Stock Trend Prediction by Classifying Aggregative Web Topic-Opinion

Li Xue[1], Yun Xiong[1], Yangyong Zhu[1], Jianfeng Wu[2], and Zhiyuan Chen[3]

[1] School of Computer Science,Fudan University,Shanghai 200433, P.R. China
[2] Shanghai Stock Exchanges,Shanghai 200120, P.R. China
[3] Department of Information Systems,University of Maryland Baltimore County, Baltimore, MD, 21250, USA
{xueli,yunx,yyzhu}@fudan.edu.cn, jfwu@sse.com.cn, zhchen@umbc.edu

Abstract. According to the Efficient Market Hypothesis(EMH) theory, the stock market is driven mainly by overall information instead of individual event. Furthermore, the information about hot topics is believed to have more impact on stork market than that about ordinary events. Inspired by these ideas, we propose a novel stock market trend prediction method by Classifying Aggregative Web Topic-Opinion(CAWTO), which predicts stocks movement trend according to the aggregative opinions on hot topics mentioned by financial corpus on the web. Several groups of experiments were carried out using the data of Shanghai Stock Exchange Composite Index(SHCOMP) and 287,686 financial articles released on SinaFinance[1], which prove the effectiveness of our proposed method.

Keywords: Opinion Mining,Aggregative Opinion,Stock Prediction.

1 Introduction

According to the EMH theory[1], many researchers believe it is a promising way to predict stock movement using the information appears on the web. Among current literature, most studies have tried to analyze the correlation between the time of web news release and that of stock movement. Although text mining based methods using news articles have achieved some success, the accuracy of prediction could hardly reach a satisfactory level[2].

Recently, inspired by the idea of opinion mining, a few studies[3,4,13] have appeared to predict stock market movement by mining sentiment information on the web, such as blog, twitter and editorial, which we call subjective analysis method. Although many studies of this research line have achieved higher prediction accuracy than before, some problems still exist. For example, most of previous studies focused on the opinions about individual events and evaluated them separately, which could hardly capture the overall opinion about hot issues.

Furthermore, most current studies cannot be directly applied to universal corpus covering multiple topics. However, the online financial corpus usually

[1] http://finance.sina.com.cn

J. Pei et al. (Eds.): PAKDD 2013, Part II, LNAI 7819, pp. 173–184, 2013.

involves many topics, just like the corpus listed in Table 1, which covers the main topics of March 15, 2011 on SinaFinance. In addition, the trends of stock market are usually determined by the hot issues on the market, which would always be written as the key topics of financial articles. Thus, it would be a promising way to make stock prediction by the overall opinion on the key topics of daily financial articles.

Table 1. A Corpus Covering Multiple Key Topics

Number of Articles	Topic Description
36	Intellectual Property Protection & Food Security
26	Nuclear Leakage of Japanese Power Station
20	Property Purchase Restriction
10	Japanese Earthquake
8	Scandal of Shuanghui Corp. on Ractopamine
6	American Financial Crisis
4	Debate on nuclear power security in Europe
3	Flood in New Jersey,America
42	Articles about Individual Securities or Company

In this paper, we introduce a novel stock prediction method using aggregative topic-opinion of web financial corpus. Firstly, we extract article opinion by lexicon-based opinion mining method, which takes intensity and polarity as the measures of article opinion. Secondly, we transform the article opinion into a multidimensional topic-opinion vector according to our proposed topic-opinion model. Thirdly, an opinion integration method by using article weight and topic weight is used to generate the Aggregative Topic-Opinion Vectors(ATOV). Finally, the stock market trend could be predicted by these ATOVs.

The rest of this paper is organized as follows. In Section II, we review some previous works on text mining based stock prediction. Next, we present the Topic-Opinion model in Section III and show details about the weighted Topic-Opinion aggregation model in Section IV. Stock trend prediction using ATOV is introduced in Section V. Then in Section VI, we perform two groups of experiments on three datasets and make comparison between the results. Finally, we summarize the paper in Section VII.

