A Case Study: News Classification Based on Term Frequency

Petr Kroha
Faculty of Computer Science
University of Technology
09107 Chemnitz
Germany
kroha@informatik.tu-chemnitz.de

Ricardo Baeza-Yates
Center for Web Research, CS Dept.
Universidad de Chile
Blanco Encalada 2120,
Santiago 6511224, Chile
rbaeza@dcc.uchile.cl

Abstract

In this paper, we investigate how much similarity good news and bad news have in context of long-terms market trends and we discuss the relation between information retrieval and text mining. We have analyzed about 400 thousand news stories coming from the years 1999 to 2002 and we argue that classification methods of information retrieval are not strong enough to solve problems like this one because the meaning of news is given not only by the used words and their frequency but also by the structure of sentences and their context. We present results of our experiments and examples of news that support this statement.

1. Introduction

These days, more and more commercially valuable business news becomes available on the World Wide Web in electronic form. However, the volume of business news is very large and there is a question how much of this kind of information moves stock markets.

In this paper, we have analyzed the relation between news and long-term market trends. We have investigated to what degree the news correspond to the long-term trends and whether the knowledge gained from news can be used as an attempt to predict long-term trends of financial markets. The novel approach is in using this technology for long-term prediction. Papers already published only investigate short-time influence of messages suitable for day-trading. We discuss them and their shortcomings in section 2. The most crucial question is how to preprocess the news before extraction and before inputting the results into the prediction engine. Our experimental results show that the techniques of information retrieval do not work very well for this purpose.

The rest of the paper is organized as follows. Related work is recalled in section 2. Section 3 introduces concerned problems and section 4 presents the methods we have used. Section 5 describes our experiments. Finally, we conclude in section 6.

2. Related Work

In related papers, the approach to classification of market news is similar to the approach to document relevance. Experts construct a set of keywords which they think are important for moving markets. The occurrences of such a fixed set of several hundreds of keywords will be counted in every message. The counts are then transformed into weights. Finally, the weights are the input into a prediction engine (e.g. a neural net, a rule based system, or a classifier), which forecasts which class the analyzed message should be assigned to.

In papers by Nahm, Mooney [8] a small number of documents was manually annotated (we can say indexed) and the obtained index, i.e. a set of keywords, will be induced to a large body of text to construct a large structured database for data mining. The authors work with documents containing job posting templates. A similar procedure can be found in papers by Macskassy [4]. The key to his approach is the user's specification to label historical documents. These data then form a training corpus to which inductive algorithms will be applied to build a text classifier.

In Lavrenko [3] we can find a method similar to our own method. To each trend, there exists a set of news that are correlated with this trend. The goal is to learn a language model correlated with the trend and use it later for prediction. A language model determines the statistics of word usage patterns among the news in the training set. Once a language model has been learned for every trend, a stream of incoming news can be monitored and it can be estimated which of the known trend models is most likely to generate the story. One difference to our investigation is that Lavrenko uses his models of trends and corresponding news only for day trading. The next difference is that we argue

1529-4188/05 $20.00 © 2005 IEEE

that this method is not suitable for the identification of market trends. The weak point of this approach is that it is not clear how quickly the market responds to news releases. Lavrenko discusses this but the problem is that it is not possible to isolate market responses for each news story. News build a context in which investors decide what to buy or sell. Fresh news occur in the context of older news and may have a different impact.

In our paper, we argue that the described methods inherited from information retrieval cannot be successfully used for the classification of news because our goal is not to find news that contain a specific set of keywords. The goal is to understand the meaning of text messages for better classification.

3. The problem

Information retrieval has motivated most of the work on text processing. Its goal is to find documents, which are most relevant with respect to a query. The content of a document is basically specified by a list of keywords that seems to describe it. To compare the query with a set of documents usually a vector space model is used. There are two main problems: how to weight occurrences of keywords and how to measure the similarity between document vectors and query vectors.

Text mining has as its goal to search for patterns in natural language text, to extract corresponding information, and to link it together to form new facts or new hypotheses. The goal is not to search for relevant documents or for something that has explicitly been written. New, previously unknown information shall be discovered by methods of text mining [1]. The fundamental limitation of text mining is that we are not able to write programs that fully interpret text. The main problem is to assign semantics, or meaning, to parts of the text.

