Text mining of securities reports for discovering reason of change

Yoshida, Shin Ichiro
Graduate School of Information, Science and Electrical Engineering, Kyushu University

Nakatoh, Tetsuya
Research Institute for, Information Technology, Kyushu University

Mitarai, Shuichi
Lafla Inc, Fukuoka System LSI 507 Technology

Hirokawa, Sachio
Research Institute for Information, Kyushu University

https://hdl.handle.net/2324/776827

Text Mining of Securities Reports for Discovering Reason of Change

Shin-ichiro Yoshida
Graduate School of Information
Science and Electrical Engineering,
Kyushu University
6-10-1 Hakozaki, Higashi-ku,
Fukuoka 812-8581, JAPAN

Tetsuya Nakatoh
Research Institute for
Information Technology,
Kyushu University
6-10-1 Hakozaki, Higashi-ku,
Fukuoka 812-8581, JAPAN

Shuichi Mitarai
Lafla Inc.,
Fukuoka System LSI 507
Momochihama 3-8-33, Sawara-ku
Fukuoka 814-0001, JAPAN

Sachio Hirokawa
Research Institute for Information
Technology, Kyushu University
6-10-1 Hakozaki, Higashi-ku,
Fukuoka 812-8581, JAPAN

Abstract

A stock market is a base of the economic activity in present-day free economy society. In order to become a listed company, there is a severe examination. Furthermore, yearly duty is attached by law for the listed company to file the fixed form report about the financial condition. In this paper, correspondence of the numeric data about the activity of a company and text data is analyzed for the financial report. All of 68 medical-supplies-related listed companies were chosen first. Then, the feature words were extracted from the reports of the companies where achievements have improved favorably. Moreover, the feature words of the financial report of the time were extracted about the companies where achievements carried out the sharp reversal to profitability.

1 Introduction

In order to examine an investment, the investor is analyzing the activity condition of various companies. Many attentions are being paid on the prediction of industry trends and the stock price by analysis of change of a stock price, topical news, etc. UFOReader [1] and UFOLenz [2] by Lafla Inc. are samples of such Web services. The stock price prediction from Twitter, Blog, a newspaper article, etc. is becoming hot. For example, if the quality and speed are improved in the analysis of Twitter, real time stock price prediction can be realistic. [2] and [10] evaluated the

[1] http://www.uforeader.com/v1/
[2] http://www.yano.co.jp/ufolenz/index.php

polarity of feeling by classifying the tweets positively and negatively, and predicted the stock price. [7] estimated the polarity of news articles. [3] and [4] analysed relationship between the number and the length of comments in Blog articles with the stock price. [11] evaluated past newspaper articles by associating them with stock price volatility at the time and past newspaper articles. They were used as the evaluation criteria for new articles.

[8] focused on the medicine related companies and compared the stock price and IR documents on the Web. [6] proposed and evaluated a method to analyse the security reports of US companies by visualizing the relationship of words. The stock price used for analysis in those researches is an index by which it is evaluated how the investor is looking at the company. It is an evaluated data from the outside, rather than a real data of the company. On the other hand, the financial report analysed in the present paper is not external evaluation but the track record of each company itself. Moreover, the contents which should be described in the financial report are prescribed by law, and a duty of the presentation for every term is imposed. That is, a financial report is an official document more reliable than an IR document on Web.

There is a research [9] which analyses financial reports, where they focused on the portion of a dividend policy among. They succeeded to clarify the difference between bankruptcy companies and going concerns. Having limited to this portion of the financial report is based on their experience. On the other hand, the feature word extraction was performed in this paper, without limiting the range. However, when the same word appears in different segments, they are

Table 1: Segments of Security Reports

segment	# companies	# documents	contents
abstract	5311	35296	Outlines, such as achievements
business	5310	35266	The enterprise contents
connected	5309	35181	An affiliated company's situation
contract	5309	35184	The important contract on management, etc.
employee	5308	35176	An employee's situation
finance	5309	35216	Analysis of financial condition and business performance
history	5309	35297	History
indicators	5308	35176	Transition of main management indexes etc.
issue	5310	35230	The issue which should be coped with
produce	5309	35180	The situation of production, an order received, and sale
rd	5309	35207	Research and development activities
risk	5310	35215	Risks of enterprise
segment	5309	35177	Enterprise information by segment according to kind

indexed as different words. As a result, the feature word can not only be found, but analysis was completed in which segment the feature words appear often. Bankrupt information has been analysed in [5] and [1] as well. They extracted reason of bankruptcy by determining the sentences that describe the cause of bankruptcy. At the point of extracting the feature word from a document group and analyzing the correlation of the feature word, it is research of the same purpose as this paper. The novelty of the present paper, compared with those researches, is that the candidates of companies and the documents for the analysis are determined by numeric data.

