# Risk Ranking from Financial Reports

**Ming-Feng Tsai**
**Department of Computer Science, National Chengchi University, Taiwan**

**Chuan-Ju Wang**
**Department of Computer Science, Taipei Municipal University of Education, Taiwan**

## Abstract

This paper attempts to use soft information in finance to rank the risk levels of a set of companies. Specifically, we deal with a ranking problem with a collection of financial reports, in which each report is associated with a company. By using text information in the reports, which is so-called the soft information, we apply learning-to-rank techniques to rank a set of companies to keep them in line with their relative risk levels. In our experiments, a collection of financial reports, which are annually published by publicly-traded companies, is employed to evaluate our ranking approach; moreover, a regression-based approach is also carried out for comparison. The experimental results show that our ranking approach not only significantly outperforms the regression-based one, but identifies some interesting relations between financial terms.

## Introduction

Information retrieval (IR) is the activity of obtaining information relevant to an information need from a collection of information sources. Due to the prevalence of IR techniques, in recent years a large amount of research has started to focus on the retrieved information for different domains, such as analyzing information in financial reports. In finance, soft information usually refers to text, including opinions, ideas, and market commentary, whereas hard information is always recorded as numbers, such financial measures in finance reports. This paper attempts to use soft information to rank the risk levels of a set of companies.

Financial risk is the amount of chance that a chosen investment instrument (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 soft information in financial reports as factors to rank the risk level of a set of companies in terms of their stock return volatilities.

We consider such a problem to be a text ranking problem: Given a collection of texts, the goal is to rank entities associated with the texts according to a real-world quantity. In this study, the texts are SEC-mandated financial reports; the quantity is the volatility of stock returns. In specific, we split the volatilities of company stock returns within a year into different risk levels, which can be considered as the relative difference of risk among the companies; after the splitting, we then use the financial reports to rank the companies in an attempt to keep them in line with their relative risk levels.

Considering the prevalence of learning-to-rank techniques, this paper attempts to use learning-to-rank techniques to deal with this problem. Unlike the previous study, 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 the exact values of volatilities should be difficult than predicting the ranks among the values. The difficulty is due to the huge amount of noise within texts and the weak connection between the predicted quantities and texts. Therefore, we turn to model the relative risk levels of the companies. Our experimental results show that in terms of two different ranking correlation metrics, our ranking approach both significantly outperforms the regression-based method with a confidence level over 95%.

## 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: $R_t = S_t/S_{t-1}$. The volatility of returns for a stock from time $t-n$ to *t* can be defined as follows:

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

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. Therefore, given a volatility *v*, we derive the risk level *r* via:

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

where $a = m - s(k+1)$ when $k \in \{1, \cdots, \ell-1\}$, $a = -\infty$ when $k = \ell$, $b = m + s(k+1)$ when $k \in \{1, \cdots, \ell-1\}$, $b = \infty$ when $k = \ell$, and *u* is a positive real number. For example, with $\ell = 2$ and $u = 1$, there are 5 risk levels (i.e., 0,1,2,3,4):

$$r = \begin{cases} 0 & \text{if } \ln(v) \in (-\infty, m-2s], \\ 1 & \text{if } \ln(v) \in (m-2s, m-s], \\ 2 & \text{if } \ln(v) \in (m-s, m+s), \\ 3 & \text{if } \ln(v) \in [m+s, m+2s), \\ 4 & \text{if } \ln(v) \in [m+2s, \infty). \end{cases}$$

Note that *r* stands for the concept of *relative* risk among *n* stocks; for instance, a stock with $r = 4$ is much more risky than another one with $r = 0$.