2 Related Work

Generally, text mining based stock prediction studies mainly follow two lines: objective analysis and subjective analysis. The objective analysis focuses on mining correlation between objective information of textual articles and stock market movement trend. There have been numerous interesting attempts including the earlier works by Wthrich, et al.[5] which started to use textual news articles for financial forecasting. Later, Lavrenko et al.[6] presented a stock prediction approach by extracting most influential news. Recently, many scholars have turned to studying the market response to financial text streams. For example, Ingvaldsen et al.[7] addressed the problem of extracting, analyzing and synthesizing valuable information from continuous financial text streams. Other than news articles, Kloptchenko et al.[8] presented a technique analyzing quantitative data from annual financial reports.

In recent years, sentiment analysis and opinion mining techniques have attracted much research attention in text mining community. For example, Ahmad et al.[9] studied on extracting multi-lingual sentiment from financial news streams, which could deal with sentiment analysis on Arabic and Chinese corpus; Devitt et al.[10] explored a computable metric of positive or negative polarity in financial news text which was consistent with human judgments and could be used in a quantitative analysis of news sentiment impact on financial markets; Sehgal et al.[3] made stock prediction using web sentiment extracted from message board; Wong et al.[4] brought out a pattern-based opinion mining method for stock trend prediction. Another amazing study that predicted stock index movement by public mood and sentiment extracted from twitter, was carried out by Bollen et al.[11].

Compared with the works mentioned above, our study is more similar to some works of the second line,such as the method proposed by Mahajan et al.[12]. However, their work made stock prediction by the major events extracted from financial news, instead of by the opinion extracted from universal financial corpus.

3 Topic-Opinion Model

3.1 Definitions

Definition 1. *(Article-Opinion) For each article d_i, PS_{d_i} and NS_{d_i} are used to represent the positive intensity and negative intensity of d_i's opinion respectively.*

In this study, the Article-Opinion of article d_i, i.e.,PS_{d_i} and NS_{d_i} are evaluated by lexicon-based method. That is, for article d_i, we determine whether d_i contains any number of negative and positive terms from the sentiment lexicon. For each occurrence, we increase the score of either negative or positive by one.

Definition 2. *(Topic-Opinion) Similar to the idea of topic model,for each article d_i, two vectors, $P\text{-}TOV_{d_i}$ and $N\text{-}TOV_{d_i}$ are defined as topic-opinion of article d_i.*

$$P\text{-}TOV_{d_i} =< ps_{d_i\text{-}topic_1}, ps_{d_i\text{-}topic_2}, ..., ps_{d_i\text{-}topic_k} >$$
$$N\text{-}TOV_{d_i} =< ns_{d_i\text{-}topic_1}, ns_{d_i\text{-}topic_2}, ..., ns_{d_i\text{-}topic_k} >$$

where, $ps_{d_i\text{-}topic_k}$ and $ns_{d_i\text{-}topic_k}$ represent positive lexicon-based opinion and negative lexicon-based opinion of article d_i on $topic_k$ respectively.

3.2 Topic-Opinion Generation

Step One: Topic Extraction

It is more reasonable to predict stock movement trend according to emerging topics from new incoming corpus slice, which capture evolutions of existing topics. Thus, we apply an online LDA model[13] rather than basic LDA model in our study to extract the latest hot issues on stock market.

In this study, we assume that the articles arrive in discrete time slices, the size of which is set to a day.Thus, the topic distribution over words on day t, $\varPhi_k^{(t)}$, is drawn from a Dirichlet distribution governed by the model on day $t-1$, which is presented below:

$$\begin{aligned}\varPhi_k^{(t)}|\beta_k^{(t)} &\sim Dirichlet(\beta_k^{(t)})\\ &\sim Dirichlet(\omega\hat{\varPhi}_k^{(t-1)})\end{aligned}$$

where $\hat{\varPhi}_k^{(t-1)}$ is the frequency distribution of a topic k over words on day $t-1$ and $0 < \omega \leq 1$ is an evolution tuning parameter introduced to control the evolution rate of the model. Thus, the generative model for day t of online LDA model can be summarized as follows:

1. For each topic $k = 1, ..., K$
   (a) Compute $\beta_k^t = \omega\hat{\varPhi}_k^{t-1}$
   (b) Generate a topic $\varPhi_k^t \sim Dirichlet(\cdot|\beta_k^t)$
2. For each document,$d = 1, ..., D^t$:
   (a) Draw $\theta_d^t \sim Dirichlet(\cdot|\beta^t)$
   (b) For each word,w_{d_i}, in document d:
      i Draw z_i from multinomial θ_d^t; $(p(z_i|\alpha_d^t))$
      ii Draw w_{di} from multinomial $\varPhi_{z_i}; p(w_{di}|z_i, \beta_{z_i}^t)$

Similar to many other applications of topic model, there is no fixed rules for setting the value of topic number K. According to our method, it is unreasonable to set K to a large number, since we mainly focuses on opinions about the hot issues in a day. Therefore, topic number K is set to a small integer, 10 in our case.

Step Two: TOV Calculation

After applying online LDA process on corpus slice D_t, the lexicon-based Article-Opinion of each article d_i could be reformulated as TOV_{d_i} by Equations (1-2):

$$ps_{d_i\text{-}topic_k} = PS_{d_i} \cdot p_{topic_i} \tag{1}$$

$$ps_{d_i\text{-}topic_k} = PS_{d_i} \cdot p_{topic_i} \tag{2}$$

Where PS_{d_i} and NS_{d_i} are positive and negative Lexicon-based Article-Opinions of d_i respectively.

4 Weighted Topic-Opinion Aggregation Model

4.1 Weighted Topic-Opinion Aggregation

Definition 3. *(Aggregative Topic-Opinion) For corpus D, two vectors, $P\text{-}ATOV_D$ and $N\text{-}ATOV_D$ are defined as aggregative topic-opinion of corpus D.*

$$P\text{-}ATOV_D = < ps_{D\text{-}topic_1}, ps_{D\text{-}topic_2}, ..., ps_{D\text{-}topic_k} >$$
$$N\text{-}ATOV_D = < ns_{D\text{-}topic_1}, ns_{D\text{-}topic_2}, ..., ns_{D\text{-}topic_k} >$$

where, $ps_{D\text{-}topic_k}$ and $ns_{D\text{-}topic_k}$ represent positive opinion and negative opinion of corpus D on $topic_k$ respectively.

Simply speaking, the ATOV of corpus slice D_t is generated by calculating the aggregate opinions on individual topic one by one. A simple way of aggregating the opinions of corpus slice D_t on topic k is showed in Equation (3):

$$ps_{D_t\text{-}topic_k} = \sum_{d_j \in D_t} \frac{ps_{d_j\text{-}topic_k}}{|D_t|} \tag{3}$$

Obviously, the opinion aggregation approach by Equation (3) assumes every article has an equal weight. However, it is unreasonable to hold this assumption in many cases. Next, we will explain this issue by the case listed in Table 1.

Among the corpus listed in Table 1, 42 articles were written about isolated events that were related to individual stock or company, the other 113 articles mainly focused on eight topics, such as Intellectual Property Protection, Japanese Nuclear Leakage, Property Purchase Restriction etc. All these topics were hot issues on March 15, 2011, which had the major influence on the stock market or some individual securities. Intuitively, the articles written about hot issues would have more impact on stock market than ordinary ones. Thus, the importance of each article should be considered when we aggregate the opinions presented by this corpus.

In addition, the number of articles on different topics varies as shown in the first column of Table 1. For example, 36 articles were written about the topic of Intellectual Property Protection and Food Security, and only 3 articles discussed the problem about the Flood happening in New Jersey. Although, both the topics were key topics on March 15, their influences on stock market were obviously not equal, so the topic importance should also be put into consideration as we aggregate the opinions on different topics. Due to these reasons, we introduce two parameters, w_{d_j} and w_{topic_k} into Equation (3), thus the opinion aggregation method is transformed into the following way:

$$ps_{D_t\text{-}topic_k} = w_{topic_k} \cdot \sum_{d_j \in D_t} \frac{w_{d_j} \cdot ps_{d_j\text{-}topic_k}}{|D_t|} \tag{4}$$

where, w_{d_j} and w_{topic_k} represent article-weight and topic-weight respectively.

4.2 Article-Weight

In our study,we acknowledge the following two assumptions.

