Text mining of bankruptcy information using formal concept analysis

Hirokawa, Sachio
RIIT, Kyushu Univ.

Nakatoh, Tetsuya
RIIT, Kyushu Univ.

Baba, Takahiro
Graduate School of ISEE, Kyushu Univ.

出版情報：Proceedings of 2011 3rd International Conference on Awareness Science and Technology, iCAST 2011, pp.527-532, 2011-12-01
バージョン：
権利関係：

Text Mining of Bankruptcy Information using Formal Concept Analysis

Takahiro Baba
Graduate School of ISEE
Kyushu Univ, Japan
2IE09092R@s.kyushu-u.ac.jp

Tetsuya Nakatoh
RIIT, Kyushu Univ, Japan
nakatoh@cc.kyushu-u.ac.jp

Sachio Hirokawa
RIIT, Kyushu Univ, Japan
hirokawa@cc.kyushu-u.ac.jp

***Abstract*—A lot of information concerning the status of companies are available on the Web. However, a simple search of documents does not explain the meaning or the cause the status. Semantical interpretation and hypotheses generation are necessary for further analysis. This paper proposes a method to analyse the cause and the situation of bankruptcy with respect to particular condition that a user can specify as a query.**

The method is based on the theory of formal concept analysis. The novelty of the method is in (a) that sentences are considered as objects and words are considered as attributes and (b) that a concise subgraph of the concept lattice is introduced and used to guess the cause. Two cases of interactive and iterative process are shown where a user proceeds from a simple query to a new hypothesis, which would not be able to found by a naive cross tabulation or keyword extraction.

I. Introduction

We can find various information concerning companies on the Web. Some web pages provide bankruptcy information. It is worth to analyze the status or any reason of bankruptcy of companies in particular area of industry.

This paper proposes a method to analyze the bankruptcy information. The co-occurrence relation of words that appear in the documents are visualized using the formal concept analysis. Moreover, a concise representation of the concept lattices are introduced and are shown to be effective to guess causes of bankruptcy.

Conventional search engine can be applied to bankruptcy information. It may return a list of documents that match the query of a user. However, this kind of simple list does not give any hints to guess the cause of bankruptcy. Interactive and iterative process of search is crucial to reach a deep comprehension of the issue.

This paper proposes a system that displays not only the list of search results but also hints for further analysis.

Fig. 1 is a screen shot of the system, where relationship of characteristic words are displayed as a directed graph as well as the ranked list of sentences that contain the query "but construction". From the graph, the user can These hints will help the user to expand or change his query. He will be able to obtain his hypothesis and confirm or deny the hypothesis. He can continue this process interactively and iteratively.

The query of the user is the trigger of the analysis process. Once the process is started, the system shows hints of further analysis. The user only have to interpret the graph of keywords and determine the direction of further search.

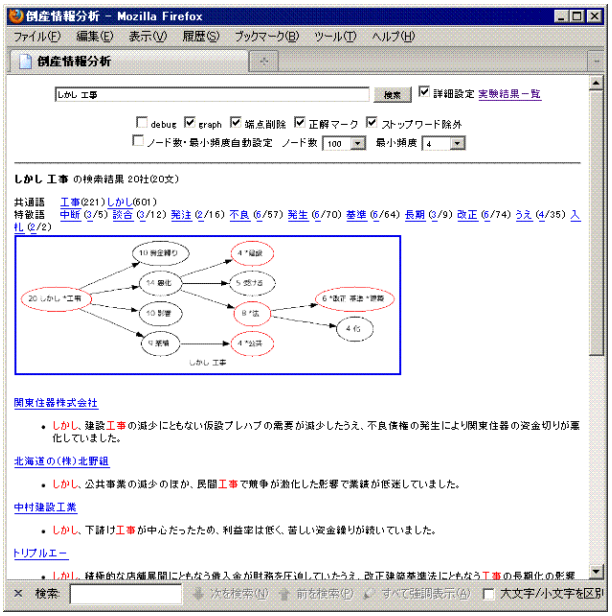

Figure 1. A Snapshot of the System

II. Basic Analysis of Bankruptcy Information

We applied the proposed method to bankruptcy information available on the Web. Each article of bankruptcy information describes the overview of the company and the situation how the company went bankruptcy.

