Visualization on Financial Terms via Risk Ranking from Financial Reports

Ming-Feng Tsai[1,2] Chuan-Ju Wang[3]
(1) Department of Computer Science, National Chengchi University, Taipei 116, Taiwan
(2) Program in Digital Content & Technologies, National Chengchi University, Taipei 116, Taiwan
(3) Department of Computer Science, Taipei Municipal University of Education, Taipei 100, Taiwan
mftsai@nccu.edu.tw, cjwang@tmue.edu.tw

ABSTRACT
This paper attempts to deal with a ranking problem with a collection of financial reports. By using the text information in the reports, we apply learning-to-rank techniques to rank a set of companies to keep them in line with their relative risk levels. The experimental results show that our ranking approach significantly outperforms the regression-based one. Furthermore, our ranking models not only identify some financially meaningful words but suggest interesting relations between the text information in financial reports and the risk levels among companies. Finally, we provide a visualization interface to demonstrate the relations between financial risk and text information in the reports. This demonstration enables users to easily obtain useful information from a number of financial reports.

KEYWORDS: Text Ranking, Stock Return Volatility, Financial Report, 10-K Corpus.

Proceedings of COLING 2012: Demonstration Papers, pages 447–452,
COLING 2012, Mumbai, December 2012.

1 Introduction

Financial risk is the chance that a chosen investment instruments (e.g., stock) will lead to a loss. In finance, volatility is an empirical measure of risk and will vary based on a number of factors. This paper attempts to use text information in financial reports as factors to rank the risk of stock returns.

Considering such a problem is a text ranking problem, we attempt to use learning-to-rank techniques to deal with the problem. Unlike the previous study (Kogan et al., 2009), in which a regression model is employed to predict stock return volatilities via text information, our work utilizes learning-to-rank methods to model the ranking of relative risk levels directly. The reason of this practice is that, via text information only, predicting ranks among real-world quantities should be more reasonable than predicting their real values. The difficulty of predicting the values is partially because of the huge amount of noise within texts (Kogan et al., 2009) and partially because of the weak connection between texts and the quantities. Regarding these issues, we turn to rank the relative risk levels of the companies (their stock returns).

By means of learning-to-ranking techniques, we attempt to identify some key factors behind the text ranking problem. Our experimental results show that in terms of two different ranking correlation metrics, our ranking approach significantly outperforms the regression-based method with a confidence level over 95%. In addition to the improvements, through the learned ranking models, we also discover meaningful words that are financially risk-related, some of which were not identified in (Kogan et al., 2009). These words enable us to get more insight and understanding into financial reports.

Finally, in this paper, a visualization interface is provided to demonstrate the learned relations between financial risk and text information in the reports. This demonstration not only enables users to easily obtain useful information from a number of financial reports but offer a novel way to understand these reports.

The remainder of this paper is organized as follows. In Section 2, we briefly review some previous work. Section 3 presents the proposed ranking approach to the financial risk ranking problem. Section 4 reports experimental results and provides some discussions and analyses on the results. We finally conclude our paper and provide several directions for future work.

2 Related Work

In the literature, most text ranking studies are related to information retrieval (Manning et al., 2008). Given a query, an information retrieval system ranks documents with respect to their relative relevances to the given query. Traditional models include Vector Space Model (Salton et al., 1975), Probabilistic Relevance Model (Robertson and Sparck Jones, 1988), and Language Model (Ponte and Croft, 1998). In addition to the conventional models, in recent years there have also been some attempts of using learning-based methods to solve the text ranking problem, such as (Freund et al., 2003; Burges et al., 2005; Joachims, 2006), which subsequently brings about a new area of learning to rank in the fields of information retrieval and machine learning. Considering the prevalence of learning-to-rank techniques, this paper attempts to use such techniques to deal with the ranking problem of financial risk.

