International Journal of Service and Knowledge Management
International Institute of Applied Informatics
2018, Vol.2, No.1, P.1 - 18

Data Mining Framework for Treating both Numerical and Text Data

Wataru Sunayama *, Tomoya Matsumoto,
Yuji Hatanaka, and Kazunori Ogohara

Abstract

In recent years, data mining and text mining techniques have been frequently used for analyzing data. Electronic data is collected in everywhere and many products and services are widely used in our daily lives. Data mining techniques such as association analysis and cluster analysis are used for marketing analysis, because those can discover relationships and rules hiding in enormous numerical data. On the other hand, text mining techniques such as keywords extraction and opinion extraction are used for questionnaire or review text analysis, because those can support us to investigate consumers' opinion in text data. However, data mining tools and text mining tools cannot be used in a single environment. Therefore, a data which has both numerical and text data is not well analyzed because the numerical part and the text part cannot be connected for interpretation. Goal of the data analysis is knowledge emergence that we find or create a new knowledge for decision making.

In this paper, a mining framework that can treat both numerical and text data is proposed. Users of the proposed system can iterate data shrink and data analysis with both numerical and text analysis tools in a unique framework. Based on the experimental results, the proposed system was effectively used to data analysis for review texts of humidifiers and fan heaters. We verified that balanced use of numerical and text analysis leads to good ideas and the users should be conscious to use both type of tools and both type of data shrink.

Keywords: text mining, data mining, data analysis support, TETDM

1 Introduction

In recent years, data mining and text mining techniques have been frequently used for analyzing data. Electronic data is collected in everywhere and many products and services are widely used in our daily lives[1]. Data mining techniques such as association analysis and cluster analysis are used for marketing analysis[2, 3], because those can discover relationships and rules hiding in enormous numerical data. On the other hand, text mining[4] techniques such as keywords extraction or opinion extraction are used for questionnaire and review text analysis, because those can support us to comprehend consumers' opinion

* The University of Shiga Prefecture, Shiga, Japan

in text data. Goal of the data analysis is knowledge emergence[5] that we find or create a new knowledge for decision making.

If we can use data mining and text mining analysis tools coincidentally, we can grasp both objective patterns or rules and subjective meanings that can be the reasons of extracted rules. For example, review texts of some products are mostly consist of numerical scores and comments for the reason why the scores are given. However, if such numerical data and text data are divided and analyzed separately, we cannot figure out which comments lead to high scores or low scores. In addition to this, we have to use two systems for realizing such analysis, because most of mining systems cannot treat both numerical and text data.

In this paper, a mining framework that can treat both numerical and text data is proposed. That is, data mining tools using R[1] are embedded to a text mining system TETDM [6], Total Environment for Text Data MIning[2]. Users of the proposed system can iterate data shrink and data analysis with both numerical and text analysis tools in a unique framework.

In the traditional data mining process [7, 8], data analysis does not end at the output of a data mining tool. As shown in Figure 1, an analyst needs to collect interpretations from outputs of data mining tools as the divergence phase. After that he/she needs to integrate those collected interpretations as the convergence phase in order to emerge new knowledge that leads to a next decision making.

That is, we should activate not only computers but human intelligence to realize knowledge emergence. As in the Figure 1, data analysis tools are utilized by computers and knowledge emergence is realized by humans. Combination of data mining and text mining leads to the collection of various interpretations especially for the divergence phase.

Though numerical data is usually arranged in a matrix with the pair of item and values, most of text data is called unstructured data because it is written in natural language. Then, a method for text mining requires some transformation to a structured data [9]. It is difficult for executing effective analysis in a circumstance that a data includes both numerical values and text. Therefore, an environment that can treat both type of data is very significant.

In the rest of this paper, TETDM, total environment for text data mining, that used for the construction of the proposed framework is described in Chapter 2. The proposed framework for combining data mining and text mining is described in Chapter 3. The evaluation experiments for the proposed framework are described in Chapter 4. Related works are described in Chapter 5 and this paper is concluded in Chapter 6.

2 TETDM

TETDM, Total Environment for Text Data Mining, is used as a basic environment for constructing the proposed framework. Figure 2 shows the interface of TETDM. This interface consists of four panels and one mining tool and one visualization tool are assigned to each panel.

TETDM has about 50 mining tools and 40 visualization tools so that users can assign one of mining tools and one of visualization tools to each panel. As for the mining tools, such as key sentences extraction, keywords extraction, text clustering, text editor, on-line dictionary, typing tools and some other analysis tools are prepared. As for the visualization tools, such as text display, html text display, bar graph, line graph, table display and some other specific visualization tools are prepared.

