Retrieving News Articles Relevant to Stock Market Fluctuations

[Final Project Report for cs498cxz]

Jai Vasanth  
University of Illinois, Urbana Champaign  
vasanth2@uiuc.edu  

Andra Ivan  
University of Illinois, Urbana Champaign  
aivan2@uiuc.edu  

ABSTRACT
This paper presents a novel approach for finding the correlation between news articles and stock prices from an information retrieval perspective. We accomplish this by ranking news articles in a very large collection based on their relevance to market price changes. The news collection is composed of six years of news and the corresponding daily stock prices. Each article in the collection is labelled as being relevant or not with respect to a significant change in stock price over a specific time window. We evaluated the performance of logistic regression, maximum entropy, rsj and compared it to a baseline. Our results show that for lower recall our methods perform significantly better than the baseline.

Categories and Subject Descriptors  
H.4 [Information Systems Applications]: Miscellaneous

General Terms  
classification, ranking

Keywords  
retrieval, text classification, language models

1. INTRODUCTION
A lot of work has been done in the field of financial analysis trying to predict market trends by using time series analysis on the stock prices. No one has been able to claim success with these methods. Furthermore, we know that what motivates people to buy or sell stocks is heavily biased by the news that they read about the company. Keeping up with today’s abundance of online news articles can be a job by itself. So we propose methods for automatically ranking news articles based on their potential impact on stock price swings. In this way, a recommendation system can be built to prioritize incoming news articles.

The closest work in this direction was done by Lavrenko, but taking a slightly different approach [4]. They create separate language models for each company, which restricts their predictions only to companies whose language models they learned. We hope to show that our methods provide a more general approach to retrieving stock-sensitive news.

The rest of the paper is structured as follows: Section 2 describes the data collection we used for testing our retrieval models. Section 3 describes the implementation of our models, and we show their performance in section 4.

2. DATA COLLECTION
Our data consists of 44533 news articles spanning over a period of six years (from 2000 to 2005). The articles were collected from CNET News (www.news.com). Also the daily stock prices of all the companies in the RUSSEL 3000 and NASDAQ index over the same period was collected from Yahoo Finance. The number of companies whose stock information we collected totaled to 6063. The daily NASDAQ index values were also collected over the same period of time.

2.1 News Articles
All the articles in our data collection have been preprocessed from html to plain text format. We also extracted the date published for each of the articles, and used it in our procedure to correlate stock prices. Furthermore, we anticipated that we would need more semantic information about each article, in terms of name entities, so we ran each article through an open source name entity tagger (Opennlp Tools: http://opennlp.sourceforge.net/) that could identify names of people, organizations, locations, numerical values, money etc. Having this information, we proceeded to extract the list of company names that were mentioned in each article and label each article as relevant or not to stock prices, based on the maximum stock fluctuation, as explained in 2.3.

2.2 Stock Prices
The Yahoo Finance website (http://finance.yahoo.com) provides daily stock price for all the companies. Their web interface is designed so as to make querying within a particular time window convenient. The various prices that were obtained were Opening Price, Highest Price, Lowest Price and Closing Price. Not much processing needed to be done. All the stock data were added into an database for efficient querying.
2.3 Defining Relevance

We search for named entities (NLTK: http://nltk.sourceforge.net/) in the news article and register all the company names in the news article. Also each news article is accompanied by a date. We picked a time window of two days and using the stock data information computed the percentage change in the stock value over the two days for all the companies that were identified in the news article. Also the corresponding percentage change in NASDAQ index price is obtained and is subtracted from the stock value changes. The NASDAQ prices represent the trend of all the tech companies. Even though the stock price of a company climbs up by a significant amount, it is not very likely that the news caused the change if the NASDAQ also went up since the entire market went up. It was just a market effect, rather than a news effect. The stock price after having NASDAQ value removed will better represent the effect news. Thus we are concerned only about stock price changes significantly above or below the NASDAQ. News articles which consist of stock price change of more than five percent are considered to be relevant (five percent is a reasonable heuristic). All our datapoints are labelled this way.

3. MODELS FOR MEASURING RELEVANCE

We suspected that there is an underlying distribution of words that distinguishes between stories related to some company’s performance or general news about the industry. The simplest and most intuitive way to model these newspaper articles is the “bag of words” model.

