Daily Stock Market Forecast from Textual Web Data

B. Wuthrich, V. Cho, S. Leung, D. Permunetilleke, K. Sankaran, J. Zhang, W. Lam*

The Hong Kong University of Science and Technology

Clear Water Bay, Hong Kong

*The Chinese University, Shatin, Hong Kong

ABSTRACT

Data mining can be described as "making better use of data". Every human being is increasingly faced with unmanageable amounts of data, hence, data mining or knowledge discovery apparently affects all of us. It is therefore recognized as one of the key research areas. Ideally, we would like to develop techniques for "making better use of any kind of data for any purpose". However, we argue that this goal is too demanding yet. It may sometimes be more promising to develop techniques applicable to specific data and with a specific goal in mind. In this paper, we describe such an application driven data mining system.

Our aim is to predict stock markets using information contained in articles published on the Web. Mostly textual articles appearing in the leading and 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. the stocks plummet) but also why it happened (e.g. because of 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 available real-time via www.cs.ust.hk/~beat/Predict daily at 7:45 am Hong Kong time. Hence all predictions are ready before Tokyo, Hong Kong and Singapore, the major Asian markets, start trading. The system's accuracy for this tremendously difficult but also extremely challenging application is highly promising.

1. INTRODUCTION

These days more and more crucial and commercially valuable information becomes available on the World Wide Web. Also financial services companies are making their products increasingly available on the Web. There are various types of financial information sources on the Web. 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), and Bloomberg (www.bloomberg.com) provide real-time news and quotations of stocks, bonds and currencies. In many cases, such information has to be purchased. The Wall Street Journal, Reuters and Dow Jones for instance charge a monthly or yearly fee for their services. Whereas newspapers are updated once or twice a day, the real-time news sources are frequently updated on the spot. All these information sources contain global and regional political and economic news, citations from influential bankers and politicians, as well as recommendations from financial analysts. This is the kind of information that moves bond, stock and currency markets in Asia, Europe and America. This rich variety of on-line information and news make it an attractive resource from which to mine knowledge.

Our research investigates how to make use of this rich online information to predict financial markets. Techniques are presented enabling 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 effect (the Dow Jones Index fell) but also why it happened (because of earnings worry for instance). 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 using information contained in articles published on the Web. In particular, articles appearing in the above named sources 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 before 7:45 am Hong Kong time, hence all predictions are available before Tokyo, Hong Kong and Singapore, the major Asian markets, start their trading day. Our techniques can complement proven numeric forecasting methods such as technical and regression analysis with technology which takes as input textual information in the form of economic and political news, analysis results and citations of influential bankers and politicians.

There are a variety of prediction techniques used by stock market analysts. Statistical techniques and regression analysis [4] provides quantitative forecasts. Pring [8] gives a comprehensive overview of recent popular technical analysis methods. The main concern of technical analysis is to identify the trend of movement from charts. Technical analysis helps to visualize and anticipate the future trend of the stock market. These techniques include peak-and-trough analysis which indicates a trend reversal when the series of rising peaks and troughs is interrupted, the moving average which reduces the fluctuations of stock prices into a smoothed trend so that the underlying trend will be more clearly visible, and so on. Technical analysis only makes use of quantifiable information. But there are also unmeasurable factors such as "general political news" which largely affect the world's stock markets. Hence, a modeling approach such as ours that can integrate also unmeasurable factors is desirable.

A multitude of promising forecasting methods for currency and stock market prediction from numeric data has been developed. These methods include statistics [6], ARIMA, Box-Jenkins and stochastic models [7,9,12,13]. These techniques 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 [10] is a successful medium- to long-term forecasting technique.

A stock market forecasting system developed by Braun [1] uses cues listed by domain experts to predict weekly movements of the stock market. Some of the cues are interpretations of trend-charting techniques. Others are ratios or statistics. Gencay [2] uses the daily Dow Jones Industrial Average Index to examine the linear and non-linear predictability of stock market using buy-sell signals generated from the moving average rule with a band between the short and long averages. The main difference of these approaches to our approach is that we use knowledge from unstructured data (web pages) instead of interpretations of numbers. In our case, even a person with little knowledge of stock markets can use our system to perform predictions because the input to our system is publicly available. We are not aware of any regular, precise and quantitative predictions about short-term stock market movements. In this respect our publicly available real-time forecasts are unique.