[1] GNU R: (URL)https://www.r-project.org/
[2] TETDM: (URL)http://tetdm.jp

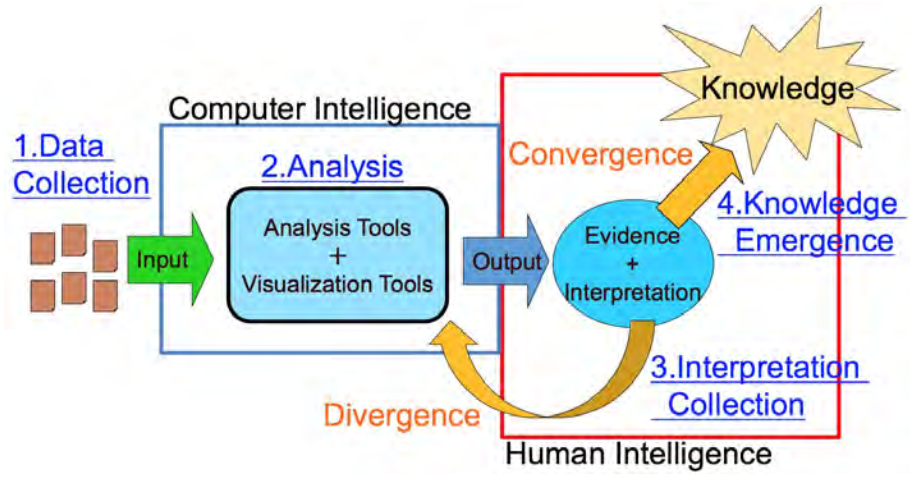

Figure 1: Procedures of Data Analysis

Users can arrange original analysis environment by deciding the number of panels and assigning mining and visualization tools to each panel. Also users can use not only tools already prepared ones but new original tools by implementation along with the TETDM specification. Since TETDM is released as an open source environment, users can customize not only tools but also the environment.

The environment can accept any kind of tools if the tools meet the TETDM specification. Therefore, we incorporate data mining tools into TETDM to realize the proposed framework.

TETDM has also an interface for knowledge emergence in the procedures of data analysis. Knowledge emergence step consists of the divergence phase and the convergence phase as described in the introduction. As for the divergence phase, TETDM has the registration window for collecting results and interpretation as in Figure 3. Each panel of TETDM has the button to call this window and users can register results and interpretations soon after he/she examines the displayed results.

As for the convergence phase, TETDM has the interface for knowledge emergence as in Figure 4. At first, as collected interpretations are arranged in the interface, users select interpretations that have some common points. After the selection, users input more general interpretation and push the combine button at the bottom of the interface. Then, the selected interpretations are combined into a single interpretation. Users iterate these procedures until all interpretations are combined into a single one that consists of one general cause and one general result.

3 Analysis Framework for Combining Data Mining and Text Mining

3.1 Target Data

In this framework, target data is required to contain both numerical/categorical data and text data. One record consists of values of items, some of them can be numerical/categorical data, and text data written in natural language. That is, what we called transaction data such

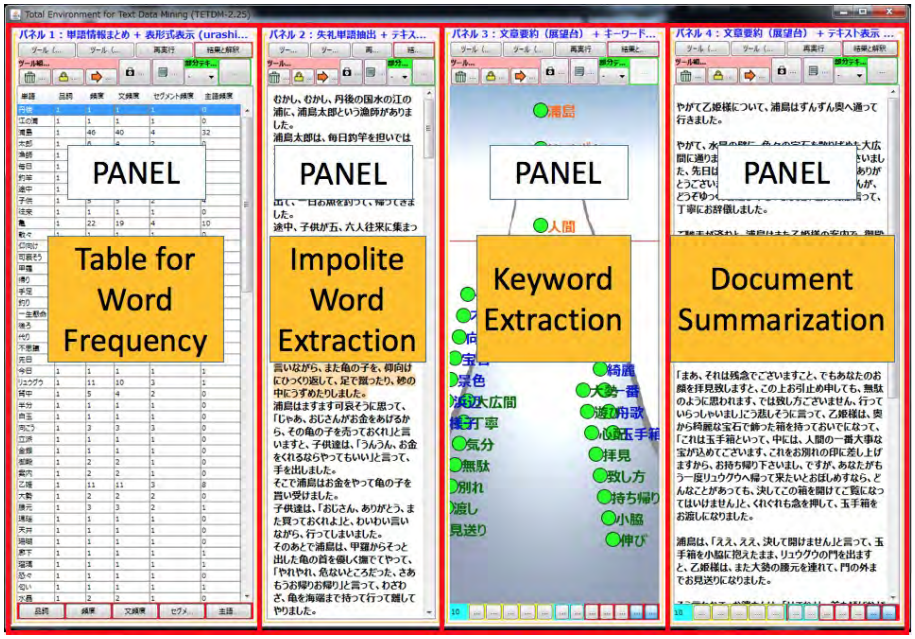

Figure 2: Outlook of Total Environment for Text Data Mining: TETDM (Users can assign a tool to each panel.)

as in Table 1 is used as a target.

