Identification of Important News for Exchange Rate Modeling

Debbie Zhang, Simeon J. Simoff and John Debenham

[1] Faculty of Information Technology,
University of Technology, Sydney, Australia
[2] {debbiez, simeon, debenham}@it.uts.edu.au

Abstract. Associating the pattern in text data with the pattern with time series data is a novel task. In this paper, an approach that utilizes the features of the time series data and domain knowledge is proposed and used to identify the patterns for exchange rate modeling. A set of rules to identify the patterns are firstly specified using domain knowledge. The text data are then associated with the exchange rate data and pre-classified according to the trend of the time series. The rules are further refined by the characteristics of the pre-classified data. Classification solely based on time series data requires precise and timely data, which are difficult to obtain from financial market reports. On the other hand, domain knowledge is often very expensive to be acquired and often has a modest inter-rater reliability. The proposed method combines both methods, leading to a "grey box" approach that can handle the data with some time delay and overcome these drawbacks.

1 Introduction

Until recently, most of exchange rate models are empirical models based on macro economic data. While these models performed reasonably well in predicting long term trends, they have little success in predicting short to middle term movements. Recent research has discovered that irrationality of the market participants, bubbles, and herd behavior are the main driven forces of the short term instability. Many market participants make their decisions before the economic data are announced, trying to beat the market. Their decisions are based on their observations and prediction of market trends. They adjust their investment strategy again after the data announcement according to the differences between their expectation and the actual data. Unarguably, news play an important role in creating such market dynamics. Recent research has confirmed that news has statistically significant effects on daily exchange rate movement. Ehrmann and Fratzscher [1] have evaluated the overall impact of macro news by analyzing the daily exchange rate responses using regression models with news variables. Three key results were found. Firstly, the news about fundamentals can explain relatively well the direction, but only a much smaller extent to the magnitude of exchange rate development. Secondly, news about US economy has a larger impact on exchange rates than news about the euro area. Thirdly,

higher degree of market uncertainty will lead to more significant effects of news releases on exchange rate movements. Prast and De Vor [2] have also studied the reaction of investors in foreign exchange markets to news information about the euro area and the United States on days of large changes in the euro-dollar exchange rates. Unlike the traditional models, daily news about economic variables as well as relevant political events in the United States and Europe were used in the regression model, which is:

$$E_t = \alpha + \sum_{i=1}^{8} \beta_i D_i + \varepsilon \tag{1}$$

where E_t is the percentage daily change in the euro-dollar exchange rate; D_{1-8} represent the following variables: 1 - real economy, euro area; 2 - inflation, euro area; 3 - change in official interest rate, ECB; 4 - statements/political events, euro area; 5 - real economy, United States; 6 - inflation, United States; 7 - change in official interest rate, United States; 8 - statements/political events, United States. It has been found that there is a strong correlation between exchange rate daily movement and the market participants' responses to the daily economy news and political events.

Motivated by their findings, Zhang, Simoff and Debenham [3] proposed a method to automate the exchange rate prediction process using text mining techniques. Nowadays, news retrieval using a computer program is effective and efficient as news data are widely available from the internet. However, to develop a program to understand the news articles and make the correct decision according to the news content is still an extreme difficult task. In this approach, the problem was broken down to multiple text classification problems and a regression modeling problem using numerical data. Relevant news are firstly identified and then being classified into the "good", "bad" or "no effect" news categories, which would have dramatically different impact on the market behaviors. "good" or "bad news" in this application are defined as the news that indicate the US dollar is destined to appreciate or depreciate versus Euro respectively. A regression prediction model is built based on these effects. Since the classification approaches are used, the training data are required to be pre-classified before the training process. The pre-classified training data contains the underlying knowledge of how data is being classified. Therefore, classification of training data is critical in any classification projects. However, many papers ignore this issue and provide no details of the methodology used in this step. This paper proposed a semi-automatic approach to assist the manual classification process.

2   News classification using a rule-based system

The training news data are required to be classified into one of the categories as shown in Figure 1 to incorporate their effects into the exchange rate model. This step can be done automatically or manually.

The automatic approach can be done by linking the news events with the changes in the exchange rate data if precise data are available. The exchange
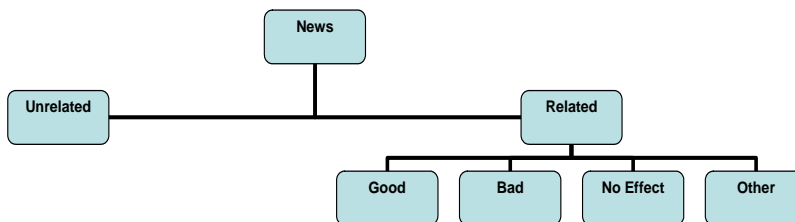
Fig. 1. News categories

