## A new graphic kernel method of stock price trend prediction based on financial news semantic and structural similarity

Wen Long1,2,3, Linqiu Song1,2,3, \*, Yingjie Tian1,2,3

1. School of Economics & Management, University of Chinese Academy of Sciences, Beijing 100190 P.R.China

2. Research Center on Fictitious Economy & Data Science, Chinese Academy of Sciences, Beijing, 100190 P.R.China

3. Key Laboratory of Big Data Mining & Knowledge Management, Chinese Academy of Sciences, Beijing, 100190 P.R.China

\* Corresponding author: cuplsonglinqiu@126.com

**Abstract**

Lots of researches try to predict the stock price trend using financial news. But almost all of them focus on the news contents while very few consider the information hiding in the relationship between different news, which may influence the stock price prediction. In this paper, we proposed a new kernel based on SVM concerning not only the inputs themselves but also the information structures among them. Medical industry financial news is used to illustrate the efficiency of our kernel. As both the news contents and the information structures are imported into our kernel, this kernel is named as semantic and structural kernel, referred to S&S kernel. By comparing the predicting accuracy of S&S kernel with other kernels, such as linear kernel, we find our method outperforms the other methods by at least 5% on accuracy, which is a quite meaningful promotion for predicting stock price trend. The result also confirms the information structure contained in daily news can offer extra information helping to predict the trend of stock price.

**Keywords:** Stock price trend prediction, financial news, information structure, S&S kernel

### Introduction

Many factors affect the stock price trend, such as market conditions, inflation, trading strategies, return on net assets and so on. According to the theory of effective market proposed by Fama (1970), asset prices should fully reflect all the information; even in the semi-strong effective market, the price should reflect all the public information in time. Therefore, if the information disclosed by the news media contains only public information, it will not bring excess returns for investors; even non-public information cannot bring any excess return under the strong effective market assumptions. In reality, many investors, especially individual investors, usually view media coverage as public information because it is easy to access and the cost is almost zero. Thus, if the media coverage can affect the stock price and bring excess return, the market is not effective.

Dyck and Zingales (2003) found that the earnings announcement published by the news media would lead to greater volatility in the stock price. Especially when less analysts pay attention to the announcement and the media is very reliable, the price fluctuations caused by the announcement are particularly evident. Shiller (2000) argued that the media played a role in fueling the rise or fall of the stock market. The cognitive bias can create an "overreaction" to good news and bad news. Therefore, how to extract the information affecting the stock price trend from media coverage to better master the market trends becomes very meaningful.

Existing researches can be broadly divided into two categories. The first category is based on the traditional method of economic analysis. Event research methods are used combining with asset pricing model or regression model to analyze the relationship between the media and the stock market. Scholars found that the release of media information can lead to the changes in stock prices, stock trading volume, stock volatility and stock returns. Some studies define the media's impact on stock returns as "media effects." The second category is based on the machine learning and other data mining methods. Combining with natural language processing and text classification, etc., researchers extract the "emotions", such as positive, negative or bullish, bearish, etc., by determining the semantics of news reports to analyze the investors’ view of stock market in a trading day and then predict the stock price trend.

Financial news has been identified having an effect on stock price. But does the relationship between financial news offer any other information? So far, there is few research considering that the relationship between financial news may influence the stock price prediction. We believe it may provide extra information helping us with the prediction, and this is the greatest innovation of this paper.

### Literature review

Investor sentiment research has been a hot spot of behavior finance. One of the most important hypothesis of classical economic theory is that investors are all rational so that sentiments have little influence on stock markets. However, studies of behavior finance show that some irrational phenomena can only be explained by investors’ sentiment factor. Most researchers tend to apply some machine learning methods on text information, such as financial news, stock comments of some social platform as well as corporate financial announcements.

Quantity methods have been applied to explore the relationship between investor sentiment and stock performance. W. Antweiler and M. Z. Frank (2004) collect more than 180,000 Wall Street Journal news stories from 1973 to 2001. Using computational linguistics method, they classify the stories according to topic, and find on average there is a reversal so that pre-event and post-event abnormal returns have the opposite sign. They also study the effect of more than 1.5 million messages posted on Yahoo Finance and Raging Bull, which shows stock messages can help predict market volatility significantly. Paul C. Tetlock (2007) quantitatively measures the media’s interactions with the stock market based on the daily news from a popular column of Wall Street Journal. It is found that pessimistic market sentiment can lead to downward pressure on the market price followed by a reversion to fundamentals. Also, extreme market sentiment is always accompanied by high market trading volume.

As most researchers use finance news or posted text on the internet to represent investor sentiment, text mining approaches are drawn into stock price prediction problem. R. P. Schumaker and H. Chen (2006) create the Arizona Financial Text System (AZFinText) by a synthesis of linguistic, financial and statistical techniques to predict the possibility of discrete stock price prediction. Xiangyu Tang, Chunyu Yang and Jie Zhou (2009) combine news mining and time series analysis to forecast inter-day stock prices. Nan Li and Desheng Dash Wu (2009) studies online forums hotspot detection and forecast using unsupervised text mining approach combing K-means clustering with SVM. Runpeng Huang, Wenming Zuo and Lingyan Bi (2015) grab data from Sina Blog and process them into an emotional tendency time series. By adding Blog sentiment into their SVM predicting model, the predicting accuracy becomes higher. Xuejing Meng, Yafei Yang and Xinquan Zhao (2016) utilize web crawler gaining massive text information from 9 well-known financial networks. Random forest is used to collect the features that influence stock index return most and a trading strategy is designed based on these features. Yong Shi, Jing Tang and Kun Guo (2017) construct investor focus index of stock forum, Xueqiu network, and financial news to analyze the impact of information from different sources on stock market.

Different materials are concerned when using text mining methods. Hua Chen, Xun Liang (2006) classify the internet stock news according to the stock characteristics. Yang Yu, Wenjing Duan and Qing Cao (2013) use a dataset contains daily media features of 824 public traded firms across 6 industries. F Ming, F Wong et al. (2014) uses daily articles from The Wall Street Journal to predict the closing stock prices. Andrew Sun, Micheal S. Lachanski and Frank J. Fabozzi (2016) investigate the textual information from user-generated microblogs to predict the stock market.