Therefore, we tried several classification and retrieval models that could learn distribution of relevant words (or features) from our training set, and evaluated the methods’ ability distinguish between relevant and non-relevant articles. Our model would most be useful to a financial reviewer, if among a few hundred articles published online daily, it would be able to select 30 or 50 most likely to significantly change some company’s stock in the next few days. The next few sections discuss in more detail the models that were implemented and evaluated, while section 4 shows their evaluation results on the test set of articles.

We have focused mainly on generative models, because we felt that binary classification is inappropriate for this problem since there is definitely a gray area between what constitutes a relevant versus non-relevant article for a future stock price change. Undoubtedly, if we were to ask humans to judge our stock collection as relevant versus non relevant, the training and test set labels would change quite a bit, and we are confident that our system’s performance would change.

3.1 Logistic Regression

Logistic regression is a technique where the posterior probabilities of the two classes (relevant and non-relevant) are modeled as linear functions of the features. [2]

\[
\log \frac{Pr(C = 1|X = x)}{Pr(C = 0|X = x)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n \tag{1}
\]

In the case of a two class problem this becomes:

\[
\log \frac{Pr(C = 1|X = x)}{1 - Pr(C = 1|X = x)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n \tag{2}
\]

Once we estimate the parameters \(\beta_i\) we can start making predictions

\[
Pr(C = 1|X = x) = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^{n} \beta_i x_i)}} \tag{3}
\]

We use a bag of words as our features. All the words are cleaned (all digits, punctuation and other symbols removed) and converted to lowercase. Also, features are normalized so that all company names are translated to <organization> feature, all currencies are translated to <money> feature, all names are translated to <person> feature and all dates to <date> feature. This would make our features more informative and yet general. The normalization procedure was possible because we tagged all the named entities in the news articles.

We used Auton Lab’s implementation of logistic regression [3]. They implemented logistic regression using truncated regularized iteratively re-weighted least squares algorithm. This made our training process pretty fast. The probability of a news article being relevant to stock price change was obtained for all test articles.

3.2 Maximum Entropy

Suppose that we are given \(n\) feature functions \(f_i\), that add the constraints that are important in the modeling process. Maximum entropy [1] produces a model \(P\) that satisfies the following two constraints

1. \(P\) is in agreement with all the \(n\) feature functions.

\[
P(f_i) = \bar{P}(f_i) \tag{4}
\]

Where \(\bar{P}\) is the empirical probability estimated from the train set.

2. Among all the models \(C\) that satisfy the previous constraint, \(P\) is the most uniform (\(P\) has the maximum entropy). A mathematical measure of the uniformity of a conditional distribution \(p(y|x)\) is provided by the conditional entropy

\[
H(p) = -\sum_{x,y} \bar{p}(x)p(y|x) \log(p(y|x)) \tag{5}
\]

Hence

\[
P = \arg\max_{p \in C} H(p) \tag{6}
\]

The feature functions that we used for learning the maximum entropy model is

\[
f_{w_i}(d) = \begin{cases} 
\mathcal{C}(w_i, d) & \text{if } w_i \text{ exists in } d \\
0 & \text{otherwise}
\end{cases} \tag{7}
\]

Where \(w_i\) is the \(i^{th}\) word and \(d\) is an article. The features that these functions generate are exactly the same ones that we used for logistic regression also. Also the words are cleaned and normalized just as in the case for logistic regression.

Maximum entropy model gave the probability of the test articles being classified as relevant to stock price change.
3.3 Robertson-Sparck Jones Model

The Robertson-Sparck Jones (RSJ) model is a generative model for measuring relevance in a collection. Given two sets of training documents (labeled each respectively relevant or non-relevant), we can learn the following two distributions for the words in the vocabulary: \( p_i \), the probability that each word \( i \) occurs in a relevant document, and \( q_i \), the probability for each term that it occurs in a non-relevant document [1]. These probabilities are directly estimated from the training set of articles, using Dirichlet smoothing.

\[
p_i = \frac{\# \text{relevant documents with } \text{term}_i + 0.5}{\# \text{total relevant documents} + 1} \tag{8}
\]

And similarly,

\[
q_i = \frac{\# \text{non-relevant documents with } \text{term}_i + 0.5}{\# \text{total non-relevant documents} + 1} \tag{9}
\]

