Long-Term Market Analysis using Text Mining*

K. Izumi

DHRC, AIST,
2-41-6 Aomi, Koto-ku, Tokyo 135-0064, Japan

T. Goto

The Bank of Tokyo-Mitsubishi UFJ, Ltd.,
2-7-3, Marunouchi, Chiyoda-ku, Tokyo 100-6417, Japan

T. Matsui

Faculty of Science and Technology, Tokyo University of Science,
2641 Yamazaki, Noda-shi, Chiba-ken 278-8510, Japan

We propose a new approach for analyzing the Japanese government bond (JGB) market using text-mining technology. First, we extracted the feature vectors of the monthly reports from the Bank of Japan (BOJ). Then, the trends in the JGB market were estimated by a regression analysis using the feature vectors. As a result, the resultant determination coefficients were over 75%, and the market trends were clearly explained using the information that was extracted from the textual data. Finally, we compared the predictive power of textual data with that of numerical data. As a result, we found that our text-mining method had better predictability than the numerical data analysis.

1. Introduction

Various types of financial information can now be found circulating on the Internet. Many studies have tried to forecast market trends by data mining numerical data, such as economic indexes and past price data[a]. However, a lot of the financial information on the net is in textual format, such as news stories, companies' reports, and experts' recommendations. Recently, some researchers have studied the market impact of on-line economic news, such

*This research was partially supported by a Grant-in-Aid for Scientific Research on Priority Areas, No. 19024071, made by the Ministry of Education, Science, Sports and Culture.

[a]For the details of this area, see the survey papers in these books.[1,2]

as Reuters, using a text-mining technique[b]. These studies mainly focused on the short-term (hourly or one-day) influence of textual data. It is implicitly assumed that a long-term market trend can be estimated by only numeric data not by textual data.

The purpose of our research is to indicate that text mining is effective for not only short-term but also long-term market trend analysis. We propose a new text-mining technology for the analysis of long-term (monthly) market trends. We estimated the monthly data on the Japanese government bond (JGB) prices from the monthly reports at the Bank of Japan (BOJ) using our approach. We then measured the prediction power of our text-mining method against that by a conventional method using numeric data.

2. Text Mining Method for Long-Term Market Analysis

There are two important points when using text mining for long-term market analysis: (1) appropriate textual data content and (2) a method to associate the textual data with time-series data.

First, we have to use textual data that contains appropriate contents and formats. We used BOJ's monthly reports on recent economic and financial developments[c]. This report analyzes the Japanese financial and economic situations from a macro point of view. Its contents are based on a discussion at the BOJ Monetary Policy Meeting, and it contains 15 to 20 A4-sized pages. The original report written in Japanese is released in the middle of the month, and it is translated into English a few days later. The following three reasons make this report suitable for a long-term market analysis. (a) It is regularly released, and therefore, we can track any temporal changes in the textual data. (b) Almost all institutional traders in Japanese markets pay attention to it and their trading behaviors are sometimes affected by it, because it reflects the basic decision making of the BOJ. (c) It has a regular format; so one document is easily comparable with another document at different points of time. All issues, such as economic growth, trade, and financial environment, appear in a fixed order. The tone of this document does not dramatically change, unless a drastic economic change occurs.

Second, in order to associate the textual data with the time-series market data, we propose the new text-mining method illustrated in Fig. 1. This

[b]See the survey paper[3] for more details about this technique.
[c]The English versions of these documents are published on `http://www.boj.or.jp/en/theme/seisaku/handan/gp/`, and the original Japanese versions of them are on `http://www.boj.or.jp/theme/seisaku/handan/gp/`.

