EXTRACTING CHARACTERISTIC OF BANKRUPT FIRMS
BY TEXT MINING

1 INTRODUCTION

Signs of the changing financial position of the companies must appear in the non-financial information earlier than we can identify the changes in the subtle financial numbers. Therefore, in this study we clarify the difference between continuing companies and bankrupt companies by paying attention to the non-financial information (qualitative information) disclosed in the annual report. We analyze the sentences which were elucidated as the definition information, which is non-financial information disclosed in Financial Reports in Japan (Japan 10-K).

In recent years, analysis of qualitative information has become remarkably important. Text mining, which is an artificial intelligence technique for solving problems, has begun to be used. Text data included in annual financial reports often tell sensitive financial information and are often written in highly sophisticated expression depending on language characteristics. People who are not financial experts have difficulties to read full texts included in financial reports because of these highly sophisticated domain-dependent expressions. Foreign investors also have difficulties in directly referring texts in financial reports for making decision of investment though they are financial experts, because these texts are usually written in local languages; Japanese 10-Ks are usually written in Japanese. Recently, foreign investor rate in Japanese stock market is getting higher, for instance, about 28 percent in 2006. Most of these investors have difficulties in reading Japanese text, and consequently in reading Japanese 10-K texts. Translation of all texts in annual financial reports costs heavily and would not be a practical solution for these difficulties.
To solve this problem stepwise and practically, we focused on extracting knowledge useful for distinguishing between bankruptcies and non-bankruptcies from Japanese 10-Ks. By referring to extracted knowledge, Japanese non-financial experts could get a clue to understanding texts in Japanese financial reports. In this paper, we reported the result of our experiments to extract expressions peculiar to bankruptcy and expressions peculiar to non-bankruptcy from Japanese 10-Ks of 90 bankrupt companies and 90 non-bankrupt companies, by using conditional probability and contextual information.

II PREVIOUS LITERATURE REGARDING TEXT MINING

It is popular to evaluate companies using financial numbers. However, accounting standards in Japan have been updated frequently in recent years, causing information users confusion and difficulty. It has become very difficult to compare financial numbers between different fiscal years or different companies. Under these circumstances, since March 31, 2004, the following sections have been required in annual reports: Uncertain Risk Information, Management’s Discussion and Analysis, and Information related to Corporate Governance. Because of the providing these new information, a new research method, text mining/content analysis, is used to evaluate companies’ condition.

Some researchers have analyzed presidents’ letters or chairmen’s statements to the shareholders [Clatworthy and Jones 2006][Clatworthy and Jones 2003] [Abrahamson and Amir 1996]. [Clatworthy & Jones 2006] examined a range of textual characteristics in the chairman's statements of 100 extremely profitable and 100 extremely unprofitable UK listed companies. They found that chairmen's statements are subject to impression management techniques as
managers' propensity to associate themselves with company’s financial results is associated with 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 listed UK companies ranked by percentage change in net profit before tax. They examined whether companies with improving and declining performance report good news and bad news differently. 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. In other noteworthy research, [Abrahamson and Amir 1996) used the content-analysis method to qualify the information contained in the president’s letter. They calculated a numerical measure of the negativity expressed by management in the president’s letter to the shareholders. Their results indicate that the information in the president’s letter is associated with performance measures based on financial information. They show that the negativity of the president’s letter affects the regression coefficient of earnings levels. [Kloptchenko et al. 2004] combined data mining methods for analyzing 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. They explained that the quantitative part of a report only reflects the past performance of the company. At the same time, the qualitative part of a report holds some 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. [Magunusson 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 predict financial performance for the future because the data expresses managers’ wishes and hopes.
[Antweiler and Frank 2004] studied the effect of messages posted on Yahoo! Finance and Raging Bull\(^1\) about the 45 companies in the Dow Jones Industrial Average and the Dow Jones Internet Index. They found that stock messages help predict market volatility both at daily frequencies and also within the trading day. Similar research has been done in Japan. [Takahashi et al. 2006] studied the effect of messages written by security analysts, headlines in newspapers, and messages from financial information companies. They found that the particular words of these messages influenced stock prices. [Henry 2006], who examined market reaction to verbal components of disclosure in firms’ earnings press releases, found that addition of verbal content and the writing style of earning-press releases results in more accurate predictions of market response.