We also concern the graphical representation of text. At present, the common text representation methods are Boolean model, vector space model, probability model and so on. These models solve the problem of text in the computer, but they do not consider the contextual structure of the text and the context order of the feature word. And the loss of this structure information will affect the full expression of the text content. Therefore, in recent years there are still many scholars engaged in text representation of the study.

Lu (1990) proposed a conceptual representation method applying to the field of information retrieval. Russell and Norvig (1995) put forward the semantic web representation of knowledge in his paper. Inderjeet and Bloedorn (1997) presented an algorithm using graph model to extract multi-document abstract. The idea is to add the semantic information of the text to the model. And then Bhoopesh and Pushpak (2002) proposed a method to construct the document feature vectors according to the UNL diagram. These methods solve the problem that the text does not adequately express the text structure and relation to a certain extent, but these models are too complicated. Method measuring the similarity of the model still deficient.

Schenker (2003) proposed the application structure of the text representation for the first time in his doctoral thesis. Schenker's diagram structure model is based on the HTML file. But the edge of the weight information is not considered. Jin Wei and Srihar Rohini K. (2007) proposed a model based on graphs to capture word order, frequency, word co-occurrence, and word meaning. This model is then applied to discover unrelated words in the text library. Zonghu Wang and Zhijing Liu (2010) proposed a Chinese text categorization method based on graph model. They first use a weighting method to select the relevant feature building graph, then improve the text representation model and design a learning algorithm to classify Chinese text through the graphs. Wu Jiangning and Liu Qiaofeng (2010) proposed a Chinese text representation method based on graph structure which considers the semantic and word order information between words and phrases in order to solve the problem of information deletion based on statistical text representation model. Antoon Bronselaer and Gabriella Pasi (2013) propose a new graph-based model which decompose the text by tokens and then attached to a vertex or an edge. S. S. Kamaruddin, A. A. Bakar et al. (2015) describe a text mining system that is able to detect sentence deviations from a collection of financial documents. The sentences are represented as graphs through a dissimilarity function.

Through the existing researches, we can see almost all the scholars focus on refining their models to improve the predicting accuracy based on news contents only. Very few of them concern both contents and the relationship between the financial news when they study the impact of financial news on stock markets. We believe daily news structures contain some information that can help us predict the stock price trend. In this paper, the information structure contained in the financial news is extracted and is represented by graphs. In order to process the graphs of daily news, we propose a new kernel, which is specified to deal with data of graphs or networks. The kernel is then embedded into SVM model as the precomputed kernel and the model is used to predict the stock price trend. As the kernel concerning both semantic and structural information, we name it as S&S kernel. Besides, the kernel we proposed is not limited to process financial news. Any text data with graph or network structure could apply this kernel to obtain more information.

The rest of this paper is arranged as follows: The construction of our kernel is proposed in section Ⅲ. Each step of deriving our kernel based on SVM is explained in detail in this section. An empirical test is illustrated in section Ⅳ, where we use financial news of a stock to show the effectiveness of S&S kernel compared with the other kernels. In section Ⅴ, we add another 3 stocks to test the robustness of our model. The last section of this paper is the conclusion.

### Method

SVM (support vector machines) is first proposed by Vladimir N. Vapnik and Alexey Ya. Chervonenkis in 1963. Simply speaking, SVM is a tool used for classification and regression. Given a set of training examples, each assigned to one of the two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier.

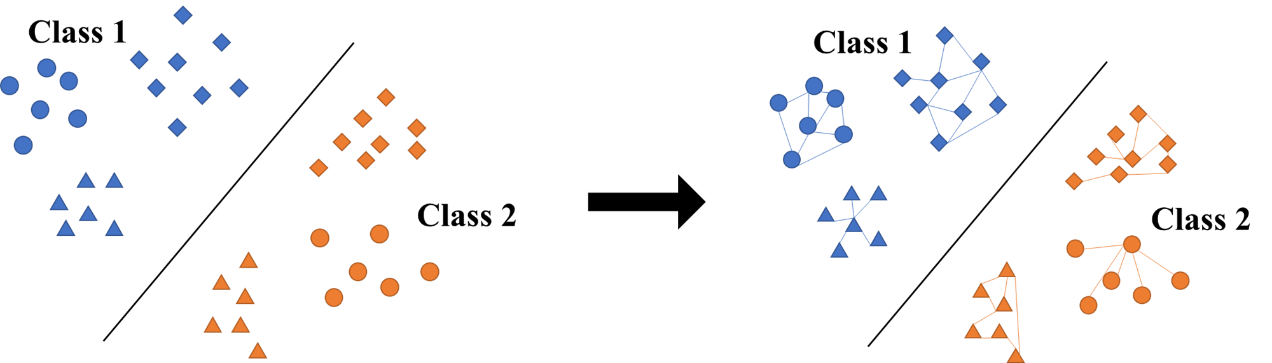
Given a training dataset of n points of the form , where the are either 1 or −1, each indicating the class to which the point belongs. Each is a p-dimensional real vector. The "maximum-margin hyperplane" is defined to divide the group of points so that the distance between the hyperplane and the nearest point from either group is maximized. The optimization of SVM with a linear hyperplane is:

(1)

The original SVM performs linear classification, however, there are some data are linearly non-separable. Bernhard E. Boser, Isabelle M. Guyon and Vladimir N. Vapnik (1992) suggested a way to create nonlinear classifiers by applying the kernel trick to maximum-margin hyperplanes. With the kernel trick, the inputs are implicitly mapped into a high-dimensional feature space, which can be divided by a hyperplane. The model is as follows:

(2)

The commonly used kernels include Polynomial kernel, Gaussian radial basis function (short as RBF) and Hyperbolic tangent kernel. These kernels have been applied to solve many classification problems. The input, however, is required to be in a vector form. For example, when using SVM to classify the sentiment of financial news, only the contents of the news are concerned. In order to consider the relationship between each two news, we need to add the structural information into the algorithm. So based on SVM, we propose S&S kernel, which concerns both semantics and structures. What should be noted is the kernel we propose here is not specific for stock price prediction via financial news. It can be applied to any other text data as long as there is a relationship between each pair of them.