Assumption 1: Articles involving similar topics are likely to have similar impacts on stork market.

Assumption 2: Articles involving hot issues are likely to have large impacts on stock market.

Following the above assumptions, the article-weight of d_i could be described by Equation (5):

$$w_{d_i} \propto \frac{|S_{d_i}|}{|D_t|} \tag{5}$$

where D_t is the corpus slice consisting of all articles issued on day t, $S_{d_i}{=}\{d_j|$ $distance(TOV_{d_j},\ TOV_{d_i}) < \delta, d_j \in |D_t|\}$, and $distance(TOV_{d_j}, TOV_{d_i})$ indicates the dissimilarity of topic distributions of d_j and d_i.

4.3 Article-Weight Calculation

By the online LDA method, a document is represented as a vector of probabilities over K topics. Compared with Euclidean distance measure, the Kullback Leibler (KL) divergence is more suitable for computing the distance between two documents distributions, p and q[14], which is given by:

$$KL(p \parallel q) = \sum_i p(i) log \frac{p(i)}{q(i)} \tag{6}$$

Since the KL divergence is not symmetric, we regard the average of $KL(p \parallel q)$ and $KL(q \parallel p)$ as KL distance(KLD) in the rest of the paper.

Taking KL distance as the distance measure for document-topic vectors, we choose K-means as the cluster algorithm. Since the clustering task is performed on the articles issued within a time slice, the article number is relatively small, which is usually less than one thousand. Thus, we can determine the optimal value of k(the number of clusters) by minimizing the Davies-Bouldin index[15] for $k = 1, 2, ..., \sqrt{n}$, where $n = min\{|D_t|, 100\}$, and $|D_t|$ is the number of articles issued within time slice t.

By K-Means cluster algorithm, the articles inside corpus slice D_t are clustered into M clusters. Based on Equation (5), we reformulate article-weight of d_i in the following way:

$$w_{d_i} = \frac{|Cluster_m|}{|D_t|} \tag{7}$$

Where, $|Cluster_m|$ represents the size of cluster m, which is equal to the number of articles inside $Cluster_m$.

4.4 Topic-Weight

The topic importance of $topic_i$ is assumed to be in direct proportion to its lexicon-based opinion intensity, and in reverse proportion to the sum of lexicon-based opinion intensity of all topics. According to this idea, for $topic_i$, positive weight $w_{P\text{-}topic_i}$ and negative weight $w_{N\text{-}topic_i}$ can be calculated by Equations (8-9):

$$w_{P\text{-}topic\text{-}i} = \frac{PS_{topic\text{-}i} + 1}{\sum_{j=1}^{K} PS_{topic\text{-}j} + K}; \tag{8}$$

$$w_{N\text{-}topic\text{-}i} = \frac{NS_{topic\text{-}i} + 1}{\sum_{j=1}^{K} NS_{topic\text{-}j} + K}; \tag{9}$$

Where, PS_{topic_i} and NS_{topic_i} represent positive and negative lexicon-based opinion of $topic_i$, respectively.

5 Stock Prediction by CAWTO

To evaluate the effectiveness of our approach, the composite index of Shanghai Stock Exchanges(SHCOMP) is chosen as the object to be predicted. In this study, SHCOMP-Trend belongs to one of the following possible cases:

- **Up:** It means the SHCOMP of the next day will rise up above one percent than current one.
- **Down:** It means the SHCOMP of the next day will drop down more than one percent than current one.
- **Stable:** It means the fluctuation of SHCOMP will be less than one percent.

The outline of our method is illustrated in Fig.1, which mainly consists of four steps.

