JOURNAL OF EMERGING TECHNOLOGIES IN ACCOUNTING
Volume 8
2011
pp. 31–44

American Accounting Association
DOI: 10.2308/jeta-10182

Extracting Key Phrases as Predictors of Corporate Bankruptcy: Empirical Analysis of Annual Reports by Text Mining

Cindy Yoshiko Shirata
University of Tsukuba

Hironori Takeuchi
Shiho Ogino
Hideo Watanabe
IBM Japan

ABSTRACT: Bankruptcy predictions have been one of the most interesting topics for accounting researchers. Most bankruptcy prediction models are developed by using financial ratios. However, signs of the changing financial position of a company may appear in the nonfinancial information earlier than we can identify the changes in the financial numbers. In recent years, analysis of qualitative information has become remarkably important, because frequent changes in accounting standards have made it difficult to compare financial numbers between years. In this study, we analyzed the sentences in financial reports in Japan and extracted key phrases/descriptions to predict bankruptcy. Our research revealed that if some particular expressions appear together with the word "dividend" or "retained earnings" in the same section of an annual report, they were effective in distinguishing between bankrupt companies and non-bankrupt companies.

Keywords: bankruptcy prediction; key phrases; nonfinancial information; text mining; Japanese annual reports.

INTRODUCTION

Bankruptcy predictions have been one of the most interesting accounting research topics. The number of bankruptcy prediction models boomed after Altman (1968) announced the Z model. Most bankruptcy prediction models are developed by using financial ratios. However, frequent updating of Accounting Standards in recent years has made it very difficult to compare financial numbers between fiscal years or between companies, creating confusion and

The paper has benefited from comments from Professor Ingrid E. Fisher, University at Albany, SUNY; Professor Brigitte W. Muehlmann, Suffolk University; Professor Miklos Vasarhelyi, Rutgers Business School; and participants of the 2009 AAA Annual Meeting. The authors also express great appreciation to Daniel O'Leary (editor) and three anonymous reviewers at the *Journal of Emerging Technology in Accounting*, whose comments improved the paper significantly.

Corresponding author: Cindy Yoshiko Shirata
Email: shirata.cindy.fe@u.tsukuba.ac.jp
Published Online: December 2011

difficulties for information users. Although Japanese Accounting Standards have not yet been completely adapted to International Financial Reporting Standards (IFRS), they are moving closer to these standards year by year. Under these circumstances, in recent years, analysis of qualitative nonfinancial information has become remarkably important, and since March 31, 2004, the following sections have been required in annual reports in Japan: Uncertain Risk Information, Management's Discussion and Analysis, and Information Related to Corporate Governance. Because of the requirement to provide this information, new research methods for evaluating a company's position, such as text mining, have become popular.

Signs of the changing financial position of companies may appear in the nonfinancial information earlier than we can identify the changes in the financial numbers. Nonfinancial variables such as the intellectual capital index can be included in bankruptcy prediction models, but such nonfinancial variables are difficult to use because they do not have enough comparability. In this paper, we demonstrate a way to differentiate between companies that became bankrupt and those that did not by paying attention to the nonfinancial information. However, rather than analyzing nonfinancial variables, we analyzed the textual data in the financial reports. Annual reports contain substantial supplementary information in narrative form, and after analyzing the sentences in this supplementary information, we were able to extract key phrases/descriptions to predict bankruptcy.

In Japan, the contents of annual reports (10-K in Japan) of all listed companies must be public and follow the same format. Text data included in annual reports are often written in highly sophisticated expressions that not only reflect language characteristics, but also reveal sensitive financial information. People who are not familiar with financial information have difficulty understanding the full meaning of financial reports because of these highly sophisticated, domain-dependent expressions. Recently, the rate of foreign investors in the Japanese stock market has been rising, up to 23.6 percent in 2008,[1] and even when they are financial experts, most of these investors have difficulty reading annual reports written in Japanese and in interpreting the text in order to make investment decisions. Signs of bankruptcy in the text of an annual report are difficult to find, and translation of the text is not a practical solution because of its high cost.

To solve this problem, we focused on extracting knowledge from Japanese public annual reports that is useful for distinguishing between bankrupt companies and non-bankrupt companies. This extracted knowledge can be useful not only to financial experts, but to all parties who are interested in understanding Japanese financial reports. In this paper, we present the results of our experiments in using conditional probability and contextual information to extract key phrases specific to bankruptcy and to non-bankruptcy from the public annual reports of 180 Japanese companies—90 bankrupt companies and 90 non-bankrupt companies.

