Stock Risk Mining by News

Qi Pan **Hong Cheng** **Di Wu** **Jeffrey Xu Yu** **Yiping Ke**

Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong
Email: {qpan, hcheng, dwu, yu, ypke}@se.cuhk.edu.hk

Abstract

Due to the fast delivery of news articles by news providers on the Internet and/or via news datafeeds, it becomes an important research issue of predicting the risk of stocks by utilizing such textual information available in addition to the time series information. In the literature, the issue of predicting stock price up/down trend based on news articles has been studied. In this paper, we study a new problem which is to predict the risk of stocks by their corresponding news of companies. We discuss the unique challenges of volatility prediction, volatility ranking and volatility index construction. A new feature selection approach is proposed to select bursty volatility features. Such selected features can accurately represent/simulate volatility bursts. A volatility prediction method is then proposed based on random walk by considering both the direct impacts of bursty volatility features on the stocks and the propagated impacts through correlation between stocks. Finally, we construct a volatility index, called VN-index, which is a time series of predicted stock volatility. Moreover, stocks are ranked based on the predicted volatility values. Such information provides investors with knowledge on how widely a stock price is dispersed from the average, as an important indicator of stock risks in a stock market. We conducted extensive experimental studies using real datasets and report our findings in this paper.

1 Introduction

Modern risk management system has been strongly criticized in the recent financial tsunami, where the numerous different arguments converge to one single theme: the current system failed to accurately estimate the risks of financial instruments, which were considered *to be isolated, but in many cases seemingly challenge human understanding* [8]. In stock markets, risk means the uncertainty of future outcomes, and is the probability that an investment's actual return is different from the expected value. The risk of a financial instrument is commonly estimated by the stock price volatility, which measures the variation or dispersion or deviation of an asset's returns from the mean value.

Several models (e.g., ARMA [7], GARCH [1]) were proposed to predict the future volatility based on historical stock price (time series information). However, these models cannot fully capture the bursty behavior of stock prices, especially when there is some breaking news hitting the market.

Several studies [17, 7, 4] have discussed the GARCH forecast errors and related the errors to the arrival of asset specific news articles, i.e., the existing models cannot interpret the change of external environment (news) and therefore could not react accordingly. In [17], a classification model is designed to detect interesting news articles that would help understand the behavior of stock price volatility. However, except for some empirical studies, none of these methods attempted to incorporate news information into risk analysis, or in other words, volatility prediction and ranking.

Although there have been many existing studies [21, 16, 17] which can predict the up/down trend of stock prices, volatility prediction and ranking from news is a new and challenging problem. We highlight the unique issues of volatility prediction, and discuss why the existing work for stock trend prediction can not be directly applied.

First, volatility carries different information from a trend. Figure 1(a) and (b) show the stock volatility and stock prices during 37 trading days (from Sept. 01, 2008 to Nov. 09, 2008), respectively. We can see that there is no obvious correlation between these two time series. Some volatility bursts occur at the turning points of stock price trends (e.g., point 13 and point 27) while others appear when there is no obvious change of stock price trend (e.g., point 21). This is because that volatility is computed as the standard deviation of stock price and therefore reflects market activities from a microscope perspective. As shown in the example, dramatic changes of daily stock prices can cause volatility bursts, but stable daily stock prices do not necessarily imply a stable volatility. A stock which has very stable daily prices may have a big fluctuation in intra-day prices, thus may produce a large volatility value.

Second, the class distribution of training text samples is very skewed if we use a text categorization approach based on news articles to predict stock volatility. Consider the daily ICBC stock prices in 37 days in Figure 1. There are 12 up trends and 25 down trends, with a ratio of 1:2 between up and down trends. On the other hand, there are only 4 volatility bursts out of 37, with a ratio of 1:8 between bursty and non-bursty volatility. Furthermore, there are 185 news articles associated with the up trend, and 363 news articles associated with the down trend, with a ratio of 1:2. On the contrary, there are only 51 news articles as positive samples, associated with bursty volatility, and 497 news articles as negative samples, associated with non-bursty volatility, with a ratio of about 1:10. We have observed similar skewed distributions in our large-scale experiments as well. The small number of positive samples related to the rare volatility bursts makes the problem of predicting volatility bursts chal-

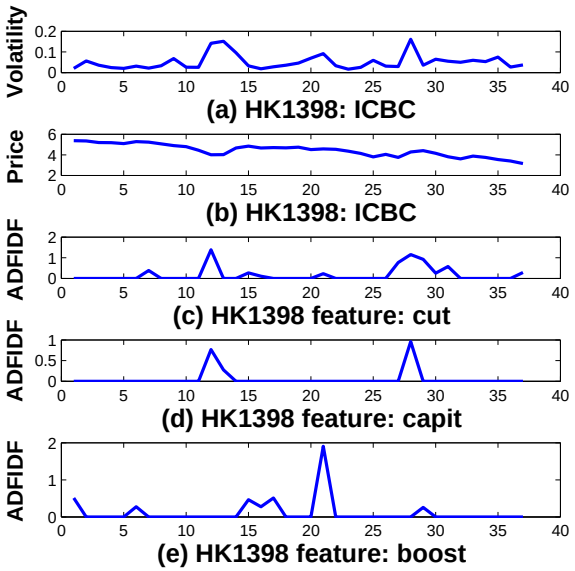

Figure 1: Volatility, Prices, and Features

lenging.

Third, volatility burst prediction shares some similarity with the prediction of the slopes of stock trends, as both problems focus on the magnitude of changes. Existing studies [21, 16, 17] can predict the up/down trend, but may not predict the slope of a trend accurately, because the available information is not sufficient. This evidence also suggests that the existing methods on trend prediction may not work on volatility prediction.

In this paper, we concentrate on volatility prediction by utilizing both time series data (stock prices) and textual information (news articles). First, the textual information is transformed into time series by using the measure ADFIDF (Adjusted Document Frequency Inverse Document Frequency). Second, representative bursty volatility features are selected based on the co-occurrences of historical stock price volatility and news articles. Third, the feature weights, which measure the degree of importance of those features for each stock are learned. Then, based on the feature weights and the incoming news, the volatility of the corresponding stock is predicted. To improve the prediction accuracy on stocks which have very limited news reports, a random walk model is used to propagate the impacts from news among stocks based on their correlation. Finally, a volatility index is constructed as a time series of predicted volatility. Stocks can be further ranked based on the predicted volatility values.