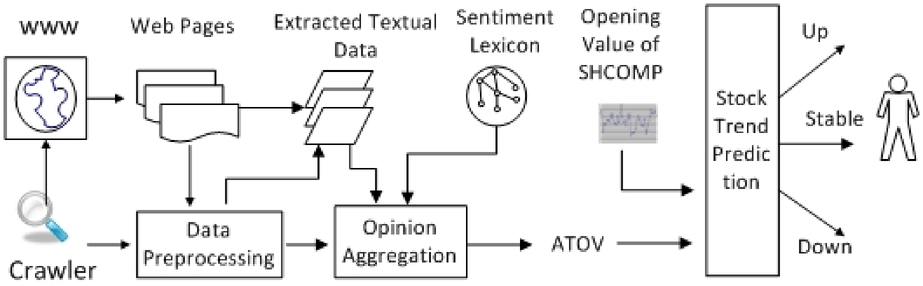

Fig. 1. Data Collection:Crawling daily financial articles from Website,such as, SinaFinance. Data Preprocessing:Performing four basic tasks, text extraction, word segmentation, stop word removal, and background word removal.Feature Extraction: Calculating ATOV according to article weight and topic weight.Making Prediction:Predicting SHCOMP movement trend by ATOV.

6 Experiments

In the experiment,directional accuracy is taken as evaluation metric.For comparison, another two groups of comparison experiments are performed:

- Making SHCOMP-Trend prediction by classifying Event-Topic-Vectors(ETVs).
- Making SHCOMP-Trend prediction by classifying Basic-ATOVs.

ETVs are generated according to the feature representation model proposed by the study[12], where each dimension represents a major event instead of an opinion. Basic-ATOV is the aggregate topic-opinion vector generated according to Equation (3), which aggregates TOVs without considering article-weight and topic-weight.

6.1 Sentiment Lexicon

A Chinese Financial Sentiment Lexicon(CFSL) is used as the sentiment lexicon for lexicon-based Article-Opinion calculation. It consists of 7409 Chinese words in total, including 631 positive words, 575 negative words, and 6203 neutral words, labeled sentiment by financial experts of SSE.

6.2 Data Setting

We crawled 287,686 financial articles issued from April 1, 2009 to September 29, 2011 from SinaFinance,a famous financial website in China,based on which we generate 912 pairs of ATOVs by applying ATOV generation method. Each pair of ATOVs consists of a P-ATOV and a N-ATOV, which represent positive aggregate topic-opinion and negative aggregate topic-opinion respectively. Among 912 pairs of ATOVs, we only select 601 pairs as the input of classifier, since only 601 days are trading days during this period time.In addition, two datasets used in comparison experiments are shown in Table 2.

Table 2. Data Sets used in comparison experiments

Data Set	Data Type	Description
Basic-ATOV	K-dimensional Vector	Contains 601 pairs of Basic-ATOVs
ETV	K-dimensional Vector	Contains 200,000 ETVs

6.3 Experiment Setting

ATOV Generation Setting

The experiment of ATOV aggregation mainly involves online LDA algorithm and k-means clustering algorithm. As to online LDA model, we employ "Matlab Topic Modeling Toolbox", authored by Mark Steyvers and Tom Griffiths in our experiment. The models were run for 200 iterations and the last sample of the Gibbs sampler was used for evaluation. As discussed in Section IV, the number of topics,

K, is set to 10 in all experiments. Following the settings in works[17], a and b are set to $50/K$, and 0.1, respectively. As to k-means clustering algorithm, the clusters number k is determined dynamic by minimizing the Davies-Bouldin index.

Classification Setting

According to current literatures, many studies take Support Vector Machine(SVM) as classifier. For comparison with the prediction accuracy by using other data features, i.e., Basic-ATOV, ETV, a SVM classifier, namely LibSVM[2] is adopted in our experiments.Besides, a classifier based on multiple data domain description(MDDD) is also adopted in the experiments, which was claimed more suitable for multi-class classification task[16].

For the SVM classifier, we use a linear kernel with default parameters (C=1). As to the MDDD classifier, the parameter β is set to 0.2 according to the settings of study[16].

6.4 Result

Result by ATOV

Table 3 shows the directional accuracy of SHCOMP-Trend prediction obtained by classifying P-ATOV and N-ATOV in two-round experiments. From the results, we can clearly find that the N-ATOV dataset achieves higher directional accuracy than P-ATOV in both rounds. In the first round experiments, the directional accuracy of N-ATOV reaches 75.1% when we take MDDD as the classifier, which is the highest directional accuracy. Judging by the two-fold cross-validation method, the average directional accuracy of N-ATOV are 70.3% and 74.7% achieved by SVM and MDDD respectively. This is a clear improvement over 65.2% and 69.5% when the P-ATOV dataset is used, which suggests the N-ATOV has higher predictive ability than the P-ATOV. In other words, our findings imply that negative opinions of hot topics usually have greater impact on stock market than positive ones. Besides, we can also find that all of the directional accuracies achieved by MDDD are higher than the corresponding ones achieved by SVM.