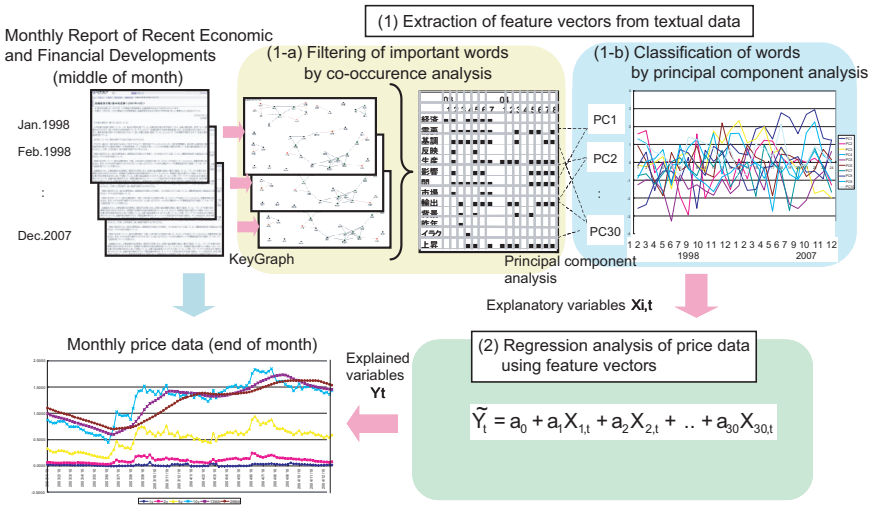

Fig. 1. Text mining framework

method consists of the following steps.

(1) Extraction of feature vectors from textual data

(a) Filtering of important words by co-occurrence analysis
(b) Classification of words by principal components analysis

(2) Regression analysis of price data using feature vectors

2.1. *Extraction of Feature Vectors from Textual Data*

At first, a morphological analysis is necessary before the extraction of the feature vectors, because we used the original Japanese documents[d]. Each sentence was divided into morphemes, the smallest meaningful units, using the Japanese language morphological analysis system ChaSen[e].

Next, important keywords were selected from the morphemes using a KeyGraph algorithm.[4,5] KeyGraph is a data-mining algorithm that finds the relationships among words or events and creates a network for the relations. In a KeyGraph analysis, the frequency of each word and any co-occurrence between words is computed from the data. In this paper, we used the Jaccard coefficient to measure the degree of co-occurrence of each

[d]All the characters are joined and not divided into words in a Japanese sentence.
[e]This software can be downloaded from `http://chasen.naist.jp/hiki/ChaSen/`.

pair of words $\{A,\ B\}$.

$$Jaccard(A,B) = \frac{p(A\,and\,B)}{p(A\,or\,B)}. \quad (1)$$

$p(A\,and\,B)$ is the ratio of the number of paragraphs where the words A and B occurred together to the number of all paragraphs in one document (a monthly report). $p(A\,or\,B)$ is the ratio of the number of paragraphs where at least either A or B occurred to the total number of paragraphs. The Jaccard coefficient increases when a pair of words frequently co-occurs and each word rarely occurs alone. Then, the KeyGraph algorithm analyzes the relative probability between the frequency and the co-occurrence of the words. Finally, it chooses keywords from the words and displays the relationships among the keywords as a network. Forty keywords were selected from every monthly report using this algorithm.

Using 120 documents from a 10-year period from January 1998 to December 2007, we created 120 networks using KeyGraph. There were 255 keywords that appeared in the networks at least once. We classified these keywords into 30 feature values according to their occurrence patterns. An occurrence matrix that was 255×120 was created. Each component of the matrix A_{ij} was registered as 1 when a word i occurred in the network at time j. Otherwise, A_{ij} was registered as a 0. Using this matrix, we extracted 30 principal components[f] that were composed from the keywords by the principal components analysis. The feature vector $\mathbf{x}_t$ from each month t consisted of the values of the 30 components $\{x_{1,t}, \cdots, x_{30,t}\}$ extracted from each monthly report.

2.2. *Regression analysis of price data using feature vectors*

Finally, the target price data y_t was estimated by a regression analysis using the time-series data of the feature vector $\mathbf{x}_t$:

$$\tilde{y}_t = a_0 + \sum_{i=1}^{30} a_i x_{i,t}, \quad (2)$$

where $\tilde{y}_t$ is the estimated value of the price at time t. Using the equation 2, not only the *in-sample* prices but also the *out-of-sample* prices can be estimated from the textual data. Since market reports are published in the middle of the month and the prices at the end of the month were estimated, the out-of-sample test was equivalent to a forecast for the next two weeks.