1.1 Main Contributions

The main contributions of the paper are summarized as follows.

- We study a new problem of predicting stock risks based on the predicted volatility by utilizing both time series information (stock price) and textual information (news articles).
- We propose a new feature selection algorithm to select bursty volatility features which have co-occurring bursty patterns with the volatility bursts of stocks. A set of such selected bursty volatility features can accurately represent the stock volatility. Feature weights are learned from historical stock prices and news articles to measure the impact of bursty keywords on stock volatility.
- We further use random walk to propagate the impacts of news among stocks based on their correlation. The random walk approach can greatly improve the volatility prediction performance for those stocks with very limited news reports. The volatility prediction and ranking methods are built on top of the random walk model.
- We conducted extensive experimental studies using real datasets and demonstrated the superiority of our approach in comparison with existing approaches.

The rest of the paper is organized as follows. The definition of stock volatility and the problem formulation are introduced in Section 2. We study bursty volatility feature selection in Section 3, and stock volatility prediction in Section 4. Section 5 presents the experimental study. Section 6 reviews some related works and background information. Finally, Section 7 concludes the paper.

2 Problem Statement

Definition 1 *Volatility is the standard deviation of the continuously compounded returns of a stock within a specific time horizon and is used to measure how widely prices are dispersed from the average as follows:*

$$\sigma = \sqrt{\sum_{i=1}^{n} [R_i - E(R_i)]^2 P_i} \quad (1)$$

where R_i is the possible rate of return, $E(R_i)$ is the expected rate of return, and P_i is the probability of R_i.

Problem Statement: Given a set of stocks $S = \{S_1, S_2, ...\}$ where S_i is a time series, and a set of documents $T = \{T_1, T_2, ...\}$ available before or at time t, we focus on predicting stock volatility and ranking stocks based on the predicted volatility at the next time unit $t + 1$, based on the available textual information.

3 Bursty Volatility Features

To predict stock volatility, we could detect breaking events from available news articles that are indicators of volatility bursts. We observe that the emergence of a breaking event is usually accompanied with a burst of features (keywords). Some features suddenly appear widely in different news articles when the event emerges whereas their occurrences drop significantly when the event fades away. By monitoring the occurrence changes of the features in news articles, we can identify whether there is any new event occurring. Then the problem is how to select a small set of features which can represent all breaking events.

As we discussed, the number of volatility bursts in a stock S_k is considerably small in comparison with the total number of up/down trends occurring in the same stock. Even though the number of news articles that are related to the volatility bursts in the stock S_k, is also observed to be small, we believe that the features in those documents can potentially predict/rank volatility bursts effectively. The desirable properties of a feature are discussed below.

Bursty Occurrences: An effective feature needs to be a bursty feature rather than a stable feature over a time interval. It is most likely that such bursty features can effectively represent volatility bursts.

High Indicative Ability: An effective feature needs to have high ability to indicate volatility bursts, i.e., the bursts of a feature need to be a good indicator of the volatility bursts of the corresponding stock. Features whose high occurrences are always accompanied with volatility bursts are more preferable than those features whose high occurrences only cause volatility bursts occasionally.

High Coverage and Low Redundancy: A minimal set of selected effective features needs to cover the volatility bursts as much as possible. By coverage we mean that the set of selected effective features, as a whole, captures all volatility bursts. By redundancy we mean that some selected features may give similar information.

In the following, we discuss bursty feature measurement and how to select bursty volatility features.

3.1 ADFIDF Measure

As each stock is representing a company, if there are some important things related to the company, the news appears immediately. Generally, the wider the news is reported, the more important the news is. If there is no bursty news, the value which measures the feature burstness should be around average. In the following, we discuss how to capture the wideness of a text feature.

Given a set of stocks $S = \{S_1, S_2, \ldots\}$, where a stock $S_k = [s_{k1}, s_{k2}, \ldots, s_{k\mathcal{I}}]$ is a sequence of stock prices in the time interval $\mathcal{I}$. In the same time interval $\mathcal{I}$, there exists a set of news/documents, $T = \{T_1, T_2, \ldots\}$, where a document $T_i \in T$ contains a set of features $\{f_{ij}\}_{j=1}^{m}$. We assume that it is known which stock S_k a document T_i is related to. The assumption is reasonable since most financial news providers do provide such information when distributing financial news articles. Then the features in the document T_i can also be identified to which stock they are related. We represent a feature f related to a stock S_k in the time interval $\mathcal{I}$ as a time series, $f(k) = [f^k(1), f^k(2), \ldots, f^k(\mathcal{I})]$, where $f^k(t)$ is defined as follows.

$$f^k(t) = \frac{DF_{k,f}(t)}{N_k(t)} \times \log\left(\frac{N_k(T)}{DF_{k,f}(T)}\right) \qquad (2)$$

where $DF_{k,f}(t)$ is the number of related documents in T containing the feature f for the stock S_k at time t, and $N_k(t)$ is the total number of documents in T related to the stock S_k at time t. Here t is a time unit defined by user, such as one month, one day, or one hour. In this paper, t is defined as one day in our evaluation. Therefore, $f^k(t)$ reflects the wideness of the feature f for the stock S_k at time t. In the following, we call $f^k(t)$ the ADFIDF (Adjusted Document Frequency Inverse Document Frequency) value of a feature f related to stock S_k at time t.

We assume that all $f^k(t)$ values, $\forall t \in \mathcal{I}$, form a normal distribution. Then we can identify the bursty features as well as the bursty time interval where such bursty features occur. Note that although there are many methods, e.g. [15], related to the bursty feature identification, there is no conclusion which is the best. Actually, we could use other distribution, and the final identified features would be similar. A typical normal distribution for $f^k(t)$, for stock $S_k \in S$, $\forall t \in \mathcal{I}$, is shown in Figure 2, where the x-axis is the $f^k(t)$ value, and the y-axis is its density. The value $f^k(t) = 0$ indicates that a feature f occurs when there are no explicit events related to stock S_k at time t. The mean value of $f^k(t)$, denoted as $\overline{f^k}$, corresponds to the maximum density value of $f^k(t)$. We say that

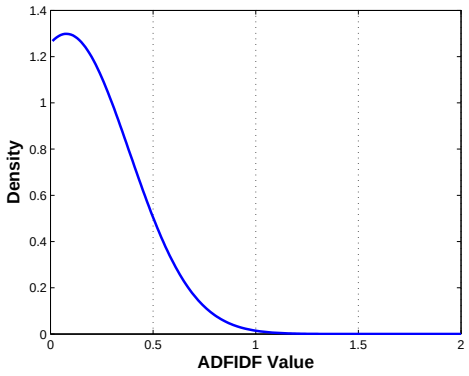

Figure 2: The f^k Normal Distribution

a feature f is a bursty feature related to stock S_k, if $\overline{f^k} \geq \delta$, for a threshold δ. In brief, the higher threshold, the better capability of filtering noise. However, if the threshold is set to be very high, it will miss many effective features; if the threshold is set to be very low, the noise will increase. We will test the threshold in Section 5.2.2. The bursty time interval, denoted as TB_f, for feature f, is a set of time intervals where f appears to be a bursty feature.