Table 3. Directional Accuracy Using ATOV

Data	Training Range	Testing Range	SVM	MDDD
P-ATOV	Apr,1,2009-Aug,1,2010	Aug,1,2010-Sep,29,2011	65%	68.3%
N-ATOV	Apr,1,2009-Aug,1,2010	Aug,1,2010-Sep,29,2011	69.5%	**75.1%**
P-ATOV	Jun,1,2010-Sep,29,2011	Apr,1,2009-May,31,2010	65.4%	70.7%
N-ATOV	Jun,1,2010-Sep,29,2011	Apr,1,2009-May,31,2010	71.1%	74.3%

Comparison Result

For comparison, the results of experiments using ETV and Basic-ATOV are shown in Table 4 and Table 5 respectively. According to Table 4, the overall directional accuracy achieved by Basic-ATOV(i.e., Basic-P-ATOV, Basic-N-ATOV) is lower than that achieved by ATOV(i.e., P-ATOV, N-ATOV), and the

[2] http://www.csie.ntu.edu.tw/cjlin/libsvm

Table 4. Directional Accuracy Using ETV

Data	Training Range	Testing Range	SVM	MDDD
ETV	Apr,1,2009-Aug,1,2010	Aug,1,2010-Sep,29,2011	59.1%	59.8%
ETV	Jun,1,2010-Sep,29,2011	Apr,1,2009-May,31,2010	59.7%	**61.9%**

Table 5. Directional Accuracy Using Basic-ATOV

Data	Training Range	Testing Range	SVM	MDDD
Basic-P-ATOV	Apr,1,2009-Aug,1,2010	Aug,1,2010-Sep,29,2011	59.1%	59.5%
Basic-N-ATOV	Apr,1,2009-Aug,1,2010	Aug,1,2010-Sep,29,2011	**62.2%**	62%
Basic-P-ATOV	Jun,1,2010-Sep,29,2011	Apr,1,2009-May,31,2010	57.5%	60.1%
Basic-N-ATOV	Jun,1,2010-Sep,29,2011	Apr,1,2009-May,31,2010	60%	61.6%

best result is 62.2%, which is attained by SVM on Basic-N-ATOV. Furthermore, negative opinions(i.e.,Basic-N-ATOV) also display higher predictive ability than positive ones(i.e., Basic-P-ATOV) according to Table 4.

In Table 5, the average directional accuracies achieved by SVM and MDDD on ETV are 59.4% and 60.8% respectively, which are very close to the result claimed by the study[14].From Fig.2, the difference between the directional accuracies obtained by Basic-ATOV and ETV is indistinct. Thus, we can hardly conclude that opinion-based information(i.e., ATOV, Basic-ATOV) has higher predictive ability than event-based information(i.e., ETV), which was claimed by Wong et al.[4].However, both types of ATOV outperform Basic-ATOV and ETV in predicting SHCOMP-Trend.