We constructed a search engine for 726 companies which went bankruptcy in recent years in Japan. We extracted the information from a web site [1]. We focus on the relationship of words. Therefore, the result of this search engine is not a list of documents but a list of sentences. Thus the target of this search engine is not the 726 documents but 4799 sentences. Each bankruptcy information is described with 6.6 sentences in average.

[1] http://news2ch33.blog108.fc2.com/

In this section, we show the cross tabulation with respect to the region of the business and the type of business as basic analysis. As another basic analysis, we show characteristic keywords for the type of business. We chose the frequent words to represent the region and type of business.

	Tokyo (616)	Osaka (209)	Fukuoka (76)	Hokkaido (73)
construction (306)	14	29	34	19
sales (281)	10	15	5	16
real estate (263)	4	11	4	8
manufacturer (257)	5	7	1	4

Table I
CROSS TABULATION OF TYPE AND REGION

Table I is the cross tabulation of the region of business and the type of business. The number in parentheses on the side of the word is the number of sentences that contain the word. We see that a lot of companies in Tokyo region went bankrupt. However, much large number of companies are in Osaka and Fukuoka region, if we focus on the companies related to construction.

From this simple analysis, we can obtain a hypothesis that construction related companies in provinces might have high probability of bankruptcy than that in the central area of Tokyo.

Type	words
construction	civil engineering, materials architecture, construction work, public, apartment house, handle, firm
sales	house, trader, establishment, articles materials, handle, apartment house selling in lots
real estate	market, edge, generate, subprime loan prime, sub, loan, random
manufacturer	parts, product, electronic, machine device, handle, lyquid crystal establishment

Table II
CHARACTERISTIC WORDS OF BUSINESS TYPE

Table II displays the feature words to the types of business. "Civil Engineering" and "Material" may be appropriate to the business related to "Construction". However, most of these words are common not only to decreased building companies but also to well managed building companies.

III. CONCEPT LATTICE AND INCLUSION GRAPH

The classification is the most basic method for analyzing the objects. If we focus on the different attributes, we would obtain different result of classification. When a set of objects and a set of attributes characterize each other, the pair is said to be a concept [2]. The set of the concepts form a lattice, which is called as the concept lattice.

In this section, we review the notion of concept lattice [2] and introduce the inclusion graph. Let D be a set of documents and W be a set of words. A subset M of $D \times W$ is called a context. Given a document $d \in D$ and a word $w \in W$, we denote $(d, w) \in M$ when d contains w. When $X \subseteq D$ and $Y \subseteq W$ satisfy the condition $doc(Y) = X$ and $word(X) = Y$, (X, Y) is said to be a concept. Here $doc(Y)$=$\{d \in D \mid \forall w \in Y,\ d$ contains $w\}$, and $word(X)$ = $\{w \in W \mid \forall d \in X, d$contains $w\}$. Given elements (X, Y) and (X', Y') of $CL(D, W)$, we define $(X, Y) >_{CL} (X', Y')$ iff $doc(Y) \subseteq doc(Y')$. The set of all concepts forms a lattice $CL(D, W)$ with respect to this order and is called a concept lattice.

Figure 2(above) is the concept lattice with respect to Table III. For the simplification of the display, the name of the objects and the attributes are displayed only once in a path that connects the left end (root) and the right end (leaf).

	a	b	c	d	e
A	1	0	0	0	0
B	0	0	0	0	1
C	1	1	1	0	0
D	0	0	0	1	1
E	1	0	1	0	0
F	1	0	0	0	1
G	0	1	0	0	0

Table III
A SAMPLE OF CONTEXT MATRIX

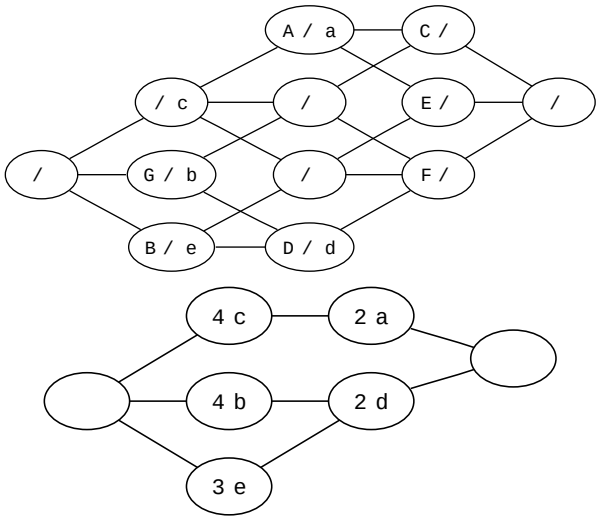

Figure 2. Concept Lattice(above) and Inclusion Graph(below) for Table 4

For example, we see only one node that contains "D/d". The simplified graph should be interpreted such that "D"

appears in all nodes which locates left to the node. Therefore, the node ”D/d” shows that the two objects B and D are characterized in terms of the two attributes ”d” and ”e”.