It is worth noting that the commonly-used measure, TFIDF (term frequency inverse document frequency) [18], cannot be used since we need the features that witness a stock in a time interval rather than the importance of the features for a document. Therefore, we use a new measure ADFIDF. The idea of DFIDF is brought from [10], but ADFIDF is different from it. The original DFIDF is computed for the whole document set, so the DF and IDF values are global. However, ADFIDF is computed for a subset of documents containing all news related to a specific stock S_k. Moreover, the ADFIDF value is computed for a specific time unit t as in $f^k(t)$.

3.2 Bursty Volatility Features

We identify all bursty features based on ADFIDF. Then we introduce a *co-occurrence rate*, denoted as $E(S_k, f)$, to measure how a bursty feature f and the volatility bursts of a stock S_k occur at the same time. The larger $E(S_k, f)$ is, the more important the feature f to the stock S_k. The idea of co-occurrence is, if a feature always bursts together with the stock volatility bursts in the same time interval, the feature is valuable for identifying volatility bursts. $E(S_k, f)$ is defined below.

$$E(S_k, f) = \frac{V(S_k, TB_f)}{|TB_f|} \Big/ \frac{V(S_k, \mathcal{I})}{|\mathcal{I}|} \qquad (3)$$

where $V(S_k, \mathcal{I})$ is the sum of the bursty volatility values regarding stock S_k in the time interval $\mathcal{I}$.

$$V(S_k, \mathcal{I}) = \sum_{t \in \mathcal{I}} V(S_k, t) \qquad (4)$$

where $V(S_k, t)$ refers to bursty volatility at time t computed using Eq.(1). Recall that TB_f is the set of bursty time intervals of the feature f, and $\mathcal{I}$ is the entire time interval. In Eq.(3), the numerator computes the average volatility over the co-occurrence time intervals of the bursty feature f and volatility bursts. The normalization by the denominator makes the co-occurrence rates of features with respect to different stocks comparable.

Algorithm 1 FeatureRank(S_k, F_k, γ)

INPUT: stock S_k, bursty feature set F_k, decay factor γ
OUTPUT: a list of pairs $(f_j, E(S_k, f_j))$ for $f_j \in F_k$ in descending order

1: compute TB_{f_i} for $f_i \in F_k$;
2: **for all** $f_i \in F_k$ **do**
3: compute $E(S_k, f_i)$ using Eq. (3);
4: **end for**
5: $\mathcal{E} \leftarrow \emptyset$;
6: **while** $F_k \neq \emptyset$ **do**
7: sort F_k in decreasing order based on $E(S_k, f_i)$;
8: let f be the first feature in the sorted F_k;
9: remove f from F_k;
10: append the pair $(f, E(S_k, f))$ into $\mathcal{E}$;
11: **for all** $f_j \in F_k$ **do**
12: $B = TB_{f_j} \cap TB_f$;
13: **if** $B \neq \emptyset$ **then**
14: $V(S_k, t) \leftarrow \gamma \cdot V(S_k, t)$ for $t \in B$;
15: update $E(S_k, f_j)$ based on Eq. (3);
16: **end if**
17: **end for**
18: **end while**
19: **return** $\mathcal{E}$;

3.3 Bursty Volatility Features Selection

In the previous subsections, we have discussed ADFIDF for feature burstness measure and the co-occurrence rate $E(S_k, f)$ for indicative ability measure for a feature f. We will discuss how to select a compact set of bursty volatility features to ensure high coverage and low redundancy.

Consider the example in Figure 1. There are three volatility bursts of the ICBC stock, denoted as S_k, shown in Figure 1(a). In addition there are three ADFIDF sequences of bursty features "cut" (denoted as f_x), "capit" (f_y) and "boost" (f_z), shown in Figure 1(c)-(e), respectively. Here, the two ADFIDF sequences f_x^k and f_y^k have 2 similar bursts corresponding to 2 out of 3 volatility bursts of S_k, and the only burst in f_z^k corresponds to the remaining volatility burst in Figure 1(a). The three bursty volatility features, f_x, f_y, and f_z, together cover the three volatility bursts in S_k. By "cover", we mean that the features jointly represent the volatility information about S_k. Assume $E(S_k, f_x) = E(S_k, f_y) \geq E(S_k, f_z)$ and the goal is to select top-2 bursty volatility features. If we select both f_x and f_y, then one of them is considered as redundant and the third volatility burst cannot be captured.

In order to select a set of representative bursty volatility features, we design an algorithm to rank all bursty volatility features such that the top-k features, to be selected from the ranking list, will be more likely to accurately capture the corresponding volatility bursts of a stock. The algorithm FeatureRank is outlined in Algorithm 1. The main idea is to reduce $E(S_k, f_y)$ if its burst time interval TB_{f_y} is overlapped with another TB_{f_x} for a higher ranked feature f_x, using a feature decay factor γ. As shown in Algorithm 1, it takes a stock S_k, a set of bursty features F_k related to S_k, and a feature decay factor γ. It computes the bursty time interval TB_{f_i} for every feature $f_i \in F_k$ (line 1), and then computes $E(S_k, f_i)$ using Eq.(3) (lines 2-4). Let $\mathcal{E}$ keep a list of pairs $(f_j, E(S_k, f_j))$ in descending order of $E(S_k, f_j)$. In a while loop (lines 6-18), in every iteration, it selects the top bursty volatility feature f from F_k and appends the pair $(f, E(S_k, f))$ to $\mathcal{E}$. It then recomputes $E(S_k, f_j)$ for all remaining $f_j \in F_k$ using the decay factor γ, if there is an overlap between the burst time interval TB_f of the selected feature f and TB_{f_j} of the feature f_j.

According to Algorithm 1, the top-2 bursty volatility features in Figure 1 would be either f_x ("cut") or f_y ("capit") plus f_z ("boost").

