FOUNDATION OF CONTROL AND MANAGEMENT SCIENCES
No Year

**Manuscripts**

Mateusz, KOBOS[*]
Jacek, MAŃDZIUK[**]

# ARTIFICIAL INTELLIGENCE METHODS IN STOCK INDEX PREDICTION WITH THE USE OF NEWSPAPER ARTICLES

Analysis of various kinds of data is an important part of information and knowledge management in a company. This paper contains an overview of literature concerning Artificial Intelligence automatic prediction systems applied to prediction of stock index values with the use of numerical time series and newspaper articles. Moreover, preliminary approach to implementation of such a system is proposed.

Key words: Artificial Intelligence, artificial neural networks, stock index prediction, automatic text analysis

## 1. INTRODUCTION

Knowledge and information management in a company requires combining and analyzing the data coming from different sources. Furthermore, a goal of purely automatic data analysis (if at all possible) is pursued. At the same time, said analysis is required to yield results reliable enough to be used to solve practical problems.

Automatic stock index prediction systems are examples of the systems which try to conform to above-mentioned requirements. Some of these systems use two main sources of information: numerical time series and textual data. The numerical time series consists of company or index stock quotations. Textual data, on the other hand, consists of articles published in specialist magazines. Such prediction systems that are related to Artificial Intelligence domain are discussed in this paper. Section 2 contains a description of different types of prediction systems and their functioning. In Section 3 results of initial experiments are presented. Section 4 concludes the paper.

[*] Warsaw University of Technology, Faculty of Mathematics and Information Science, e-mail: M.Kobos@mini.pw.edu.pl

[**] Warsaw University of Technology, Faculty of Mathematics and Information Science, e-mail: J.Mandziuk@mini.pw.edu.pl

## 2. FUNCTIONING OF A PREDICTION SYSTEM

Below we present a general scheme of a prediction system that uses both asset stock quotations and newspaper articles. The main component of the system is a prediction algorithm (approximator or classifier). The input data consists of published articles and stock quotations time series. Before the data is used in the system it has to be preprocessed by e.g. turning the quotations series into returns series (i.e. relative differences series), removing erroneous data or filling the missing data. Afterwards, textual data is converted into numerical representation which can be used directly by the prediction algorithm. The result of application of the prediction algorithm is a predicted asset value in the case of an approximator or a category of the predicted value (increase, decrease or no change in the asset's value) in the case of a classifier. The resulting prognosis can be used by investing algorithm assisting a human investor or by an algorithm that makes autonomous investments on the market.

### 2.1. Types of prediction algorithms

There are many kinds of prediction algorithms described in the literature. They can be divided into two main classes:

1. algorithms that use Artificial Intelligence methods,
2. algorithms based on mathematical models.

The most commonly used Artificial Intelligence methods are: artificial neural networks (e.g. multilayer perceptron, probabilistic network) [6, 4], Support Vector Machines [7, 8], genetic algorithms [11], decision trees, decision rules [9], naive Bayes models [3], k-Nearest Neighbours [8]. Among methods belonging to the second group, the most popular are time series prediction models stemming from ARIMA or GARCH [2] and linear discriminant analysis models [1].

### 2.2. Input data

As we mentioned in the introduction section, there are two kinds of input data used by the prediction algorithm:

1. textual data,
2. numerical time series.

The most commonly used textual data come from popular financial magazines ("The Wall Street Journal", "Financial Times"), press announcements, news agencies (e.g. "Dow Jones Newswire", "Reuters" or "Bloomberg") and press reports. On the other hand, the most popular time series are: companies stock quotations, indices quotations, foreign exchange rates, corporate bonds returns.

### 2.3. Numerical representations of text

Numerical representations of text, which can be used to represent press articles, can be divided into two groups:

1. automatically-generated representations,

2. representations based on *a priori* knowledge.

Among automatically-generated representations the most popular is the "bag-of-words" representation also known as "vector space" representation. The representation assumes that each document corresponds to a numerical vector. Each vector's coordinate corresponds to one word appearing in the set of all documents. The value of vector's coordinate indicates how important a given word in a document corresponding to the vector is. Despite being simple, "bag-of-words" representation is useful in practice. This representation will be presented more thoroughly in the next subsection.

Among representations based on *a priori* knowledge (which often require that article analysis is made by a human) the following approaches can be distinguished: article analysis by a human and application of predefined expert rules concerning article content [6], using predefined key words in "bag-of-words" representation [9, 8], using expert knowledge saved in "cognitive map" [4], using regular expressions to identify predefined document category [11].