**Fig. 1.** Standard SVM classification and the new proposed SVM classification

3.1 The construction of input structure

The construction of S&S kernel is a process of building multiple kernels into one. Basically, the process can be separated into two parts: semantic part and structural part. Visually, we can regard the semantic part as vertices and the structural part as edges as we construct the structure based on the relationship between each pair of vertices (which mean the semantics of the contents). We extract features from contents as vertices and then calculate the similarity among them. This part is defined as the semantic kernel part. Our main innovation is the structural kernel part. To define the structural kernel, we construct the edges between the vertices first.

In order to represent the relationship comprehensively, we develop two types of graph structures: text points graph and key words graph. One of the differences between these two graphs is the definition of the vertices: text points graph uses each piece of data as vertices while key words graph uses key words. The key words here can be obtained manually or automatically through algorithms like bag-of-words or any other text to vectors methods.

1. Text points graph

In text points graph, we regard each piece of text as vertices. As for the edge, the key words co-occurrence of each pair of texts are used as the weight. In detail, for each pair of texts (texts *s* and text *t*), we pick the key words appear in both texts and sum up the minimum frequency of each key words as the weights between the pair of texts:

(3)

For example, in stock price prediction via financial news, we take each financial news as vertices. The key words are manually defined by expert knowledge. The frequency of each key word *i* in each financial news is denoted as and the weight between the pair of *news s* and *news t* is the sum of the minimum co-occurrence which is . The text points graphs represent the connection among all the texts and the dimension of the graphs should be equal to the number of texts.

1. Key words graph

We use the key words information to supplement the text points graph, especially when there is only one piece of text under certain conditions, which would lead to the construction of text point graph becoming impossible. Extracted key words here are used as vertices. The key words can be obtained from expert knowledge or can be generated from certain algorithms. For each pair of key words and , we use the sum of the minimum co-occurrence of the words in each pair of text as the weight between the pair of key words:

(4)

For example, in stock price prediction via financial news, the frequencies of key word p and key word q in each piece of financial news are counted first, denoted as and respectively. The minimum co-occurrence of these two words in each financial news is calculated, . Then the frequency of each two words is the sum of the minimum co-occurrence through all the financial news. The key words graph is constructed and the dimension of these graphs should be equal to the number of the key words.

3.2 The definition of S&S kernel

Based on the graph construction, we can define kernels on each part of the graph. According to the segmentation between vertices and edges, the measurement is also divided into two parts: the similarity among the contents features and the similarity among the structures.

For the similarity among the contents features of each pair of texts, which is equivalent to the similarity among the vertices of each two graphs, we calculate the 2-norm distance between them and go through all pairs of the texts. Then use to define the similarity between these two graphs. What should be noted here is that there may be two pieces of texts having nothing in common, meaning the distance between them cannot be calculated. Considering such situation, we define the similarity of this situation to be zero. So, for *graph s* and *graph t*, the contents similarity between them is:

(5)

where denotes the contents features of *text s* and denotes the contents features of *text t*.

When calculating the similarity among the structures, we also need to calculate the distance in two parts based on our construction separately. For text points graphs, we need to deal with the unequal dimension problem as the dimensions of these graphs equal to the numbers of data each graph contains. Here we apply PCA to take the principal components of each piece of data first. Then the lower dimension of each pair of graphs is applied as the number of principal component we take. For example, and stand for the two graphs respectively, where and are the number of vertices in each graph correspondingly. For each row of and , we take their first principal components as their feature vectors and calculate the 2-norm distance traversing all pairs of vectors. The average of all vector distances is taken as the distance between and . Then is used to define the edge similarity between these two graphs.

(6)

For key words graphs, the calculation is much simpler as the dimension of each graph is all the same, which is the number of key words. We traverse the 2-norm distance between the corresponding key words of two graphs and use the average of all the distances as the distance of the two days’ graphs. Again, we use to define the edge similarity between these two graphs.

(7)

To denote the structural similarity, we use a hyper parameter to combine and together.

(8)

The parameter is determined by Grid search.

To combine the contents and edges together, we use the same idea as we denote the edge similarity. Another hyper parameter is applied to combine matrix and together.

(9)

Also, parameter is determined by Grid search. The kernel we proposed is based on the similarity we define above.

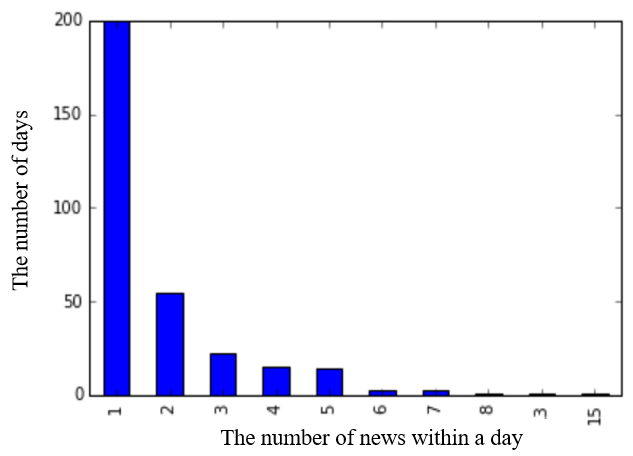
### Experimental test based on financial news and stock price movement

As we have demonstrated in the literature review, financial news is tested to have certain effects on stock price and can be used to predict the stock price movement ([1],[2],[3],[4], etc.). And SVM has been widely used to predict the movement by classifying moving up or down ([4], [5], [6], [7], [8], [9]). Based on the relationship between financial news and stock price movement, we now apply SVM with proposed S&S kernel to predict the stock price movement.

1. Data source and statistical features

The news text materials we use are all from the financial channel of ifeng.com, which is a well-known professional financial portal in China. Considering the popularity and the development of the industry in these years, we choose medical and health industry as our object. The news is obtained by web crawler. We collect the financial news of stocks from September of 2012 to March of 2017. In order to demonstrate our model in detail, we take a random sample of stock SZ002424 to exhibit the results. More samples’ results will be found in Section V and the appendix.

For SZ002424, there are 581 news spreading over 315 days. This means the news distribute quite sparsely. Fig 2 shows that there are nearly of the days has only one piece of news related to stock SZ002424, while there is also one day having 15 pieces of news. For days with only one piece of news, as our research shows, the matrix Y which is corresponding to the similarity of news graph may have little effect for our prediction.