According to this simplified display, there may be some nodes that have empty object and empty attribute. Actually, imagine a case where n be the max of the number of objects and the number of attributes. Then, the number of concepts may be 2^n. However, we have at most n names of objects and attributes, if we use simplified display. This implies that there are at most $2n$ labeled nodes and that most of the nodes have empty label. As the result, we would not be able to learn any hints from these ”almost empty” visualization.

This paper introduces a notion of ”Inclusion Graph” that erases all these empty nodes from the concept lattice. In an inclusion graph, moreover, only the attributes, i.e., words, are labeled on the nodes so that the graph can be interpreted intuitively.

Given two words $u, v \in W$, we define $u >_{IG} v$ iff $doc(u) \subseteq doc(v)$. The ordered set $(W, >_{IG})$ with this order is called an inclusion graph and is denoted as $IG(D, W)$.

Figure 2(below) is the inclusion graph made from Table 4. We can see that ”b” and ”e” are bridged with ”d”.

The inclusion graph displays an essence of the concept lattice. However, the structure of the inclusion graph is completely embedded in the concept lattice.

Koester [10] used the concept lattice as an interface of search engine, where the title, a snippet and URL of search result form an object and the extracted feature terms as attributes. We can apply the formal concept lattice to documents and can be construct such a search engine [1]. A concept lattice can be drawn as a directed graph where each node represents a concept that is determined a pair of a set of documents and a set of keywords. Adjacent nodes of a concept display the subclasses of the concept.

However, the co-occurrence relation of words in a document does not capture the causal relation. In this paper, we consider sentences as objects instead of documents. Each sentence that describes the situation of bankruptcy of a company is considered as a separate document. Li et al. [11] uses sentences for opinion extraction. They used a fixed set of words. On the other hand, any word can be an attribute in our approach.

IV. Case Analysis - Revised Building Standard Law

The inclusion graph (Figure 3) was made from the document #545(”The Daiei Farm Inc. and the Shimizu House went bankruptcies”). The document includes 7 sentences in which the overview and the bankruptcy reason are written for the enterprises.

The graph itself is generated automatically. Note that some words are marked with asterisk in the figure. Independent to the system, we have carefully read the documents and listed the words that are considered as the causes of

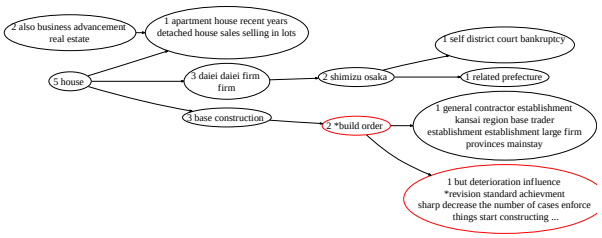

Figure 3. Inclusion Graph of Daiei Farm Inc.

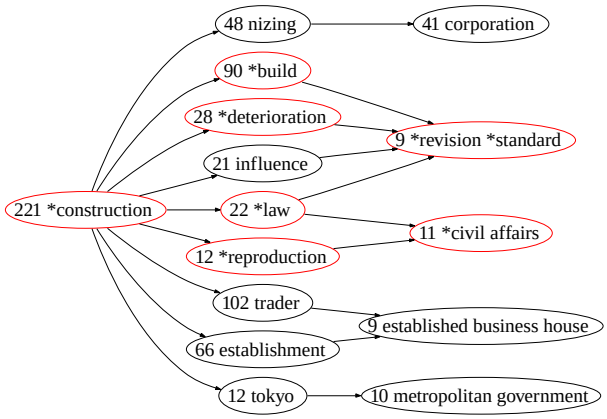

Figure 4. Inclusion graph of “construction”

bankruptcy of each company. We think that we can evaluate the usefulness of the method by analyzing how these correct words would appear in the inclusion graph. Note also that the nodes that contain the correct words are drawn with red circle.

