Daily Prediction of Major Stock Indices from textual WWW Data

B. Wüthrich, D. Permunetilleke, S. Leung, V. Cho, J. Zhang, W. Lam*

The Hong Kong University of Science and Technology
Clear Water Bay, Hong Kong
*The Chinese University of Hong Kong, Shatin, Hong Kong
beat@cs.ust.hk

Abstract

We predict stock markets using information contained in articles published on the Web. Mostly textual articles appearing in the leading and the most influential financial newspapers are taken as input. From those articles the daily closing values of major stock market indices in Asia, Europe and America are predicted. Textual statements contain not only the effect (e.g., stocks down) but also the possible causes of the event (e.g., stocks down because of weakness in the dollar and consequently a weakening of the treasury bonds). Exploiting textual information therefore increases the quality of the input. The forecasts are available real-time via www.cs.ust.hk/~beat/Predict daily at 7:45 am Hong Kong time. Hence all predictions are available before the major Asian markets start trading. Several techniques, such as rule-based, k-NN algorithm and neural net, have been employed to produce the forecasts. Those techniques are compared with one another. A trading strategy based on the system's forecast is suggested.

Introduction

An increasing amount of crucial and commercially valuable information is becoming available on the World Wide Web. Today, with financial services companies bringing their products onto the Web various types of financial information have also come online. Among many others, the Wall Street Journal (www.wsj.com) and Financial Times (www.ft.com) maintain excellent electronic versions of their daily issues. Reuters (www.investools.com), Dow Jones (www.asianupdate.com), CNN (www.cnnfn.com) and Bloomberg (www.bloomberg.com) provide real-time news and quotations of stocks, bonds and currencies.

Our research investigates ways to make use of this rich online information in predicting financial markets. Techniques are presented enabling viewers to predict the daily movements of major stock market indices from up-to-date textual financial analysis and research information. Unlike numeric data, textual statements contain not only the event (e.g., the Dow Jones Indus. fell) but also why it happens (e.g., because of earnings worries). Therefore, exploiting textual information, especially in addition to numeric time series data, increases the quality of the input. Hence improved predictions are expected from this kind of input.

We predict stock markets by using information contained in articles published on the Web. In particular, the lead articles appearing in the mentioned newspapers are taken as input. From those articles, the daily closing values of major stock markets in Asia, Europe and America are predicted. The prediction is publicly available at 6:45 pm Eastern time, hence all predictions are available before the major Asian markets, Tokyo, Hong Kong and Singapore, start their trading day.

There is a wide variety of prediction techniques (see Fayyad et al (1996)), some also used by stock market analysts. Very popular among financial experts is technical analysis (Pring 1991). The main concern of technical analysis is to identify the trend of movements from charts. Technical analysis helps to visualize and anticipate the future trend of the stock market. Technical analysis only makes use of quantifiable information in terms of charts. But charts or numeric time series data only contain the event and not the cause why it happened. A multitude of promising forecasting methods have been developed to predict currency and stock market movements from numeric data. Among these methods are statistics (Iman and Conover 1989, Nazmi 1993), ARIMA (Wood et al. 1996), Box-Jenkins (Reynolds and Maxwell 1995) and stochastic models (Pictet 1996). These techniques as well as the successful Quest system (Agrawal et al. 1996) take as input huge amounts of numeric time series data to find a model extrapolating the financial markets into the future. These methods are mostly for short-term predictions whereas Purchasing Power Parity is a successful medium- to long-term forecasting technique.

The rest of the paper is organized as follows. Section 2 describes the techniques and architecture on which the system is based. Section 3 presents results using various forecasting engines. Section 4 concludes the paper.

Prediction Techniques

Our system predicts daily movements of five stock indices: the Dow Jones Industrial Average (Dow), the Nikkei 225 (Nky), the Financial Times 100 Index (Ftse), the Hang Seng Index (His), and the Singapore Straits Index (Sti). Every morning Web pages from www.wsj.com containing financial analysis and information about what happened on the world's stock, currency and bond markets are downloaded. This most recent news is stored in *Today's news*, see Figure 1. *Index value* contains the latest closing values, they are also downloaded by the agent who is active only on stock trading days.

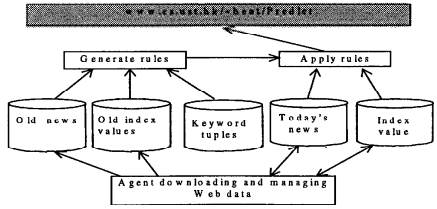

Figure 1: architecture and main component of the prediction system.