**Fig. 2.** The distribution of the news number for SZ002424

We aim to predict the trend of stock price based on financial news, so the return of a stock within a trading day is processed into two signs:

(10)

1. Text preprocessing

As for text preprocessing, we choose bag-of-words model to extract the features out of financial news. The bag-of-words model is commonly used to process the raw text: the occurrence of each word is used as a feature for training a classifier. In this model, a text (such as a sentence or a document) is represented as the bag of its words, disregarding grammar and even word order but keeping multiplicity. It shows great efficiency and lead to relatively reasonable prediction accuracy. Though there are other text processing methods which may outperform bag-of-words, the point of our paper is the S&S kernel. Thus, we apply bag-of-words model here to preprocess the original financial news and use the outputs of the model as features of each news.

1. Graph construction

Based on our primary thought, financial news within a day can be related to each other. The information behind the relationship has long been ignored. But the structural information may contribute a lot to our stock price prediction.

Before we construct the graph, we need to obtain the key words first so that we can take the frequency of them to build the edges. We first list the key words for a specific industry. For example, in our experiment, the key words we listed for bio-medicine industry include *biology, medicine, health, nursing, intermediate* and so on. Besides, the abbreviation of the listed companies are also included in these key words. The reason is that the companies of the same sub-industry may share similar business or produce similar products, leading they are always mentioned together in the same news. Whereas a company usually run multiple businesses, so different news may have different emphasis and different companies may be mentioned in different news. In this paper, we focus on the medical industry and list 323 key words including both industrial terminology and stock abbreviations for the empirical test. Then the weight on each edge can be calculated and the graph can be built.

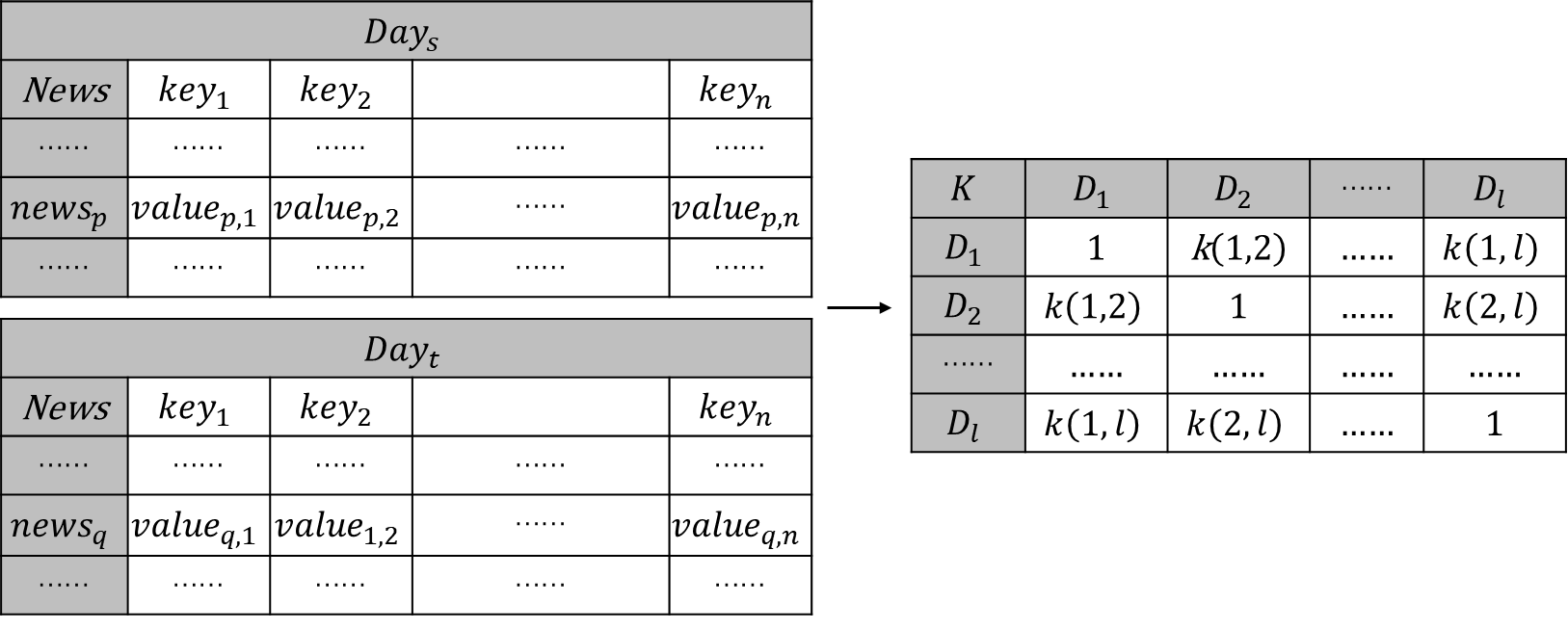
For text points graphs, we regard each news as a vertice. The key words graphs, which use each key word as a vertice, is then applied the to supplement the text points graphs, especially when there is only one news within a day (This would cause the construction of news graph becoming impossible). The graph construction process is shown in Fig 3. In this paper, we divide the news construction by dates, meaning each graph corresponds to the financial news within a day.