TEXT MINING LITERATURE

The literature includes several articles that discuss the analysis of a company's financial position using text mining. Clatworthy and Jones (2006) examined a range of textual characteristics in the chairmen's statements of 50 extremely profitable and 50 extremely unprofitable U.K.-listed companies. They found that chairmen's statements are subject to impression management techniques, as managers' propensity to associate themselves with their company's financial results is related to the firm's underlying financial performance. They also found that unprofitable companies focus more on the future than on past performance. Clatworthy and Jones (2003) focused on chairmen's narratives in the top 50 and bottom 50 U.K.-listed companies ranked by

[1] Tokyo Stock Exchange market announces updated statistical information, including the ratio of foreign investors (http://www.tse.or.jp/english/index.html).

percentage change in net profit before tax. They examined whether companies report good news and bad news differently when their performance is improving and when it is declining. The noteworthy finding in this research is that both groups prefer to take credit for good news themselves, while blaming the external environment for bad news. Bryan (1997) analyzed Management Discussion and Analysis (MD&A) disclosure information and found that future operation and planned capital expenditure explanations are associated with short-term performance. However, MD&A is not mandatory yet in Japan.

In other research, Abrahamson and Amir (1996), using the content-analysis method to qualify the information in the president's letter to the shareholders, calculated a numerical measure of the negativity in the president's letter. Their results indicate that the negativity of the president's letter is associated with performance measures based on financial information, in particular the regression coefficient of earnings levels. Kloptchenko et al. (2004) combined data mining methods to analyze quantitative and qualitative data from financial reports in order to see whether the textual part of the report contains some indication of future financial performance. The quantitative part of a report, they pointed out, only reflects the past performance of the company, while the qualitative part holds a message about the company's future performance. They concluded that the tone of a written report tends to change some time before the actual financial changes occur. Magnusson et al. (2005) also concluded that a change in the textual data usually indicates a change in the financial data in the following quarter. Textual data in the quarterly report can help to predict financial performance for the future.

Some research verifies that qualitative information can predict companies' going concern status. Cormier et al. (1995), who developed a model that also includes qualitative corporate governance characteristics suggested by current audit practice, showed that some qualitative variables provide consistent signals about going concern failures. According to Kleinman and Anandarajan (1999), the fact that qualitative factors have power to predict a going concern report suggests that companies can evaluate other companies even if the auditor, for political or other reasons, has chosen not to render a qualified going concern report. Ponemon and Schick (1991) also concluded that nonfinancial characteristics associated with organizational decline could be used by auditors to corroborate judgments regarding the financial condition of client companies. Back (2005) shows that it is possible to explain financial difficulties in small and medium-sized firms based on nonfinancial variables. Smith and Taffler (2000), who explored the relationship between a firm's narrative disclosures and bankruptcy using a content analysis approach, provide evidence that the chairman's statement alone is highly associated with the event of firm failure, reinforcing the argument that such un-audited narrative disclosures contain important information associated with the future of companies and are not just reporting on past performance.

Shirata and Sakagami (2009) analyzed the text data in the annual reports of 21 bankrupt Japanese companies and 24 non-bankrupt Japanese companies. Using morphological analysis, they extracted keywords to discriminate between the two groups. They found that the Dividend Section of the annual report contained a unique explanation of the company's financial position. The results of their analysis reveal that terms such as "dividends," "profit appropriation," and "retained earnings" are among those with prominent differences in appearance frequencies between the two groups. In particular, clear disparities between bankrupt and non-bankrupt groups were found in the frequency of appearance of such terms as "dividends" and "dividend propensity." Their evidence supports the conclusion that continuing entities consistently incorporate messages in their financial reports that are directed at shareholders. However, it is difficult to predict the bankruptcy of a particular company merely by counting the word frequencies of something like "dividends" and "retained earnings," because all companies use the words "dividends" and "retained earnings." In this paper, we explore text mining, a method that is more reliable than morphological analysis to predict bankruptcy.

TEXT MINING OF CORPORATE ANNUAL REPORTS

Issues Related to Previous Studies

In most previous research using text mining for corporate evaluation, researchers analyzed text data based on word frequency calculated by morphologically analyzed text. However, single extracted words or phrases might not reveal important information that may be captured by looking at word-to-word dependencies and the contexts around high-frequency words. When Shirata and Sakagami (2009) extracted high-frequency words in the annual reports of a group of bankrupt companies and a group of non-bankrupt companies, they found that the occurrence of "dividend" was totally different between the two groups. However, calculating only the word frequency is not enough to specify whether a company will go bankrupt, because the word "dividend" can be used in many different contexts, for example "paid a dividend" or "could not pay a dividend." This observation revealed some issues with previous text analyses. The texts of corporate annual reports were usually syntactically correct and not difficult to analyze morphologically; however, the sentences were so long that it was often difficult to analyze word-to-word dependency relationships. In addition, it was often difficult to analyze how the words were used in the contexts of these reports because the Japanese text often contained indirect and sophisticated expressions. Therefore, the word-frequency information did not extract sufficient knowledge from the text. The key issue is how to extract more detailed information for corporate evaluation using text analysis.