### 2.3.1. "Bag-of-words" text representation

Below we introduce "bag-of-words" text representation in more detail. This representation was applied in our initial experiments. Let $a(D,w)$ be a value of vector's coordinate corresponding to word *w* in document *D*. Let *T* be the set of training documents used to train prediction algorithm, i.e. to adjust algorithm's parameters.

The most popular "bag-of-words" representations are:

- binary representation: $a(D,w)=\begin{cases}1 & w\in D\\ 0 & w\notin D\end{cases}$,
- TF representation: $a(D,w)=TF(D,w)$,
- TF-IDF representation: $a(D,w)=TF(D,w)IDF(T,w)$,

where:

- $TF(D,w)$ - frequency of word *w* appearance in document *D*,
- $IDF(T,w)=\log(\frac{|T|}{|T_w|})$ - logarithm of inverted occurrence frequency of documents containing word *w* among documents from set *T*,
- $T_w$ - set of documents containing word *w* (it is a subset of *T*).

The word which is associated with a high coordinate value in a vector corresponding to the document can be interpreted as significant. On the other hand, the word associated with a value which is close to zero can be interpreted as insignificant. Using this interpretation, TF representation considers as important the words which occur frequently in a document. The TF-IDF representation considers as important the words which occur frequently in a given document and at the same time occur rarely in other documents.

### 2.3.2. Dictionary in the "bag-of-words" text representation

Set of words which correspond to vector's coordinates in "bag-of-words" representation is called *dictionary*. Initially, the dictionary consist of all words occurring in all documents. During the preprocessing stage, some of the words are discarded. To remove the words which are considered redundant, the following methods of dictionary reduction are frequently used:

- methods related to linguistic properties of words:
  a) discarding "stop words" (prepositions, pronouns, conjunctions),
  b) synonyms identification,
  c) *stemming* – identification of linguistically related words and converting them into common stem
- methods related to features of the whole documents set:
  a) discarding the most frequently or most rarely occurring words,
  b) discarding words occurring in similar quantities in all predefined classes of documents (words with large information entropy),
  c) discarding words with the smallest CTF-IDF coefficient.

The CTF-IDF coefficient can be defined for each word $w$ as: $CTF - IDF(w) = \frac{\text{count}(w)}{|T_w|}$, where $\text{count}(w)$ is a number of word $w$ occurrences in all documents. It is worth noting, that words which are considered significant (i.e. correspond to large CTF-IDF coefficient) are the ones which occur frequently in small number of documents.

## 3. RESULTS OF EXPERIMENTS

This section contains a description of initial experiments conducted by the authors. The goal of the experiments was to verify the possibility of stock index prediction using selected index quotations and newspaper articles.

### 3.1. Input data

In the experiments, abstracts of newspaper articles and time series of returns calculated based on index quotations, both from the period between 2006.04.01 and 2007.04.01, were used.

The number of 46623 articles from "The Wall Street Journal" newspaper was used in total. The articles were made available to the authors by the publisher of "The Wall Street Journal" and the distributor of the magazine – "ProQuest" company. Basic statistical indicators describing distribution of the number of articles published every day are as follows: minimum = 1, maximum = 180, mean = 104, standard deviation = 48. There are 5 days with visibly smaller numbers (less than 6) of published articles (this anomaly is present in the original data source).

The returns time series describes everyday returns of S&P500 index quotations. The data comes from http://finance.yahoo.com service. Basic statistical

indicators describing distribution of returns are as follows: minimum = -0.0038, maximum = 0.022, mean = 0, standard deviation = 0.007.

### 3.2. Prediction algorithm and text representation

Multilayer perceptron neural network with RProp learning algorithm was used as a prediction algorithm. The network's input layer size varied. In each experiment it was adjusted to the number of past days used in the prediction. The network had one hidden layer with the number of neurons equal to the half of the input layer size. The output layer was composed of one neuron. The goal was to predict the next day's return of an index value. During the learning process the "early stopping" technique was used. Early stopping validation set consisted of 20% randomly chosen vectors from the training set.

As a text representation method, TF-IDF was chosen. "Stop words" were discarded from the dictionary and Porter algorithm [10] was used to perform stemming. To further reduce dictionary size, words with small CTF-IDF coefficient were discarded.

The time series was split into two parts. The first 70% of the series composed the training set whereas the remaining 30% of the series was used as a testing set.

