Weblog Analysis for Predicting Correlations in Stock Price Evolutions

Milad Kharratzadeh[1], **Mark Coates**[2]
Department of Electrical and Computer Engineering
McGill University, Montreal, Canada
[1]milad.kharratzadeh@mail.mcgill.ca, [2]mark.coates@mcgill.ca

Abstract

We use data extracted from many weblogs to identify the underlying relations of a set of companies in the Standard and Poor (S&P) 500 index. We define a pairwise similarity measure for the companies based on the weblog articles and then apply a graph clustering procedure. We show that it is possible to capture some interesting relations between companies using this method. As an application of this clustering procedure we propose a cluster-based portfolio selection method which combines information from the weblog data and historical stock prices. Through simulation experiments, we show that our method performs better (in terms of risk measures) than cluster-based portfolio strategies based on company sectors or historical stock prices. This suggests that the methodology has the potential to identify groups of companies whose stock prices are more likely to be correlated in the future.

Introduction

We investigate whether the content of weblogs provides useful information about the future evolution of stock prices and whether this is complementary to the information embedded in historical stock prices. Our focus is on the correlations between stock prices — the ability to predict such correlations can be used to reduce the risk of an investment portfolio.

The paper is divided into two sections. First, we describe the application of a graph clustering approach to weblog data. We cluster 342 companies[1] from the S&P 500 index using a "similarity" metric we define based on the number of times that the names of the companies co-appear in blog articles. We analyze the resultant clusters and show that the clustering approach can capture interesting relations between companies, including those arising from joint business ventures and external world events.

In the second part, motivated by the fact that stock prices are affected by business fundamentals, company and world events, human psychology, and other factors, we propose a cluster-based portfolio selection method which uses both weblog data and historical prices. By combining these data sources we take into account more factors that affect future prices and can better identify clusters of companies whose prices are likely to evolve in a correlated fashion. We can reduce investment portfolio risk by avoiding investment in multiple companies whose stock prices are likely to decrease at the same time, i.e., companies in the same cluster. We compare our cluster-based portfolio-selection method with other techniques based on company sectors or historical prices.

The rest of the paper is organized as follows. We provide a brief discussion of related work, then describe the blog-based clustering procedure and provide a qualitative analysis of the clusters. We then propose our cluster-based portfolio selection method and compare it to other methods through simulations. Finally, we summarize results and conclude.

[1]Of the 500 companies in the S&P index, we eliminate 158 from our analysis because of reasons such as unavailability of historical price data, rare appearance of the company's name in blogs, or confusion of the name with other instances of the word.

Related Work

A substantial body of work explores the application of clustering techniques to identify companies with highly-correlated stock prices. The data employed has been predominantly historical stock prices. For example, in (Gavrilov et al. 2000) Gavrilov et al. clustered stocks according to measures derived from historical prices and compared the clusters to a "ground truth" based on the S&P 500 sectors. In (Dorr and Denton 2009) Dorr and Denton introduced a novel algorithm for identifying common patterns in different time series and applied it to stock price data to detect companies with correlated price evolutions.

There has also been research into the analysis of news items, blogs and twitter feeds with the goal of predicting future stock prices; see e.g., (Wuthrich et al. 1998), (Mittermayer 2004), (Kaya and Karsligil 2010), (Bollen, Mao, and Zeng 2011). These methods pre-process the collected text documents and then categorize each article based on keywords and feature extraction. Many of the techniques employ sophisticated textual and sentiment analysis algorithms. Different learning procedures are applied to the outputs of these algorithms to predict stock price evolutions.

Our work is novel in its application of a graph clustering technique that employs similarity metrics constructed from a *combination* of historical price data and blog data. We believe our work is the first to highlight the potential of blog data to provide complementary information that can be used to reduce portfolio risk.

Blog-based Clustering of Companies

We now explain the clustering procedure we employ and analyze the resultant clusters. The goal is to illustrate that even an extremely simple procedure for extracting information from blog text can reveal interesting company relationships.