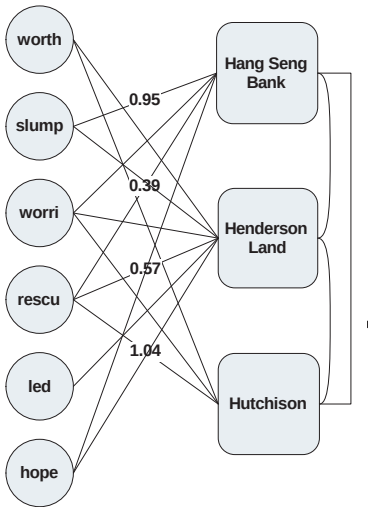

Figure 3: Volatility Prediction Based on Random Walk

4 Volatility Prediction

We have discussed how to select bursty volatility features in Section 3. Such bursty volatility features are selected based on how the burst features in documents co-occur with the volatility bursts in stocks. In this section, we discuss how such selected bursty volatility features are used to predict the stock volatility. The bursty volatility features can have both direct impacts on stock volatility and propagated impacts on stock volatility through stock-stock correlation, as volatility of a stock may affect and be affected by others.

To predict the stock volatility at time t, we use news articles which arrive before t. For example, to predict stock volatility on a particular day, we collect news articles which appear before 10:00AM on that day (market opening time) for prediction. The news articles which appear after 10:00AM will be used for next day prediction.

4.1 Graph Construction

We construct an edge-weighted node-labeled graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ where $\mathcal{V} = \mathcal{V}_F \cup \mathcal{V}_S$ is a set of nodes, $\mathcal{V}_F$ represents the set of bursty volatility features, and $\mathcal{V}_S$ represents the set of stocks. $\mathcal{E} = \mathcal{E}_{FS} \cup \mathcal{E}_{SS}$ is a set of edges, where $\mathcal{E}_{FS}$ represents a set of edges from a node in V_F to a node in V_S, and $\mathcal{E}_{SS}$ represents a set of edges from a node in V_S to another node in V_S. A node in $\mathcal{V}$ is associated with a unique label, so we treat labels as node identifiers. The edge weight on an edge $(v_f, v_s) \in E_{FS}$ represents the impact of a bursty volatility feature $v_f \in \mathcal{V}_F$ to a stock $v_s \in \mathcal{V}_S$. The higher the weight, the larger impact of the feature on the stock. The edge weight on an edge $(v_{s_1}, v_{s_2}) \in E_{SS}$ represents the degree of co-occurrences of volatility bursts between two stocks v_{s_1} and v_{s_2} in $\mathcal{V}_S$. The higher the weight, the more co-occurrences of the volatility bursts of two stocks.

Figure 3 shows a simple graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$. Here, $\mathcal{V}_F$ contains 6 bursty volatility features, "worth", "slump", "warri", "rescu", "led", and "hope". $\mathcal{V}_S$ contains 3 stocks, "Hang Seng Bank", "Henderson Land", and "Hutchison". Table 1 shows the edge

Feature	Hang Seng Bank	Henderson Land	Hutchison
worth	0	0.88213	1.2434
slump	0.94723	0.61459	0
worri	0.98096	0.72786	0.33304
rescu	0.38712	0.56985	1.0362
led	0	0.61459	0
hope	0.63762	0.68553	0

Table 1: Impacts of Bursty Volatility Features

weights from a feature (a node in $\mathcal{V}_F$) to a stock (a node in $\mathcal{V}_S$) for Figure 3. The feature "rescu" is linked to all three stocks with different weights, which means that all these stocks are influenced by the feature "rescu". Some features may only have impacts on a subset of stocks. For example, the feature "led" does not have any impacts on "Hang Seng Bank" or "Hutchison", so there is no edge from "led" to "Hang Seng Bank" or "Hutchison".

Based on the graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, we perform random walk and calculate the volatility of a stock at time $t+1$ based on the available bursty feature information at time t, as well as predicted volatility of correlated stocks, as in Eq.(5).

$$\mathbf{V}(S_k, t+1) = \alpha \sum_{(f_i, S_k) \in \mathcal{E}_{FS}} f_i^k(t) \cdot E(S_k, f_i) + (1-\alpha) \sum_{j \neq k} \rho(S_j, S_k) \cdot \mathbf{V}(S_j, t+1) \quad (5)$$

Here, the first part measures the accumulated direct impacts from bursty volatility features to a stock. This is the information we captured from news to stocks. Recall that $f_i^k(t)$ is the ADFIDF value to indicate how the feature f_i is related to stock S_k at time t (Eq.(2)), and $E(S_k, f_i)$ is the co-occurrence rate to measure how feature bursts of f_i and volatility bursts of S_k occur at the same time. The second part captures the propagated volatility bursts from correlated stocks S_j based on random walk, as stocks may affect each other in the stock market. This part can also improve volatility prediction for stocks which have very little related news. The correlation factor $\rho(S_j, S_k)$ is computed as follows.

$$\rho(S_j, S_k) = \frac{\sum_{\tau=1}^{t} (V(S_j, \tau) - \overline{V(S_j)})(V(S_k, \tau) - \overline{V(S_k)})}{\sigma_{V(S_j)} \sigma_{V(S_k)}} \quad (6)$$

Here, $V(S_k, \tau)$ is the bursty volatility at time τ computed using Eq.(1). $\overline{V(S_j)}$ and $\overline{V(S_k)}$ are the mean volatility values of the two stocks S_j and S_k, respectively. $\sigma_{V(S_j)}$ and $\sigma_{V(S_k)}$ are the standard deviation of volatility for the two stocks in the time interval $[1, t]$.

4.2 Volatility Prediction

Based on the graph and random walk model, we discuss the procedure of volatility prediction. Volatility prediction involves two phases, namely a training phase and a testing phase. The training phase is done based on a set of documents (news articles) T, and a set of stocks S, obtained in the time interval $[1, \mathcal{I}]$. The testing phase is, given a set of new documents T', on a time step t, to predict stocks volatility on the next time unit $t+1$.