Sample Data

In order to extract specifically Japanese explanations that can discriminate between bankrupt companies and non-bankrupt companies, we chose the annual reports of 90 companies that went into bankruptcy between 1999 and 2005, and the annual reports of 90 companies that did not go into bankruptcy. For the bankrupt companies, we collected annual reports from one year prior to their bankruptcy. The non-bankrupt companies were extracted from among all companies listed on the Tokyo Stock Exchange market during the same period as the bankrupt companies.

One way to compare bankrupt companies and non-bankrupt companies is to use all listed firms as non-bankrupt companies. However, data text files for these annual reports were not available in Japan, and it would have taken much time to retype all the textual data. Therefore, we reduced the number of sample firms to save time. In order to give the smaller sample the same financial position distribution as all listed firms in Japan, we first used the SAF (Simple Analysis of Failure) model[2] to rank the companies (Shirata 2003).

The SAF2002 model was developed by analyzing the financial data of 1,436 bankrupt companies and 3,434 non-bankrupt companies extracted by a systematic sampling method from 107,034 companies. The variables for the SAF2002 model were selected by using a Classification and Regression Tree (CART) to analyze the financial data of companies that entered bankruptcy between 1992 and 2001 in Japan. Shirata (1999) had concluded, after comparing the accuracy of variables selected by conventional statistical methods and those selected by CART, that the variables selected by CART have a stronger discriminant power than the variables selected by conventional statistical methods. The prediction power of variables selected by CART was also compared with variables in the previous literature (Altman 1968; Deakin 1972; Blum 1974; Altman et al. 1977; Ohlson 1980; Palepu 1986; Lennox 1999). The best set of variables for the SAF2002

[2] The SAF2002 model is the most powerful and popular bankruptcy prediction model in Japan. It has been used by many financial institutions and rating companies for credit control, and many individual investor Blog sites refer to the SAF model. The score calculated by the model is used as the corporate valuation index for the research paper by the Ministry of Finance.

TABLE 1

Variables for SAF2002 Model

Var.	Classification Value	Partial R^2	F-value	Classification Value
X7	Retained Earnings to Total Liabilities and Owners' Equity	0.1671	830.00	$<$8.86175
X10	Net Income Before Tax to Total Liabilities and Owners' Equity	0.0114	47.56	$<$0.5857
X37	Inventory Turnover Period	0.0593	260.73	$>$2.00055
X26	Interest Expenses to Sales	0.0186	78.31	$>$1.05925

model was analyzed, and the final indices for the model are shown in Table 1. When the model was built, different types of models with selected variables were compared for their discriminant power: a linear model, a logistic model, a quadratic function model, and a normal kernel model. The linear model was found to be the most stable and to produce less miss-classification than other models. In addition, the liner model was easy to handle, with strong discriminate power in rating the companies. The model equation is presented as Formula 1:

$$SAF\ \ Value = 0.01036X7 + 0.02682X10 - 0.06610X37 - 0.02368X26 + 0.70773. \qquad \text{Formula 1}$$

A company's bankruptcy risk is measured by inserting the indices shown in Table 1 into each variable in Formula 1 and computing the SAF value: an SAF value of 0.7 and below quickly raises the bankruptcy risk.

After calculating the SAF values of all listed firms, we ranked all companies from the highest to the lowest and performed systematic extraction at equal intervals so that the total came to 90. The distribution of the 90 non-bankrupt companies that were selected was assumed to be similar to the distribution of all the companies in the real stock exchange market.

Corporate annual reports contain various parts such as financial statements, a company profile, a list of directors, and the auditor's opinion. For this study, we analyzed the Dividend Policy section in accordance with the finding of Shirata and Sakagami (2009) that the Dividend Policy section of the annual report contained the specific explanation of the company's financial position.

Methodology

Text mining technology for the Japanese language is still in a developmental stage. Fundamentally, text mining involves applying data mining techniques to text data. If we compare text mining to the data mining process used by Fayyad et al. (1996), we see major differences from ordinary data mining: the data gathered in "data acquisition" are text data rather than quantitative data; manipulation is performed in many cases by using such word classes as nouns, adjectives, verbs, and adverbs as clues for "data selection"; and in order to subject the text to the data mining process, text data have to be converted to some sort of quantitative indicators in "data conversion." Text mining technology is also called natural language processing or statistical natural language processing (Manning and Schütze 1999).