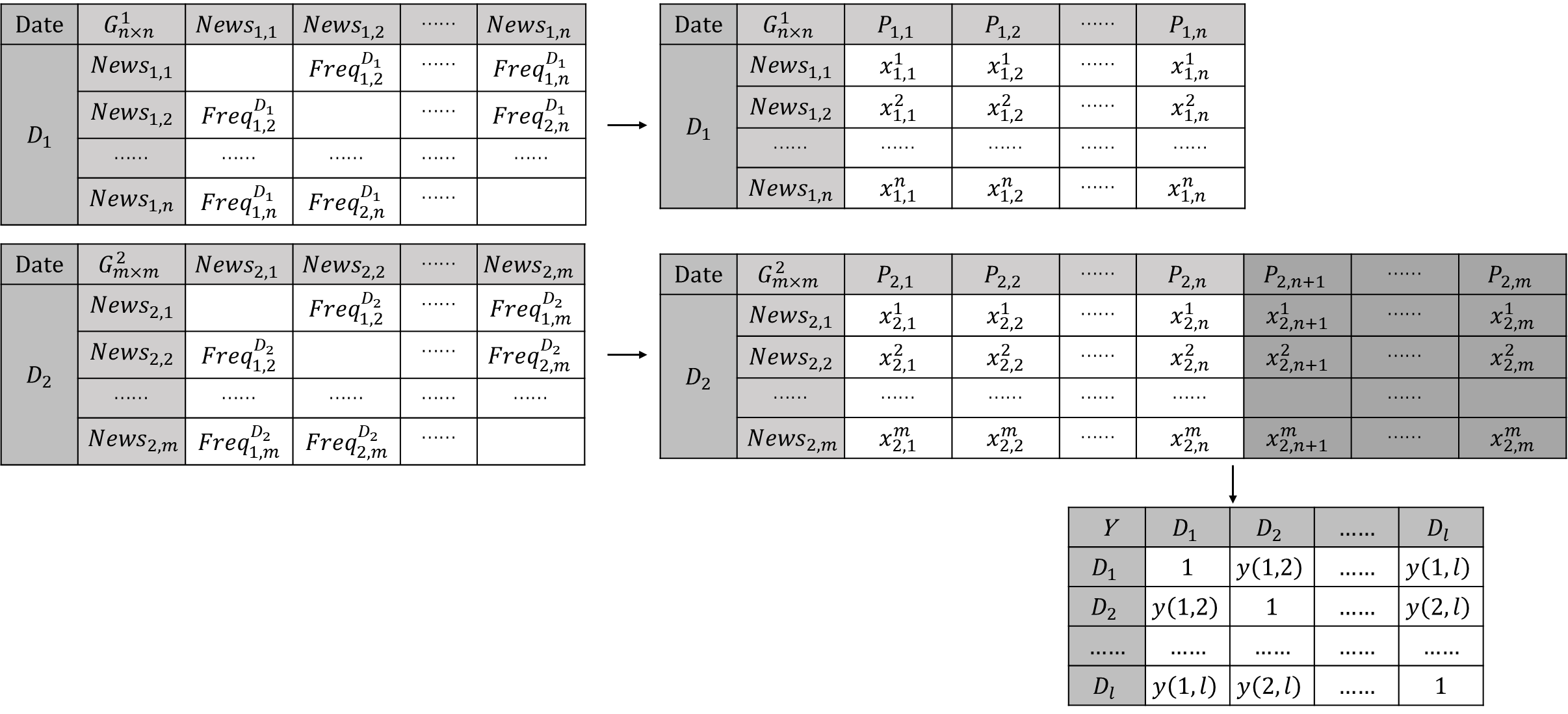
**Fig. 3.** Construction of news text points graph and key words graph

1. Similarity Measurement

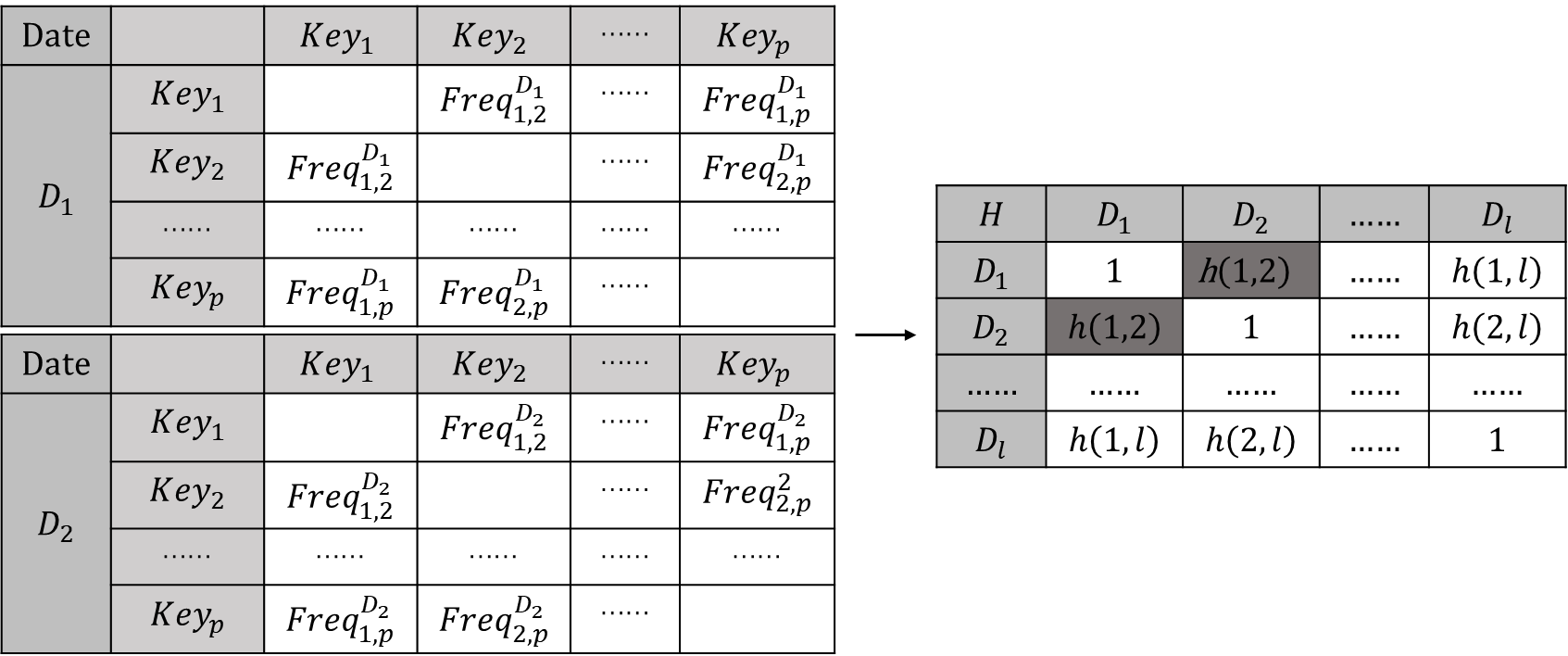
After preprocessing the financial news as well as constructing daily graphs, we need to measure the similarity between each two graphs. The measurement is divided into two parts: the similarity between the news features that we gain from bag-of-words model and the similarity between the daily graphs. The calculating process is shown in Fig 4, Fig 5 and Fig 6. We traverse all pairs of news in two days and use the average distance as the distance between the two days. So the distance matrix of news features can be acquired. Function k(s,t), y(s,t) and h(s,t) corresponds to function (5), (6) and (7) respectively.