Datasets and clustering methodology

We use two different data sets: weblog contents and historical stock prices. The weblogs were obtained using Spinn3r[2], a web service for indexing the blogosphere. The contents of 133,683,918 blogs were collected between Jan. 13th and Feb. 14th, 2011, resulting in almost 3 TB of data. In later sections, we also analyze blogs from the period Nov. 3rd to Dec. 16th, 2011. The blog data include the original HTML, annotations, and metadata (e.g., author information and time of publication). We processed only the main article text. Historical stock prices were obtained from Yahoo finance[3].

Graph clustering strives to group "similar" objects based on a distance (or similarity) metric. The aim is to maximize the intra-cluster similarity while minimizing the inter-cluster similarity. We first need to define a suitable similarity measure for the objects, in this case the 342 companies from the S&P 500 index. The similarity measure that we define between two companies is the number of mutual appearances of the company names in blog articles in a given time period. This is a very simple metric; more sophisticated methods could be applied to develop more meaningful metrics. Our goal is to highlight the potential information provided by blog data regarding future stock price correlation; the simple similarity metric is sufficient for this purpose.

We can apply any graph clustering algorithm to the coappearance matrices achieved above for each individual day, for several days (by adding the similarity matrices for those days), or in a dynamic way in which the similarity matrix is updated in an adaptive manner. Here, we add up the coappearance matrices for nine days (Jan 13 to Jan 21). If the number of days is very small, then our data is too noisy and the resultant clustering is meaningless. If the similarity matrix is aggregated over many days, then interesting local events go undetected. Nine days is a compromise (relatively similar results are achieved for the range of 7-12 days).

We apply the Greedy-Agglomerative Normalized Cut (GANC) (Tabatabaei, Coates, and Rabbat 2012) graph clustering algorithm which aims to minimize the normalized cut criterion. We chose this algorithm because it delivered the most meaningful results when applied to the data compared to other state-of-the-art graph clustering algorithms. In this case, we chose to cluster the companies into 24 groups. There are 24 sub-sectors (clusters) in the S&P 500 classification of companies; building a clustering with the same number of clusters makes comparison more meaningful.

Analysis of the results

An overview of the clusters identified by GANC is depicted in Fig. 1. In this figure each node represents a cluster, and its size is linearly related to the number of companies in that cluster. The width and darkness of the edges between two nodes show the sum of the weights of the edges between the companies of the two clusters, i.e., thicker and darker edges between clusters indicate stronger ties between them.

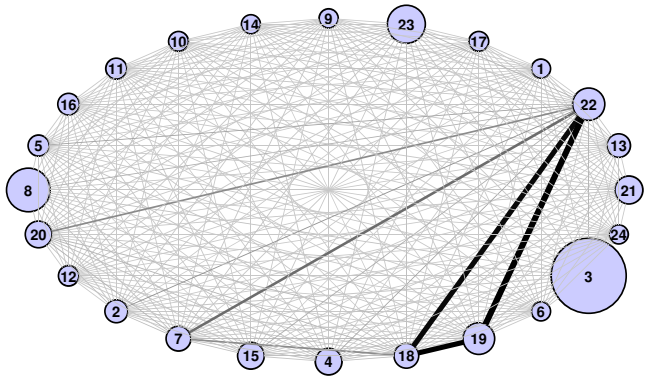

Figure 1: Overview of the clusters. For clarity purposes, we linearly mapped cluster sizes into a specific range.

There are many weak ties, with just a few notably thicker and darker edges. Cluster 22 has a central role and has stronger ties with other clusters. This cluster consists of companies mainly in the fields of "IT" & "Telecom": companies such as Amazon, Apple, and Google. These company names appear often in blogs, leading to strong (and perhaps artificial) ties with other companies. Figure 1 also highlights the triangle of high correlation between clusters 18, 19, and 22. Clusters 18 and 19 contain companies mainly in "Consumer Discretionary" and "Consumer Staples" sectors: companies such as Coca Cola, Nike, and Starbucks. These consumer-oriented companies are also prevalent in blogs.