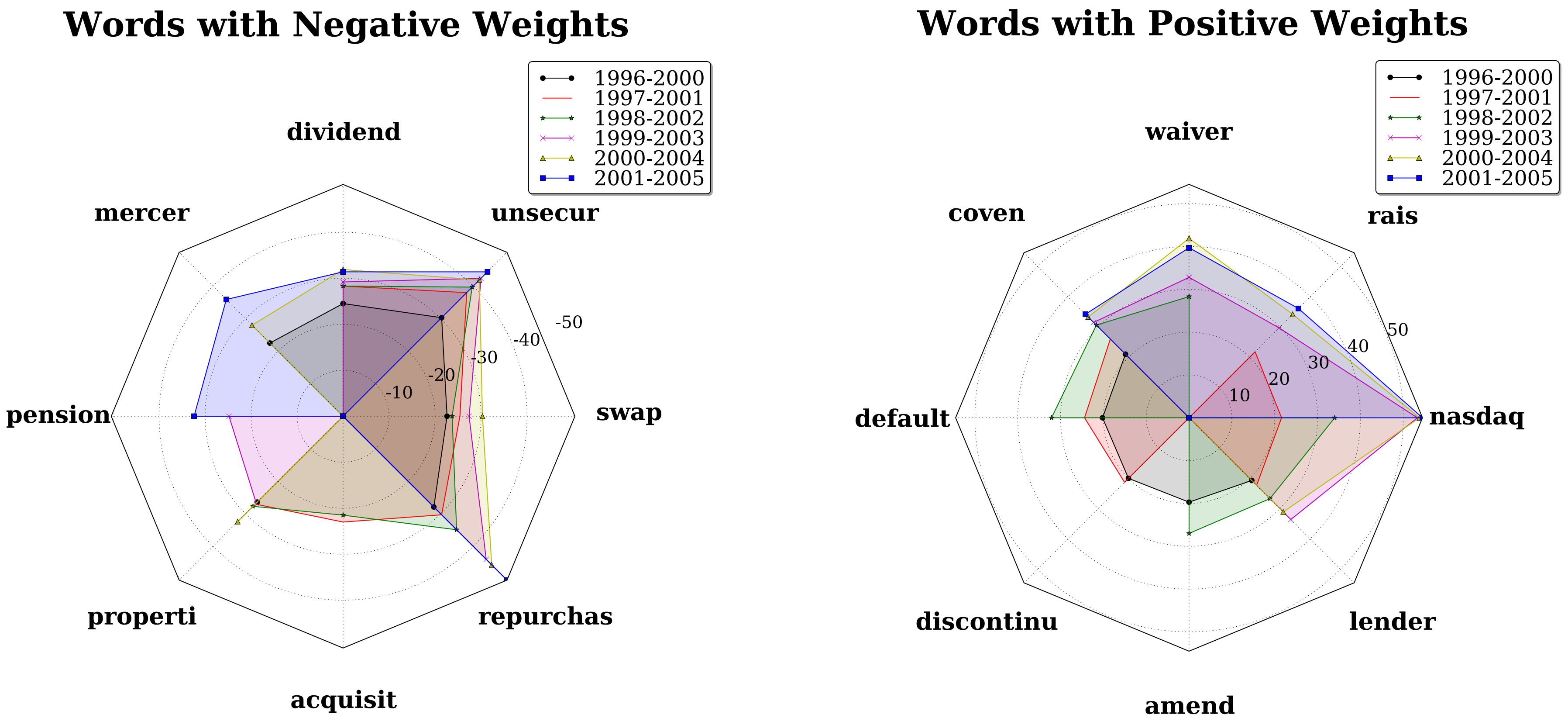
After classifying the volatilities of stock returns (of companies) into different risk levels, we formulate our text ranking problem as follows: Given a collection of financial reports $D = \{\boldsymbol{d}_1, \boldsymbol{d}_2, \boldsymbol{d}_3, \cdots, \boldsymbol{d}_n\}$, in which each $\boldsymbol{d}_i \in \boldsymbol{R}^d$ and is associated with a company $c_i$, we aim to rank the companies via a ranking model $f : \boldsymbol{R}^d \rightarrow \boldsymbol{R}$ such that the rank order of the set of companies is specified by the real value that the model *f* takes. In specific, $f(\boldsymbol{d}_i) > f(\boldsymbol{d}_j)$ is taken to mean that the model asserts that $c_i > c_j$, where $c_i > c_j$ means that $c_i$ is ranked higher than $c_j$; that is, the company $c_i$ is more risky than $c_j$. In this paper, we adopt Ranking SVM for our text ranking problem.