Natural language processing consists of two steps: (1) analyzing the context of sentences using morphological analysis or systematic analysis, and (2) counting the frequencies of selected important phrases/words in each report. Then the differences in frequencies between the two groups, bankrupt companies and non-bankrupt companies, are calculated. We also looked at the effect of the location of the phrases/words within the sentence.

To extract specific phrases to discriminate between the bankrupt group and non-bankrupt group, we analyzed the context of sentences using (1) morphological analysis and (2) the conditional probability method. We used IBM OmniFind Analytic Edition (OAE) because of the lack of other tools for analyzing Japanese text data using conditional probability (Yoshida and Takuma 2007; Zhu et al. 2008).

Following is an explanation of the conditional probability method:

If "E" is appearing in the texts, then:

$Na =$ the total number of companies included in the data;
$Nb =$ the number of bankrupt companies;
$Ns =$ the number of non-bankrupt companies;
$Ea =$ the total number of companies that mentioned E;
$Eb =$ the number of bankrupt companies that used E; and
$Es =$ the number of non-bankrupt companies that used E.

The probability of all companies that used E is:

$$\mathrm{P}(E) = Ea/Na.$$

The probability of bankrupt companies that used E is:

$$\mathrm{P}(Bankrupt, E) = Eb/Na.$$

The probability of non-bankrupt companies that used E is:

$$\mathrm{P}(Non\text{-}Bankrupt, E) = Es/Na.$$

The conditional probabilities—P($Bankrupt|E$), used to extract the expressions specific to the bankrupt group, and P(Non-$Bankrupt|E$), used to extract the expressions specific to the non-bankrupt group—are calculated as follows:

$$\mathrm{P}(Bankrupt|E) = \mathrm{P}(Bankrupt, E)/\mathrm{P}(E),$$

$$\mathrm{P}(Non\text{-}Bankrupt|E) = \mathrm{P}(Non\text{-}Bankrupt, E)/\mathrm{P}(E).$$

If ($Bankrupt|E$) is reasonably large compared to the probability of the bankrupt group P($Bankrupt$), then E can be regarded as an expression specific to the bankrupt group. In the same way, if (Non-$Bankrupt|E$) is reasonably large in comparison to P(Non-$Bankrupt$), then E can be regarded as an expression specific to the non-bankrupt group.

In order to observe whether the conditional probability tool was effective in extracting expressions specific to the bankrupt group and the non-bankrupt group, we also used the following metrics in our experiment:

(1) the differences between the probabilities of the companies that used E within the bankrupt group and within the non-bankrupt group,
(2) the Kullback-Leibler distance (Kullback and Leibler 1951), and
(3) $\chi 2$ where the theoretical probability was assumed to be P(E).

These metrics were used in the document categorization to identify the expressions that were specific to a certain document collection and that contributed to the categorization performance (Yang and Pedersen 1997).

Extracting Representative Key Phrases

Simple syntactic word-to-word dependency patterns are often insufficient for mining long sentences such as those in annual reports, because a word and its modifier—in other words a governor and its dependent—often appear a long distance from each other in such documents. To overcome this insufficiency, we prepared a list of technical terms to focus on, and then extracted co-occurring words as candidates for the words distinctive to each corporate group within a phrase governed by the technical terms, in other words within the region of the topic word. The technical term list, which follows, is based on previous research (Shirata and Sakagami 2009).

Technical terms:

dividend
retained earnings
stockholder
revenue
fund

These technical terms are often used as topic words within a certain context, for instance "As for dividends" and "To talk about retained earnings"; therefore, it is important for our analysis to focus on words that co-occur with these technical terms. In such contexts, detailed information about the topic words is often found in phrases directly or indirectly governed by the topic words. For instance, the following topic word/information pairs can be extracted from the following sentence: "As for retained earnings this year, we are planning to use it for effective investment for future business development."

(1) retained earnings . . . future;
(2) retained earnings . . . business development;
(3) retained earnings . . . effective investment.

We extracted detailed information about topic words by using the method described above. Based on this detailed information on topic words, we investigated in what contexts these topic words were specific to bankruptcy and non-bankruptcy groups.

Experiments Using Word-Frequency Statistical Metrics

Tables 2 and 3 show the top 10 words specific to the bankrupt group and to the non-bankrupt group, respectively, according to the metric described earlier. The amount (Yen) and number of shares in the tables represent "digits + monetary units" and "digits + shares," respectively. These word sequences were matched and merged into the representative forms, [amount in Yen] and [number of shares], in preparing the information for the OAE.