There are researches that verify whether qualitative information has the prediction power of companies’ going concern. [Cormier et al. 1995] developed a model that also includes qualitative corporate governance characteristics suggested by current audit practice. He pointed out that some qualitative variables provide consistent signals about going concern failures. According to [Kleinman and Anandaraian 1999] the fact that qualitative factors have power in predicting the 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 non-financial characteristics associated with organizational decline could be used by auditors to corroborate judgments regarding the financial condition of client companies. [Back 2005] shows evidence that it is possible to explain

\(^1\) http://ragingbull.quote.com
financial difficulties in small and medium-sized firms based on non-financial variables. [Smith and Taffler 2000] explored the relationship between the firm's narrative disclosures and bankruptcy using a content analysis approach. Their result is 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 the companies and are not just reporting on past performance. [Shirata & Sakagami 2006] analyzed text data of Annual Reports of 21 Japanese bankrupt companies and 24 Japanese non-bankrupt companies. They extracted key words to discriminate between two groups, bankrupt and non-bankrupt by using morphologically analysis. They found that the Dividend Section of the annual report showed the peculiar explanation to show their financial position. Observation of the results of their analysis reveals that such terms as “dividends,” “profit appropriation” and “retained earnings” are among the terms with prominent differences in appearance frequencies between the two groups. In particular, clear disparities are found in the frequencies of appearance of such terms as “dividends” and “dividend propensity” between the two groups. Their evidence supports the conclusion that continuing entities consistently incorporate messages that are directed at shareholders in their financial reports.

III TEXT MINING OF CORPORATE ANNUAL REPORTS

Issues on previous studies

In most of previous researches about corporate evaluation by using text mining, researchers analyzed text data based on word frequency calculated by morphologically analyzed text. However, extracted word might be lack of some important information that was included in original text, such as word-to-word dependencies and the contexts around high-frequency words.
For instance, word “dividend” could be used in two different contexts; “pay a dividend” and “pass a dividend”. When we only calculated word frequency and extracted high-frequency word in each group, we found that “dividend” appeared in both groups frequently. The difference between the context of “dividend” in bankrupt group and context of “dividend” in non-bankrupt group could not be found only by watching word frequency. This observation showed us that previous text analysis had some issues and we should consider such detailed information more than word occurrence; for instance, like word-to-word dependencies and word co-occurrence in short distance, for extracting expressions peculiar to certain financial position of companies. Texts included in corporate annual reports were usually syntactically correct and were not difficult to be morphologically. On the other hand, the length of each sentence was so long that it was often difficult to analyze word-to-word dependency relations. In addition, it was often difficult to analyze how the words used in the contexts of the Japanese annual report, because the Japanese text was often written in indirect and sophisticated expressions. Therefore, the word frequency information was not sufficient for getting knowledge from indirect and context-dependent texts, as we described in above in this section. How to extract more detailed information than word frequency is one of a key issue to be solved in text analysis for corporate evaluation.

**Sample Data**

In order to investigate peculiar Japanese explanation that can discriminate between bankrupt companies and non-bankrupt companies, we prepared 90 annual reports of bankrupt companies and 90 annual reports of non-bankrupt companies for our experiment. Bankruptcy samples consist of 90 companies that had gone into bankruptcy during the period between 1999 and 2005.
The 90 non-bankrupt companies were extracted from among all companies listed on the Tokyo Stock Exchange Market. We extracted the companies in accordance with the following procedures; a) a bankruptcy discriminate analysis was performed on all companies that had been listed on the Tokyo Stock Exchange Market using the SAF (Simple Analysis of Failure) 2002 bankruptcy prediction model [Shirata 2003], and b) all companies’ SAF values, which are bankruptcy discriminant values, were ranked from the highest to the lowest, and systematic extraction were performed at equal intervals. Selected 90 non-bankrupt companies could be assumed that their distribution was similar to the real distribution of the companies in real stock exchange market.