## Experiments and Analysis

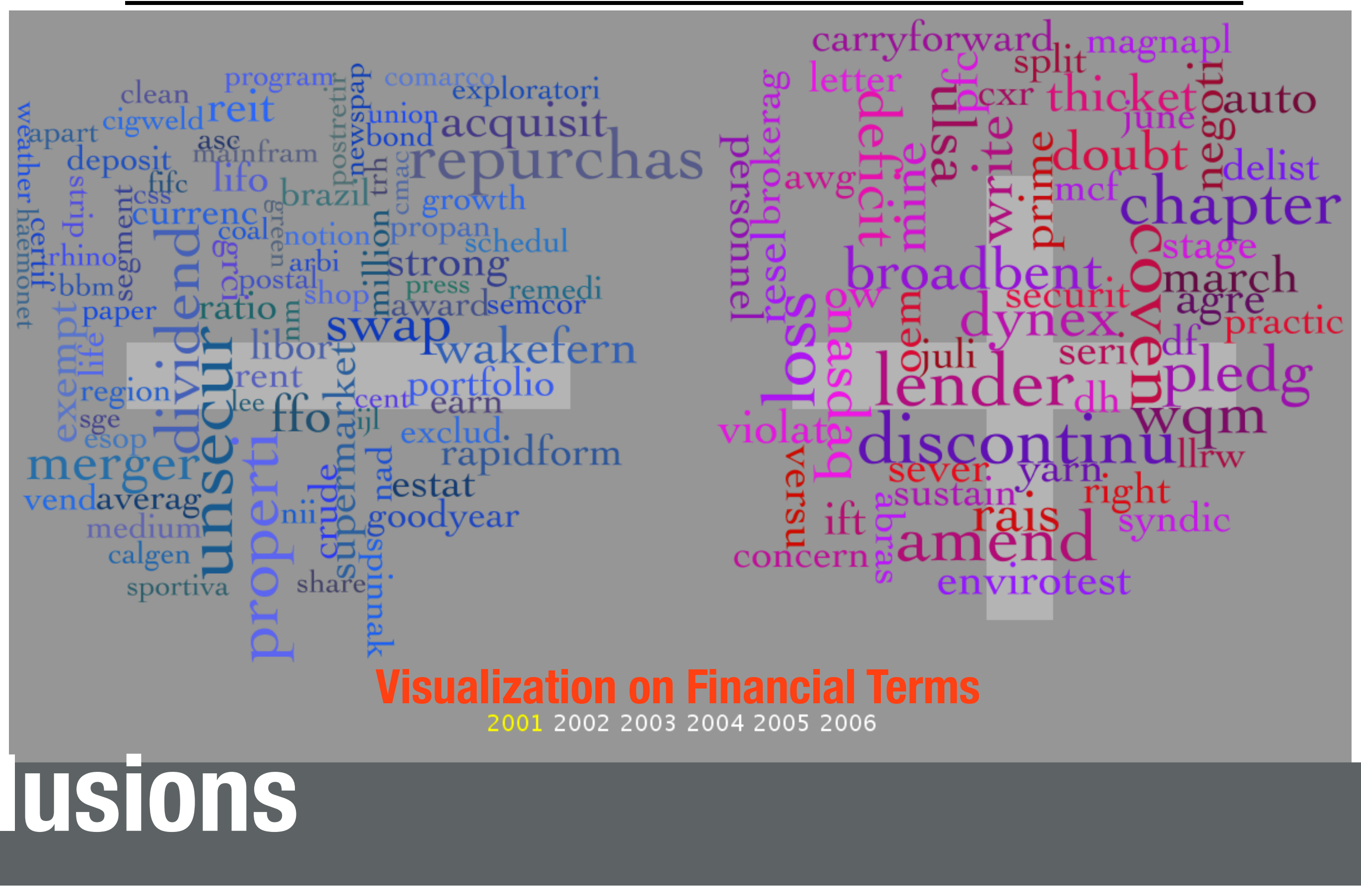
This paper uses the 10-K Corpus to conduct the experiments; only Section 7 "management's discussion and analysis of financial conditions and results of operations" (MD&A) is included in the experiments since typically Section 7 contains the most important forward-looking statements.

Table 1 tabulates the experimental results, in which all reports from the five-year period preceding the testing year are used as the training data (we denote the training data from the *n*-year period preceding the testing year as $\boldsymbol{T}^n$ hereafter). For example, the reports from year 1996 to 2000 constitute a training data $\boldsymbol{T}^5$, and the trained model is tested on the reports of year 2001. Since the goal is to rank the risk levels by using text information, only word features are considered; in our experiments, the unigram TF and TF-IDF features are used.

In addition, via the trained models, we also observe some interesting phenomena: for instance, some financial terms usually appear together and indicate high-risk levels. In our dataset, the term "default" usually co-occurs with the term "debt." In finance, a company defaults when it cannot meet its legal obligations according to the debt contract; as a result, the terms of "default" and "debt" are usually associated with a relative high-risk level.


**Words with Negative Weights** **Words with Positive Weights**

| Method | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | Average |
|---|---|---|---|---|---|---|---|
| **Feature: TF** | | | | **Kendall's Tau** | | | |
| SVR (baseline) | 0.522 | 0.516 | 0.523 | 0.496 | 0.500 | 0.474 | 0.505 |
| Ranking SVM | 0.508 | 0.504 | 0.519 | **0.498** | **0.515** | **0.483** | 0.505 (0.460) |
| **Feature: TF** | | | | **Spearman's Rho** | | | |
| SVR (baseline) | 0.553 | 0.548 | 0.555 | 0.525 | 0.528 | 0.500 | 0.535 |
| Ranking SVM | 0.540 | 0.536 | 0.551 | **0.527** | **0.544** | **0.509** | 0.535 (0.474) |
| **Feature: TFIDF** | | | | **Kendall's Tau [2]** | | | |
| 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 [4]** | | | |
| 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)** |



## Conclusions

This paper uses the soft information in financial reports to rank the risk levels of a set of companies. Specifically, we tackles a ranking problem with a collection of financial reports and apply learning-to-rank techniques to rank the companies to keep them in line with their relative risk levels specified by their stock return volatilities. Our experimental results show that our ranking approach significantly outperforms the regression-based method with a confidence level over 95% in terms of two different metrics. Future directions include how to incorporate Standard Industrial Classification (SIC) into our ranking approach and how to develop a hybrid model consisting of both soft and hard information in finance for improving the ranking performance.

ECIR 2013