In Figure 1, *Old news* and *Old index values* contain the training data, the news and closing values of the last one hundred stock trading days. *Keyword records* contains over four hundred individual sequences of words (those sequences are the equivalent of phrases in Lent, Agrawal, and Srikant (1997)) such as "bond strong", "dollar falter", "property weak", "dow rebound", "technology rebound strongly", etc. These are sequences of words (either pairs, triples, quadruples or quintuples) provided once by a domain expert and judged to be influential factors potentially moving stock markets.

Given the downloaded data described, the prediction is done as follows:

1. The number of occurrences of the keyword records in the news of each day is counted.
2. The occurrences of the keywords are then transformed into weights (a real number between zero and one). This way, for each day, each keyword gets a weight.
3. From the weights and the closing values of the training data, probabilistic rules are generated Wüthrich (1995), Wüthrich (1997).
4. The generated rules are applied to today's news. This predicts whether a particular index such as the Dow will go up (appreciates at least 0.5%), moves down (declines at least 0.5%) or remains steady (changes less than 0.5% from its previous closing value).
5. From the prediction whether the Dow goes up, down or remains steady, and from the latest closing value also the expected actual closing value such as 8393 is predicted.
6. The generated predictions are then moved to the Web page www.cs.ust.hk/~beat/Predict where each day at 7:45 am local time in Hong Kong (6:45 pm Eastern time) the daily stock market forecast can be followed, see Figure 2.

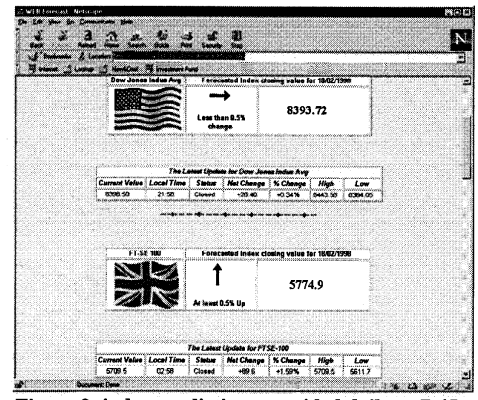

Figure 2: index predictions provided daily at 7:45 am Hong Kong time.

In what follows, we describe the individual steps of this process. The counting of keyword records is case insensitive, stemming algorithms are applied and the system considers not only exact matches. For example, if we have a keyword record "stock drop", and a web page contains a phrase "stocks have really dropped", the system does still count this as a match.

In a next step, a weight (i.e. a real number between zero and one) for each keyword record is computed. Figure 4 depicts this situation.

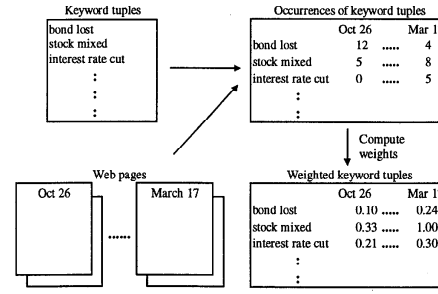

Figure 3: weights are generated from keyword record occurrences.

There is a long history in the text retrieval literature on using keyword weighting to classify and rank documents. Keen (1991) and Salton & Buckley (1988) give an overview on term weighting approaches in automatic text retrieval. In contrast to these approaches, however, we consider not a single keyword but pairs, triples, quadruples or quintuples of keywords. There are many approaches to conducting term weighting. One commonly used approach is to use three components: term frequency, document discrimination, and normalization.

Term frequency (TF) is the number of occurrences of a keyword record in a day's web pages. Keyword records that are mentioned frequently are assigned a larger weight. Term frequency factor alone is not a good indicator of the strength or importance of a keyword record. This is due to the fact that if a keyword record appears on each day's web pages, the keyword record is not a characteristic for a particular day.

Therefore, *category frequency* (CF) is introduced. For each possible category: stock index up, down, or steady, the CF of a keyword record is the number of training days containing the keyword record in that particular category at least once, see Table 1.

Keyword Record	Hsi up	Hsi down	Hsi steady
Bond lost	**13**	**23**	**18**
Stock mixed	**2**	**31**	**1**
Interest rate cut	**20**	**18**	**9**

Table 1: category frequency of keyword records with respect to an index.