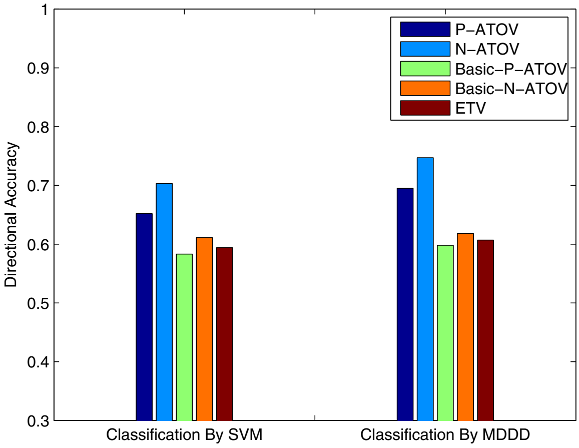

Fig. 2. Prediction Results Obtained By SVM and MDDD

7 Conclusion

In this study, we have explored a new way of predicting stock trend according to the overall opinions on hot topics discussed in web financial corpus. To achieve this goal, a weighted topic-opinion aggregation method is first proposed, by which the aggregative topic-opinion vector(ATOV) can be generated according to article weight and topic weight. By classifying such ATOVs, the stock market movement trend can be predicted with high accuracies. To prove the effectiveness of this method, several groups of experiments on real world data have been carried out, among which the highest directional accuracy of SHCOMP-Trend prediction is up to 75.1%. Furthermore, based on the outcomes of comparison experiments, the ATOV gains a notable advantage over Basic-ATOV and ETV, which are the aggregative opinion generated by a basic integration method and the event-based information extracted by the topic model respectively. In addition, the negative aggregative topic-opinions are found to have higher predictive ability than positive ones. Finally, we also find that different classifiers could lead to relatively large variations of prediction accuracy. Consequently, how to select a suitable classifier according to the type of specific prediction task, i.e., binary classification, multiple classification, becomes one issue of our future works.

Acknowledgment. The research was supported by the National Science Foundation Project of China under Grant.No.61170096 and National Science and Technology Support Program of China under Grant. No.2012BAH13F02.

References

1. Fama, E.F.: The behavior of stock-market prices. The Journal of Busines 38(1), 34–105 (1965)
2. Koppel, M., Shtrimberg, I.: Good news or bad news? Let the market decide. In: Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications, pp. 86–88 (2004)
3. Sehgaland, V., Song, C.: SOPS: Stock Prediction Using Web Sentiment. In: Proceedings of the 7th IEEE International Conference on Data Mining 2007, Data Mining in Web 2.0 Environments Workshop, Omaha, U.S, pp. 21–26 (2007)
4. Wong, K.-F., Xia, Y., Xu, R., Wu, M., Li, W.: Pattern-based Opinion Mining for Stock Market Trend Prediction. International Journal of Computer Processing of Oriental Languages 21(4), 347–362 (2008)
5. Wthrich, B., Cho, V., Leung, S., Peramunetilleke, D., Sankaran, K., Zhang, J., Lam, W.: Daily Prediction of Major Stock Indices from Textual WWW Data. In: Proceedings of the 4th ACM SIGKDD, NY, pp. 364–368 (1998)
6. Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D., Allan, J.: Language models for financial news recommendation. In: Proceedings of the 9th International Conference on Information and Knowledge Management (2000a)
7. Ingvaldsen, J., Gulla, V., Læreid, T., Sandal, P.: Financial News Mining: Monitoring Continuous Streams of Text. In: Proc. IEEE/WIC/ACM International Conference on Web Intelligence (2006)

8. Kloptchenko, A., Eklund, T., Back, B., Karlsson, J., Vanharanta, H., Visa, A.: Combining data and text mining techniques for analyzing financial reports. In: Proc. Eighth Americas Conference on Information Systems (2002)
9. Ahmad, K., Cheng, D., Almas, Y.: Multi-lingual sentiment analysis in financial news streams. In: Proc. of the 1st Intl. Conf. on Grid in Finance, Italy (2006)
10. Devitt, A., Ahmad, K.: Sentiment Polarity Identification in Financial News: A Cohesion-based Approach. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, pp. 984–991 (June 2007)
11. Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. Journal of Computational Science (2011)
12. Mahajan, A., Dey, L., Haque, S.M.: Mining Financial News for Major Events and Their Impacts on the Market. Presented at the WI-IAT 2008. IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, pp. 423–426 (2008)
13. AlSumait, L., Barbar, D., Domeniconi: Online LDA: Adaptive topic model for mining text streams with application on topic detection and tracking. In: Proceedings of the IEEE International Conference on Data Mining (2008)
14. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer (2006)
15. Davies, D.L., Bouldin, D.: A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1:224-1:227 (1979)
16. Xue, L., Chen, M., Xiong, Y., Zhu, Y.: User Navigation Behavior Mining Using Multiple Data Domain Description. In: IEEE/WIC/ACM, International Conference on Web Intelligence and Intelligent Agent Technology (2010)
17. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. The Journal of Machine Learning Research 3, 993–1022 (2003)