The relevance score is defined as follows, where \( Q \) is the query (in our case a full text article), and \( D \) is the document model represented by \( p_i \) and \( q_i \):

\[
\log O(R = 1 \mid Q, D) = \sum \log \frac{p_i(1 - q_i)}{q_i(1 - p_i)} \tag{10}
\]

This score assesses the relevance score of a given article by evaluating how likely a term is to appear in the relevant model versus how likely it is to appear in the non-relevant model. Thus, although common terms and stopwords are not filtered out from the plain text articles, they will have a very small contribution to the final document score. The relevance score will be influenced by the appearance or lack of discriminant terms in the document evaluated for relevance.

There are certain variations on this scoring scheme that have been tried and reported in the evaluation section. The RSJ model doesn’t have an explicit way of modeling term frequency in the query document, so it remains to be decided how to weight the frequency of each word in the document to be scored. One possibility is to use the raw term frequency in the scoring scheme as just the count of each word in the query document. It is also possible to use BM25/Okapi frequency weighting, which means it might be the case that even common words can contribute to the relevance model.

\[
W_i = \frac{TF_i(k_i + 1)}{k_i + TF_i} \log \frac{p_i(1 - q_i)}{q_i(1 - p_i)} \tag{11}
\]

4. EVALUATION

In this section we show the performance for each of the models discussed previously. Precision-Recall curves best summarize the performance of our retrieval scores and we are interested to see whether our methods perform better than randomly retrieving articles from the collection. Therefore, we create a precision-recall baseline for our data set by selecting \( k \) documents at random from the collection and plotting the precision/recall values for different choices of \( k \).


Even this simpler RSJ model outperforms the baseline for the entire recall spectrum. We find it strange that raw term frequency weighting achieves better results then Okapi term frequency weighting, which means it might be the case that even common words can contribute to the relevance model.

The best performance over the baseline was achieved with logistic regression and maximum entropy. Both methods tend to produce good precision at low recalls. Maximum entropy beats logistic regression at lower recall, but logistic regression tends to perform better than maximum entropy on an average.

Although not shown in the figures, we have also evaluated...
three additional methods, namely KL-divergence, feedback 2-mixture model and SVMs. The performance is either not significant or below the baseline. SVM provides hard class labels for the points thus treating it completely as a classification problem, whereas we need relevance scores for each article in order to be able to rank them or be able to choose different threshold to evaluate performance. KL-divergence is almost expected to perform poorly, because it just measured similarity between the word probability distribution of each test document with the distribution of the set of relevant articles. This measure is probably heavily weighted by common terms. That is why we also implemented a feedback retrieval model, where the probabilities for each word appearing in the relevant model is evaluated in an iterative EM approach as to maximize the log likelihood of the feedback model. The results were again disappointing, but this may be an inherent problem with EM for problems where the distribution of the data is unknown (we have assumed two mixture models for generating a document: background and relevant, but this may not accurately model the news article data).

In Appendix A we list excerpts from the highest and lowest ranked documents for each of the methods we tried for scoring relevance. We notice that some of these articles appear across the results from all the retrieval methods. Also, we notice that the highest scoring documents tend to contain specific news about a company’s performance, product releases and forecasting quarterly reports. On the other hand, the lowest scoring documents are retrospective comments about the industry.

All these methods provide a score for each feature (a higher score indicates more influence on relevant article). We tried observing the features but unfortunately found no real noticeable trend in them. It could be because of the fact that our features are very simple and that no real pattern can be inferred based on single words. It could also be due to the fact that there are many patterns in the text that are not obviously noticeable that influence the market prices.

5. CONCLUSIONS AND FUTURE WORK
In this paper we showed that treating the problem of financial news recommendation as a classification problem gives relatively high precision compared to the baseline at low recalls. Therefore, modeling news content of financial articles can be more useful when treated as a ranking problem.

There are many direction to explore from here, including a better labeled collection, better features and we also hope that boosting over the train set might improve retrieval performance. In fact all of the models that we have tried will suffer from the same shortcomings: we are ignoring many of the semantic relations between words (so using synonyms) or relative proximity of words could change drastically the top retrieved documents.