In extracting descriptions specific to bankrupt companies, we were unable to extract clear differences in the descriptions between bankrupt and non-bankrupt companies using conditional probability. All of the retained earnings of bankrupt companies had already been lost one year prior to their bankruptcy; therefore, they had finished the fiscal year without any dividends. The Dividend Policy sections in the annual reports usually consisted of four or five sentences, and there were very few ways to select the explanation indicating "no dividend." Sentences expressing the lack of a dividend tended to be quite short and similar to each other, so the results from extracting these expressions using different metrics did not differ significantly.

On the other hand, in extracting expressions specific to non-bankrupt companies, we did find many differences by using four different metrics. Keywords such as "dividend," "execute," or "dividend policy" were used more frequently by non-bankrupt companies than by bankrupt

TABLE 2

Top 10 Content Words Specific to Non-Bankrupt Group (Verb, Adjective, Noun)

Difference of Occurrence Probability			Distance between Each Probability			Chi-Square Test			Conditional Probability		
	N	B		N	B		N	B		N	B
[# of stocks]	79	14	[# of stocks]	79	14	[amount of money]	79	25	throughout year	28	0
[amount of money]	79	25	interim dividend	50	4	[# of stocks]	79	14	research and development	10	0
interim dividend	50	4	throughout year	28	0	interim dividend	50	4	allot	17	1
executive board	41	4	[amount of money]	79	25	executive board	41	4	consolidated result	8	0
decision	41	5	executive board	41	4	decision	41	5	middle	15	1
execute	53	21	decision	41	5	throughout year	28	0	corporate value	14	1
throughout year	28	0	allot	17	1	execute	53	21	acquisition	7	0
dividend	48	21	middle	15	1	dividend policy	89	69	stock repurchase	7	0
fund	33	8	fund	33	8	fund	33	8	hold	7	0
go	42	18	execute	53	21	dividend	48	21	interim dividend	50	4

B: Number of occurrences in documents of bankrupt group.
N: Number of occurrences in documents of non-bankrupt group.

TABLE 3

Top 10 Content Words Specific to Bankrupt Group (Verb, Adjective, Noun)

Difference of Occurrence Probability			Distance between Each Probability			Chi-Square Test			Conditional Probability		
	N	**B**		**N**	**B**		**N**	**B**		**N**	**B**
no dividend	5	46	no dividend	5	46	no dividend	5	46	regret	1	32
regret	1	32	regret	1	32	regret	1	32	continue	0	10
intend	17	35	intend	17	35	loss	1	17	loss	1	17
loss	1	17	do	17	33	resumption of dividend	2	18	recovery	0	8
resumption of dividend	2	18	environment	12	27	recording	4	20	drastic	0	8
recording	4	20	recording	4	20	net loss	0	18	fully	0	7
do	17	33	severe	6	21	continue	3	10	quick	0	7
severe	6	21	resumption of dividend	2	18	severe	6	21	pass	0	6
net loss	3	18	net loss	3	18	intend	17	35	lead to	1	10
environment	12	27	loss	1	17	dividend	0	8	according to	0	5

B: Number of occurrences in documents of bankrupt group.
N: Number of occurrences in documents of non-bankrupt group.

TABLE 4

Expressions Appearing with "Dividend"

Bankrupt Companies	no dividend, regret, stop, severe
Non-Bankrupt Companies	interim dividend, dividend, year-end dividend, including [number of shares], basic strategy, add up, [amount in Yen], consider, decide, stockholder, turnover, additional, increasing dividend, based on, retained earning

companies, but more than 20 percent of the bankrupt companies also used these expressions. In contrast, although the top 10 expressions extracted by using conditional probability were frequently used by non-bankrupt companies, no bankrupt companies used these words. "Research and development" and "corporate value" were extracted using conditional probability as descriptions specific to the non-bankrupt group. These phrases were rarely used by bankrupt companies. This result leads us to conclude that non-bankrupt companies are continuing to pursue "research and development" and are trying to emphasize their "corporate value" more than companies headed for bankruptcy. Based on these results, we believe that conditional probability was effective in extracting infrequent expressions that are specific to a certain company group.

Experiments with Topic Words and Their Context

Table 4 shows some extracted words/phrases specific to bankrupt companies and non-bankrupt companies that appear together with the topic word "dividend." The use of the topic word "dividend" along with "[number of shares]" and "[amount in Yen]" was specific to non-bankrupt companies only in the context governed by the topic word "dividend," not in the full text of the annual reports.