rate data are recorded regularly over a time period, often referred to as "time series". At the same time, news about the events that may cause the data fluctuation are also recorded. The correlation between these two observations plays an important role in developing the prediction models. Methodologies like: Apriolike Discovery of Association Rules, Template-Based Mining for Sequences, and Classification of Temporal Data can be used to discover their correlations. Recently, Ting etc [4] proposed an approach to classify the company announcements into three categories according to changes of the share prices. Lavrenko etc [5] also presented a method for identifying news stories that influence the behavior of financial markets using language models. However, these approaches require precise data which are recorded with high sampling frequencies [6]. Testing the influence of the information requires precise data on both the information available prior to a public announcement and the actual released information. Recording the information flow accurately can be very difficult, particularly for the exchange rate data. Although the times at which scheduled macroeconomic releases are available, identification of the pre and post announcement market behavior is often very difficult. This is due to the decentralized nature of currency exchange market and leaking of inside information to some of the market participants. Delays in publication of the announcements on the internet news sources further contribute to the difficulty. These are the typical issues in the data retrieved from online news web sites including www.bloomberg.com, which are used in this study. At the time the news is published on the web site, some market operators already have the information and have reacted correspondingly.

More common approaches of data sorting relies on analysts' knowledge of the domain. News data are manually classified into several categories according to a set of rules that are specified by the domain experts. Rule based systems such as expert systems can be employed to help collecting and maintenance of the rules. However, the knowledge acquisition and rule generation process from domain experts is time consuming, expensive, and sometimes has only a modest inter-rater reliability. The rules derived from the domain experts can only a good starting point. It would be a sensible approach to use the data patterns to improve the rules. Therefore, a "grey box" method using domain knowledge in

conjunction with time series features of exchange rate data to refine the rules is proposed. The domain knowledge to form the initial rules is the white part of the model. However, these rules are not complete and can only be used to describe the data partially. Therefore, the rules need to be fine tuned using the data.

The intention of rule-based classification is to apply a validated, fully disclosed and understandable set of rules to perform the classification. The classification algorithm includes following steps:

– Specify a set of rules that reasonable to experts.
– Associate the text data with the exchange rate trends and classify the news according to the tread.
– Refine the rules if they are contradictive to the classified data in the previous step.
– Reclassify the data according to the new rules.

2.1   Obtaining Information for Developing Rules

Without an on-site domain expert, rules can be developed using published models that represent economists' understanding of the system. The rules used in this paper were developed using regression models provided by [1] and [2]. According to the regression result tables in the literature, the independent variables represent the factors that affect the exchange rate. The sign on the coefficient (positive or negative) indicates the direction of the effect. Although the size of the coefficient for each independent variable gives the size of the effect that variable is having on the exchange rate, they are ignored in the rules as they are highly sensitive to the data sets. It is difficult to quantify how the information is assessed by market participants since they have different responses for the same set of data when they have different expectations.

2.2   Association the News Data with the Exchange Rate Data

Each news item can also be classified according to its association with a particular trend in the exchange rate data. Considering most of the news web sites have time delay in publishing latest news and their refreshing frequency is about twice a day, the time window is set to 12 hours before and 12 hours after the exchange rate data. Figure 2 illustrates the alignment.

The trends are then aligned with news stories for classification. The results are used to examine the rules derived in Section 2.1. Adya etc [7] developed and automated heuristics to detect six features of the time series based on statistics: outliers, level shifts, change in basic trends, unstable recent trends, unusual last observation and functional form. However, to simplify the analysis process, only the upward trend and downward trend are identified. The exchange rate changes of each day are calculated. The days with large changes are selected, such as t1, t2 and t3 shown in Figure 2. The news articles that associated with these days are also selected and put into either "good" news or "bad" news categories
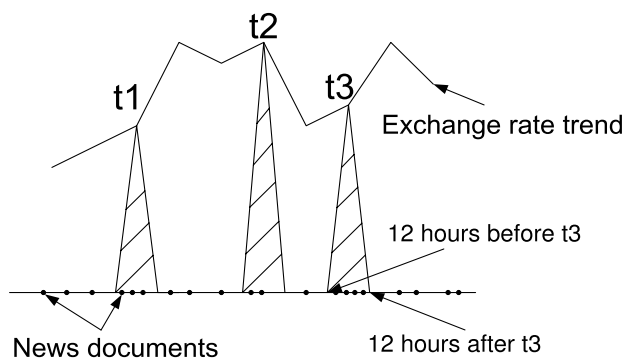
Fig. 2. News Alignment

according to their associated trend. At this stage, both categories contain many unrelated news since many news items of that day may have no impact on the exchange rate.

2.3 Refining the Rules and Classification Results

The rules derived in the section 2.1 can be validated by the classification results obtained in the section 2.2. The news in the "good" or "bad" categories are examined and summarized manually. Some simple heuristics are used to refine the rules. For example, if a rule describes an event has positive impact on the exchange rate, but in the data, both "good" and "bad" news contains multiple times of this event, it suggests that this event has no impact in the exchange rate at least within the period the data were collected. The rule should be removed. If multiple news items in the same category indicate that an event has impact in the exchange rate data, a new rule is added to link the event to the exchange rate movement. After the rules are being refined, the whole data set is manually classified according to the rules. In the final classification results, both "good" and "bad" news items may occur in the same day. This is reasonable as the changes of exchange rate are modeled as the the combined effect of multiple news items.