The training phase is done as follows. First, for each stock $S_k \in S$, we compute the volatility over the time interval $[1, \mathcal{I}]$, denoted as $V(S_k) = [\sigma_1, \sigma_2, \ldots, \sigma_{\mathcal{I}}]$, where σ_i is computed using Eq.(1). Then we determine a set of bursty features $F_k = \{f_1, f_2, \ldots\}$, where $f_i \in F_k$ corresponds to a time series of ADFIDF $f_i^k = [f_i^k(1), f_i^k(2), \ldots, f_i^k(\mathcal{I})]$ in the time interval $[1, \mathcal{I}]$. $f_i^k(t)$, $t \in [1, \mathcal{I}]$, is computed using Eq.(2). Second, we compute the co-occurrence rate $E(S_k, f_i)$ for every $f_i \in F_k$ using Eq.(3). Third, we obtain a list of pairs $(f_i, E(S_k, f_i))$ using Algorithm 1 to rank the features with a decay factor γ. Finally, we compute the correlation $\rho(S_j, S_k)$ between two stocks S_j and S_k using Eq. (6).

The testing phase is done as follows. Suppose that we obtain a set of new documents T', at time t. First, we compute $f_i^k(t)$ for every f_i in a document in T', that is related to S_k. Second, we construct an edge-weighted graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$. $\mathcal{V} = \mathcal{V}_F \cup \mathcal{V}_S$, where $\mathcal{V}_F$ is the set of features that both appear in T' and are burst volatility features obtained in the training phase. The edge weight for an edge (f_j, S_k), from a bursty volatility feature f_i to a stock S_k is assigned as $E(S_k, f_i)$ which is computed in the training phase. The edge weight for an edge (S_j, S_k), between two stock nodes, is assigned as $\rho(S_j, S_k)$ computed in the training phase. An example is illustrated in Figure 3. Third, we compute $\mathbf{V}(S_k, t+1)$, for every $S_k \in S$, using Eq.(5) iteratively, until it converges based on random walk.

4.3 Volatility Index and Volatility Ranking

Based on the predicted stock volatility, we could perform two analytical tasks: volatility index construction and stock volatility ranking.

A volatility index for stock S_k in the time interval $\mathcal{I}$ is a time series of predicted volatility values: $NI(S_k) = [\mathbf{V}(S_k, 1), \mathbf{V}(S_k, 2), \ldots, \mathbf{V}(S_k, \mathcal{I})]$. We call it VN-index since it is a volatility index constructed from news. If the predicted volatility is accurate, the correlation between VN-index and the real volatility sequence in the testing period should be large. The correlation is quantitatively measured using the Pearson correlation coefficient.

$$\rho(NI(S_k), V(S_k)) = \frac{\text{cov}(NI(S_k), V(S_k))}{\sigma_{NI(S_k)} \sigma_{V(S_k)}} \quad (7)$$

where $NI(S_k)$ is the volatility index sequence, $V(S_k)$ is the real volatility sequence, and cov means covariance. We will evaluate the quality of the constructed VN-index in Section 5.1.

Besides the volatility index construction, we can further rank stocks based on their predicted volatility values $\mathbf{V}(S_k, \mathcal{I}+1)$ for each stock $S_k \in S$. The ranking quality will also be evaluated based on the ground truth from the real volatility value in Section 5.2.

5 Experimental Study

In this section, we evaluate our proposed volatility prediction approach through two groups of experiments: volatility index construction and volatility ranking.

We archive the minute-level intra-day stock prices and the news articles from the Hong Kong Exchange Market and Don Jones Factiva database[1] from Jan. 1, 2008 to Dec. 31, 2008, respectively. All 42 component stocks for Hang Seng Index (HSI) are selected, which are the most influential and most widely held public stocks in Hong Kong. At each day t, the daily realized volatility is computed by applying Eq.(1) on the 1-minute time series.

[1] http://www.factiva.com/

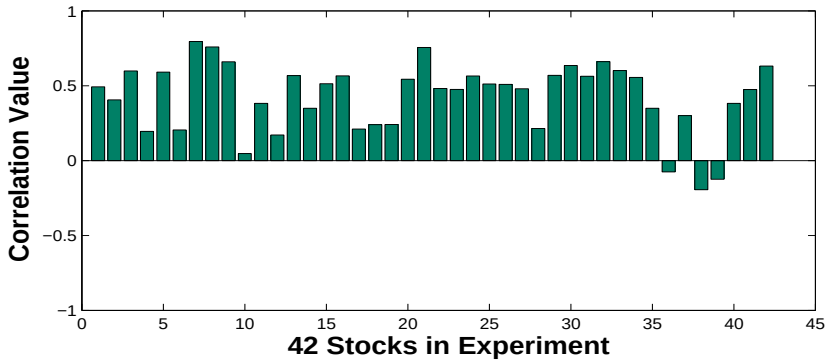

Figure 4: Prediction based on Bursty Features

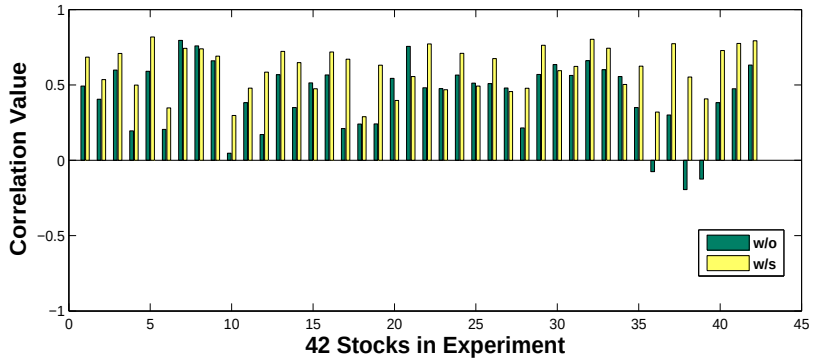

Figure 5: Prediction based on Random Walk

In total, over 150,000 news articles are collected. Each news article is related to a specific stock according to Factiva's classification system. Besides, we only tag news articles which appear before 10:00AM as the news article at that day for prediction, since most newspapers will release their news story before the market opens. The news articles which appear after 10:00AM will be labeled as the news articles of next day for prediction, e.g., the news release at 7:00PM will be used for next day prediction.

For the preprocessing of these news articles, all features are stemmed using the Porter stemmer. Features are words that appear in the news articles, with the exception of digits, web page address, email address and stop words. Features that appear more than 80% of the total news articles in a day are categorized as stop words. Features that appear less than 5% of the total news articles in a day are categorized as noisy features. Both the stop words and noisy features are removed. All data from Jan. 1, 2008 to Nov. 30, 2008 are used for training, and the data from Dec. 1, 2008 to Dec. 31, 2008, are used for evaluation.