Expressions about no dividends, for instance, "to our regret, we decided to close this fiscal year without any dividends" or "the severe economic environment does not allow any dividends this year," were used only by bankrupt companies. These expressions, "to our regret" or the indirect phrase "does not allow," could be considered excuses frequently found in Japanese reports. If these excuse phrases appear together with "dividend," the possibility that the company is headed for bankruptcy is high.

Shown in Table 5 are extracted words/phrases specific to bankrupt companies and non-bankrupt companies that occur together with the topic word "retained earnings." The phrases "capital investment," "plant and equipment," and "new business" were used frequently by non-bankrupt companies in the "retained earnings" section of the report, but were not specific to non-bankrupt companies in the full text of the annual reports.

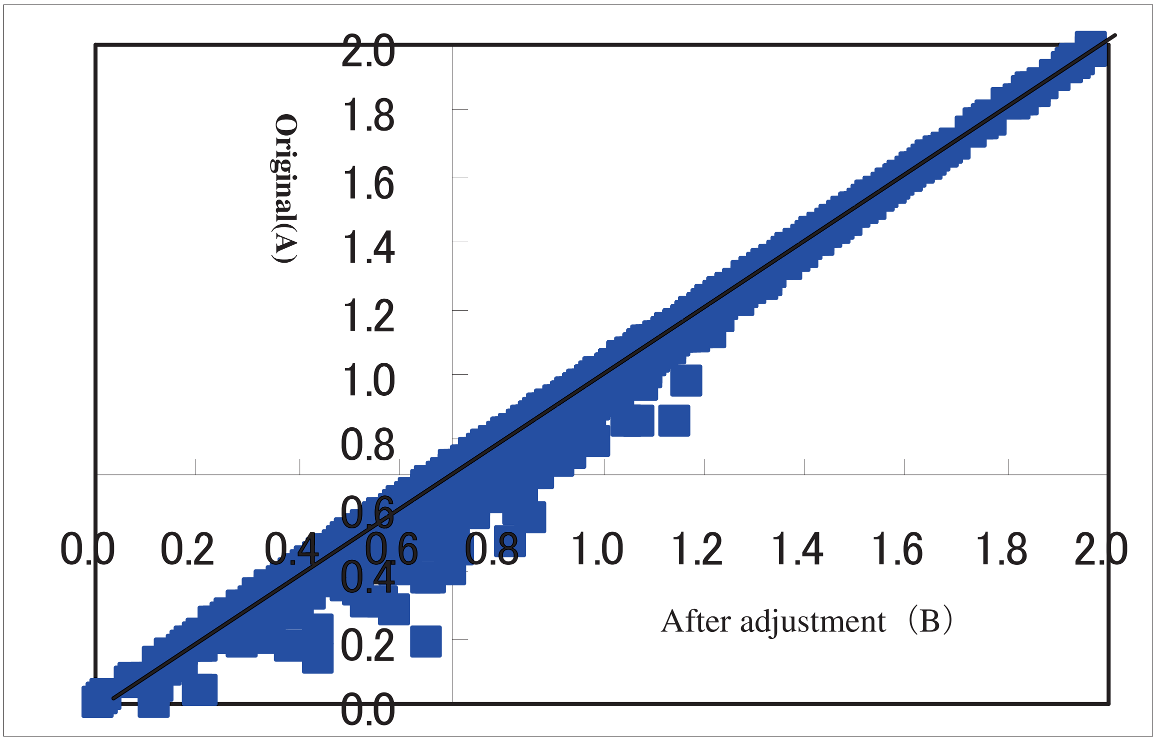
Both "dividend" and "retained earnings" are specific to neither bankrupt companies nor non-bankrupt companies. However, when the words/phrases shown in Table 4 and Table 5 appear

TABLE 5

Expressions Appearing with "Retained Earnings"

Bankrupt Companies	basic, enrich, react, stable, revenue, reimbursement, status
Non-Bankrupt Companies	allot, grow, capital investment, investment in R&D, competency, rationalization, corporate value, plant and equipment, new business, develop, invest, respond, leverage, extend, execute, long-term, improvement, usage, business, future, management environment

FIGURE 1
SAF Value Before and After Adjustment of Accounting Treatment upon Introducing "Deferred Tax Assets" Accounting and Revaluation of Invested Securities and Lands



together with the word "dividend" or/and "retained earnings," they effectively distinguish between bankrupt companies and non-bankrupt companies.

When extracting descriptions that accompanied the topic words, it was also useful to know the context in which the topic words were used. For instance, we could assume that non-bankrupt companies have some retained earnings and therefore would continue to invest in "plant and equipment" (i.e., "capital investment") and the development of "new business." This assumption was supported by our experimental results for extracting descriptions within the "retained earnings" section.