Corporate annual reports consist of various parts such as financial statements, company profile, a list of directors, auditor’s opinion, and so on, and therefore have tens of pages. In this paper, we analyzed the sections of Dividend Policy Section, because [Shirata & Sakagami 2006] had found that the Dividend Policy Section of the annual report showed the peculiar explanation to show their financial position.

**Text Mining Tools**

We tried conditional probability to extract specific expressions to discriminate between the bankrupt company group and non-bankrupt company group using IBM OminiFind Analytic Edition (OAE)

If $E$ is an expression appearing in the data texts, then $N_a$ is the total number of companies included in data, $N_b$ is the number of bankrupt companies, $N_s$ is the number of non-bankrupt companies, $E_a$ is the total number of companies who mentioned $E$, $E_b$ is the number of bankrupt
companies who used $E$, and $E_s$ is the number of non-bankrupt companies who used $E$.

Probability of all companies who used $E$ is:

$$P(E) = E_a / N_a$$

Probability of bankrupt companies who used $E$ is:

$$P(\text{Bankrupt}, E) = E_b / N_a$$

Probability of non-bankrupt companies who used $E$ is:

$$P(\text{Non-bankrupt}, E) = E_s / N_a$$

The conditional probability $P(\text{Bankrupt}|E)$ used to extract the expressions specific to the bankrupt group and $P(\text{Non-bankrupt}|E)$ used to extract the expressions specific to non-bankrupt group are calculated as:

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

$$P(\text{Non-bankrupt}|E) = P(\text{Non-bankrupt}, E) / P(E)$$

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

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

- The differences between the probabilities of the companies that used $E$ within the bankrupt group and the probabilities of the companies that used $E$ under the non-bankrupt group
- The Kullback-Leibler distance
- $\chi^2$ 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 peculiar key words**

Simple syntactic word-to-word dependency patterns are often insufficient for mining such long sentences in annual reports, because a word and its modifier, in other words, a governor and its dependent, often appear in long distance in such documents. To solve this insufficiency, we prepared technical term list to be focused, and then extracted co-occurred words as candidates of peculiar words to each corporate group within a phrase governed by listed technical terms, in other words, within regions of topic words. We defined technical term list as follows based on the previous research [Shirata and Sakagami 2006].

Technical terms: dividend
  retained earnings
These technical terms often used as topic words in context, for instance, “As for dividend…” and “To say about retained earnings…”, and it is important for our analysis to focus on co-occurred words with these technical terms. In such context, detailed information about topic words often described in a same phrase directly or indirectly governed by topic words. We intend to detailed information about topic words by using our method described above. For instance, the following topic word – information words pair could be extracted from a sentence “As for retained earnings this year, we are planning to use it for effective investment for future business development.”

✓ retained earnings ...future…
✓ retained earnings….business expansion
✓ retained earnings… effective investment

Based on this detailed information on topic words, we investigated in what context these topic words were peculiar to bankruptcy and non-bankruptcy.

**Experiments by using word-frequency statistical metrics**

Tables 1 and 2 show the top 10 words specific to the bankrupt group and the non-bankrupt group, respectively, according to the metric described before. The [amount (Yen)] and the [number of shares] in the tables are the representative forms of “digits + monetary units” and “digits + ‘shares’” respectively. These word sequences were matched and were merged into the
representative forms, [amount] and [number of shares], in OAE preprocess.