**Fig. 4.** Calculating the content similarity between the news features



**Fig. 5.** Calculating the edge similarity between text points graphs



**Fig. 6.** Calculating the edge similarity between key words graphs

As matrix Y and matrix H are used to denote the structure of the news, we use a hyper parameter to combine them together.

(13)

The parameter is determined by Grid search. Also, hyper parameter is imported to combine matrix E and matrix K.

(14)

Also, parameter is determined by Grid search.

1. About the hyperparameters

As we construct our kernel using two hyperparameters: , which is used to determine our daily news graph, and , which is used to determine the weight between the news content and the news graph within a day, grid search method is applied here to test different parameters. The step gap we use here is 0.1.

### Results and discussions

In this part, we analyze the prediction result of stock SZ002424 using SVM with S&S kernel. The prediction result of other socks can be found in Appendix.

First of all, we use the heatmap to test the influence of our hyper parameters and . In Fig 7, the horizontal axis represents and the vertical axis represent . When we apply our method to predict the price movement with the news in the same day, which is shown in Fig. 7(a), we can see the color becomes darker and darker with becomes greater. This means the greater the weight of the daily news edge is, the higher the predicting accuracy is. However, there is no clear gradual change along with the horizontal axis, meaning hyper parameter may have less effect than . So, the weight assignment within the daily news graph construction may, somehow, provides similar information.

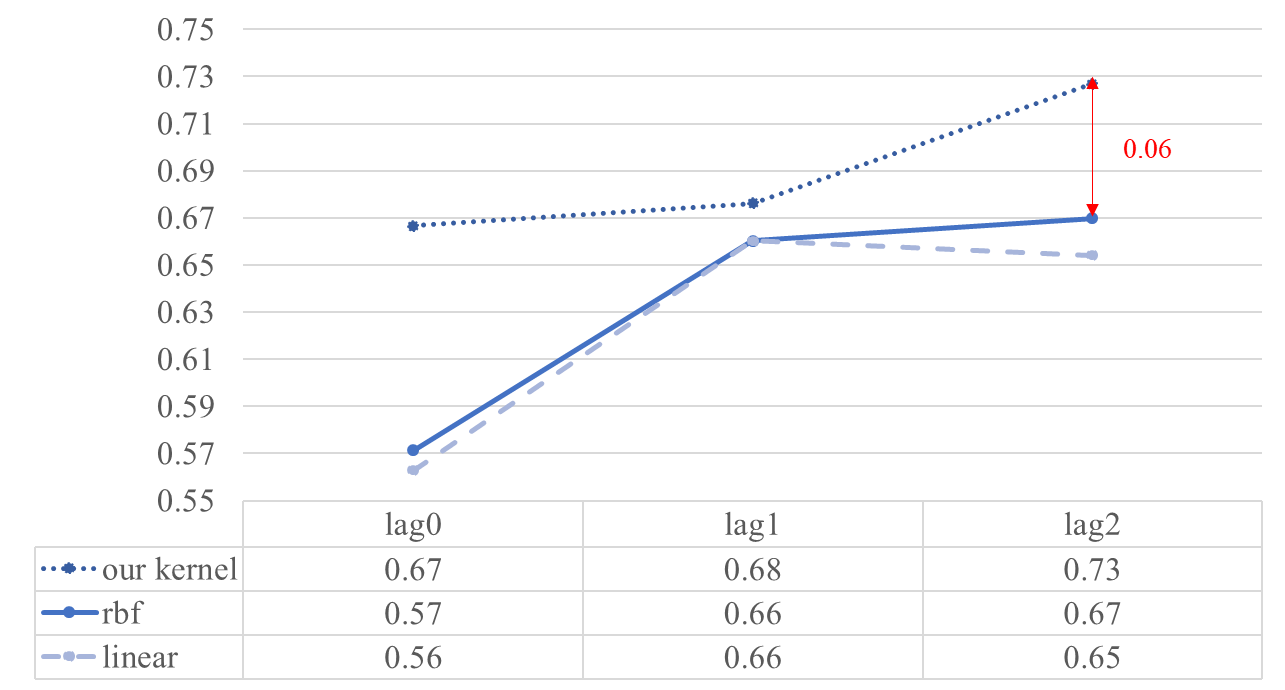
We also predict the price movement with 1 day to 3 days lag after the news, which is shown in Fig. 7(b)(c)(d). The same result can be found as the color becomes darker from top to bottom. Besides, the highest predicting accuracy is quite high and a 73% accuracy is reached when we use our kernel to predict the stock price movement with 2 days lag. All these results strongly support our point of view that the relationship between the structure of daily news from day to day can provide us more information to predict the stock price movement which can further help investors make right decisions.

|  |  |
| --- | --- |
|  |  |
| (a) Predicting the stock price of | (b) Predicting the stock price of |
|  |  |
| (c) Predicting the stock price of | (d) Predicting the stock price of |

**Fig. 7.** The heatmap of predicting accuracy using different hyper parameters

Note: The horizontal axis represents and the vertical axis represents . The color bar on the right is accuracy of prediction.

But can our method provide higher accuracy compared with traditional SVM? Or the news contents have provided enough information to help predict the stock price movement? Here, we apply SVM with linear kernel, sigmoid kernel and radial basis function kernel (rbf kernel) to predict the stock price movement compared with our kernel. Considering the number of our sample, we apply a five-fold cross validation to gain the average predicting accuracy. The result is shown in Fig 8. The accuracy of SMV with sigmoid kernel is the same as rbf kernel. We only show the result of rbf kernel here for simplicity. From the accuracy of price trend prediction, our kernel outperforms SVM on both rbf kernel and linear kernel. Fig 8 is the predicting accuracy of the three different kernels of SVM with 0, 1, 2 days lag. We can see the kernel we propose always has higher accuracy than the other two kernels. The price movement predicting accuracy using news contents only is 57.14% with no day lag, while our kernel’s accuracy is 66.66%, which is 9.52% higher than rbf kernel. Besides, our kernel can reach nearly 73% accuracy when predicting the price movement with 2 days lag.