The rest of the paper is organized as follows. Section 2 presents the techniques and architecture on which the system is based. Section 3 presents the results, the prediction accuracy of the system for most recent period, 16th Dec 1997 to 17th Feb 1998. Its performance is compared with random prediction and some other possible forecasting techniques. Section 4 summarizes the results, opens further research issues and concludes the paper.

2. 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 (Hsi), and the Singapore Straits Index (Sti). Every morning an agent is downloading fifteen Web pages from the indicated news sources containing financial analysis reports and information about what happened on world's stock, currency and bond markets. 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, see Figure 2.

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 tuples* contains

over four hundred individual sequences of words 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.

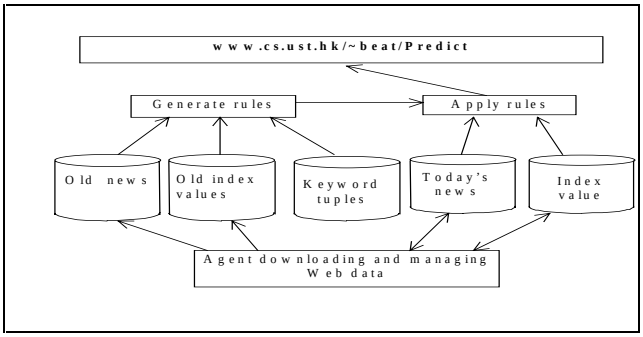

Figure 1: architecture and main component of the prediction system.

Figure 2:agent daily downloading Web pages and index values.

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

1. The number of occurrences of the keyword tuples 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 [14,15].
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 before at 7:45 am local time in Hong Kong (6:45 pm Eastern time) the daily stock market forecast can be followed, see Figure 3.

Figure 3: index predictions provided daily at 7:45 am Hong Kong time.

3. 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 6th 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 1. The third column in Table 1 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 whereas Singapore is the smallest in terms of market capitalization and liquidity, hence it is the least predictable.

	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 1:performance in the period 6th Dec 97 to 6th March 98.

Table 2 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 soon be found on the indicated Web page.

	Distribution of actual outcome (in percentage)			Distribution of the forecast (in percentage)		
	up	Steady	down	up	steady	down
Dow	**48.3**	**26.7**	**25**	**45**	**33.3**	**21.7**
Nky	**33.5**	**41.7**	**25**	**33.3**	**30**	**36.7**
Ftse	**35**	**43.3**	**21.7**	**51.7**	**33.3**	**15**
Hsi	**35**	**21.7**	**43.3**	**26.7**	**28.3**	**45**
Sti	**40**	**20**	**40**	**31.7**	**26.6**	**41.7**
average	**38.3**	**30.7**	**31**	**37.7**	**30.3**	**32**

Table 2:distribution in the period 6th Dec 1997 to 6th March 1998.

Using k-NN learning algorithm [5] 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 [3]. 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 forward 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%.*

Though comparing classifiers is full of pitfalls [11] we consider for the time being our probabilistic rule-based approach to be most reliable for this specific application. It has also to be noted that for shorter periods of a few weeks, our rule based prediction system achieves sometimes over *60%* accuracy.

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. For example, in the considered period the average change on both sides was well over *1.5%* for the Hsi and Sti. If we take *1.5%* instead of *0.5%* as average change then our trading profit would actually triple as the trading profit increases with the volatility of the market, not with the direction of the movement. On the other hand, 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. In Hong Kong, trading cost are typically *0.25%* to *0.5%* for retail investors (see South China Morning Post, 8 March 98) but much lower for institutional fund managers.

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 2 we would have bought on *27* days and short sold on *13* days. Looking at Table 1, 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 and the extrapolation to a year period are presented in Table 3. It has to be 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 is independent from the actual long or medium term movement of the index but it depends on the daily volatility of the markets (average change when it goes up or down) and the forecasting performance of the system.