We see the words that related to the job of the company, such as ”house”, ”fundamental construction” and ”apartment house”. We also see another kind of words that are related to bankruptcies, such as ”order”, ”deterioration” and ”sharp decrease”. The words ”construction” and ”revision” appear as related words to ”deterioration” and ”sharp decrease”. We guess that the revision of the building standard law gave a negative influences to the enterprises. In fact, the document for this company says that the enforcement of the revised building standard law and the expansion of the company to to the real estate business caused the increase of the dept and went wrong. We can see that the cause and effect links in the inclusion graph.

From these analysis, we consider a hypothesis that the revision of building standard law might be one of the causes for the bankruptcies of building related companies, not only of this company. Then, we paid attention to the word ”construction” that appeared in Figure 3. Figure 4 is an inclusion graph of ”construction”. It is seen that the words ”construction”, ”revision”, ”standard” appear as related words of ”building”. This analysis strengths our hypothesis.

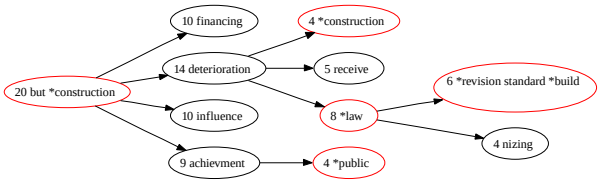

Figure 5. Inclusion graph of “but+construction”

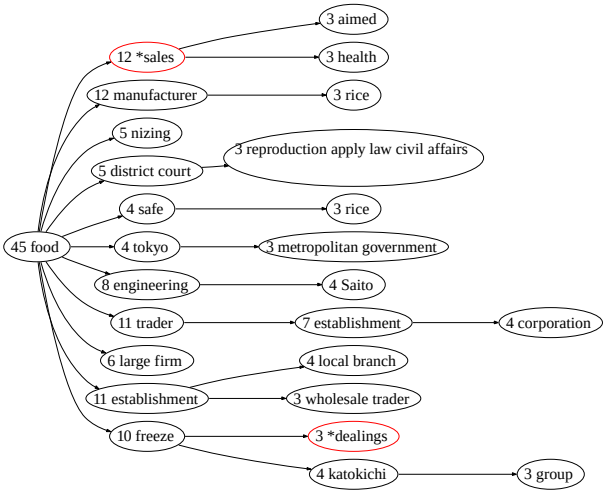

Figure 6. Inclusion graph of “food”

In Figure 4, we note that the red nodes that contain the right words of bankruptcy cause contains the word ”but”. The word ”but” was often seen in the nodes that contain other correct words, in other samples of inclusion graph as well. We considered that this word should be important to the analyze. In fact, it turned out that the documents are written in two segments. The first segment describes the general description of a company. The second segment describes the situation of bankruptcy of the company, where the second segment starts with the word ”but”. This implies that ”but” tends to co-occur with other words of bankruptcy cause.

From these observation, we think that the bankruptcy factor would be able to be extracted by restricting the sentence including ”but” and ”construction”. Figure 5 is the inclusion graph with respect to the query ”but construction”, which have a fewer nodes than that of Figure 4 and is easy to guess the relation of bankruptcy and the revision of building standard law.

V. Case Analysis - Food Related Companies

There were 20 companies whose document contain the word ”food”. There were 45 sentences with the word. Figure 6 is the inclusion graph of the word ”food”.

Figure 6 has 3 features of words – (a) sales and healthy, (b) rice and safe and (c) freezing and ”Katokichi”. We tried a refinement search with the words ”food freeze” and

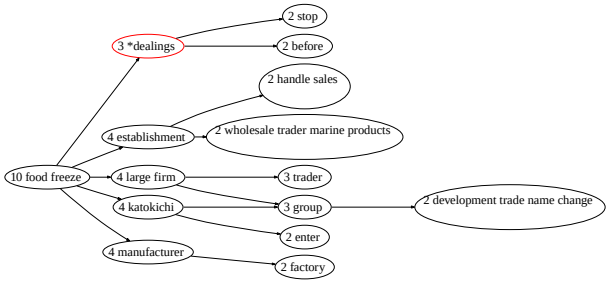

Figure 7. Inclusion graph of “food+freeze”

obtained 5 companies and 10 sentences. The inclusion graph is Figure 7 , where we found that (c) has something to do with the companies which has business relations with ”katokichi” company.