For example, the keyword record "bond lost" appeared on twenty three days when the His index went down. Based on the CF, *category discrimination factor* (CDF) is computed:

$$CDF_i = \frac{\max(CF_{i,up}, CF_{i,down}, CF_{i,steady})}{t_i}$$

where CF_{ic} is the category frequency of a keyword record i in category c, and t_i is the number of days that the keyword record i appears. Taking CDF into account assigns keyword records that concentrate in one category alone a higher weight. They are calculated by multiplying the term frequency with category discrimination (TF×CDF). Table shows such weights.

	Mar 15	Mar 16	Mar 17
bond lost	**1.26**	**0.42**	**1.70**
stock mixed	**0.86**	**0.0**	**6.88**
interest rate cut	**0.85**	**1.70**	**2.13**
day's Maximum	**1.26**	**1.70**	**6.88**

Table 2: maximum values used to do normalization.

The third weighting term is the *normalization factor*. For each day, we find the maximum weight of a keyword record (*day's maximum*) and divide the weights for that day by this maximum. This assures that the final weight is a real number between zero and one. We tried various other weighting schemes (see Leung (1997) for more information on different weighting schemes) but the one described yields the highest forecasting accuracy.

Once the keyword counts are transformed into weights three rules sets are generated for each index to forecast, see Figure 5. Wüthrich (1997) describes how such probabilistic rules (unlike other rule based approaches these rules can also deal with weights) with conditional probabilities are generated.

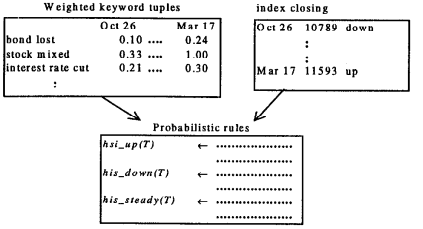

Figure 4: rules are generated from weighted keywords and closing values.

The rule bodies consist of the keyword records and their evaluation yields a probability saying how likely the particular index is going up, down or remains steady. The following is the sample rule set generated on 6th March computing how likely the Nky is going to move up today.

```
STOCK_UP(T)  <--STOCK_ROSE(T-1),
                NOT[NASDAQ_DROP(T-1)],
                NOT[PROPERTY_SURGE(T-2)],
                NOT[INTEREST_HIKE(T-2)]

STOCK_UP(T)  <--STERLING_ADD(T-1),
                NOT[TECHNOLOGY_SELL(T-2)]

STOCK_UP(T)  <--YEN_PLUNG(T-1)

STOCK_UP(T)  <--TOKYO_RETREAT(T-2),
                NOT PROPERTY_DOWN(T-2)
```

Once these rules are generated, they are applied to the most recently collected textual news and analysis results. So the likelihood of the Nky going up on 6th March depends for instance on the weight computed for `stock rose` on 5th March, and the of `property surge` valid for 4th March. From those probabilities, i.e. how likely the Nky is going up, down or remains steady, the final decision is taken. For example, the final decision is that the Nky moves up. Though maximum likelihood yields already good results for making this final decision, we found a slight improvement over maximum likelihood for this application (this is described in detail in Cho and Wüthrich (1998)).

After the direction of the stock market movement is determined (up, down or steady), the closing value of the stock index is computed. Suppose the generated rules expect an upward movement for the next day and the Tokyo's major stock index closed yesterday at 16865. The generated rules are now also applied to the training data and the average real percentage change x for those days for which the rules indicate an upward movement is

determined. The forecasted closing value for Nky is therefore 16865*(1+x) = 17148 (see figure 2).

Performance

The reported performance is achieved in the three months period, 6th Dec 1997 to 6th March 1998, this includes 60 stock trading days or test cases. As training period serve always the most recent 100 stock trading days. That is, to forecast on the 6th March the system is first trained on the period 5th Dec 97 to 5th March 98. A good yardstick is the accuracy, i.e. what percentage of the predictions are correct. For instance, if the system predicts up and the index moves indeed up then it is correct, otherwise, if the index is steady or down it is taken as wrong. The accuracy is shown in the second column of Table 3. The third column in Table 3 indicates how many times the system predicts up or down and it was actually steady; or, the system predicts steady and it was actually up or down. The last column indicates the percentage of totally wrong predictions. That is, the system expects the index to go up and it moves down, or vice versa. It is not surprising that the results for the Dow and Ftse are the best as most of the news we download is about these mature markets.

	accuracy	slightly wrong	wrong
Dow Jones Indus.	**45%**	**46.7%**	**8.3%**
FT-SE 100	**46.7%**	**36.7%**	**16.6%**
Nikkei 225	**41.7%**	**38.3%**	**20%**
Hang Seng	**45%**	**26.7%**	**28.3%**
Singapore Straits	**40%**	**38.3%**	**21.7%**
average	**43.6%**	**37.4%**	**19%**

Table 3: performance in the period 6rd Dec 97 to 6th March 98.