In recent year, there have been some studies conducted on mining financial reports, such as (Lin et al., 2008; Kogan et al., 2009; Leidner and Schilder, 2010). (Lin et al., 2008) use a weighting scheme to combine both qualitative and quantitative features of financial reports together, and

propose a method to predict short-term stock price movements. In the work, a Hierarchical Agglomerative Clustering (HAC) method with K-means updating is employed to improve the purity of the prototypes of financial reports, and then the generated prototypes are used to predict stock price movements. (Leidner and Schilder, 2010) use text mining techniques to detect whether there is a risk within a company, and classify the detected risk into several types. The above two studies both use a classification manner to mine financial reports. (Kogan et al., 2009) apply a regression approach to predict stock return volatilities of companies via their financial reports; in specific, the Support Vector Regression (SVR) model is applied to conduct mining on text information.

3 Our Ranking Approach

In finance, volatility is a common *risk* metric, which is measured by the standard deviation of a stock's returns over a period of time. Let S_t be the price of a stock at time t. Holding the stock for one period from time $t-1$ to time t would result in a simple net return: $R_t = S_t/S_{t-1}$ (Tsay, 2005). The volatility of returns for a stock from time $t-n$ to t can be defined as

$$v_{[t-n,t]} = \sqrt{\frac{\sum_{i=t-n}^{t}(R_i - \bar{R})^2}{n}}, \tag{1}$$

where $\bar{R} = \sum_{i=t-n}^{t} R_i/(n+1)$.

We now proceed to classify the volatilities of n stocks into $2\ell+1$ risk levels, where $n, \ell \in \{1, 2, 3, \cdots\}$. Let m be the sample mean and s be the sample standard deviation of the logarithm of volatilities of n stocks (denoted as $\ln(v)$). The distribution over $\ln(v)$ across companies tends to have a bell shape (Kogan et al., 2009). Therefore, given a volatility v, we derive the risk level r via:

$$r = \begin{cases} \ell - k & \text{if } \ln(v) \in (a, m - sk], \\ \ell & \text{if } \ln(v) \in (m - s, m + s), \\ \ell + k & \text{if } \ln(v) \in [m + sk, b), \end{cases} \tag{2}$$

where $a = m - s(k+1)$ when $k \in \{1, \cdots, \ell - 1\}$, $a = -\infty$ when $k = \ell$, $b = m + s(k+1)$ when $k \in \{1, \cdots, \ell - 1\}$, and $b = \infty$ when $k = \ell$. Note that r stands for the concept of *relative* risk among n stocks; for instance, the stock with $r = 4$ is much more risky than that with $r = 0$.

After classifying the volatilities of stock returns (of companies) into different risk levels, we now proceed to formulate our text ranking problem. Given a collection of financial reports $D = \{\mathbf{d_1}, \mathbf{d_2}, \mathbf{d_3}, \cdots, \mathbf{d_n}\}$, in which each $\mathbf{d_i} \in \mathbb{R}^d$ and is associated with a company c_i, we aim to rank the companies via a ranking model $f : \mathbb{R}^d \to \mathbb{R}$ such that the rank order of the set of companies is specified by the real value that the model f takes. In specific, $f(\mathbf{d_i}) > f(\mathbf{d_j})$ is taken to mean that the model asserts that $c_i \succ c_j$, where $c_i \succ c_j$ means that c_i is ranked higher than c_j; that is, the company c_i is more risky than c_j in this work.

This paper adopts Ranking SVM (Joachims, 2006) for our text ranking problem. Within a year, if the ground truth (i.e., the relative risk level) asserts that the company c_i is more risky than c_j, the constraint of Ranking SVM is $\langle \mathbf{w}, \mathbf{d_i} \rangle > \langle \mathbf{w}, \mathbf{d_j} \rangle$, where $\mathbf{w}, \mathbf{d_i}, \mathbf{d_j} \in \mathbb{R}^d$, and $\mathbf{d_i}$ and $\mathbf{d_j}$ are two word vectors. Then, the text ranking problem can be expressed as the following constrained