[f]The number of principal components was determined so that the accumulated contribution was over 60%.

3. Analysis of Monthly Price Data of JGB

We used our text-mining method to analyze the monthly price data from the Japanese government bond (JGB) market. The sample data were textual (BOJ's monthly reports) and price data (closing prices of 1-, 2-, 5-, and 10-year JGBs every month) over a 10-year period (120 months) from January 1998 to December 2007.

3.1. *Market Analysis using BOJ's Monthly Reports*

First, the KeyGraph algorithm and principal components analysis was used to extract 30 components from the 10 years of monthly reports mentioned above. There were two types of components. The first one represents *movement*. For example, component 1 consisted of keywords such as "yokobai (sideways)", "kennai (within the range)", and "yuruyaka (slow)", and component 5 consisted of "joushou (rise)", "atamauchi (hit the ceiling)", "nannka (weaken)", among others. The other is a component that represents *economic fundamentals*. For example, component 2 consisted of keywords related to the interest-rate level such as "risuku (risk)", "kokusai (government bond)", and "rimawari (yield)", and component 3 represents business activities from keywords such as "juyou (demand)", "kaizen (improvement)", and "seisan (production)".

Second, using the time-series data of the 30 principal component scores, we carried out a regression analysis of the JGB market data. Irrelevant variables were dropped by step-wise selection using AIC criteria. There were 23-25 variables selected for the 1-, 2-, 5-, or 10-year JGB prices. The coefficients of determination R^2 had sufficient results: 75.24% (JGB 1Y), 78.47% (JGB 2Y), 76.76% (JGB 5Y), and 74.65% (JGB 10Y). Then, we made out-of-sample forecasts using the textual data from January to June 2008. Figure 2a-d shows the estimated price paths for the in-sample and out-of-sample periods. In the out-of-sample period, the estimated paths were similar to the actual ones. They dropped first, then rose, and then dropped again.

3.2. *Comparison with Numerical Data Analysis*

In order to test the effectiveness of the textual data, we compared the predictive power of the textual data with that of the numerical data. We conducted a regression analysis on the JGB data using 19 standard nu-

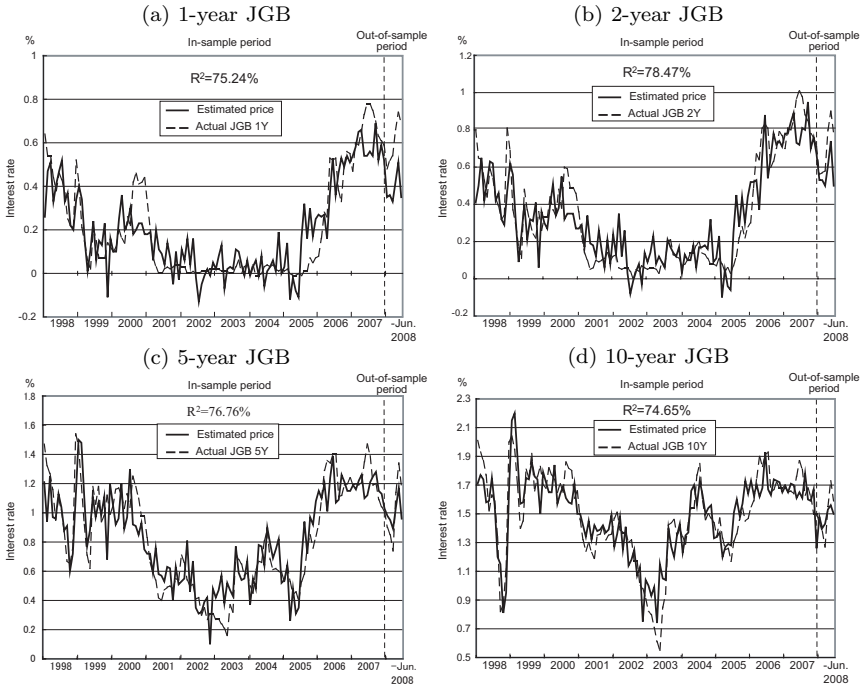

Fig. 2. Estimation of 1-, 2-, 5-, and 10-year JGBs.