**Fig. 8.** The predicting accuracy of three kernels with different lags

Based on this phenomenon, we propose a hypothesis that the financial news may have lag effect on the stock price. In fact, sometimes a breaking favorable news can drive the stock price up for consecutive days. In order to test this effect, we use to separately to train our model. The period we choose here is 5 days as the trading days within a week is 5 days.

As there are 5 trading days within a week, we test our kernel with 1 day to 5 days lag and the result is shown in Fig 9. There is a clear inverse U shape where the highest predicting accuracy is reached on 2 days lag while the accuracy of 5 days lag is the lowest. The experimental result can verify our hypothesis which means the financial market needs some time to digest the information from the news and this effect mainly lasts for two to three days after the news reported.

**Fig. 9.** The predicting accuracy along with time lags.

As the accuracy reaches the highest level when predicting the stock price 2 days after, we analyze the effect of our two hyper parameters again. We use the average predicting accuracy of all values of when is given to represent the accuracy of the given . The same is done to each value of . The result is shown in Fig 10. For the average accuracy of different , we can see a clear upward trend, meaning the higher weight given to the graph we constructed, the higher accuracy can be acquired. The result here is as same as what we gain from the heat map. While for parameter, there is a sudden high spot when and with becoming greater (except ), the average accuracy moves downwards. Therefore, matrix H should be assigned a higher weight meaning the relationship between the key words structures contains more effective information.

|  |  |
| --- | --- |
|  |  |
| (a) Average accuracy of different | (b) Average accuracy of different |

**Fig. 10.**  Average predicting accuracy with 2 days lag

### Stability test

In order to further strengthen our point, we add another 3 stocks to test the universality of our kernel. These three stocks are stochastically chosen from our news data set, which are SZ300049, SZ300142 and SZ000661. A descriptive statistic of the 3 stocks as well as SZ002424 is shown in Table 1.

**Table 1**

The descriptive statistics of the testing stocks.

|  |  |  |  |
| --- | --- | --- | --- |
| Stocks | Number of news | Days of news releasing | Max number of news per day |
| SZ002424 | 581 | 315 | 15 |
| SZ300049 | 780 | 390 | 15 |
| SZ300142 | 870 | 420 | 18 |
| SZ000661 | 855 | 405 | 16 |

We first use the daily news contents to predict the stock price on the same day, meaning we use ’s news to predict the trend of ’s stock price. Linear kernel, rbf kernel, sigmoid kernel and poly kernel are used as benchmark here. As we can see from Table 2, for these four stocks, our kernel outperforms the other three kernels. Our kernel can reach the highest accuracy for 300142.SZ, even using the other kernels have reached a 64.94% accuracy, which is quite high and also means the news really influence this stock’s price trend. A clear result from this table is that the kernel we propose can always outperform the other kernels by 5%.

**Table 2**

The comparison of predicting accuracy different kernels.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 002424.SZ | 300049.SZ | 300142.SZ | 000661.SZ |
| SVM-linear kernel | 56.30% | 53.85% | 51.84% | 49.00% |
| SVM-rbf kernel | 57.14% | 55.13% | 64.94% | 52.51% |
| SVM-sigmoid kernel | 57.14% | 55.13% | 64.94% | 52.51% |
| SVM-poly kernel | 57.14% | 55.13% | 64.94% | 52.51% |
| SVM- S&S kernel | **64.76%** | **58.73%** | **69.30%** | **59.01%** |

We also need to test whether our kernel performs best when predicting the stock price with 3 days’ lag as the result we gain from 002424.SZ. We calculate the predicting accuracy of the 3 stocks price trend with 1 day, 2 days and 3 days lag, which means we use ’s news to predict the trend of ’s stock price respectively. The results are summarized in Table 3. Different from 002424.SZ, the accuracy of the three stocks all reaches highest with 2 days lag. Therefore, financial market needs some time to digest the information from the news and this effect is always reflected in the stock price with 2 to 3 days lag, which is consistent with the actual situation.

**Table 3**

Predicting accuracy using five-fold cross validation.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 002424.SZ | 300049.SZ | 300142.SZ | 000661.SZ |
| Lag 0 day | 64.76% | 58.73% | 69.30% | 59.01% |
| Lag 1 day | 66.66% | 61.28% | 73.79% | 65.67% |
| Lag 2 day | 67.62% | 73.34% | 80.44% | 71.61% |
| Lag 3 day | 72.70% | 73.33% | 79.51% | 71.11% |

The time effect leads us thinking about that history information may contain some useful information for future predicting. So we use the earliest 80% of our news to train the model and use the rest 20% to test. The result is shown in Table 4. Compared with Table 3, we can see that the predicting accuracy improves except 000661.SZ with 1-day lag and 3-days lag. For SZ300142, the accuracy is quite astonishing and we think a deterministic promotion with some contingency factors can lead to the result.

**Table 4**

Predicting accuracy using the earliest 80% to train and the rest to test.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 002424.SZ | 300049.SZ | 300142.SZ | 000661.SZ |
| Lag 0 day | 66.67% | 71.79% | 83.33% | 61.73% |
| Lag 1 day | 71.43% | 62.82% | 83.33% | 65.43% |
| Lag 2 day | 71.43% | 74.36% | 90.48% | 74.07% |
| Lag 3 day | 79.37% | 74.36% | 89.29% | 67.90% |

### Conclusion

In this paper, we propose a new kernel, S&S kernel, to deal with the problem of stock price trends prediction through financial news. This problem has drawn a lot of attention these years along with the popularity of machine learning. But almost all the scholars focus on digging the information contained in the financial news contents, and the internal structure of daily news, also containing some information, is not considerable. The semantic and structural relevance between two days’ news can help with the prediction of the stock price trend. Our empirical test shows that S&S kernel outperform the other four commonly used kernels by at least 5% on predicting accuracy. When we prolong the lag days of prediction, a clear inverse U shape can be found. It is helpful to show that financial news has an effect on stock price and this effect usually last for 2 to 3 days.

From the construction of our kernel, we can see that a higher weight is given to the graph we constructed instead of news content. This is a very novel finding as researches have paid their attention to the news content only and tried to reach a higher accuracy through adjusting their model. The daily news graph we imported plays a more important role when predicting the accuracy. What else should be noticed here is that we only apply bag-of-words method and SVM here, other models can be applied to predict the stock price trend with importing the structural information in the future and hopefully a better result can be acquired.