Method	2001	2002	2003	2004	2005	2006	Average
Feature: TFIDF	**Kendall's Tau (Kendall, 1938)**						
SVR (baseline)	0.517	0.536	0.531	0.515	0.515	0.514	0.521
Ranking SVM	**0.539**	**0.549**	**0.543**	**0.526**	**0.539**	**0.525**	**0.537* (6.57E-4)**
Feature: TFIDF	**Spearman's Rho (Myers and Well, 2003)**						
SVR (baseline)	0.549	0.567	0.562	0.545	0.544	0.540	0.551
Ranking SVM	**0.571**	**0.580**	**0.575**	**0.556**	**0.568**	**0.551**	**0.567* (6.97E-4)**

Numbers in brackets indicate the p-value from a paired t-test. Bold faced numbers denote improvements over the baseline, and * indicates that the entry is statistically significant from the baseline at 95% confidence level.

Table 1: Experimental Results of Different Methods.

optimization problem.

$$\min_{\mathbf{w}} V(\mathbf{w},\xi) = \frac{1}{2}\langle \mathbf{w},\mathbf{w}\rangle + C\sum \xi_{i,j,k}$$
$$\text{s.t.} \begin{cases} \forall(\mathbf{d_i},\mathbf{d_j}) \in Y_1 : \langle \mathbf{w},\mathbf{d_i}\rangle \geq \langle \mathbf{w},\mathbf{d_j}\rangle + 1 - \xi_{i,j,1} \\ \cdots \\ \forall(\mathbf{d_i},\mathbf{d_j}) \in Y_n : \langle \mathbf{w},\mathbf{d_i}\rangle \geq \langle \mathbf{w},\mathbf{d_j}\rangle + 1 - \xi_{i,j,n} \\ \forall i \forall j \forall k : \xi_{i,j,k} \geq 0, \end{cases} \tag{3}$$

where $\mathbf{w}$ is a learned weight vector, C is the trade-off parameter, $\xi_{i,j,k}$ is a slack variable, and Y_k is a set of pairs of financial reports within a year.

4 Experiments and Analysis

In this paper, the 10-K Corpus (Kogan et al., 2009) is used to conduct the experiments; only Section 7 "management's discussion and analysis of financial conditions and results of operations" (MD&A) is included in the experiments since typically Section 7 contains the most important forward-looking statements. In the experiments, all documents were stemmed by the Porter stemmer, and the documents in each year are indexed separately. In addition to the reports, the twelve months after the report volatility for each company can be calculated by Equation (1), where the price return series can be obtained from the Center for Research in Security Prices (CRSP) US Stocks Database. The company in each year is then classified into 5 risk levels ($\ell = 2$) via Equation (2). For regression, linear kernel is adopted with $\epsilon = 0.1$ and the trade-off C is set to the default choice of SVM^{light}, which are the similar settings of (Kogan et al., 2009). For ranking, linear kernel is adopted with $C = 1$, all other parameters are left for the default values of SVM^{Rank}.

Table 1 tabulates the experimental results, in which all reports from the five-year period preceding the test year are used as the training data (we denote the training data from the n-year period preceding the test year as $\mathbf{T}^n$ hereafter). For example, the reports from year 1996 to 2000 constitute a training data $\mathbf{T}^5$, and the resulting model is tested on the reports of year 2001. As shown in the table, with the feature of TF-IDF, our results are significantly better than those of the baseline in terms of both two measures. In addition to using $\mathbf{T}^5$ as the training data, we also conduct other 4 sets of experiments with $\mathbf{T}^1, \mathbf{T}^2, \mathbf{T}^3, \mathbf{T}^4$ to test the reports from

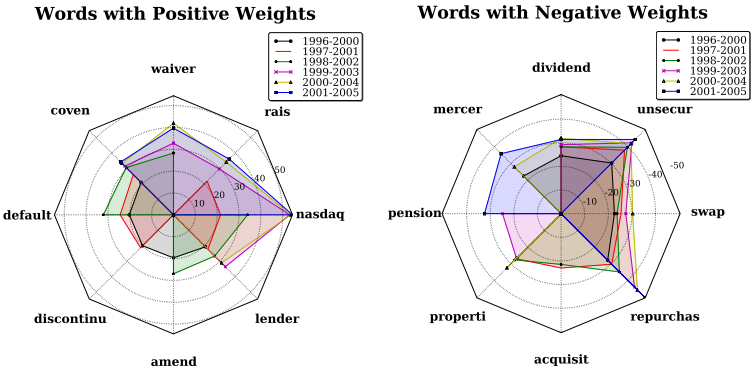

Figure 1: Positive and Negative Weighted Terms Across Different Models.