VI. Related Work

The corporate comparison research has been a key object of cognitive study in business administration and organizational science [7] from early days. They analyzed financial reports with respect to the numerical data of the companies.

The numerical analysis became a main theme of the research thanks to the development of the computer. Analysis and prediction of bankruptcy is one of the hot topics in these area [17], [12], [20]. Shin et al. [17] applied SVM for bankruptcy prediction with the 10 controlling parameters, such as, total asset growth, contribution margin, operating income to total asset. Li et al. [12] compared the data mining methods and the statistical methods for the business failure prediction. They used financial ratios of business status of companies and categorical variables. Tsai [20] compared the prediction performance of five well-known feature selection methods in bankruptcy prediction. These analysis are based on numerical features and are not easy for non-experts to utilize in practice.

Visualization is known to be effective in many area. Goda et al.[6] visualized the relationship of defaulted companies as KeyGraph, where a node represents a defaulted company and an edge represents a pair of defaulted companies in the same area and in the same period of the year. A limit of visualization is that we need further textual explanation of the result obtained through visual analysis.

Text mining methods, such as keyword extraction and keyphrase extraction [3], [5], [21], [13] , are expected to be applicable to annual reports and bankruptcy information. Kida[9] used a text mining tool to analyze the part of business situation of the annual reports. Takahashi et al.[18] confirmed the strong effect of the occurrence of keywords, such as ”upward/downward surprise in forecast”, in the titles of analysts’ reports and the stock price return.

Takeuchi et al.[19] focused on particular regions in documents to capture contextual information. They used the phrases, such as ”as regards (ni tsukimashite ha)”, ”because

of(no tame)" etc, to determine the regions. Then they extracted pairs of a topic word and a keyword in a region to compare the annual reports of bankrupt companies those of sound companies. Li et al. [11] captured a topic as the co-occurrences of words in one sentence and the occurrences of the same topic in different sentences. Their target and the data was not in analysis of corporate reports but in opinion retrieval. But they are summarized as "word-context" and "pair-pattern" approaches in Turney & Pantel [22]. The method of the present paper belongs to the "word-context" approach, where a sentence represent a context.

Visualization and graph-base model have been extensively studied in document summarization. Ouyang et al. [14] used undirected graph to measure the similarity of sentences and generated document summaries. Uchida et al. [23] analyzed the free text of consumers' questionnaires with hierarchical keyword graph, where the relationship among words are displayed based on the co-occurrence in the same sentence. Yamamoto & Orihara [24] considered the word co-occurrence graph based on co-occurrence of two words within sentences and applied SVM to characterize feature words from term frequency, degree of word node, average path length and clustering coefficient. All of these approaches consider undirected graph where only relative position, i.e. being close or far, has meaning. On the other hand, the present paper construct directed graphs, where a word in the left-side are supposed to have general meaning and the word in the right-side is used in narrow contexts. Similar directed graph, named Concept Graph, are used in [16] for a constructing word hierarchical from a dictionary, in [8] for analyzing the group structure of researchers activities and in [15] for investigating financial reports. The relation of words and the Concept Graph [16], [8], [15] are determined with a threshold. On the other hand, the relation of words and the graph (Inclusion Graph) proposed in the present paper are based on the formal concept lattice and does not require such parameters.

VII. Conclusion and Further Work

This paper proposed a method of inclusion graph which can be embedded in the concept lattice. Based on the method, an interactive and iterative analysis system is constructed and applied to the bankruptcy information available on the Web. Characteristic words as drawn as directed graph by which the cause of bankruptcy were obtained as hypothesis.

It is not always the case that the words in the inclusion graph represent the cause of bankruptcy, because the documents describe not only the situation of bankruptcy but also the history and outline of the company in general.

For example, the "rice" and "safe" in Figure 7 might remind us some incident concerning food. But, in this case, it comes simply from the name of a company "Rice Safe Food(Okome-Anshin-Shokuhin)". We consider that this kind of difficulties would be solved by considering the semi-structure of the documents where the general description and the description of bankruptcy are separated as different substructures.

The qualitative evaluation of the proposed method is necessity as further work. We selected the correct word manually. This list of words should be used for qualitative evaluation.