6. TASK ASSIGNMENT
Andra Ivan: Parsed the html and retrieved the news, performed named entity tagging and identified the company names, implemented and tested RSJ, KL-divergence, Feedback model
Jai Vasanth: Obtained the news dataset, obtained the stock data, created database and loaded stock information, implemented automatic labeling of the data set (relevant or not), implemented and tested logistic regression, svm and maximum entropy.

7. REFERENCES
APPENDIX

Appendix A: Excerpts from Highest and Lowest Ranked Articles

Top Relevant Articles returned by RSJ

- New top-end Sun chip arriving by October. Unable to meet an earlier deadline, the company now says the 900MHz UltraSparc III has passed tests and will ship in workstations in the next 90 days.
- Earnings season hits Wall Street this week, with hundreds of companies giving investors a glimpse of their business results from the third quarter. Of the 1,262 companies that issued earnings pre-announcements this quarter, 782 companies, or 62 percent, warned of lower earnings.
- Intel’s P6 chip architecture not dead yet. Banias, a low-power chip for notebooks and Net devices, will be based on the older Pentium processor design that’s being phased out in other Intel product lines.
- Amazon.com, Compaq Computer, Lucent and a slew of telecommunications companies are slated to report their quarterly results. With the exception of Amazon, which said it would top previous projections, all major companies reporting this week have lowered the bar for themselves.

Least Relevant Articles returned by RSJ

- Last week’s terrorist attacks marked a significant turning point in the debate over computer and Internet privacy, giving new weight to calls for broader government surveillance powers.
- Software worms have become the weapons of choice for vandals to spread their latest creations, surpassing all malicious-code predecessors in popularity.
- Proposed anti-spam measures before Congress this year are facing opposition from all sides, despite longstanding angst over junk e-mail and high hopes that 2001 would be the year a federal law would pass.
- And during the past year, several companies including Microsoft, IBM, 24/7 Media, EarthLink, Excite@Home and DoubleClick have appointed chief privacy officers, apparently to articulate and enforce privacy policies.

Top Relevant Articles returned by Logistic Regression

- Lucent Technologies has given Wall Street indications that there may be light at the end of the tunnel for the financially besieged company, but some analysts wonder if the worst is yet to come. After a year marked by a series of financial gaffes, Lucent executives have spent the past few months performing a massive makeover of a company that was once the darling of Wall Street investors.
- The PC industry will begin its quarterly health check later today, when Gateway reports its second-quarter earnings. The direct PC seller is expected to report earnings of 36 cents per share and revenue around 2.1 billion when it releases its results after the close of regular trading. For the same quarter a year ago.
- Lucent Technologies’ financial woes continued Tuesday as the telecommunications equipment maker said it expects next quarter’s profits will be lower than expected. In after-hours trading, at 4:40 p.m. PT, Lucent’s stock plummeted 7.25 dollars, from 31.38 dollars to 24.13 dollars, according to the Island ECN Web site.
- Microsoft at last introduced Windows 2000, an ambitious attempt to make up ground in the high-end computing world. The business-use software, which the giant hopes will make inroads on the server computers that power Web sites while staving off fast-growing Linux, marks an industry milestone because numerous other companies are expecting to benefit.

Least Relevant Articles returned by Logistic Regression

- The truth is out there. Just don’t go looking for much of it surrounding “Majestic,” the new online adventure game launched recently by video and computer games giant Electronic Arts. The game, which draws obvious inspiration from “The X-Files” and the 1997 thriller “The Game,” guides players through an interactive mystery based on actual conspiracy theories involving alien contacts and government cover-ups.
- Venture capitalists recovering from dot-com scare. Though venture capitalists say their business has undergone a significant chill, they agree that the drop in temperature is unlikely to threaten—and may even improve—the overall health of the industry. Adam Zong could almost taste the money. He had just wrapped up a presentation to a potential investor and was told that a check for 1.5 million enough to propel his young company.
- Virtual shareholder meetings flop. A certain law lets companies hold shareholder meetings solely online, but critics complain it coshareholders’ rights in jeopardy. You can research stocks, review them, and even purchase them completely online. But once you’ve actually bought the stock, you’re back to the real world. Annual reports, proxies and other corporate documentation are still shipped out by paper mail to shareholders every year.