SUMMARY

Japanese accounting principles have undergone dramatic changes in recent years, spurring various discussions on how the changes in accounting standards affect financial analysis. These changes in accounting standards cause difficulty in predicting bankruptcy from a purely financial analysis point of view.

One can confirm that the comparability between the financial figures applied under the new accounting standards and the financial figures of the previous fiscal years is being undermined by the former. In order to verify the extent to which the financial numbers have been influenced by the changes in the accounting standards, we calculated (1) the SAF values without making the adjustments in accordance with the changes and (2) the SAF values after making adjustments based on *revaluation excess* + *deferred tax asset* using the 2005 financial data, and then confirmed their distributions. Figure 1 shows SAF values before (A) and after the (B) adjustment. The amount of retained earnings is influenced by the revaluation excess and the deferred tax asset. If the company had many “Mochiai (cross-holding)” securities, changing the value of these securities influenced the retained earnings of the company, even though the company had not gone through the P/L.

In Figure 1, the intersection of the X- and Y-axes was set to 0.7, which is the cut-off point for the bankruptcy classification. The lower limit was set to zero because of limited display space, even though the SAF values of many companies were negative. Companies plotted below the 45-degree line are companies whose respective SAF values improved as a result of the change in accounting standards. That means they seem to be in a better financial position after introducing the new accounting treatment. A close observation of the results reveals that many companies that experienced strong adjustment effects are located around the bankruptcy line at 0.7. On the other hand, companies with an SAF value of 0.4 or lower before adjustment (rated “BBB” or lower) experienced no substantial adjustment effects, as their retained earnings had already been exhausted. Blue-chip companies with an SAF value of 1.3 or higher (rated “AAA” or higher) may be deemed to have also experienced almost no adjustment effects (either they did not perform the procedures at all or none of the procedures substantially affected the financial figures). These results showed us that introducing new accounting standards sometimes leads information users to misunderstand a company’s financial position.

Sixteen of the 90 non-bankrupt companies in the sample showed less than a 0.7 SAF value. On the other hand, only six companies in the 90 bankrupt sample showed higher than a 0.7 SAF value one year prior to their bankruptcy, and only two companies showed higher than a 0.7 SAF value two years prior to their bankruptcies. The miss-classification rates were 21.1 percent of the non-bankrupt companies and 0.7 percent of the bankrupt companies. For instance, the SAF value of Kotobuki Kogyo, which went into bankruptcy in October 2002, was 0.8007 two years prior to the bankruptcy, and the SAF value of Awamura Seisakusho was 0.83004 two years prior to bankruptcy in September 2004. On the other hand, although both companies’ SAF values showed very stable financial positions, the key phrase “to our regret” appeared in sentences including “dividend” and “retained earnings” two years before their bankruptcies. We also found that another six miss-classified companies expressed the excuses “to our regret” in the Dividend Policy section, and they actually had not paid any dividends even though their SAF value showed more than 0.7. Thus the appearance of particular expressions like “to our regret” in the narrative information can improve the bankruptcy prediction power.

CONCLUSION

In this paper, we report our analysis of sentences in financial reports in Japan and the extraction of key phrases/descriptions to predict bankruptcy. Both Shirata (2003) and Altman et al. (1977) showed that the “retained earnings to total assets ratio” had the strongest power to discriminate between bankrupt companies and non-bankrupt companies, even though these two studies were done in different countries and in different periods. Also, Shirata and Sakagami (2009) revealed that the keywords, “dividend” and “retained earnings,” show prominent differences in appearance frequency between the bankrupt group of companies and the non-bankrupt group. However, because “dividend” and “retained earnings” appear in all companies’ annual reports, it is difficult to predict the bankruptcy of a particular company by only counting the word frequencies of “dividend” and “retained earnings.” Our research revealed that if some particular expressions appear together with the words “dividend” or “retained earnings” in the same section of the annual report, these co-occurring words can be effective in discriminating between bankrupt companies and non-bankrupt companies. We found that for non-bankrupt companies, “research and development,” “capital investment,” and “new business” will appear in sentences including “dividend” and “retained earnings.” This result could indicate that high revenue and enough retained earnings will lead to investment for research and development.

Phrases specific to bankruptcy were highly ranked using each metric we examined. These expressions were descriptions about no dividends, for instance, "to our regret, we decided to close this fiscal year without dividends" or "the severe economic environment does not allow any dividends this year." The expression "to our regret" and the indirect phrase "does not allow," which could both be considered excuses, were found frequently in the annual reports of Japanese companies before their bankruptcy. We conclude that if these expressions appear together with "dividend," the possibility that the company is headed for bankruptcy is high, even if the financial numbers still show a stable position.