Numerical data used in data mining can be collected in a circumstance that a human does not conscious to produce data such as logs of a system or sensors. On the other hand, text data used in text mining is produced by human description. In this study, absolutely separated numerical and text data do not become the target of the proposed system. That is, numerical and text data that have some connection become the target.

For example, review data that includes numerical evaluations and text comments, submitted comments to a social network service that include comments and numerical data and user profiles, examination data that includes text answers and those scores, and a data that has correspondence between a numerical part and a text part become the target.

In addition to this, the current system can cope with one text part only. Therefore, a target data must consist of numerical data set and a single text part. However, the system can be extended to cope with multiple text parts by switching each text part in the future.

3.2 Framework of Data Analysis

Figure 5 shows the framework for combining data mining and text mining. The purpose of analysis is that users acquire features or tendencies of the input data. In the process of the analysis, users iterate analysis and shrink data because most of knowledge comes from a part of data with some conditions.

Therefore, in the first step, users of the system input data that contains both numerical and text data. In the second step, users analyze the data by using data mining or text mining tools. In the third step, users shrink data by numerical or words conditions. After that, the shrunk data set is given to the tools as input again. In this loop of the analysis, both data mining and text mining are available in this framework.

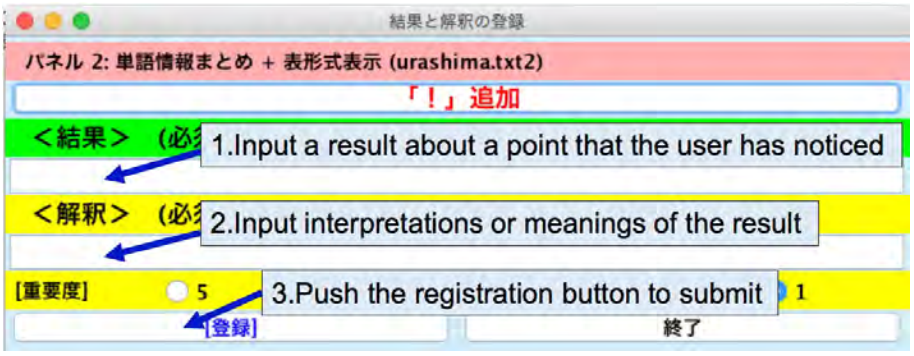

Figure 3: Registration Window for Results and Interpretation of TETDM

Table 1: Example of Target Data (Reviews of fan heaters)

ID	Age	Gender	Total	Outlook	Size	Comment
1	50	1	4	4	4	It's light, small, and convenient to move.
2	30	2	4	5	4	It's silent because wind power is weak.
3	40	2	4	5	5	I bought for replacing one that I had bought ten years ago.
4	40	2	5	5	5	It's best for me because I'm sensitive to cold. I'm satisfied with it.
5	50	1	3	3	4	I should be careful for missing to switch off.
6	50	1	5	5	5	It's powerful instead of its compact size. I'm satisfied with it.
...	...	...	...	...	...	...

3.3 Input Data Format

Input data is given as transaction data that consists of pairs of items and those values. Because TETDM can not treat such input data currently, only text part of the data is given to TETDM as an ordinary input. Then, numerical part of the data is prepared as csv format and given to the data mining tool directly [3].

In TETDM, an input text is divided into segments at the point where the specific word such as "sunaribarafuto" is inserted. A segment is also divided into sentences where periods exist. When a set of review data or questionnaire data is input, "sunaribarafuto" is inserted in order to distinguish each person.

3.4 Tools for Analysis

In this subsection, tools for data mining and tools for text mining are described.

[3]The input data format of text data and numeric data will be considered to become more convenient one.