We perform the experiments on a PC with a Pentium IV 3.4GHz CPU and 2GB RAM.

5.1 Volatility Index Construction

In this part, we construct the VN-index based on predicted volatility values and evaluate the quality. We focus on the following questions:

(1) What are the effects of the proposed techniques (e.g., direct impacts from bursty volatility features versus propagated impacts based on random walk) in our algorithm? How much improvement can each of them contribute respectively?

(2) What is the overall quality of the VN-index compared with the ground truth?

5.1.1 Prediction based on Bursty Features

First, the VN-index is constructed purely based on the news information without taking the stock-stock correlation into consideration. That means the predicted volatility is computed by setting $\alpha = 1$ in Eq.(5).

The result is show in Figure 4. Each column in the figure represents a correlation value between the real stock volatility and the VN-index for a stock. The average correlation value is 0.4252, the maximum one is 0.7951, and the minimum one is -0.1944. Although the overall performance looks good (note that the average value of correlation between stocks is only 0.4094), the performance varies dramatically for different stocks.

We further analyze the result for those stocks whose correlation is very low, i.e., the predicted volatility is inaccurate. We find that for those stocks, their related features are much less than the average number. When a stock's related features are not sufficient to describe the stock price changes, the volatility prediction based on news is inaccurate.

5.1.2 Prediction based on Random Walk

To improve the prediction for stocks which have very insufficient news reports, we exploit the stock-stock correlation through random walk to propagate the news impacts. In this experiment, the VN-index is constructed based on Eq.(5).

As shown in Figure 5, the left columns are the correlation results based on volatility prediction from news only, while the right columns are the correlation results based on both news direct impacts and propagated impacts from random walk. When random walk is added to the prediction model, the average correlation is 0.6021, which improves a lot from 0.4252. In addition, the correlation value for every stock is

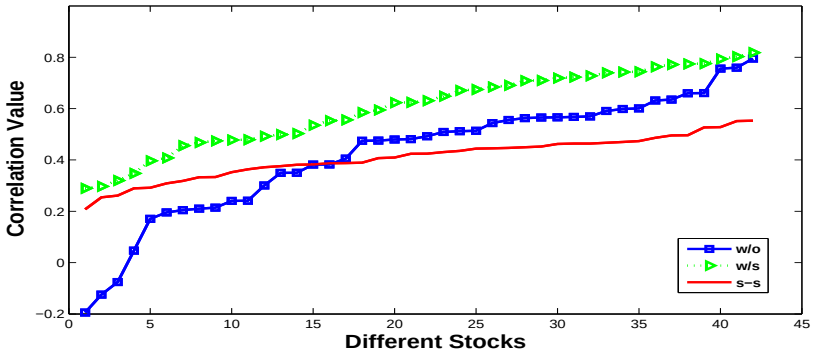

Figure 6: Comparison with Correlation between Stocks

positive.

5.1.3 Comparison with Stock-Stock Correlation

We further compare the correlation of the predicted volatility and the true volatility and the correlation between stocks. As shown in Figure 6, 's-s' means the correlation value between one stock and the other 41 stocks in the whole year of 2008. 'w/s' and 'w/o' represent the results with random walk and without random walk, respectively.

The mean value of correlation between stocks is 0.4094. For our approach without random walk, the average correlation between stock volatility and VN-index is 0.4252. When random walk is applied, the performance is even better. That means the co-movement of VN-index and stock volatility is better than the co-movement of volatility of different stocks in the stock market.

5.1.4 The Effect of Feature Selection

In this section, we evaluate the effectiveness of feature bursts and the impacts of the threshold δ on the correlation value. As shown in Figure 7, the influence of the threshold for a single stock is large. When $\delta = 0$, all features are used. When δ is large enough, no feature is selected. The y-axis is the correlation between the predicted volatility by news and the real daily volatility. When $\delta = 5$, the correlation achieves the maximum, while at both ends (i.e., when δ is very small or large), the correlation is much lower. The experimental results show that, selecting too many features (i.e., when δ is very small) actually decreases the correlation, as many selected features are not bursty features that are indicative of volatility bursts. In an extreme when $\delta = 0$, all possible features from the news articles are used. We can see that the correlation value is actually very low. Similarly, selecting too few features (i.e., when δ is very large) also downgrades the correlation as some of the truly bursty features are not selected. In this paper, we set the threshold δ for each stock as the value which provides the highest correlation value of bursty feature and stock price volatility in the training dataset.

5.2 Volatility Ranking

In this section, we evaluate the quality of ranking stocks based on their predicted volatility values. We focus on the following questions:

(1) How does our approach compare with other approaches in volatility ranking?

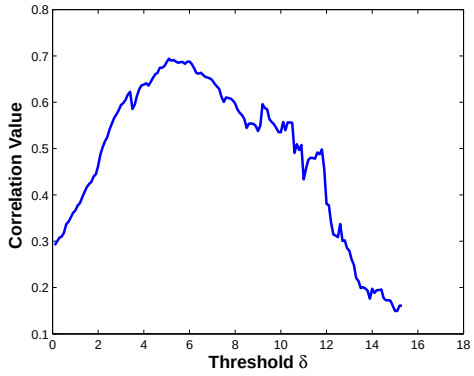

Figure 7: The Influence of δ for A Single Stock

(2) What are the effects of the proposed techniques (e.g., bursty feature, random walk) in our algorithm? How much improvement can each of them contribute respectively?

(3) If we combine our approach with traditional approach purely based on historical stock price data (e.g., GARCH), can we achieve any further improvement?

In this experiment, we use a much smaller training set for evaluation. Specifically, the data from Sept. 01, 2008 to Oct. 24, 2008 are used for training, and the data from Oct. 25 to Nov. 09, 2008 are used for evaluation.

5.2.1 Ranking Quality Comparison

We compare our proposed volatility ranking approach, denoted as VbN, for Volatility-by-News, with the following approaches.

- **Random Selection**: The volatility rank list is formed based on random selection. The accuracy is the statistical mean accuracy value for ranking.
- **Baseline Model**: The volatility rank list is formed based on average volatility on the training set.
- **GARCH**: We apply GARCH model to predict the volatility of next day and rank the stocks based on the predicted volatility. We use a five year daily stock data for training GARCH model, because if the time series is not long enough, the performance will be bad. The UCSD Garch toolbox is used in the experiments.