merical data[g] from January 2000 to December 2007. Like the textual data analysis, 8-11 variables were selected using the step-wise selection with AIC criteria. As a result, the numerical data could clearly explain the movement of the sample data. The coefficients of determination R^2 were 84.61% (JGB 1Y), 87.96% (JGB 2Y), 82.22% (JGB 5Y), and 71.39% (JGB 10Y).

However, the numerical data had inferior prediction power to the textual data. We carried out out-of-sample forecasts using the numerical and textual data from January to April 2008. Figure 3 shows that the forecast errors of the numerical data were about three times larger than those of

[g]Monetary based (y-t-y); leading index of business conditions; loans and discounts (y-t-y); money supply M2+CD (y-t-y); machinery orders (m-t-m); current accounts (s.a.); machine tool orders (y-t-y); number of bankruptcies (y-t-y); indices of tertiary industry activity (m-t-m); corporate service price index (y-t-y); trade balance on customs clearance basis (s.a.); indices of industrial production (m-t-m); unemployment rate; ratio of job offers to applicants; consumer price index - Tokyo (y-t-y); consumer price index - Japan (y-t-y); new dwellings started (y-t-y); consumer confidence index; living expenditure (y-t-y), where y-t-y means "year-to-year basis", m-t-m means "month-to-month basis", and s.a. means "seasonally adjusted".

the textual data for the 2-, 5-, and 10-year JGBs.

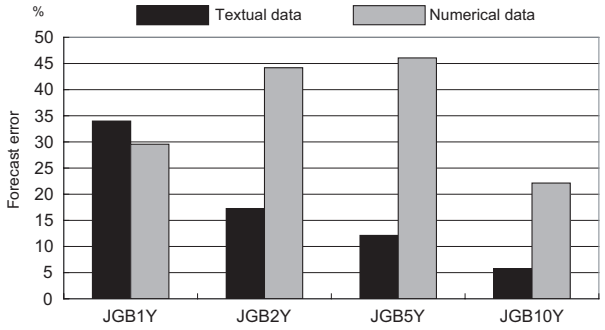

Fig. 3. Comparison of forecast errors where Y-axis is percentage of root mean square error to average value of actual prices in out-of-sample period.

4. Conclusion

In this paper, we proposed a new text-mining method for analyzing monthly market data. We tested it using BOJ's monthly reports and the JGB market data from the past 10 years. As a result, our method performed well for both the in-sample and out-of-sample periods. It also had better predictability than the numerical data analysis. These results indicate that textual data can effectively be used for not only short-term market analysis but also long-term market analysis.

References

1. S.-H. Chen and P. P. Wang (eds.), *Computational Intelligence in Economics and Finance* (Springer, 2003).
2. S.-H. Chen, P. P. Wang and T.-W. Kuo (eds.), *Computational Intelligence in Economics and Finance: Volume II* (Springer, 2007).
3. M.-A. Mittermayer and G. F. Knolmayer, *Text Mining Systems for Market Response to News: A Survey*, Tech. Rep. No 184, University of Bern (2006).
4. Y. Ohsawa, N. E. Benson and M. Yachida, Keygraph: Automatic indexing by co-occurrence graph based on building construction metaphor, in *ADL '98: Proceedings of the Advances in Digital Libraries Conference*, (IEEE Computer Society, 1998).
5. Y. Ohsawa, Keygraph: Visualized structure among event clusters, in *Chance Discovery*, eds. Y. Ohsawa and P. McBurney (Springer Verlag, 2003) pp. 262–275.