**TABLE 1: TOP 10 CONTENT WORDS PECULIAR TO NON-BANKRUPT GROUP**

<table>
<thead>
<tr>
<th>Difference of occurrence probability</th>
<th>Distance between each probability</th>
<th>X**2</th>
<th>Conditional probability</th>
</tr>
</thead>
<tbody>
<tr>
<td># of N</td>
<td># of B</td>
<td># of B</td>
<td># of N</td>
</tr>
<tr>
<td>79</td>
<td>14</td>
<td>79</td>
<td>14</td>
</tr>
<tr>
<td>[# of stocks]</td>
<td></td>
<td>50</td>
<td>4</td>
</tr>
<tr>
<td>[amount of money]</td>
<td>79</td>
<td>25</td>
<td>interim dividend</td>
</tr>
<tr>
<td>interim dividend</td>
<td>50</td>
<td>4</td>
<td>throughout year</td>
</tr>
<tr>
<td>executive board</td>
<td>41</td>
<td>4</td>
<td>[# of stocks]</td>
</tr>
<tr>
<td>decision</td>
<td>41</td>
<td>5</td>
<td>executive board</td>
</tr>
<tr>
<td>execute</td>
<td>53</td>
<td>21</td>
<td>decision</td>
</tr>
<tr>
<td>throughout year</td>
<td>28</td>
<td>0</td>
<td>allot</td>
</tr>
<tr>
<td>dividend</td>
<td>48</td>
<td>21</td>
<td>middle</td>
</tr>
<tr>
<td>fund</td>
<td>33</td>
<td>8</td>
<td>fund</td>
</tr>
<tr>
<td>go</td>
<td>42</td>
<td>18</td>
<td>execute</td>
</tr>
</tbody>
</table>

Remarks: # of B … number of occurrences in documents of bankrupt group

# of N … number of occurrences in documents of non-bankrupt group

In extracting expressions specific to bankrupt companies, we were unable to find clear differences between the results 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 dividend. The sections of “Dividend Policy” in annual report usually consisted of 4 or 5 sentences, and there were very few ways to select the explanation saying “no dividend”. Sentences expressing the lack of a dividend tend to be quite short and similar to each other, so that the results of extracting these expressions using different metrics did not differ significantly.

We did find many differences among the results of extracting expressions specific to non-bankrupt companies by using 4 different metrics. Some keywords, such as “dividend”, “execute”, or “dividend policy”. were used more frequently by non-bankrupt companies than by bankrupt companies, but more than 20% of the bankrupt companies also used these expressions. In contrast, the top 10 expressions extracted by using conditional probability were frequently used by non-bankrupt companies, but no bankrupt companies used these expressions. From these results, we believe that the conditional probability was effective in extracting infrequent expressions that are specific to a certain company group.

“Research and development” and “corporate value” were extracted using conditional probability as expressions specific to the non-bankrupt group. These expressions were rarely used by bankrupt companies. By considering this result, we concluded 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.
### TABLE 2: TOP 10 CONTENT WORDS PECULIAR TO BANKRUPT GROUP (VERB, ADJECTIVE, NOUN)

<table>
<thead>
<tr>
<th>Difference of occurrence probability</th>
<th>Distance between each probability</th>
<th>X**2</th>
<th>Conditional probability</th>
</tr>
</thead>
<tbody>
<tr>
<td># of N</td>
<td># of B</td>
<td># of N</td>
<td># of B</td>
</tr>
<tr>
<td>No dividend</td>
<td>5</td>
<td>46</td>
<td>No dividend</td>
</tr>
<tr>
<td>regret</td>
<td>1</td>
<td>32</td>
<td>regret</td>
</tr>
<tr>
<td>intend</td>
<td>17</td>
<td>35</td>
<td>intend</td>
</tr>
<tr>
<td>loss</td>
<td>1</td>
<td>17</td>
<td>do</td>
</tr>
<tr>
<td>resumption of dividend</td>
<td>2</td>
<td>18</td>
<td>environmen t</td>
</tr>
<tr>
<td>recording</td>
<td>4</td>
<td>20</td>
<td>recording</td>
</tr>
<tr>
<td>do</td>
<td>17</td>
<td>33</td>
<td>severe</td>
</tr>
<tr>
<td>severe</td>
<td>6</td>
<td>21</td>
<td>resumption of dividend</td>
</tr>
<tr>
<td>net loss</td>
<td>3</td>
<td>18</td>
<td>net loss</td>
</tr>
<tr>
<td>environme nt</td>
<td>12</td>
<td>27</td>
<td>loss</td>
</tr>
</tbody>
</table>

**Experiments by using context topic words and their regions**

Listed below are some extracted expressions specific to bankrupt companies and non-bankrupt companies that appear together with the topic word “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
Listed below are extracted expressions specific to bankrupt companies and non-bankrupt companies that occurred together with the topic word “retained earnings”.