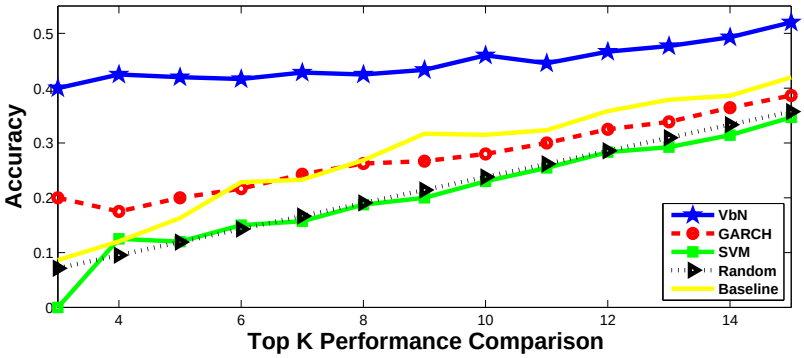

Figure 8: Volatility Ranking Comparison

Normalized Real Volatility	VbN	GARCH	SVM
CHALCO	Li & Fung	Esprit Hldgs	Esprit Hldgs
Li & Fung	CHALCO	Li & Fung	FIH
New World Dev	Esprit Hldgs	HK & China Gas	CITIC Pacific
Esprit Hldgs	FIH	New World Dev	CHALCO
COSCO Pacific	Henderson Land	Sino Land	Sinopec Corp
MTR Corporation	COSCO Pacific	Hutchison	CNOOC
Sino Land	Hutchison	China Shenhua	CLP Hldgs
China Mer Hldgs	HK & China Gas	Henderson Land	Hutchison
Cathay Pac Air	China Mer Hldgs	FIH	Hang Seng Bank
Hang Lung Prop	Sino Land	Cheung Kong	Li & Fung
China Unicom	New World Dev	HKEx	BOC Hong Kong
CITIC Pacific	Cathay Pac Air	Hang Seng Bank	Bank of E Asia
China Resources	Yue Yuen Ind	China Mer Hldgs	China Resources
Yue Yuen Ind	Hang Seng Bank	Cathay Pac Air	SHK Prop
Henderson Land	Cheung Kong	CHALCO	ICBC

Table 2: Ranking Result Comparison

- **SVM**: we label news articles as positive and negative based on whether the volatility bursts occur after the news release, using a similar approach as in [17]. We use the most promising text classification model support vector machine (SVM) [12], to train the text classifier. Based on the classifier and the news features in the testing phase, the volatility of stocks is ranked.

In this experiment, since stocks have volatility in different scales, all the predicted and real volatility values are normalized by their mean value and are transform into relative volatility. The accuracy is measured by overlap-similarity [19], $OS(\tau_1, \tau_2)$, which indicates the degree of overlap between the top n volatility stocks of the two rankings τ_1 and τ_2, where τ_1 is the ranking computed by a model, and τ_2 is the actual ranking from ground truth. The overlap of two stock sets A and B (each of size k) is defined as $\frac{|A \cap B|}{k}$. As a case study, Table 2 shows a comparison among VbN, GARCH, and SVM against the ground truth, using top-15 stocks that have high volatility bursts in the next time unit. In Table 2, stocks are ranked in descending order of the corresponding volatility value. For results in Table 2, the overlap-similarity between VbN and the ground truth (normalized real volatility) is 0.67, larger than that between GARCH/SVM and the ground truth, which are 0.53 and 0.33, respectively.

We further test VbN in comparison with the other four methods by varying k= 3–15 in the top-k ranking list. Figure 8 shows the mean value of accuracy comparison between different methods over the entire testing period. From Figure 8, when k is small ($k = 3$), the accuracy of VbN is 40% higher than SVM, 25% higher than Random Selection, and 20% higher than GARCH. When k increases, the accuracy of all methods increase, but VbN outperforms the other methods in all cases. When $k = 15$, VbN achieves an accuracy of 50% which is around 15% higher than other approaches.

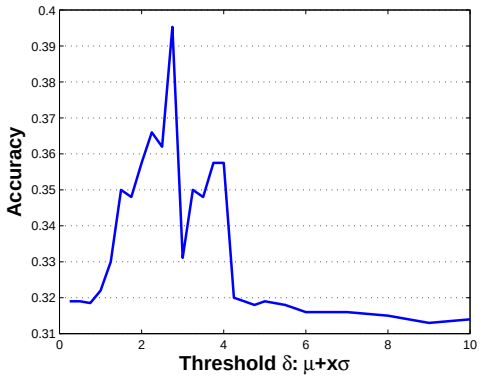

Figure 9: Volatility Ranking with Different δ

5.2.2 Volatility Ranking based on Bursty Features

In this experiment, we evaluate the effectiveness of bursty features and the impacts of the threshold δ on volatility ranking. If $\delta = 0$, all related features are included in the bursty feature set. On the other hand, if δ is set to a large value, there may not be any bursty feature being selected. Figure 9 shows the results. The x-axis is in a range of $\mu + x\sigma$, where μ is the mean of ADFIDF, σ is its deviation, and x is an integer in the range of $[0, 10]$. y-axis is the average accuracy of top-k results for k= 1–15. As shown

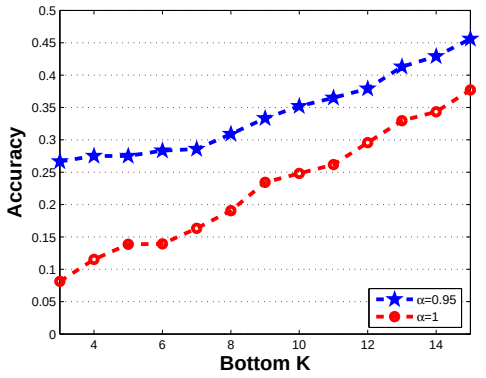

Figure 10: Volatility Ranking by Random Walk

in Figure 9, capturing bursty features is a very important factor for ranking stocks based on volatility. When $\delta = \mu + 2.75\sigma$, the ranking accuracy is the highest (39.5%), which is 7.7% higher than using all the bursty features (31.8%) and 8% higher than using no news information (31.5%). As the purpose of this experiment is to measure the effect of bursty features, we set $\alpha = 1$ in Eq.(5).