Table 4 shows the distribution of the actual outcome and the distribution of the forecast. The judgement of the predicted numeric value of an index is best done by comparing the chart of the actual value with the chart of the predicted value. This can be found on the indicated Web page.

	Distr actual outcome (%)			Distr forecast (%)		
	up	steady	down	up	steady	down
Dow	**48.3**	**26.7**	**25**	**45**	**33.3**	**21.7**
Ftse	**35**	**43.3**	**21.7**	**51.7**	**33.3**	**15**
Nky	**33.3**	**41.7**	**25**	**33.3**	**30**	**36.7**
Hsi	**35**	**21.7**	**43.3**	**26.7**	**28.3**	**45**
Sti	**40**	**20**	**40**	**31.7**	**26.6**	**41.7**
Avg	**38.3**	**30.7**	**31**	**37.7**	**30.3**	**32**

Table 4: distribution from 6th Dec 97 to 6th March 98.

Using k-NN learning algorithm (Michie, Spiegelhalter and Taylor 1994) the best accuracy was achieved by k-NN with k=9 and the Euclidean similarity measure: Ftse 42%, Nky 47%, Dow 40%, Hsi 53% and Sti 40%. The test period was, however, shorter. We also tried forward neural net work with 423 input nodes, for each keyword one input node, one layer of 211 hidden nodes and three output nodes using Back-propagation training algorithm. After optimizing some parameters, we achieved the following accuracy on 40 test days in the period 16th Dec to 17th Feb 1998: Hsi 43.9%, Ftse 35.4%, Dow 36.8%, Nky 34.1% and Sti 32.5%.

We also tried regression analysis on a twenty day moving average of the closing value of each index. This method does not yield forty percent accuracy for any of the indices. Another way to forecast is to just look at the outcome (up, steady or down) of a particular index over the n last days and to predict from there the next day. Feed foreward neural net was used to train on 60 days and n was varied between 4 to 10. The average accuracy achieved on forty test days is Dow 36%, His 40%, Fts 42%, Nky 27.5% and Sti 35%.

In fact, the performance of the system already enables to construct a simple and money making trading strategy (Note that one can buy the index in the futures markets; alternatively, one can buy the largest stocks to simulate the index as for example the eight largest stocks in Hong Kong account for over 90% of the Hsi; one can also leverage this trading strategy by buying in the options market). To keep the calculation comprehensible we make some assumptions.

Whenever the market goes up it appreciates on average by 0.5%; when it is steady there is on average 0% change and when the market goes down it slumps on average by 0.5%. Note that by definition of up (down) the market appreciates already at least 0.5% (slumps at least 0.5%). Hence, in reality, markets move on average by much more than 0.5% when it goes up or down. This pessimistic assumption is meant to compensate the next assumption which is to our advantage.

There are no trading costs involved when buying or selling. Trading costs actually depend on the amount traded, the specific futures exchange and the brokerage.

When the market opens we can on average buy or sell at yesterday's closing price.

We trade as follows.

- Suppose the system predicts up, we buy when the opening bell rings and sell when the market is about to close.
- Suppose the system predicts steady, we don't trade.
- Suppose the system predicts down, we short-sell (selling without having yet bought) when the opening bell rings and buy back when the market is about to close.

In summary, after each day we have closed out all positions, that is, we are neither long nor short on anything.

We can calculate the profit when for instance trading the Dow Jones Industrial Average during the considered 60 trading days in the period 6th Dec to 6th March. According to

table 5 we would have bought on 27 days and short sold on 13 days. Looking at table 4, we would have made 0.5% profit on 12 days by buying into the market; and we would have made 0.5% profit by short selling the index on 6 days. On 19 of the 40 days when we actively traded we would have been slightly wrong, hence neither a profit nor loss would have been booked. On the remaining 3 days when we traded we would have booked a loss of 0.5% as our system predicted wrongly. Overall, assuming to have bet each day the same amount of money, the profit is (12+6-3)* 0.5% or 7.5% over three months. This equals 30% capital appreciation over one year. In the same period, 6th Dec to 6th March, the Dow itself appreciated by only 5.1%. The similar results for the other indices are: trading strategy for Ftse yields 5.5% (11% actual movement of the index), Nky 5% (4.3%), His 3.5% (-4.6%), and Sti 4.5% (-8.8).

