun sets can be computed out by NLP approach.

Prior to explain the filter method, intuitively, it is more likely to have relationship depending on the peculiarity of the noun set. So the unique word set is extracted from every original set. However, we do not deny that the mutual words relate to the target word either. To interpret the word filter, some set representations are used in this section. Let $S_j = \{w_1, w_2, \cdots w_{n-1}, w_n\}$ be the original noun set of stock *j*. Therefore $SN = \{S_1, S_2, \cdots S_{n-1}, S_n\}$ is the set of all stocks. Then we can get the final keyword set of stock *j*, which is defined as USE_j.

$$USE_j = \{x_m \in S_j, x_m \notin (\sum_{k=0}^{n} S_k - S_j) \mid m = 1,2...n\}$$

This method is extensible friendly. For instance, we can set a similarity threshold of noun sets to control the word filter procedure.

2.3 Multiple Linear Regression(MLR)

Many previous researches use VAR(vector auto regression) as their regression model. However it is difficult to use primordial data in this procedure, in virtue of non-stationary stock volatility. They generally preprocess the raw trade data before regression. Consequently, we prefer employ the MLR with lag parameter considered.

MLR is a extending variable statistical method built on linear regression model which is an an approach to model the relationship between a scalar dependent variable y and one or more explanatory variables denoted X.

Since MLR is capable to estimate the correlation or importance of pertinent factors, it has been applied in many fields for a very long term. Stock market is no exception. Lawrence Harris[6] has used MLR to examine individual stock price clustering distribution. MLR is an efficient approach to evaluate the result of word selection and the importance of various words.

In our model, the independent variables are lagged and normalized GT time series that is expressed as xi. Daily trade volume and price are the dependent variables in this regression. Two MLR procedures can be presented as following equations.

$$price = a_0 + \sum_{i=1}^{n} a_i x_i + \varepsilon \quad \text{when} \quad x_i \in USE_j$$

$$tradeVolume = a_0 + \sum_{i=1}^{n} a_i x_i + \varepsilon \quad \text{when} \quad x_i \in USE_j$$

By this way, the elementary interaction can be estimated. We have considered about lags during our experiments.

2.4 Granger causality test

The Granger causality test is a statistical hypothesis test for determining whether one time series is useful in forecasting another.[7] Specifically in this paper, Granger test give a further probable interpretation to the correlation we captured from the previous regression step. We take logarithim of all parameters to reduce the skewness and kurtosis, which is the same normalizing approach as T Dimpfl et al used when they built VARmodel.

Let y be the first order difference of stock price time series and x be the first order difference of search volume series. the tow steps of Granger test are presented as following equations:

$$\log(y_t) = a_0 + \sum_{m=0}^{t-1} a_m \log(y_m) + \varepsilon$$

Next, the autoregression is augmented by including lagged values of x:

$$\log(y_t) = a_0 + \sum_{m=0}^{t-1} a_m \log(y_m) + b_0 + \sum_{n=0}^{t-1} b_n \log(x_n) + \varepsilon$$

If the series of x contribute to explain significantly the future value of y, x is called the Granger cause of y. However, this test is taken for the purpose of adminicle to regression, since Granger causality is time-sensitive.

3. Experimental result

3.1 Development setting

A number of development tools have been used in our experiment. Five corporations in chemical industry (Shin-Etsu Chemical Co., Ltd.(Shin'esu), Kao Corporation, Takeda Pharmaceutical Co., Ltd., Nippon Steel & Sumitomo Metal Corp. (Shin'ni), and Astellas Pharma Inc. (Astellas)) are chosen, of which the top 5 pages search results in Google, about 60 to 70 links included, are obtained as the initial text collection. Then we employ Lucene-gosen (http://code.google.com/p/lucene-gosen/) in relevant sentence segmentation. The noun segment set, in which the words can be used to present some objects, containing geographic name, organization name and common noun. In this experiment, we use the last kind of nouns, while we do not deny that there exist correlations between those nouns and the trade series.

Except segmenting, the rest procedures are developed by Python. Data analysis is processed in R platform.

3.2 Word filter

There are about 800~1500 words in the noun set that the

frequence is more than five times. While the amount of words rapidly declines to about 200~300 when the threshold is set into ten times. After the final filter the amount drops to about 5~30, while increase the precision much more.

The primary terms in this set boil down to raw materials, products, advertisement, customer-related terms. Here we show some word in the final set.

Shin'etsu: Silicone, Quartz, Derivative, Cellulose, Pheromone.....
Kao: Cat, Detergent, Hamming, Housework, Soap, Hair, ...
Takeda: Pharmacy, Vitamin, Alinamin, Kampo....
Shin'ni: Steel plate, Volleyball, Wind power, Stainless...
Astellas: Renal, Syndrome, Bladder, Hardware....