References

[1] T. Baba, L. Liu, S. Hirokawa, Formal Concept Analysis of Medical Incident Reports, Springer LNCS 6278, pp. 207–214, 2010

[2] C. Carpineto, G. Romano, Concept Data Analysis Theory and Application, John Wiley and Sons, 2004

[3] S.W.K. Chan, Extraction of salient textual patterns: Synergy between lexical cohesion and contextual coherence, IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans. 34 (2), pp. 205–218, 2004

[4] R.Chau, A.C.Tsoi, M. Hagenbuchner, V.C.S Lee, A ConceptLink Graph for Text Structure Mining, Proc. the 32nd Australasian Computer Science Conference, pp. 129–137, 2009

[5] W.T. Chuang, J. Yang, Extracting sentence segments for text summarization: A machine learning approach, , SIGIR Forum (ACM Special Interest Group on Information Retrieval) , pp. 152–159, 2000

[6] S. Goda, Y. Ohsawa, Chance discovery in credit risk management - Time order method and directed KeyGraph for estimation of chain reaction bankruptcy structure, Springer LNAI 4914 , pp. 247–254, 2008

[7] Huff, A.S.(ed.), Mapping Strategic Thought, Chichester, Wiley 1990

[8] Y. Iino, S. Hirokawa, Time Series Analysis of R&D Team Using Patent Information, Springer LNCS 5712, pp. 464–471, 2009

[9] M. Kida, Cognitive research of Asahi's organizational renewal – textmining of annual reports organizational science, Academic Journal, Vol.39. No.4, 2006

[10] B. Koester, Conceptual Knowledge Retrieval with FooCA: Improving Web Search Engine Results with Contexts and Concept Hierarchies, Proc. ICDM2006, Springer LNAI 4065, pp.176–190,2006

[11] B. Li, L. Zhou, S. Fen, K.-F. Wong, A Unified Graph Model for Sentence-based Opinion Retrieval, Proc. 48th ACL, pp.1367–1375, 2010

[12] H. Li, J. Sun, J. Wu, Predicting business failure using classification and regression tree: An empirical comparison with popular classical statistical methods and top classification mining methods, Expert Systems with Applications 37 (8), pp. 5895–5904, 2010

[13] X. Liu, J. Webster, C. Kit, An extractive text summarizer based on significant words, Springer LNAI 5459, pp. 168–178, 2009

[14] Y. Ouyang, W. Li, F. Wei, Q. Lu, Learning similarity functions in graph-based document summarization, Springer LNAI 5459, pp. 189–200, 2009

[15] K. Qian, S. Hirokawa, K. Ejima, X. Du, A Fast Associative Mining System based on Search Engine and Concept Graph for Large-Scale Financial Report Texts, Proc. The Second IEEE International Conference on Information and Financial Engineering, pp. 675–679, 2010

[16] Y. Shimoji, T. Wada, S. Hirokawa, Dynamic Thesaurus Construction from English-Japanese Dictionary, Proc. The Second International Conference on Complex, Intelligent and Software Intensive Systems, pp. 918–923, 2008

[17] K.-S. Shin, T.-S. Lee, H.-J. Kim, An application of support vector machines in bankruptcy prediction model, Expert Systems with Application 28 (1), pp. 127–135, 2005

[18] S. Takahashi, M. Takahashi, H. Takahashi , K. Tsuda, Analysis of Stock Price Return Using Textual Data and Numerical Data Through Text Mining, Springer LNCS 4252, pp. 310-316, 2006

[19] H. Takeuchi, S. Ogino, H. Watanabe, Y. Shirata, Context-based text mining for insights in long documents, Springer LNAI 5345, pp. 123–134, 2008

[20] C.-F. Tsai, Feature selection in bankruptcy prediction, Knowledge-Based Systems 22, pp.120–127, 2009

[21] P.D. Turney, Learning algorithms for keyphrase extraction, Information Retrieval 2 (4), pp. 303–336, 2000

[22] P.D. Turney, P. Pantel, From frequency to meaning: Vector space models of semantics, Journal of Artificial Intelligence Research 37, pp. 141–188, 2010

[23] Y. Uchida, T. Yoshikawa, T. Furuhashi, E Hirao, H. Iguchi, Extraction of important keywords in free text of questionnaire data and visualization of relationship among sentences, IEEE International Conference on Fuzzy Systems, art. no. 5277332, pp. 1604–1608, 2009

[24] Y. Yamamoto, R. Orihara, Keyword extraction using the word co-occurrence network properties that is independent of languages and document types and its evaluation by prediction of headline words (in Japanese), Transactions of the Japanese Society for Artificial Intelligence 24 (3), pp. 303–312, 2009