### 3.3. Experiments description

Among other things, a relation between "memory length" of prediction algorithm and prediction error was examined during the experiments. *Memory length* means a number of previous days from which data (textual and numerical) is made available to the prediction algorithm. Additionally, dictionaries of different sizes were used. The results were compared to a simple heuristical prediction rule stating that tomorrow's index value will be the same as today's (i.e. the return will be equal to zero). It is worth noting that this heuristic is quite robust. It stems from Random Walk Model, which is related to Effective Market Hypothesis. This rule is often applied to verify the effectiveness of financial time series prediction. Root of Mean Square Error (*RMSE*) was used as a measure of prediction error.

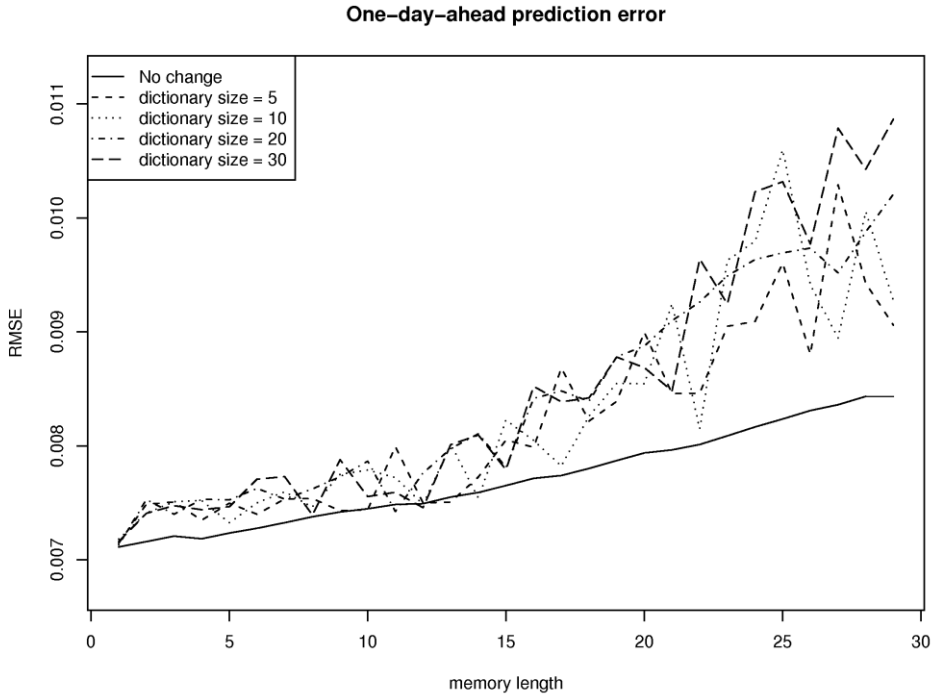


Figure 1. Dependency between memory length and one day index return prediction error. The solid line corresponds to the rule stating that tomorrow's index value will be the same as today's one.

### 3.4. Results analysis and possible system extensions

The following two conclusions were reached on the basis of conducted experiments. First, application of the prediction algorithm yielded worse results than "no change" heuristic. Furthermore, the larger memory length, the larger prediction error.

The main cause of obtaining unsatisfactory results is probably a problem with selection of significant words that compose a dictionary. Coefficient CTF-IDF applied to the given data causes the words which are insignificant but popular (occuring in all documents) to be selected to the dictionary. Because of it, after applying TF-IDF representation, the majority of the vectors corresponding to the documents have zero value. As a result, according to the representation interpretation, they are meaningless. The top-10 words that were selected for the dictionary are a good example of such insignificant words: "in", "on", "Mr", "sai", "he", "lear", "compani", "U.S", "hi", "Nuv". To solve this problem, a research concerning selection of more adequate coefficient that measures word importance throughout all documents is planned. Improvement can also be achieved by a better words preprocessing which would eliminate very popular words such as: "in", "on", "he" from the dictionary.

The low prediction quality can also stem from applying text representation which is too simple for a given problem. Hence, research concerning occurrences of whole expressions instead of single words is planned. Expressions occurrences can be measured approximately i.e. we could assume that given expression appears in a given document if, for example, two words forming the expression are placed close enough in the text.

The next reason of obtaining unsatisfactory results is probably not taking into consideration the overtone of a given article. It seems that using the information if the article has positive, negative or neutral overtone could considerably increase effectiveness of the system. Thus, automatic classification of a document based on its overtone is planned. To achieve this goal, an attempt to search for predefined expressions related to positive or negative opinions stated in the document is planned. Examples of such expressions include: "investors are anxious", "markets values decrease", "financial problems", "good financial results", "bull market" etc.