2 Selecting Target Companies by Numeric Analysis

Figure 1 shows the movement of sales and profits of each company. Note that the values are normalizing to see the change instead of the values themselves. In this paper, we focused two types of movements in these graphs. The first targets are the companies which are developing achievements favorably. Other targets are the companies which seem to have carried out the sharp reversal to profitability.

2.1 Companies with continuous profit growth

We selected 10 companies (Table 2) from Figure 1 which show a continuous growth of profit.

We analysed 79 reports from these 10 companies and extracted feature words. Table 4 displays the top 20 words and the segments where each word appears.

Table 2: Growth Corporation

Code	Name of company
E00923	Shionogi & Co., Ltd.
E00924	Mitsubishi Tanabe Pharma Co.
E00935	Kaken Pharmaceutical Co., Ltd.
E00942	Rhoto Pharmaceutical Co.,Ltd.
E00944	Hisamitsu Pharmaceutical Co.,Inc.
E00947	Mochida Pharmaceutical Co.,Ltd.
E00960	Biofermin seiyaku Co.,Ltd.
E00974	Towa Pharmaceutical Co.,Ltd.
E00975	Fuji Phama Col, Ltd.
E00976	Sawai Pharmaceutical Co.,Ltd.

2.2 Companies with sharp reversal to profitability

We selected 6 companies (Table 3) from Figure 1, which seems to have achieved sharp reversal to profitability. We selected 15 reports that correspond to the period of sharp reversal to profitability and extracted feature words. Table 5 shows the top 20 words and the segments where each word appears.

Table 3: Sharp-reversal-to-profitability company

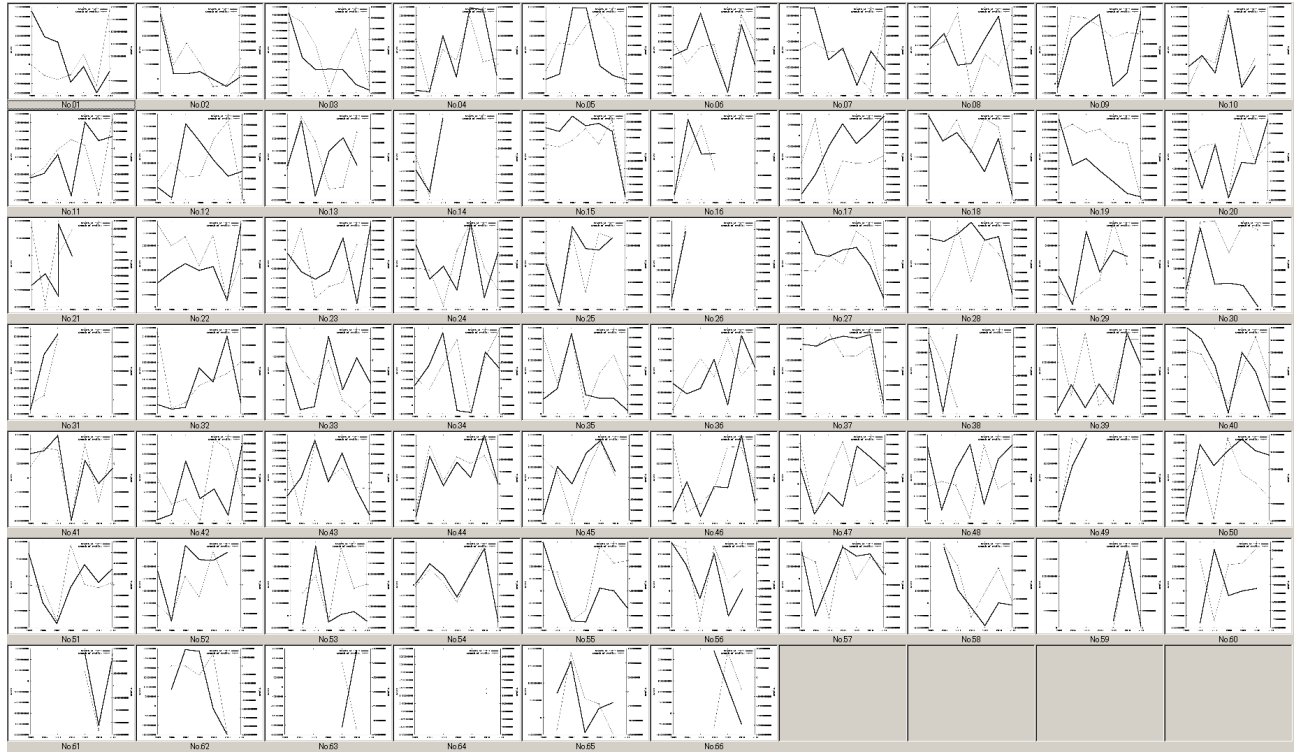| Code	Name of company
E00926	TOC Co., Ltd.
E00938	Toyama Chemical Co., Ltd.
E00978	Medical & Bilogical Laboratories Co.,Ltd.
E00982	LTT Bio-Pharma Co.,Ltd.
E00984	Daiich Sankyo Co., Ltd.
E22560	Taiko Pharmaceutical Co., Ltd.