3.3 Result of Regression

In order to test whether it is time-sensitive for the regression analysis, we run 8 times, which lag ranges from 0 to 7 weekdays. As well as the original data set contains nonstationary series, especially for the stock price volatility, VAR model cannot be applied into the original ones. The search volume data is download from the homepage of Google Trends (http://www.google.com/trends). Due to statistical reasons for the system, GT claims that the trends data may be a little bit different from time to time.

The study phase lasted three years from 2010 to 2012. As employing the normalized daily GT data, we aim at looking for a shorter term correlation within seven weekdays. Historical daily trade volume and price are captured from Yahoo finance (http://finance.yahoo.co.jp/).

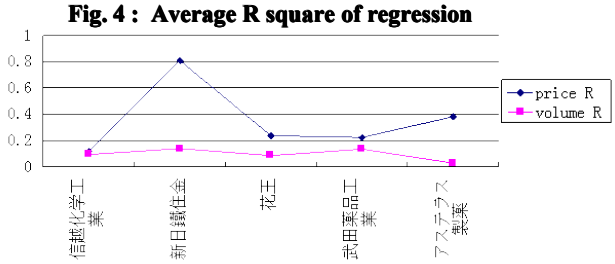
Table 1 is the percent of significant correlated elements to the trade volume and trade price. In this table, the left column of each company is the regression result of volume, while the other side is that of price. The last line of this table is the percent of non-obvious related words, among which the optimal result is boldfaced.

Table 1: Proportion of related terms

	Shin'etsu		**Kao**		**Takeda**		**Shin'ni**		**Astellas**	
	Pri	Vol	Pri	Vol	Pri	Vol	Pri	Vol	Pri	Vol
***	0.5	0.25	0.35	0.06	0.5	0.53	0.35	0.13	0.8	0.03
**	0.21	0.08	0	0.06	0	0.09	0	0.09	0	0.08
*	0.13	0.23	0.06	0.09	0.06	0.13	0.06	0.07	0	0.03
.	0.08	0.13	0.06	0.08	0.06	0.06	0.06	0.05	0	0.1
no	**0.08**	0.31	**0.53**	0.71	0.38	**0.19**	**0.53**	0.66	**0.2**	0.78

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Fig. 4 : Average R square of regression



Besides, parts of fitting results are presented in Fig2 and Fig.3 that the lag is two weekdays, because the data is two days latter in the actual conditions.

Fig.2 is the Residuals vs Fitted of LRM of five corporations. The black points in these graphs stand for residual and the red lines present every fitting linear model, while this figure gives a visualized expression to the relation between residual and model. If points distribute on the either sides of red line equably, it becomes an efficient regression. Fig.3 is Normal Q-Q graph of regression, in which the black points are residuals as well. The five sub-graphs in Fig.2 and Fig.3 are listed by the order of Shin'etsu, Kao, Takeda, Shin'ni and Astellas .

Fig.4 shows the average R squares of all regression, in which the blue points are the R square of trade price regression, well the pink ones are R square of trade volume regression.

4. Analysis

4.1 Regression effect

It is obvious that the regression of trade price is much better than the regression of trade volume, for the reason that all proportions of related term of price regression exceed that of volume, except the corporation of Takeda. Percent differed most significantly in the company of Astellas from about 20% sharply rising to 78% in the Table 1. This result provide evidence that we indeed extract the relevant words by filter.

From the following two Figures, it is feasible to analyze the regression effect of five corporations. In Fig.2 it is evident that the residuals if Buta is not distributed equally to the two sides of fitting red line, especially for the right side of this fitted line. What's more, there exist more outliers than other regression. Inversely, the residuals distribution of Shin'etsu, Kao, Shin'ni are satisfactory, the black points are balanced on either side.

According to our experiment, there are not too many residual outliers for these regressions, expect Buta, which we can infer from Fig. 3. As we can see in this graph the residuals mainly distribute along with a line, although there are some points dispersed in both ends. According to Fig. 3, we can infer that they follow normal distribution, because majority points distribute at the interval from -1 to 1.

4.2 Regression credibility

Almost the same inference can be deduced by the difference of R square. In Fig. 4 the R square of price regression is higher than R square of volume regression no matter for which corporation. So we infer that the regression based on our framework perform better for the price trend instead of trade volume. It is worth noting that the average R square of formers is about 35.38%, relatively, the other one is just less than 10%. Particularly, the average R square of Shin'ni even reach to about 0.8, which is the top place of our experiment. The R square measures the reliability of regression result. There is no doubt that the performance of R square is not only caused by increasing correlation terms, but also the representativeness for a company. For example, Kao even contains more relevant words than Shin'ni, however the R square of Shin'ni significantly precede the former.

Through granger test, we found causality relation between the

log of search volume and trade series, which is similar with the previous researches. This test avoids regression from the risk of spurious regression. However, just as we have known, the correlation is hard to be complete and direct causality. It is apparent that the high search volume shows the highlight of a specific object, while the origin may come from many aspects. Analogously, stock price is affected by diverse of factors. As a result, our framework is defined as an unforeseen factors detection method.