The improvement of prediction results could also be achieved by taking into consideration parts of the articles other than abstracts. We plan to test robustness of the prediction with the use of the title and the first paragraph of article's text. These parts of an article often contain summary of the whole publication, so they are especially important for article understanding task. To improve the results, greater weights (corresponding to greater relevancy) will be assigned to the words appearing in the title than words occurring in the abstract or in the first paragraph of the article.

Moreover, prediction of stock quotations of a company (or companies) from a given market segment is planned. The prediction will make use of articles related to this selected company (or companies). Prediction results could be better than the results of index quotation prediction because of potentially stronger relationship between an article referring to a given company and a price of company's share.

In case of company stock quotations prediction, a positive influence on prediction quality can come from including the subject matter of a given article in our model – e.g. article describing financial report of a company can have greater influence on a share price than a description of company's new product. To classify an article, we can use one of the following approaches: use classification predefined by the newspaper distributor, or search for key words related to given subject matter. Similar approach to analysis of scientific articles was proposed in [5].

Also a comparison of results yielded from the use of an automatically generated dictionary and results yielded from the use of a dictionary predefined by a human seems to be interesting and worth testing. The predefined dictionary would contain words or expressions recognized as important for assessing the meaning of the document. In the case of prediction of stock quotations of a company from a given industry, dictionary of words related to this industry could be used. For example, for computer industry significant expressions could be: "security flaw", "bug", "hacker attack" etc.

## 4. CONCLUSIONS

The paper contains a description of structure and functioning of automatic asset quotations prediction system. Popular "bag-of-words" texts representation types are presented. Furthermore, results of initial experiments are shown which, unfortunately are far from being satisfactory. Several possible reasons for low efficacy of the system are pointed out and suggestions of how to overcome these inefficiencies are proposed. The work is yet at preliminary stage and some progress is expected after implementation of modifications proposed in section 3.4.

## 5. ACKNOWLEDGEMENTS

The authors would like to thank ProQuest company and the publishers of "The Wall Street Journal" and "Financial Times" for granting access to textual data used in the initial experiments.

### REFERENCES

[1] Aasheim C., Kohler G.J., Scanning World Wide Web documents with the vector space model, Decision Support Systems, vol. 42, issue 2, pp. 690-699, 2006.

[2] Edmonds R.G., Kutan A.M., Is public information really irrelevant in explaining asset returns?, Economics Letters, vol. 76, issue 2 , pp. 223-229, 2002.

[3] Gidófalvi G., Elkan. C., Using News Articles to Predict Stock Price Movements, (Technical Report). Department of Computer Science and Engineering, University of California, San Diego (USA), 2003.

[4] Hong T., Han I., Knowledge-based data mining of news information on the Internet using cognitive maps and neural networks, Expert systems with applications, vol. 23, pp. 1-8, 2002.

[5] Klincewicz K., Miyazaki K., Software Sectoral Systems of Innovation in Asia. Empirical Analysis of Industry-Academia Relations. Proceeding of the IEEE International Conference on Management of Innovation and Technology, Singapore, 2006, vol. 1, pp. 494-498, IEEE, Singapore (China), 2006

[6] Kohara K., Ishikawa T., Fukuhara Y., Nakamura Y., Stock price prediction using prior knowledge and Neural Networks, Intelligent systems in accounting, finance and management, vol. 6, pp. 11-22, 1997.

[7] Mittermayer M.-A., Forecasting Intraday Stock Price Trends with Text Mining Techniques, Proceedings of the 37th Hawaii International Conference on System Sciences (HICSS'04), Hawaii, vol. 3, IEEE Computer Society, Washington, DC (USA), 2004.

[8] Mittermayer M.-A., Knolmayer G.F., NewsCATS: A News Categorization And Trading System, Proceedings of the Sixth International Conference on Data Mining (ICDM'06), Hong Kong (China), pp. 1002-1007, IEEE Computer Society , Washington, DC (USA), 2006.

[9] Peramunetilleke D., Wong R.K., Currency Exchange Rate Forecasting from News Headlines, Proceedings of the 13th Australasian database conference (ADC2002),

Melbourne, vol. 5, pp. 131-139, Australian Computer Society, Melbourne (Australia), 2002.

[10] Porter M.F., An algorithm for suffix stripping, Program, vol. 14(3), pp. 130−137, 1980.

[11] Thomas D., News and Trading Rules (Ph.D. thesis). Carnegie Mellon University, 2003.

[12] Wuthrich B., Cho V., Leung S., Permunetilleke D., Sankaran K., Zhang J., Lam W., Daily Stock Market Forecast from Textual Web Data, Systems, Man, and Cybernetics, 1998. 1998 IEEE International Conference on, San Diego, CA, (USA), vol. 3, pp. 2720-2725, 1998.