Historical information is concerned in the last part of our paper. As the traditional econometrics believes the stock price is affected by the previous price and uses the time series to predict the stock price in the future, we believe the information in the past, both contents and structures, can help with price predicting. The experimental result, again, helps to verify our view and gains a remarkable result. Still, contingency factors should be considered.

Our main idea is to provide a new thought for the stock price predicting through financial news problem. There are more works can be done to refine this method. For example, stock dictionary can be used to preprocessing the text and more macro news can be added to supply the news data set. Besides, we only focus on the stocks from one industry. Stocks from other industries can be tested and even the stock index can be taken as an object. Finally, the idea of the information structure similarity actually provides a wider space and we hope this idea can be applied to more fields in the future.

### Reference

1. Dyck A., Zingales L. The Media and Asset Prices, [C]. Harvard Business School Ehrlich D, Guttman I, Schonback P, Mills J. 2003.
2. Robert Shiller, Irrational Exuberance, [M]. Princeton, 2000
3. EF Fama, Risk, Return, and Equilibrium, [J]. Journal of Political Economy, 1971, 79 (1) :30-55
4. Werner Antweiler and Murray Z. Frank, Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards, [J]. The Journal of Finance, 2004, 59 (3) :1259–1294.
5. Paul C. Tetlock, Giving Content to Investor Sentiment: The Role of Media in the Stock Market, [J]. The Journal of Finance, 2007, 62 (3) :1139-1168
6. Robert P. Schumaker and Hsinchun Chen, Textual Analysis of Stock Market Prediction Using Financial News Articles [C]. 12th Americas Conference on Information Systems (AMCIS), 2006.
7. Xiangyu Tang, Chunyu Yang, Jie Zhou, Stock Price Forecasting by Combining News Mining and Time Series Analysis [C]. 2009 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology – Workshops, ISSN 978-0-7695-3801-3/09, pp. 279-282.
8. Nan Li and Desheng Dash Wu, Using text mining and sentiment analysis for online forums hotspot detection and forecast [J]. Decision Support Systems 48, 2010, pp. 354–368.
9. Huang Run-peng, Zuo Wen-ming and Bi Ling-yan, Predicting the Stock Market Based on Microblog Mood, [J]. Journal of Industrial Engineering and Engineering Management, 2015, (01):47-52+215.
10. Xuejing Meng, Yafei Yang, Xinquan Zhao, The research on financial news and stock market investment strategy, [J]. Review of Investment Studies, 2016, (08):29-37.
11. Yong Shi, Jing Tang, Kun Guo, The impact of investors focus and sentiment of social media on China stock market [J]. Journal of Central University of Finance & Economics, 2017, (07):45-53.
12. Hua Chen, Xun Liang, Classification and Industry Analysis of Stock News in Cyberworld, [J]. Computer Development & Applications, 2006, (11):2-3+6.
13. Yang Yu, Wenjing Duan, Qing Cao, The impact of social and conventional media on firm equity value: A sentiment analysis approach, [J]. Decision Support Systems, 2013, 55 (4): 919-926
14. Felix Ming, Fai Wong, Zhenming Liu and Mung Chiang, Stock Market Prediction from WSJ: Text Mining via Sparse Matrix Factorization, [C]. IEEE International Conference on Data Mining, 2015 :430-439
15. Andrew Sun, Micheal S. Lachanski, Frank J. Fabozzi, Trade the Tweet: Social Media Text Mining and Sparse Matrix Factorization for Stock Market Prediction, [J]. [International Review of Financial Analysis](https://www.researchgate.net/journal/1057-5219_International_Review_of_Financial_Analysis), 2016.10
16. Lu X. Document Retrieval: a Structural Approach [J]. Information Processing and Management, 1990, 26 (2):209-218.
17. Russell S. J., Norvig P. Artificial Intelligence: a Modern Approach [M]. Prentice Hall, 1995.
18. Inderjeet M., Bloedorn E. Multi-Document Summarization by Graph Search and Matching [C]. In:Proceedings of the Fifteenth National Conference on Artificial Intelligence, 1997, 622-628.
19. Bhoopesh, Pushpak. Text Clustering Using Semantics [C]. The 11th International World Wide Web Conference, 2002.
20. Schenker A. Graph-Techniques for Web Content Mining [D]. Florida: University of South Florida, 2003.
21. J Wei, RK Srihari, Graph-based text representation and knowledge discovery, [J]. Acm Symposium on Applied Computing, 2007: 807-811
22. Zonghu Wang , Zhijing Liu, Graph-Based Chinese Text Categorization, [C]. International Conference on Electrical & Control Engineering, 2010: 1092-1095
23. Wu Jiangning, Liu Qiaofeng, Research on Graph Structure Based Method for Chinese Text Representation, [J]. Journel of the China Society for scientific and technical information, 2010, 29 (4): 618-624
24. Antto Bronselaer, Gabriella Pasi, An approach to graph-based analysis oftextual documents, [C]. 8th Conference of the European Society for Fuzzy Logic and Technology, 2013.
25. Siti Sakira Kamaruddin, Azuraliza Abu Bakar and Abdul Razak Hamdan et al., A text mining system for deviation detection in financial documents, [J]. Intelligent Data Analysis 2015,19: 9-44
26. Vapnik V N, Chervonenkis A Y. On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities[M]. Measures of Complexity. Springer International Publishing, 2015:264-280.

### Appendix

The heatmap of predicting accuracy using different hyper parameters for stock 300049.SZ, 300142.SZ and 000661.SZ.

|  |  |
| --- | --- |
| 300049.SZ | |
|  |  |
| (a) Predicting the stock price of | (b) Predicting the stock price of |
|  |  |
| (c) Predicting the stock price of | (d) Predicting the stock price of |

|  |  |
| --- | --- |
| 300142.SZ | |
|  |  |
| (a) Predicting the stock price of | (b) Predicting the stock price of |
|  |  |
| (c) Predicting the stock price of | (d) Predicting the stock price of |

|  |  |
| --- | --- |
| 000661.SZ | |
|  |  |
| (a) Predicting the stock price of | (b) Predicting the stock price of |
|  |  |
| (c) Predicting the stock price of | (d) Predicting the stock price of |