5.2.3 Volatility Ranking by Random Walk

We evaluate the effectiveness of VbN based on random walk with $\alpha = 0.95$ in Eq.(5). We observe that the ranking of all component stocks in Hang Seng Index is not noticeably affected by varying α. It is because that the component stocks of Hang Seng Index are reported intensively. Therefore, the improvement based on propagated impacts by random walk is not obvious. In this experiment, we test another set of 42 stocks including 11 HSI-component stocks and 31 non HSI-component stocks that only receive few news articles from time to time. Figure 10 shows the mean value accuracy comparison for the bottom-k out of the 42 stocks, when $\alpha = 1$ (without random walk) and $\alpha = 0.95$ (with random walk) over the entire testing period. As seen in Figure 10, when $\alpha = 0.95$ (with random walk), its accuracy becomes 27.5% which is 20% higher than the accuracy when $\alpha = 1$ (without random walk). When we increase the k value of the bottom-k stocks from 1 to 15, the smallest accuracy margin is still as large as 10%, which indicates random walk is effective to improve the accuracy for ranking stocks that do not frequently receive news articles.

6 Related Works

The first systematic examination on the impacts of textual information on the financial markets was conducted in [13], which compared the movements of Dow Jones Industrial Average with general news during the period from 1966 to 1972. [5] formulated an *activity monitoring task* for predicting the stock price movements, which issued alarms based on the content of the news articles. [21] developed an online system for predicting the opening prices of five stock indices, where by combining the weights of the keywords from news articles and the historical closing prices of a particular index, some probabilistic rules were generated using the approach in [20]. [6] proposed a model for mining the impact of news stories on the stock prices, by using a t-test based split and merge segmentation algorithm for time series preprocessing and SVM [12] for impact classification. [17] discovered a relationship between the news and abnormal stock prices behavior. But it focused more on how to detect these influential news using text categorization.

As aforementioned, the problem of volatility prediction is different from trend prediction. In this work, instead of studying the news articles, we attempt to capture the breaking events by finding a set of representative bursty features which can describe the bursts of stock price volatility.

So far, there have been many studies related the topic of bursty feature detection [15, 11, 10]. For example, [15] proposed an algorithm for constructing a hierarchical structure for the features in the text corpus by using an infinite-state automaton. Similar to these approaches, our proposed algorithm is based on the bursty features. Different from these approaches, we also require the bursty features to be concurrent and have a good coverage over the bursty periods of stock price volatility.

For the web graph, the traditional link analysis methods PageRank [2] and HITS [14] attempted to calculate the importance of a Web page based on the scores of the pages pointing to that page. The rank vector can be computed by repeatedly iterating over the web graph structure until a stable assignment of page importance is obtained. [3] used a bipartite graph to model the process of news generation and built a model to rank the news articles and the sources that generate them. [9] used a set of biased initial restart probability vectors in computing PageRank. They attempted to capture a more accurate importance score with respect to a particular topic. In order to consider indirect influence of features through correlated stocks, this work constructs a graph based on the correlation between stocks and takes the influence from bursty features as the initial starting probability in computing the final volatility rank.

7 Conclusion

In this paper, we studied a new research problem of predicting and ranking stock volatility based on news, where volatility is an important stock risk measure. We discussed the unique challenges of volatility prediction/ranking, and showed that the existing approaches on stock trend prediction cannot effectively solve our problem. We defined the bursty volatility features and proposed an algorithm to select a set of highly indicative bursty volatility features to represent volatility bursts. The main idea is to utilize features in news articles to strengthen the prediction and ranking of volatility. In addition, we proposed a random walk based approach that propagates the news impacts through correlated stocks. We conducted extensive performance study on volatility index construction and stock volatility ranking using real datasets and demonstrated the effectiveness of our proposed approach.

8 Acknowledgements

The work was supported by grants of the Research Grants Council of the Hong Kong SAR, China No. 419008 and 419109, and the Chinese University of Hong Kong Direct Grant No. 2050446.

References

[1] T. Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327, April 1986.

[2] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer*

Networks and ISDN Systems, 30(1–7):107–117, 1998.

[3] G. M. D. Corso, A. Gullí, and F. Romani. Ranking a stream of news. In *Proc. of WWW'05*, pages 97–106, 2005.

[4] L. H. Ederington and J. H. Lee. The short-run dynamics of the price adjustment to new information. *The Journal of Financial and Quantitative Analysis*, 30(1):117–134, 1995.

[5] T. Fawcett and F. J. Provost. Activity monitoring: Noticing interesting changes in behavior. In *Proc. of KDD'99*, pages 53–62, 1999.

[6] G. P. C. Fung, J. X. Yu, and H. Lu. The predicting power of textual information on financial markets. *IEEE Intelligent Informatics Bulletin*, 5(1):1–10, 2005.

[7] R. Gencay, M. Dacorogna, U. A. Muller, O. Pictet, and R. Olsen. *An Introduction to High-Frequency Finance*. Academic Press, 2001.

[8] Greenspan. Excerpts from greenspan speech on global turmoil. In *The Markets, reprinted in The New York Times, November 6*, 1998.

[9] T. H. Haveliwala. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE Transactions on Knowledge and Data Engineering*, 15:784–796, 2003.

[10] Q. He, K. Chang, and E.-P. Lim. Analyzing feature trajectories for event detection. In *Proc. of SIGIR'07*, pages 207–214, 2007.

[11] Q. He, K. Chang, E.-P. Lim, and J. Zhang. Bursty feature representation for clustering text streams. In *SDM'07*, 2007.

[12] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *Proc. of ECML'98*, pages 137–142, 1998.

[13] F. Klein and J. A. Prestbo. *News and the Market*. Chicago: Henry Regenry, 1974.

[14] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, 1999.

[15] J. M. Kleinberg. Bursty and hierarchical structure in streams. In *Proc. of KDD'02*, pages 91–101, 2002.

[16] V. Lavrenko, M. D. Schmill, D. Lawire, P. Ogivie, D. Jensen, and J. Allan. Mining of Concurrent Text and Time Series. In *Proc. of KDD'00 Workshop on Text Mining*, 2000.

[17] C. Robertson, S. Geva, and R. C. Wolff. Can the content of public news be used to forecast abnormal stock market behaviour? In *Proc. of ICDM'07*, pages 637–642, 2007.

[18] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., 1986.

[19] P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison-Wesley, 2006.

[20] B. Wüthrich. Probabilistic knowledge bases. *IEEE Transactions on Knowledge and Data Engineering*, 7(5):691–698, 1995.

[21] B. Wuthrich, D. Permunetilleke, S. Leung, V. Cho, J. Zhang, and W. Lam. Daily prediction of major stock indices from textual www data. In *Proc. of KDD'98*, pages 364–368, 1998.