Figure 1: List of Corporate-earnings Graph

3 Feature Words of Target Companies

3.1 Feature Words of Companies with Continuous Profit Growth

We see feature words of progressing companies in the segments of finance and RD.

Feature words extracted in the finance segments, such as 後発 (late-coming), 加算 (addition), 新設 (establishment) and 揃え (arrange), concern the revision of the prescription form and the revision of the Regulations for NHI Pharmacies' and NHI Pharmacists' Responsibilities in April 2008.

The government is promoting the policy for mitigation of payment by the patient and the improvement of medical insurance finances. Those companies with continuous profit growth are receiving benefits by arranging generic drugs.

The word 総合 (synthesis/collective) is used in the phrases such as "synthetic research organization", "collective strength of a group" and "product collective strength". The target companies know their strength in synthesis organization.

The words of セロトニン (serotonin), 取り込み (uptake), 選択 (select) and 不安 (anxiety) concerns the antidepressant drugs, such as SSRI(Selective Serotonin Reuptake Inhibitors) and SNRI(Serotonin & Norepinephrine Reuptake Inhibitors). Profits can be in-terpreted as increasing by developing and marketing such medicine.

3.2 Feature Words of Companies with Sharp Reversal to Profitability

Most of the characteristic words of reports of companies with sharp reversal to profitability appear in the section of enterprise risk.

The word 類似 (similar/imitation) concerns the risk against similar or imitation drug. Users might buy similar drug of the other company. The low quality of imitation drug might cause damage to the company. They recognize those damages caused by rumors as their risk.

The words 薬品 (medicine), 府 (prefecture), 局 (bureau), 衛生 (health), 大阪 (Osaka) concern the permission and the risk by legal restrictions, such as 香港衛生署薬剤業及毒薬管理局許可 (permission by Hong Kong Administration Bureau of Health, Medicine and Poison) 中国国家食品薬品監督管理局許可 (permission by China Administration Bureau of Food and Medicine) and 大阪府知事許可 (permission by Osaka Prefecture Governor).

Table 4: Feature Words in IR Documents (Continuous Profit Growth)

rank	weight	TF_d	TF	DF	word	abstract	business	connected	contract	employee	finance	history	indicators	issue	produce	RD	risk	segment
1	0.444	4	14	11	点 (point)						√							
2	0.409	4	31	28	総合 (synthesis)						√							
3	0.378	4	12	7	山形 (YAMAGATA)							√						
4	0.359	2	6	5	セロトニン (serotonin)											√		
5	0.358	3	8	5	目 (view)						√							
6	0.333	2	15	13	基づい (based)					√								
7	0.330	2	9	7	職能 (function)					√								
8	0.330	2	9	7	生じれ (produce)												√	
9	0.329	9	77	43	後発 (late-coming)						√							
10	0.328	2	18	16	取り込み (uptake)											√		
11	0.315	2	3	2	山形 (YAMAGATA)	√												
12	0.314	2	13	10	揃え (arrange)						√							
13	0.310	2	21	18	加算 (addition)						√							
14	0.310	3	47	42	選択 (select)											√		
15	0.304	3	46	39	見直し (reexamination)						√							
16	0.301	2	17	13	飛躍 (leap)						√							
17	0.300	2	20	16	新設 (establishment)						√							
18	0.300	2	20	16	不安 (anxiety)											√		
19	0.299	3	37	26	所 (place)	√												
20	0.298	2	22	18	数量 (quantity)						√							

4 Conclusion and Future Work

In this paper, we analysed the financial reports of medicine related companies, firstly by specifying the companies and the periods using numeric data and then extracted feature words from the financial reports of target companies in the target periods. Case studies are shown concerning 68 medicine-related companies. 10 good companies with continuous profit growth are analysed. It turned out that the strength of those companies is in their collective organization and their positive correspondence toward the government policy which promotes generic medicine. Those descriptions are found in the financial segments of the reports. Another strength is found in the research and development segment of the report where they focus on antidepressant drugs. Companies with sharp reversal to profitability are analysed as well. The extracted feature words can be clues to interpret the change found in the numeric data.