We now briefly examine some of the more interesting clusters. Cluster 1 consists of two companies, namely Allegheny and FirstEnergy, both belonging to the "Electric Utilities" sub-section. These companies merged on 25th Feb. 2011. The data we used to create the clustering was gathered in January, i.e. prior to this combination. The event could be detected because the news and gossip about the merger started much earlier. This illustrates that our clustering can detect the impact of future market events even though it employs very coarse analysis of weblog text. Cluster 9 consists of Halliburton and Transocean which are two "Energy" companies. These two companies were involved in the Gulf oil spill disaster along with BP and the Presidential Oil Spill Commission blamed them for the disaster.

Some clusters are formed based on company sectors. Examples are cluster 5 (Technology Hardware), cluster 12 (Defense Technology), cluster 13 (Health Care), and cluster 22 (IT & Telecom). A large cluster (#3) consisting of 91 companies from different sectors and with no clear relationship with each other and also with low similarity measures compared to other clusters. The clustering algorithm we used (GANC) commonly produces a large cluster consisting of elements which have weak ties with other objects and do not clearly belong to any of the other formed clusters.

[2]http://www.spinn3r.com/

[3]We employed an EXCEL add-in developed by Randy Harmelink (http://finance.groups.yahoo.com/group/smf_addin/).

Portfolio Selection

We now introduce a cluster-based portfolio selection method that uses both the weblog data and historical stock prices. Stock prices are affected by business fundamentals, company and world events, human psychology, and much more. News from the company and other world events plays an important role in the price fluctuations. The news and analyses presented in weblogs can psychologically affect traders and influence their decisions to buy or sell a stock.

Clustering based on a similarity metric that draws on both weblog data and historical prices allows us to detect correlations induced by news, world events, and psychology (from the weblog data) and correlations induced by business fundamentals (from the historical prices). We thus conjecture that using the weblog data can lead to better prediction of market correlations with reduced uncertainty (risk).

Risk management, portfolios, and risk measures

In the stock market, investing in only one company can be risky, because if the stock price goes down, for any reason, then the entire investment is in danger. The alternative is to invest in a collection of companies, called a portfolio. The dependence on individual stock performance is then much lower. By asssembling a portfolio of companies with non-correlated prices, the risk can be further reduced. In order to compare different portfolios in terms of risk, we need to employ a risk measure which quantifies the risk in some way. Here, we introduce some notation and then explain three different measures that we employ in our comparison.

Consider a given portfolio such as a collection of stocks. We denote the *value* of of this portfolio at time t by V_t. We can model this as a random variable, which is observable at time t. For a given time horizon Δ, such as 1 or 10 days, the *loss* of the portfolio over the period $[t, t+\Delta]$ is defined as $L_{[t,t+\Delta]} = -(V_{t+\Delta} - V_t)$. Although $L_{[t,t+\Delta]}$ is observable at time $t+\Delta$, it is random from the viewpoint of time t. The distribution of $L_{[t,t+\Delta]}$ is called the *loss distribution*. In this work we compute the statistics of loss distribution in an empirical way. Based on the loss distribution, we can define several risk measures (McNeil, Frey, and Embrechts 2005):

The *variance* of the loss distribution is the most basic and well-known risk measure but does not distinguish between gains and losses. The *value-at risk (VaR)* of the portfolio at confidence level $\alpha \in (0,1)$ focuses on losses:

$$\text{VaR}_\alpha = \inf\{l \in \mathbb{R} : P(L > l) \leq 1 - \alpha\} \quad (1)$$

The VaR measure does not take into account the size of the loss. For a continuous loss L with $E(|L|) < \infty$ and differentiable CDF, F_L, the *expected shortfall* at confidence level $\alpha \in (0,1)$ is defined as:

$$\text{ES}_\alpha = E(L | L \geq \text{VaR}_\alpha). \quad (2)$$

Clustering based on blogs and historical prices

We define the following similarity matrix:

$$\mathbf{S} = \lambda \mathbf{W} + (1-\lambda) c |\mathbf{P}| \quad (3)$$

where $\mathbf{W}$ is the coappearance matrix derived using weblogs data, $\mathbf{P}$ is the historical price correlation matrix, and $0 < \lambda < 1$ and $0 < c$ are weighting coefficients. Here, $\mathbf{W}$ is the matrix that results from adding the coappearance matrices for several days. Assume that we have the historical prices of M companies for the past N days and denote the price time-series of company i as $t_k^{(i)}; k = 1, \ldots, N$. Then the elements of matrix $\mathbf{P}$ are defined as the Pearson correlation coefficient of the price time-series:

$$P_{i,j} = \frac{\sum_{k=1}^{N} (t_k^{(i)} - \mu_i)(t_k^{(j)} - \mu_j)}{\sqrt{\sum_{k=1}^{N} (t_k^{(i)} - \mu_i)^2 (t_k^{(j)} - \mu_j)^2}}, \qquad 1 < i, j < M \quad (4)$$

where $\mu_i = (\sum_{k=1}^{N} t_k^{(i)})/N$ is the average stock price of company i over the past N days. The Pearson correlation coefficient represents the correlation of normalized time-series and thus enables us to identify stocks which follow similar trends but are valued very differently, e.g., because of stock splits.The elements of $\mathbf{P}$ are between -1 and $+1$. In forming the similarity matrix we use the absolute value of $\mathbf{P}$ to ensure positivity; there are very few negative elements of $\mathbf{P}$ and they have small values. We scale with a coefficient c so that the values of $c|\mathbf{P}|$ and $\mathbf{W}$ have the same range.

In our experiments we chose $\lambda = 0.5$. The results are very similar for values in the range $0.3 < \lambda < 0.7$ and degrade for larger and smaller λ. For example, the variance increases 3% and 10% respectively for $\lambda = 0$ and 1 compared to $\lambda = 0.5$. Similar results are observed for VaR and ES. Note that $\lambda = 0$ corresponds to the GANC clustering using only the historical prices, and $\lambda = 1$ uses just the weblogs data. This shows that mixing the information is better than using either data source in isolation.

We use the GANC clustering algorithm, for 24 clusters, the same as the number of sectors in the S&P index. Once we have a clustering we make the portfolio by choosing one company from each cluster uniformly at random.

Experimental Results

To build the similarity matrix, $\mathbf{S}$, for each day, we add the coappearance matrices of the last five days to construct $\mathbf{W}$. We use one year ($\sim$250 business days) of historical stock prices for calculating $\mathbf{P}$ and for the K-means algorithm. The choice is dictated by the trade-off between increased noise for short windows and over-smoothing for large windows. We compare portfolios constructed based on three clustering strategies: K-means clustering using historical price data, clustering based on S&P sectors, and GANC clustering based on weblog data and historical prices. Each approach generates a clustering of 24 clusters and we select one company uniformly at random from each cluster.

Assume that we have \$100 to invest in a portfolio. We invest in each of the companies in the portfolio equally (i.e. \$100/24). We perform the clusterings and portfolio-selection for the working days over the period of Jan 18 to Feb 11 (four business weeks). For each day we follow the procedure explained in the previous part (building similarity matrix, performing clustering, making the portfolio), and then empirically compute the three introduced risk measures for the next 100 days (long-term) or 50 days (mid-term). This involves forming an empirical distribution from the actual

future prices and using it to estimate the variance, expectation, and tail-probability in the three risk measures. The same experiment is repeated for 500 different portfolios and the results are averaged to reduce the effects of randomness.

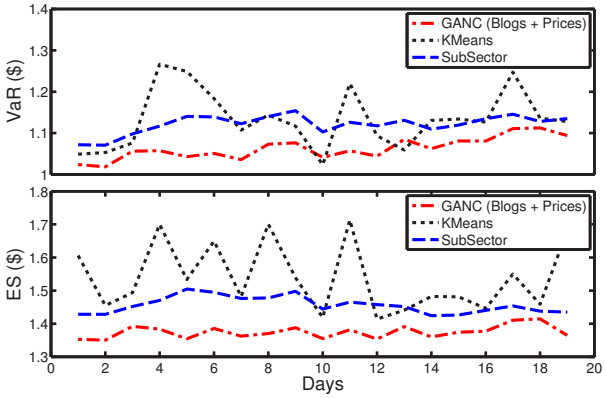

Figure 2: Comparison of Value-at-Risk (top) and Expected Shortfall (bottom)