REFERENCES

Abrahamson, E., and E. Amir. 1996. The information content of the president's letter to shareholders. *Journal of Business Finance & Accounting* 23 (8): 1157–1182.

Altman, E. I. 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance* 23 (4): 589–609.

Altman, E. I., R. G. Haldeman, and P. Narayanan. 1977. ZETA analysis: A new model to identify bankruptcy risk of corporation. *Journal of Banking and Finance* 1: 29–54.

Back, P. 2005. Explaining financial difficulties based on previous payment behavior, management background variables and financial ratios. *European Accounting Review* 14 (4): 839–868.

Blum, M. 1974. Failing company discriminant analysis. *Journal of Accounting Research* 12 (Spring): 1–25.

Bryan, H. S. 1997. Incremental information content of required disclosures contained in management discussion and analysis. *The Accounting Review* 72 (2): 285–302.

Clatworthy, M., and M. J. Jones. 2003. Financial reporting of good news and bad news: Evidence from accounting narratives. *Accounting and Business Research* 33 (3): 171–185.

Clatworthy, M., and M. J. Jones. 2006. Differential patterns of textual characteristics and company performance in the chairman's statement. *Accounting, Auditing & Accountability Journal* 19 (4): 493–511.

Cormier, D., M. Magnan, and B. Morard. 1995. The auditor's consideration of the going concern assumption: A diagnostic model. *Journal of Accounting, Auditing & Finance* 1 (2): 201–222.

Deakin, E. B. 1972. A discriminant analysis of predictors of business failure. *Journal of Accounting Research* 10 (Spring): 167–179.

Fayyad, U. M., G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. 1996. *Advances in Knowledge Discovery and Data Mining*. Menlo Park, CA: AAAI/MIT Press.

Kleinman, G., and A. Anandarajan. 1999. The usefulness of off-balance-sheet variables as predictors of auditors' going concern opinions: An empirical analysis. *Managerial Auditing Journal* 14 (6): 273–285.

Kloptchenko, A., T. Eklund, B. Back, J. Karlsson, H. Vanharanta, and A. Visa. 2004. Combining data and text mining techniques for analyzing financial reports. *International Journal of Intelligent Systems in Accounting, Finance, and Management* 12 (1): 29–41.

Kullback, S., and R. A. Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics* 22 (1): 79–86.

Lennox, C. 1999. Identifying failing companies: A re-evaluation of the logit, probit and DA approaches. *Journal of Economics and Business* 51: 347–364.

Magnusson, C., A. Arppe, T. Eklund, B. Back, H. Vanharanta, and A. Visa. 2005. The language of quarterly reports as an indicator of change in the company's financial status. *Information and Management* 42 (4): 561–574.

Manning, C. D., and H. Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Ohlson, J. 1980. Financial ratio and the probabilistic prediction of bankruptcy. *Journal of Accounting Research* 18 (Spring): 109–131.

Palepu, K. 1986. Predicting takeover targets. *Journal of Accounting and Economics* 8: 3–35.

Ponemon, L. A., and A. G. Schick. 1991. Financially distressed companies and auditor perceptions of the twelve characteristics of decline: A research note. *Auditing: A Journal of Practice & Theory* 10 (2): 70–83.

Shirata C. Y. 1999. *Predictable Information of Bankruptcy in Japan*. Tokyo, Japan: Chuokeizai-Sha.

Shirata, C. Y. 2003. *Predictors of Bankruptcy after Bubble Economy in Japan: What Can You Learn from Japan Case?* Proceedings of the 15th Asian-Pacific Conference on International Accounting Issues, Bangkok, Thailand.

Shirata, C. Y., and M. Sakagami. 2009. An analysis of the "going-concern assumption": Text mining from Japanese financial reports. *Journal of Emerging Technologies in Accounting* 6: 1–16.

Smith, M., and R. J. Taffler. 2000. The chairman's statement: A content analysis of discretionary narrative disclosures. *Accounting, Auditing & Accountability Journal* 13 (5): 624–647.

Yang, Y., and J. O. Pedersen. 1997. *A Comparative Study on Feature Selection in Text Categorization*. Proceedings of the 14th International Conference on Machine Learning, Nashville, TN.

Yoshida, I., and D. Takuma. 2007. Software architecture for interactive text mining process. *IBM Japan Provision* (52): 71–78.

Zhu, W.-D., S. Chitiveli, K. Cole, S. Harms, and R. Muraleedharan. 2008. Introducing OmniFind Analytics Edition: Customizing for text analytics. Available at: http://www.redbooks.ibm.com/redbooks/pdfs/sg247568.pdf