<table>
<thead>
<tr>
<th>Bankrupt companies</th>
<th>basic, enrich, react, stable, revenue, reimbursement, status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-bankrupt companies</td>
<td>allot, grow, capital investment, investment in R &amp; D, competency, rationalization, corporate value, plant and equipment, new business, develop, invest, respond, leverage, extend, execute, long-term, improvement, usage, business, future, management environment</td>
</tr>
</tbody>
</table>

In the region of the topic word “dividend”, “[number of shares]” and “[amount in Yen]” were specific to non-bankrupt companies, while these expressions were used by both bankrupt companies and non-bankrupt companies in the full texts of the financial reports. In other words, these expressions were specific to non-bankrupt companies only in the context governed by the topic word “dividend”, and were not specific to non-bankrupt companies in the full text of the annual reports. Similarly, “capital investment”, “plant and equipment” and “new business” were used frequently by non-bankrupt companies in the region of “retained earnings”, but these expressions were not specific to non-bankrupt companies in the full text of the annual reports. The topic words “dividend” and “retained earnings” themselves were specific to neither bankrupt companies nor non-bankrupt companies. However, when these expressions appeared together with these topic words they were effective in distinguishing between bankrupt companies and non-bankrupt companies.

When extracting expressions from the regions of the topic words it was also useful to know about the context in which the topic words were used. For instance, we could assume that
non-bankrupt companies had some retained earnings and therefore will continue to invest in 
“plant and equipment” (i.e. “capital investment”) and the development of “new business”. This 
was supported by our experimental results for extracting expressions within the region of a topic 
word such as “retained earnings”.

IV  CONCLUSION

In this paper, we extracted peculiar expressions to bankruptcy companies and non-bankruptcy 
companies by using few different metrics. Bankruptcy group and non-bankruptcy group did not 
make any big difference, and high frequency words among all of the texts were extracted without 
using conditional probability. This result led us to consider conditional probability would be 
useful to distinguish bankruptcy and non-bankruptcy companies.

Expressions peculiar to bankruptcy were highly ranked by using each metric we examined. 
These expressions were description about no dividend, for instance, “to our regret, we decided to 
finish this year without dividend” or “severe environment does not allow any dividend this year”. 
These expressions could be considered as excuses frequently found in Japanese report by using 
“to our regret” or indirect phrase “does not allow”.

We then prepared some technical terms as topic words governing context, and extracted 
co-occurred peculiar patterns to bankruptcy and non-bankruptcy with these topic words. As the 
result of this experiment, we found that "research and development”, “capital investment” and 
“new business” will appear in sentence including "dividend” and “retained earnings” of 
non-bankrupt companies. This result could be considered to indicate that high revenue and
enough retained earnings could lead to investment for research and development.

This knowledge could not be inferred from the analysis by using only words and word-to-word dependencies. Our method would be effective for long-sentence text mining rather than short-sentence texts. Important information would distribute various parts of a long sentence, and it would difficult to get highly accurate result of syntactic analysis for long-sentence text. We found that using Co-occurrence around the topic words could be one of effective solutions to this difficulty.

**REFERENCE**


Ferrucci, D., and Lally, A., Building an example application with the Unstructured Information Management Architecture, IBM System Journal, Vol. 43, No. 3, 455-476, 2004


IBM OmniFind Analytics Edition

http://www-06.ibm.com/jp/software/data/products/omnifind-analytics/


Salton, G. 1989. Automatic text processing: The transformation, analysis, and retrieval of
information by computer. Reading, MA: Addison-Wesley.


Yoshida, I., and Takuma, D., Software architecture for interactive text mining process, IBM