3 Experiment results

The experiment was conducted using the news articles and the US dollar versus Euro exchange rate data, both collected from www.bloomberg.com between 7/2/2005 to 4/7/2005.

Rules are derived from Ehrmann and Prast's models (see [1] and [2]), which are summarized in Table 1.

Their models also include "statements/political events", the area in which the rules are mainly expanded.
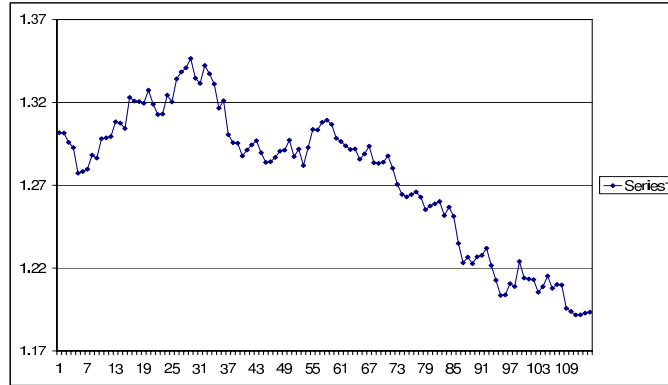
Fig. 3. News Alignment

Table 1. Summary of Ehrmann and Prast's models of positive and negative factors

	Positive factors	Negative factors
US announcement surprised	Inflation, Retail sales Nonfarm payrolles, Industrial production, N.A.P.M, Trade balance, Advance GDP, Average workweek Consumer confidence	Unemployment rate, Housing starts, CPI, PPI
German announcement surprises	CPI, Unemployment rate	Ifo Business Climate Industry production Manufacturing orders Retail sales Trade balance Business confidence PPI

To align trends with news data, a predefined threshold is used to define the trend. If Δg, which is defined in equation , is within the positive and negative range of the predefined threshold, it is considered that there is no changes in the price.

$$\Delta g_t = \frac{p_t - p_{t-1}}{p_{t-1}} \times 100\% \qquad (2)$$

where, Δg_t denotes the change of the exchange rate of day t, p_t and p_{t-1} denotes the exchange rate of day t and t-1.

In the experiment, the threshold of 0.1% is used. Within the experiment period, there were 40 days of upward trend and 49 days of downward trend. The news were compared on a daily base manually. Data have shown that Greenspan has profound inferences in the exchange rate. Although this is a well known fact, the original rules do not include this due to the models' limitation. Therefore, Greenspan's indication of positive or negative economic trend becomes one of the rule. The other rule worth to mention is the effect of U.S. 10 year treasury. The news of U.S. 10 year treasury appear very often. The first thought was that it could be an indicator of the rate trend. After carefully examining the data, no direct association was found. The final classification results using the refined rules are listed in Table 2.

Table 2. Classification results

Total number of News: 2589
      Number of "Unrelated" News: 1885
      Number of "Related" News: 704
            Number of "Good" News: 200
            Number of "Bad" News: 113
            Number of "No_Effect" News: 230
            Number of "Other" News: 161

As the experiment is still in an early stage, having data collected in a longer period would be better to evaluate the rules. More data are being collected using a program that can retrieve news data regardless the news web page format [8]. However, the preliminary results show that this approach provides better classification results than the classification solely based on the domain knowledge.

4 Conclusions

Exchange rate prediction using news article requires high quality training data. To achieve this, the news articles that have impact on exchange rate should be correctly associated with one of its trends, namely the upward and downward trends. Classification solely based on time series data requires precise and timely data, which are difficult to obtain for this application. On the other hand, domain

knowledge is often very expensive to be acquired and often has a modest inter-rater reliability. The proposed method combines both methods, leading to a "grey box" approach, which well suits this particular application.

References

1. Ehrmann, M., Fratzscher, M.: Exchange rates and fundamentals: new evidence from real-time data. Journal of International Money and Finance **24** (2005) 317–341
2. Prast, H.M., de Vor, M.P.H.: Investor reactions to news: a cognitive dissonance analysis of the euro-dollar exchange rate. European Journal of Political Economy **21** (2005) 115–141
3. Zhang, D., Simoff, S., Debenham, J.: Exchange rate modelling using news articles and economic data. In: 18th Australian Joint Conference on Artificial Intelligence. (2005)
4. Yu, T., Jan, T., Debenham, J., Simoff, S.J.: Incorporate domain knowledge into support vector machine to classify price impacts of unexpected news. In: Australasian Data Mining Conference. (2005)
5. Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D., Allan, J.: Language models for financial news recommendation. In: Ninth International Conference on Information and Knowledge Management (CIKM), Washington (2000) 389–396
6. Hautsch, N., Hess, D.: Bayesian learning in financial markets - testing for the relevance of information precision in price discovery. Journal of Financial and Quantitative Analysis (2005)
7. Adya, M., Collopy, F., Armstrong, J.S., Kennedy, M.: Automatic identification of time series features for rule-based forecasting. International Journal of Forecasting **17** (2001) 143–157
8. Zhang, D., Simoff, S.: Informing the curious negotiator: Automatic news extraction from the internet. In: Australasian Data Mining Conference, Cairns, Australia (2004)