Even though we can observe how much ambiguous the news about markets and stocks are, we will formulate the following hypothesis.

Hypothesis: Statistically, during growing markets,
news about stocks and markets have contents
that are different from those during falling
markets.

If this hypothesis is true we can find templates for news occurring in good times and bad times, and use them for forecasting the movement of the current market.

When the market is going up then it should follow from the assumptions described above:

- The relative frequency of positive and negative keywords in news sets is typical, i.e. positive keywords should be in majority compared with negative keywords in growing market and vice versa in falling market. This assumption will be investigated by diagnostic methods.
- The probabilistic profile of news sets is typical. This assumption will be investigated by classification methods.

Keyword	Up1999	Down2000	valid?
steig	20.01	19.68	yes
positiv	11.68	11.90	no
Gewinn	27.96	24.39	yes
erhöht	9.49	10.66	no
wachsen	3.29	3.84	no
Keyword	Down2002	Up2003	valid?
steig	11.39	17.15	yes
positiv	9.30	15.05	yes
Gewinn	18.34	29.79	yes
erhöht	5.79	10.39	yes
wachsen	2.35	3.17	yes

Table 1. Positive and negative words

In the sequel we describe how we have tested this hypothesis and the results we have obtained. First, we investigate the frequency of substrings in section 4.2 and the probability of keywords in section 4.3. Second, we describe how we used a classifier for finding out how similar the sets of weekly news are in section 4.4, 4.6, and 4.7.

3.1. Diagnostic methods

The text processing scheme here is based on keyword counting. The keyword table of positive and negative words (example in Tab. 1) has been created by hand. Additionally, we have constructed a sorted file of words probabilities for all sets of news and we have extracted positive and negative keywords from the first 1.000 words in each set.

3.2. Classification methods

One common approach to the problem of document classification is to find typical distribution of word probabilities for each class during the training phase, which uses a set of labeled documents. These probabilities are calculated directly from word frequencies and stored in the database for later use. Once a sufficient number of training documents have been processed, we can start asking the classifier to classify new documents that it has not seen before. The classifier returns an ordered list of the most probable classes for a new document.

Class	Time interval	News
Up1999:	13.11.1999 - 13.03.2000	32299
Down2000:	14.03.2000 - 14.07.2000	30228
Down2002:	05.12.2002 - 05.03.2003	23875
Up2003:	06.03.2003 - 06.07.2003	35998

Table 2. Set of news and market trends

This concept is simple but it has the disadvantage that it analyzes any document only as a bag of words ignoring sentence structure. We have used this method to support experimentally our suspicion that methods based on keyword frequency are not suitable for text mining. As we have shown in examples, the sentence structure may be important.

4. Experiments

4.1. Experimental data

To test the hypothesis formulated in section 3, we have used historical data of the German market index DAX30. As experimental data we collected news from only one subscribed source. They have a volume of about 12.000 news stories per month. Further, we have used the commonly accepted assumptions:

- Markets move in trends.
- News influence trends.
- Only a small minority of investors follow the rule "Sell on good news".

We collected about 400.000 text messages containing financial and political news from October, 1999 to the end of September, 2003. They are about 8.000 in a month. The actual outcomes of the index DAX30 are collected for the same period. We manually approximated the trends and found two points when long-term trends changed. It was on March 13, 2000 when the trend changed from UP to DOWN, and on March 6, 2003 when the trend changed from DOWN to UP. We divided the news according to the time intervals into four sets. Each set contains 16 files (about 30.000 news) corresponding to 16 weeks Tab. 2.

4.2. Inverse document frequency of substrings

In the first experiment, we constructed a table of substrings of positive and negative keywords (five positive, five negative) and tested the inverse document frequency of these substrings (in percents) 4 months before and 4 months after the point of change in both cases, i.e. for the news collections Up1999, Down2000, Down2002, and Up2003. The usage of substrings is advantageous because the German language has rich possibilities (declination, conjugation) how to derive words from a stem, i.e "steig" catches "steigen" (in English: to rise), "steigt, "steigten", "steigte", "steigend", "steigende", "steigenden", "steigendes", "steigender", also "Steigerung", "Steigerungen", "ansteigen" etc. We have used it instead of some stemming algorithm. Since the used software was not able to count words but news we have computed the inverse document frequency IDF as the number of news where the given substring occurs divided by the total number of news.