year 2001 to 2006; [1] there are in total 30 testing instances including the experiments with $\mathbf{T}^5$. The results show that in terms of both measures, our results with TF-IDF are significantly better than the baseline.[2]

Figure 1 illustrates the top positive and negative weighted terms appearing more than twice in the six $\mathbf{T}^5$ models trained on TF-IDF; these terms (8 positive and 8 negative) constitute the radar chart in Figure 1. Almost all the terms found by our ranking approach are financially meaningful; in addition, some of highly risk-correlated terms are not even reported in (Kogan et al., 2009).

We now take the term *defaut* (only identified by our ranking approach) as an example. In finance, a company "defaults" when it cannot meet its legal obligations according to the debt contract; as a result, the term "default" is intuitively associated with a relative high risk level. One piece of the paragraph quoted from the original report (from AFC Enterprises, Inc.) is listed as follows:

> *As of December 25, 2005, approximately $3.0 million was borrowed under this program, of which we were contingently liable for approximately $0.7 million in the event of* **default**.

Conclusion

This paper adopts learning-to-rank techniques to rank the companies to keep them in line with their relative risk levels via the text information in their financial reports. The experimental results suggest interesting relations between the text information in financial reports and the risk levels among companies; these findings may be of great value for providing us more insight and understanding into financial reports. Finally, we provide a visualization interface to demonstrate the relations between financial risk and text information in the reports. This demonstration enables users to easily obtain useful information from a number of financial reports.

[1] Due to the page limits, some of the results are not listed in the paper, but they are available from the authors upon request.
[2] The *p*-value from a paired *t*-test for Spearman's Rho is 1.21E-4 and for Kendall's Tau is 7.27E-5.

Future directions include how to reduce the noise within texts, and how to incorporate Standard Industrial Classification (SIC) into our ranking approach. In addition, a hybrid model consisting of both financial and text information may be also one of our future directions.

References

Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., and Hullender, G. (2005). Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, ICML '05, pages 89–96. ACM.

Freund, Y., Iyer, R., Schapire, R. E., and Singer, Y. (2003). An efficient boosting algorithm for combining preferences. *J. Mach. Learn. Res.*, 4:933–969.

Joachims, T. (2006). Training linear svms in linear time. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, pages 217–226. ACM.

Kendall, M. (1938). A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Kogan, S., Levin, D., Routledge, B., Sagi, J., and Smith, N. (2009). Predicting risk from financial reports with regression. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 272–280. ACL.

Leidner, J. L. and Schilder, F. (2010). Hunting for the black swan: risk mining from text. In *Proceedings of the ACL 2010 System Demonstrations*, ACLDemos '10, pages 54–59. ACL.

Lin, M.-C., Lee, A. J. T., Kao, R.-T., and Chen, K.-T. (2008). Stock price movement prediction using representative prototypes of financial reports. *ACM Trans. Manage. Inf. Syst.*, 2(3):19:1–19:18.

Manning, C., Raghavan, P., and Schutze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

Myers, J. and Well, A. (2003). *Research design and statistical analysis*, volume 1. Lawrence Erlbaum.

Ponte, J. M. and Croft, W. B. (1998). A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 275–281. ACM.

Robertson, S. E. and Sparck Jones, K. (1988). Relevance weighting of search terms. In *Document retrieval systems*, pages 143–160. Taylor Graham Publishing.

Salton, G., Wong, A., and Yang, C. S. (1975). A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620.

Tsay, R. (2005). *Analysis of financial time series*. Wiley-Interscience.