	Actual performance		Trading strategy	
	3 month	12 month	3 month	12 month
Dow	5.1%	20.4%	7.5%	30%
Ftse	11.4%	45.6%	5.5%	22%
Nky	4.3%	17.2%	5%	20%
Hsi	-4.6%	-18.4%	3.5%	14%
Sti	-8.8%	-35.2%	4.5%	18%
Average	1.48%	5.9%	5.2%	20.8%

Table 3: performance of the index versus an active trading strategy based on the system's forecast, 6th Dec 97 to 6th March 98

From Table 3 can be seen that the trading strategy is also less risky in the sense that it is able to yield positive and steady returns in bull as well as bear markets. Our strategy would have also beaten almost all mutual fund managers over the considered period. The strategy beats the Hang Seng Index by over *30%,* but the Hang Seng Index alone already beats *27* of *28* actively managed mutual funds investing in Hong Kong stocks over the same three months period (see Money Magazine of the South China Morning Post, 8th March 98). Hang Seng Index also outperformed 18 of these *28* mutual funds in the year ending on 6th March 98. The situation is not much different for the other indices (e.g. see "Dart beats Wall Street Pros" The Wall Street Journal, Jan 15, 1998; available via the archive of www.wsj.com).

4. CONCLUSIONS

We presented techniques and developed facilities for exploiting especially textual financial news and analysis results. A prediction system has been built that uses data mining techniques and sophisticated keyword tuple counting and transformation to produce periodically forecasts about stock markets. Our techniques complement proven numeric forecasting methods such as technical and regression analysis with technology taking 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 effect (the stocks plummet) but also why it happened (because of 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 on the Web.

Various ways of transforming keyword tuple counts into weights have been investigated and several learning techniques, such as rule-based, nearest neighbor and neural net, have been employed to produce the forecasts. Those techniques are compared to each other. Rule based technique proves most reliable and yields encouraging results for this tremendously difficult but also extremely challenging application. The prediction accuracy is significantly above random guessing and is absolutely comparable to what can be expected from human expert predictions. In summary, the performance results reveal that our system can potentially serve as a decision support tool to help for instance portfolio managers time the market. 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 pretty reliable predictions, then 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 data into account. One might consider combining our techniques with conventional 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. Furthermore, when taking continuously updated information into account then also intra-day prediction of stock markets or even individual stocks becomes feasible.

5. REFERENCES

[1] H. Braun, "Predicting stock market behavior through rule induction: an application of the learning-from-example approach," *Decision Sciences*, vol. 18, no. 3, pp. 415-429, 1987.

[2] R. Gencay, "Non-linear prediction of security returns with moving average rules," *Journal of Forecasting*, vol. 15, no.3, pp. 165-174, 1996.

[3] J. Hertz, A. Krogh, and R.G. Palmer, *Introduction to the Theory of Neural Computation*, Addison Wesley, 1997.

[4] R. L. Iman, W. J. *Conover, Modern Business Statistics*, Wiley, 1989.

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

[6] N. Nazmi, "Forecasting Cyclical Turning Points with an Index of Leading Indicators: A Probabilistic Approach," *Journal of Forecasting*, Vol 12, No. 3&4, pp. 216-226, 1993.

[7] O.V. Pictet et al., *Genetic Algorithms with Collective Sharing for Robust Optimization in Financial Applications*, TR Olsen & Associates Ltd., Seefeldstr. 233, 8008 Zurich, 1996.

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

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

[10] F.L. Rivera-Batiz and L.A. Rivera-Batiz, *International Finance and Open Economy Macroeconomics*, 2nd edition, MacMillan, 1994.

[11] S.L. Salzberg, "On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach", *Data Mining and Knowledge Discovery*, Vol. 3: 317-328, 1997

[12] H. Wong, W.C. Ip, Y.H. Luan and Z.J. Xie, *Wavelet Detection of Jump Points And an Application to Exchange Rates*, TR The Hong Kong Polytechnic University, Dept. of Applied Mathematics, 1996.

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

[14] B. Wüthrich, "Probabilistic Knowledge Bases," *IEEE Transactions of Knowledge and Data Engineering*, Vol. 7, No. 5, pp. 691-698, 1996.

[15] B. Wüthrich, "Discovering Probabilistic Decision Rules," *Int. Journal of Intelligent Systems in Accounting Finance and Management*, Vol 6, pp 269-277, 1997.