The morphological analysis has to be improved in further work. We encountered cases where crucial words are split into two words and was not able extracted as a result. In other situation, the same words in different phrases are identified where the statistics analysis failed for the word.

References

[1] Takahiro Baba, Tetsuya Nakatoh and Sachio Hirokawa, "Text Mining of Bankruptcy Information using Formal Concept Analysis," Proc. of *3rd International Conference on Awareness Science and Technology (iCAST2011)*, pp.527–532, 2011.

[2] Johan Bollen, Huina Mao and Xiao-Jun Zeng, "Twitter mood predicts the stock market." *Journal of Computational Sci- ence*, Volume 2, Issue 1, March 2011, pp. 1–8, 2011.

[3] De Choudhury, M., Sundaram, H.,John, A.Seligmann, D.D. "Can blog communication dynamics be correlated with stock market activity?", *Proceedings of the 19th ACM Conference on Hypertext and Hypermedia, HT'08 with*

Table 5: Feature Words in IR Documents (Sharp Reversal to Profitability)

rank	score				word	segments												
	weight	TF_d	TF	DF		abstract	business	connected	contract	employee	finance	history	indicators	issue	produce	RD	risk	segment
1	0.339	3	6	4	二酸化 (2 oxidization)						√							
2	0.325	2	8	7	B												√	
3	0.321	5	36	23	発生 (generating)										√			
4	0.321	2	6	5	保管 (storage)	√												
5	0.319	4	33	24	薬品 (medicine)												√	
6	0.318	12	41	12	致し (do)	√												
7	0.318	4	24	15	府 (prefecture)												√	
8	0.314	2	5	4	止 (stop)		√											
9	0.314	4	18	10	比し (compared)	√												
10	0.303	2	4	3	シーエス (CS)										√			
11	0.299	2	11	9	丹平中田 (Tanpei Nakata)										√			
12	0.298	3	36	31	財務 (financial affairs)													√
13	0.294	6	52	29	出荷 (shipment)	√												
14	0.290	6	54	30	類似 (similar)												√	
15	0.290	2	8	6	労働省 (Ministry of Labor)							√						
16	0.289	3	41	36	出荷 (shipment)						√							
17	0.288	4	34	20	局 (bureau)												√	
18	0.288	5	55	37	衛生 (health)												√	
19	0.284	8	57	22	大阪 (Osaka)												√	
20	0.284	2	7	5	地道 (honest)						√							

Creating' 08 and WebScience' 08, pp. 55–59, 2008.

[4] Eric Gilbert and Karrie Karahalios, "Widespread Worry and Stock Market", *4th International AAAI Conference on Weblogs and Social Media(ICWSM)*, 2010.

[5] Sachio Hirokawa, Takahiro Baba and Tetsuya Nakatoh, "Search and Analysis of Bankruptcy Cause by Classification Network," *Model and Data Engineering, Lecture Notes in Computer Science*, Volume 6918/2011, pp.152–161, 2011.

[6] Kun Qian, Sachio Hirokawa, Kenji Ejima and Xiaoping Du, "A fast associative mining system based on search engine and concept graph for large-scale financial report texts," Proc. *2nd IEEE ICIFE (Information and Financial Engineering)*, pp.675–679, 2010.

[7] Hiroyuki Sakai, Shigeru Masuyama, "Estimation of Impact Contained in Articles about each Company in Financial Articles," *Information Processing Society of Japan (IPSJ)*, vol 94, pp.43–50, 2006. (in Japanese)

[8] Toshihiko Sakai, Masashi Matsushita, Brendan Flanagan, Jun Zeng and Sachio Hirokawa, "Analysis of influence of Investor Relation Documents to stock price", Proc. of *FSKD2012 : 9th International Conference on Fuzzy Systems and Knowledge Discovery*, pp. 1280–1284, 2012.

[9] Cindy Y. Shirata and Manabu Sakagami, "An Analysis of the "Going-Concern Assumption": Text Mining from Japanese Financial Reports," The Journal of *Emerging Technologies in Accounting, Strategic and Emerging Technologies Section of the American Accounting Association*, pp.1–16, 2009.

[10] Mike Thelwall, evan Buckley and Georgios Paltoglou, "Sentiment in Twitter Events," Journal of the *American Society for Information Science and Technology*, Volume 62, Issue 2, pp. 406–418, 2011.

[11] He Zhang and Shigeki Matsubara, "Quantitative Analysis of Relevance between News Articles and Stock Price Change", *Information Processing Society of Japan (IPSJ)*, pp.183–184, 2008. (in Japanese)