The hypothesis that positive keywords, precisely their stems, are in a majority when the market is going up and reversely that the negative keywords are in a majority when the market is going down could not be proven. The validity of the result is 50 % in the first experiment.

In the second experiment, we used the first 1000 words with the largest probability in classes Up1999 (T1) and Down2000 (T2) (see section 4.3), filtered them intuitively to find positive and negative keywords (25 positive, 19 negative), compressed to substrings, and computed their IDF. The results have shown that the hypothesis is not valid for positive keywords (valid only for 1 case from 25 in Tab. 3) but it is valid for negative keywords (valid for 16 cases from 19). Detailed tables can be found in [2].

We could formulate another hypothesis and start a bigger experiment in this direction because inverse document frequency of subsets of negative substrings seems to correspond with the falling trend. We could perhaps try to prove a weak hypothesis saying that negative keywords are in a majority during falling markets. But instead of that we performed the next experiment with term probabilities.

4.3. Term probabilities

In this experiment, we were investigating the two classes Up1999 and Down2000 using the statistical toolkit BOW [5] for diagnostics of the lexical model. We were looking for the probability of positive and negative keywords. The hypothesis that in good times the probability of positive words in stock exchange news is greater than the probability of the same positive words in bad times (and vice versa for the negative words) could not be proven. We found 17 out of 43 words which did indeed fit with our hypothesis - the rest (26), however, did not.

4.4. Basic classification

In this experiment, we were investigating the four classes (T1=Up1999, T2=Down2000, T3=Down2002, T4=Up2003) using again the statistical toolkit BOW [5] but now for the classification of documents. Additionally, we built a class T5=Now2003 containing 8 documents with

Keywords	IDF-Up1999	IDF-Down2000	Valid?
neu	35.06	39.26	no
gewinn	22.69	24.39	no
mehr	17.59	20.16	no
stieg	16.59	19.68	no
best	14.43	16.37	no
gut	13.32	14.55	no
fest	11.19	12.52	no
positiv	9.67	11.90	no
besser	9.52	11.15	no
wachstum	9.46	12.57	no
erreich	9.21	10.17	no
sehr	8.75	10.37	no
erfolg	7.69	9.57	no
profit	6.13	6.51	no
erzielt	5.52	7.69	no
wichtig	4.62	6.07	no
kletter	4.06	4.72	no
zuleg	2.92	2.82	yes
hohen	2.89	3.63	no
optimist	1.13	1.26	no
expand	0.90	1.36	no
intensiv	0.85	1.19	no
geschaffen	0.75	1.11	no
umsatzplus	0.69	0.95	no
etablier	0.68	1.02	no

Table 3. Inverse document frequency of positive substrings with largest probability

messages of the last 8 weeks of the year 2003 (at the point the research was running). We wanted to find out to which class these messages would be assigned, i.e. in what a trend we just were. As a result (Naive Bayes method) all documents were assigned to their classes with exception of 3 of 4 documents on NOW2003 that were assigned to UP2003.

Using the method of probabilistic indexing, we got another classification. Five documents of Down2000 were assigned to Up1999, 5 documents of Up2003 were assigned to Down2002 and 3 documents of Now2003 were assigned to Down2002.

4.5. Classification of the current trend

After these experiments we used all documents of classes Up1999, Up2003, Down2000, and Down2002 as training sets. As a testing set we used all documents of the class Now2003.

Using the Naive Bayes method, all 8 documents of class Now2003 were classified as being members of class Up2003. Class Now2003 was not in the training set and the class Up2003 was the nearest one. This result has been proven by 25 trials and corresponds with the current reality because trend Up2003 seems to continue.

This would be a promising result but by using the probability indexing method for the same classification we obtained a completely different result, in which all documents of Now2003 had been assigned to class Down2002.

4.6. Similarity of classes

The next investigated question was how documents would be classified when their classes were not a part of the training set. One could expect that e.g. documents of class Up1999 have much more in common with class Up2003 than with classes Down2000 and Down2002. We can generalize and state a hypothesis that documents from Up-classes have enough common features to belong to a common class Up. The same we could expect from the documents of the Down-classes.