It is emphasized that the trading strategy also yields positive returns when the index is actually going down over the medium to longer term. The performance of the trading strategy depends on the daily volatility of the markets and the forecasting accuracy.

Conclusions

We presented techniques and developed facilities for exploiting textual financial news and analysis results. A prediction system has been built that uses data mining techniques and sophisticated keyword record counting and transformation to produce periodical forecasts of stock markets. Our techniques take as input textual information in the form of economic and political news, analysis results and citations of influential bankers and politicians.

Textual statements contain not only the event (stocks plummet) but also the possible causes of the event (weakness in the dollar and consequently a weakening of the treasury bonds). Exploiting textual information in addition to numeric time series data increases the quality of the input. Hence improved predictions are expected. The forecasts are real-time available via www.cs.ust.hk/~beat/Predict.

The system can potentially also serve as a decision support tool to help portfolio managers or traders of options on futures time the market. For instance, portfolio managers of mutual funds or institutional pension funds have to invest millions of dollars over a period as short as one week. They typically invest an equal amount of money each day (this is known as dollar cost averaging). However, if they have a prediction that the stocks are rising and they may themselves have the hunch that stocks are appreciating today, then they can try timing the market. On certain days they would delay their investment (the stocks are expected to weaken but the market starts steady or strong); whereas on other days they might invest more and earlier in the day (when the closing value is expected to be up and the market starts weak).

There are various directions in which this research can be extended. First, we do not yet take much numeric time series data into account. One might consider combining our techniques with specialized numeric time series forecasters. Second, as more and more information becomes available on the Web, other input sources might also be considered and prove to be of higher quality.

References

Agrawal, R., et al. 1996. The Quest Data Mining System. *Proc. KDD96.*

Cho, V. and Wüthrich, B. 1998. Towards Real-time Discovery from Distributed Information Sources. *Proc. PAKDD 98.*

Fayyad, U. M.; Piatetsky-Shapiro G.; Smyth P. and Uthurusamy R. 1996. *Advances in Knowledge Discovery and Data Mining*, AAAI Press / The MIT Press, pp625.

Hines, W.W. 1990. *Probability and Statistics in Engineering and Management Science*, 3rd edition, Wiley.

Iman, R. L.; and Conover, W. J. 1989. *Modern Business Statistics*, Wiley.

Keen, E. M. 1991. Query term weighting schemes for effective ranked output retrieval. *15th International Online Information Meeting Proceedings*, 135-142.

Lent, B.; Agrawal, R.; and Srikant, R. 1997. Discovering Trends in Text Databases, *Proc. 3rd Int Conf. On Knowledge Discovery and Data Mining*, California.

Leung, S. 1997. Automatic Stock Market Predictions From World Wide Web Data, Mphil thesis, HKUST.

Michie, D.; Spiegelhalter, D.J. and Taylor, C.C. 1994. *Machine Learning, Neural and Statistical Classification*, Englewood Cliffs, N.J., Prentice Hall.

Nazmi, N. 1993. Forecasting Cyclical Turning Points with an Index of Leading Indicators: A Probabilistic Approach, *Journal of Forecasting*, 12:(3&4), 216-226.

Pictet, O.V. et al., 1996. Genetic Algorithms with Collective Sharing for Robust Optimization in Financial Applications, TR Olsen & Assoc Ltd., Zurich.

Pring, M. J. 1991. *Technical Analysis Explained*, McGraw-Hill.

Reynolds, S.B. and Maxwell, A. 1995. Box-Jenkins Forecast Model Identification. *AI Expert*, 10(6):15-28.

Salton, G.; and Buckley, C. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24:(5)513-523.

Wood, D. et al., 1996. Classifying trend movements in the MSCI USA Capital Market Index – a Comparison of Regression, ARIMA and Neural Network. *Computers & Operations Research*, 23(6):611-622.

Wüthrich, B. 1995. Probabilistic Knowledge Bases. *IEEE Transactions of Knowledge and Data Engineering* 7(5):691-698.

Wüthrich, B. 1997. Probabilistic Knowledge Bases. *Int. Journal of Intelligent Systems in Accounting Finance and Management* 6:269-277.