5. Conclusion

Above all, we can glean two main conclusions: firstly, there exist correlations between the keyword set and the stock trade data; Then the regression of the stock price is better than that of volume. Our framework finally proved to be contributing to building stock price regression model in the Japanese stock market. Despite the fact that we did not use the VAR, the time factor has been taken into account by lag parameters.

The reason why we choose the Japanese market is that some researchers questioned English tweets stock market forecasting methods because there are many countries in the world using of English. However, if we perform text mining confined in Japanese, this result is likely to be more close to the local market in fact.

In the technical melioration, an efficient method has been proposed. We will consider optimization approaches such as vocabulary filter method or VAR-based model. This method may predict the stock trend in an advantageous way. If possible, this framework will be integrated into the SNS based some text mining technology.

For market application, this method is not limited to the stock market, while we attempt to apply this method into option market or the foreign exchange market.

References

[DA 2011] Da, Z., Engelberg, J. and Gao, P., In Search of Attention, The Journal of Finance, 2011

[Dimpfl 2012]T Dimpfl , Thomas, Jank, Stephan, Can internet search queries help to predict stock market volatility?, CFR Working Papers, University Koln EconStor, 2012

[Foucault 2011]Foucault, T., Sraer, D. and Thesmar, D. J.: Individual Investors and Volatility, The Journal of Finance, 2011

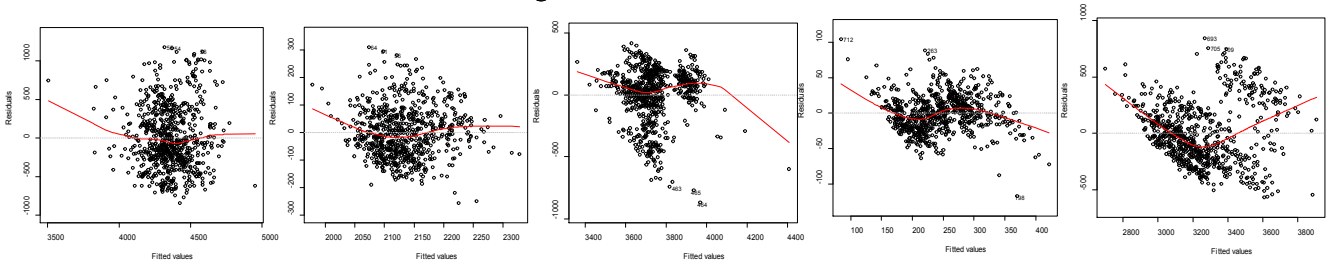
[Choi 2011] Hyunyoung Choi, Hal Varian, Predicting the Present with Google Trends, THE JOURNAL OF FINANCE, 2011

[Radinsky 2008]Kira Radinsky, Sagie Davidovich and Shaul Markovitch, Predicting the News of Tomorrow Using Patterns in Web Search Queries, International Conference on Web Intelligence and Intelligent Agent Technology, 2008

[Harris 1991]Lawrence Harris, Stock Price Clustering and Discreteness,The Review of Financial Studies, 1991

[Granger 1969] Granger, C. W. J., Investigating Causal Relations by Econometric Models and Cross-spectral Methods, Econometrica, 1969

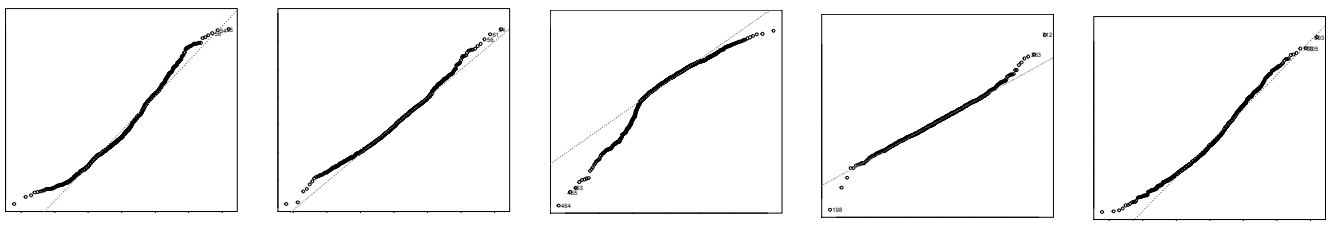
Fig.2 : Price LRM Residuals vs Fitted



Shin'etsu Kao Takeda Shin'ni Astellas

The abscissa is the ticket price, the vertical axis is the residual.

Fig.3: Price LRM Normal Q-Q



Shin'etsu Kao Takeda Shin'ni Astellas

The abscissa is the theoretical quantiles, the vertical axis is standardized residuals.