The means (average loss over the period) are almost the same for the three methods (slightly negative, indicating that we have a negative loss, i.e. benefit). The average values of the risk measures are presented in Table 1. Over the long term (100 days) our method has 24% and 19% less variance than the portfolio-selection methods based on K-Means and sub-sector clusterings, respectively. The long-term empirical risk measures $\text{VaR}_{0.9}$ and $\text{ES}_{0.9}$ are compared over time in Fig. 2. The $\text{VaR}_{0.9}$ of our method, on average, is 5% and 6% less than K-Means and sub-sector clusterings respectively. Also, on average, our method has 6% and 10% lower $\text{ES}_{0.9}$ than K-Means and sub-sector clusterings respectively. Since K-Means clustering is solely based on the historical stock prices, it is prone to high variability. By combining historical prices with weblogs data, we increase the stability as well as decreasing the risk.

	Variance		VaR		ES	
#Days	100	50	100	50	100	50
S&P Sectors	1.07	1.13	1.12	1.28	1.45	1.99
Prices	1.14	1.22	1.13	1.39	1.53	2.10
Blogs+Prices	0.87	1.00	1.06	1.20	1.37	1.79

Table 1: Averaged risks for long-term (100 days) and mid-term (50 days) for two different periods (Jan/Feb and Nov/Dec, respectively).

To verify the above results, we repeated the same experiment for the period of Nov. 3rd 2011 until Dec. 16th 2011. The plots (not shown) for Value-at-Risk, and Expected Shortfall have similar characteristics to those depicted in Fig. 1. The average values of the empirical risk measures, calculated over 50 days, are shown in Table 1. Considering both blog and historical stock price data results in portfolios with substantially lower risk.

Conclusion

In the first part of this paper, we introduced a new way of clustering 342 companies in S&P 500 index based on the data collected from many weblogs. Our clustering uses the similarity matrix which is formed based on the coappearances of the company names in the blog articles. Through analysing the resultant clusters, we showed that our method is able to capture some interesting relationships between the companies and some financial, company, and world events.

In the second part, motivated by the fact that the stock market prices are affected by business fundamentals, company and world events, human psychology, and other factors, we introduced a cluster-based portfolio-selection method which uses both the data collected from weblogs and historical stock prices. We showed through simulation experiments that our portfolio-selection method performs better (in terms of risk measures) than the other ones both in the mid-term and long-term. We did not discuss how to choose companies from the clusters; however, there is a freedom for the investor to choose the companies from the clusters in a more intelligent way which suits its investment strategy.

We did not compare our method to any real-world portfolio selection method, and we do not claim that our method would be competitive. In this work, our goal is to illustrate that we can have a better prediction of the correlations of future stock price evolutions by taking into account the data collected from weblogs. We introduced a methodology for extracting and incorporating this information and showed the potential improvements for one application.

References

Bollen, J.; Mao, H.; and Zeng, X. 2011. Twitter mood predicts the stock market. *J. Computational Science* 2(1):1–8.

Dorr, D. H., and Denton, A. M. 2009. Establishing relationships among patterns in stock market data. *Data and Knowledge Eng.* 68:318–337.

Gavrilov, M.; Anguelov, D.; Indyk, P.; and Motwani, R. 2000. Mining the stock market (extended abstract): which measure is best? In *Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*.

Kaya, M., and Karsligil, M. 2010. Stock price prediction using financial news articles. In *Proc. IEEE Int. Conf. Information and Financial Eng.*

McNeil, A.; Frey, R.; and Embrechts, P. 2005. *Quantitative risk management: concepts, techniques and tools*. Princeton Series in Finance. Princeton University Press.

Mittermayer, M. A. 2004. Forecasting intraday stock price trends with text mining techniques. In *Proc. Hawaii Int. Conf. System Sciences*.

Tabatabaei, S. S.; Coates, M.; and Rabbat, M. G. 2012. Ganc: Greedy agglomerative normalized cut for graph clustering. *Pattern Recognition (to appear)* 45(2):831–843.

Wuthrich, B.; Cho, V.; Leung, S.; Permunetilleke, D.; Sankaran, K.; Zhang, J.; and Lam, W. 1998. Daily stock market forecast from textual web data. In *Proc. IEEE Int. Conf. Systems, Man, and Cybernetics*.