The hypothesis stated above could not be proven. For example, all 16 documents of the class Down2002, which were not in the training set, had been assigned to the class Up2003, all 16 documents of Up2003 were assigned to Down2002.

The hypothesis that documents of classes Down2000 and Down2002(resp. Up1999 and Up2003) have enough similarity so that documents of class Down2003 will be assigned to class Down2000 when class Down2003 was not in the training set could not be proven. It has been found that in such a case documents will be assigned to a class which is the nearest in time not in the features of the market.

4.7. Up-Down classification

For this experiment, we built a new model, in which we formed a new class Up (32 files) from the documents of classes Up1999, Up2003 and and a new class Down (32 files) from the documents of classes Down2000 and Down2002. The class Now2003 (8 files) was not modified.

Using a training set with 50 % of data (method Naive Bayes, resp. method of probability indexing)) we obtained the following results. From 16 documents of Down were 10, resp. 11 classified as Up. From 16 documents of Up were 2, resp. 4 classified as Down.

As we defined all data of set Now2003 as test data, all 8 files of Now2003 were assigned to the Up class in all trials, which corresponds with the results obtained before. When using the Up class as a testing set (all others as training set), 28 files were assigned to Down and 4 files assigned to Now2003. When using the Down class as a testing set (all others as training set), all 32 files were assigned to Up class.

5. Conclusions

This paper is the first attempt to investigate the relation between market news and long-term trends of the market. We have found that:

- After we had reduced the number of classes to two (Up and Down) the classifier classified news with an average accuracy of about only 70 %.

- The little similarity between classes Up1999, Up2003 resp. Down2000, Down2002 found and described in section 4.6 supports our statement that methods based on term frequency are not suitable for text mining. In this case the similarity as neighbor in time seems to have more influence than the similarity as neighbor in market trends. The following example contains two news stories having identical term frequency an illustrates that simple statistics of term frequency are not strong enough to distinguish good news from bad news in all cases.

  Example:

  News story 1: ”XY company closed with a loss last year but this year will be closed with a profit”.

  News story 2: ”XY company closed with a profit last year but this year will be closed with a loss. (End of example)

- Often authors are using ambiguity and common phrases intentionally because the market situation is not clear and they have to generate some news. Sometimes authors are following interests of their employers or some investor groups and therefore their messages cannot be taken very seriously.

The short-term influence of a news story depends not only on its content but very much on the current state of the market, on the current mood of investors, and on others news. This means that the same news story could cause quite different market reactions depending on the time point it appears. That is why we have analyzed large sets of news and long-term market trends. Exploiting textual information can not be seen as the only method of market prediction but it may potentially increase the quality of the prediction process.

To get farther though we need more sophisticated language models and analysis. The correspondence between classes of news and market trends could practically be used for market forecasting if we would be able to classify news more exactly.

References

[1] M. Hearst. What is Text Mining? http://www.sims.berkeley.edu/ hearst/text-mining.html.

[2] P. Kroha and R. Baeza-Yates. Classification of Stock Exchange News. Chemnitzer Informatik-Berichte CSR-04-02, TU Chemnitz, Nov. 2004. ISSN 0947-5125.

[3] V. Lavrenko, M. Schmill, D. Lawrie, P. Ogilvie, D. Jensen, and J. Allan. Language models for financial news recommendation. In *Proceedings of the Ninth International Conference on Information and Knowledge Management*, pages 389–396, 2000.

[4] S. Macskassy and F. Provost. Intelligent Information Triage. In *Proceedings of SIGIR'01, New Orleans, USA*, Sept. 2001.

[5] A. McCallum. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/ mccallum/bow.

[6] U. Y. Nahm and R. Mooney. A Mutually Beneficial Integration of Data Mining and Information Extraction. In *Proceedings of the 7th National Conference on Artificial Intelligence AAAI 2000, Austin, USA*, pages 627–632, 2000.

[7] U. Y. Nahm and R. Mooney. Mining Soft-Matching Rules from Textual Data. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence IJCAI-01*, 2001.

[8] U. Y. Nahm and R. Mooney. Text Mining with Information Extraction. In *Spring Symposium on Mining Answers from Texts and Knowledge Bases AAAI 2002, Stanford*, 2002.