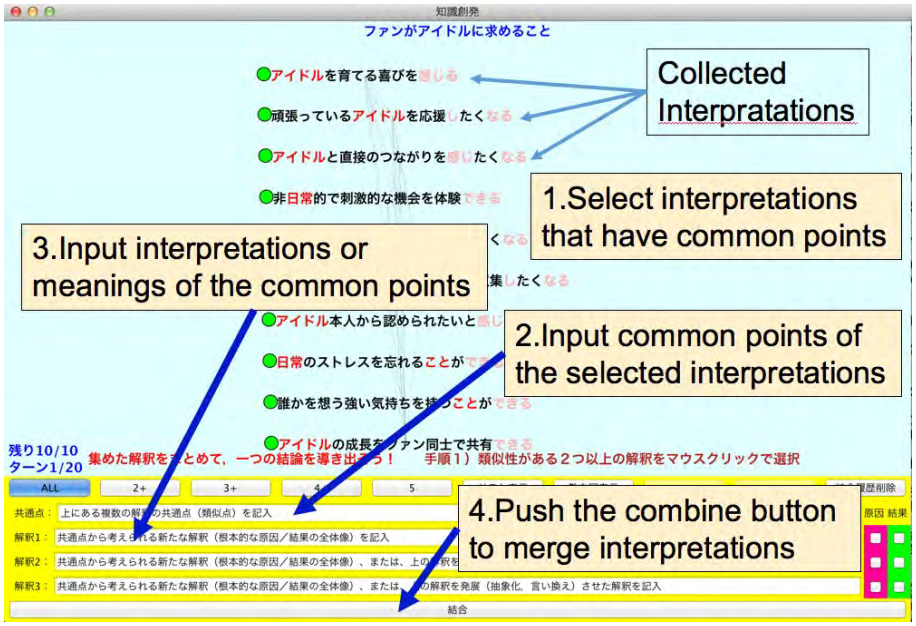

Figure 4: Interface for Knowledge Emergence of TETDM

3.4.1 Tools for Data Mining

In this study, R, a statistical analysis software, is used as a text mining tool. R has many functions for data mining and can be called from JAVA language, because TETDM is coded by JAVA. Two tools for data mining, "DataMining" and "DataMining Table" are implemented and embedded into TETDM.

Figure 6 shows the display of two data mining tools, "DataMining" and "DataMining Table." In the right part of the interface, an input data is displayed, and users can select a function to use. 14 functions are available such as average, minimum, maximum, median, variance, standard deviation, correlation, association analysis and so on. For the basic statistic values such as average and variance, uses can select the part of data table in the upper side of the panel as for the input to the functions. Correlation calculates relationships between two columns, columns mean items, for all combinations. Association analysis outputs rules of data with conditional probabilities.

Outputs of the R functions are displayed in the left panel of the interface as in "DataMining Table." That is, users select a function in the right panel, then the results are displayed in the left panel.

3.4.2 Tools for Text Mining

TETDM contains about 50 text mining tools. Though all of text mining tools supplying in TETDM are available, novice users are not easy to use those tools instantly. Therefore, in this framework, five text mining tools that can be used with data mining tools are selected for the prototype system. By using these text mining tools, users can grasp the tendencies of the whole or the part of input data.

- Word Extraction: This tool extracts input words from input texts and highlights the words.

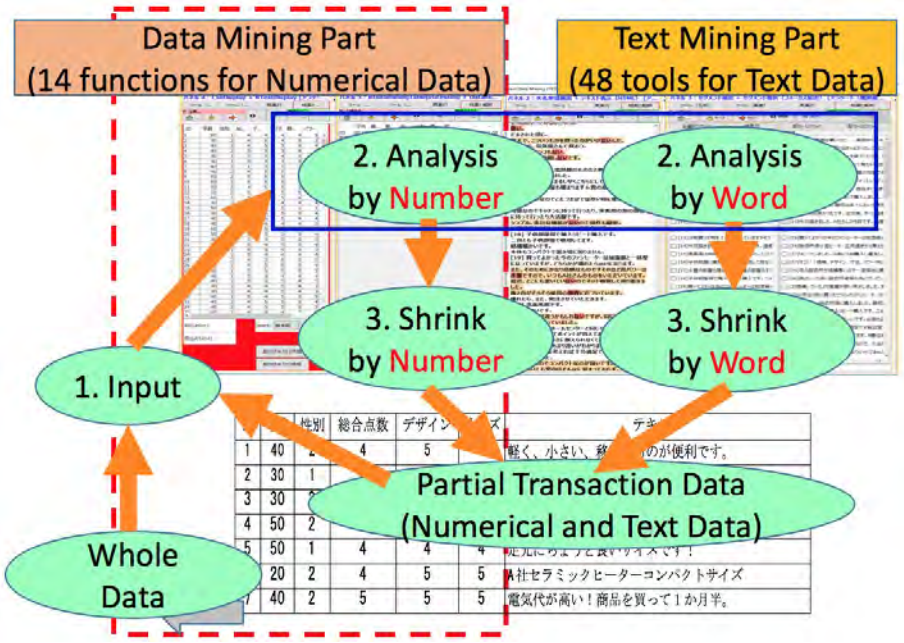

Figure 5: Framework for Combining Data Mining and Text Mining

- Text Summarization: This tool summarizes input texts by extracting important sentences [10].
- Impolite Words Extraction: This tool extracts impolite words from input texts and highlights the words.
- Text Clustering: This tool classifies input texts hierarchically by the words relationship among texts (segments) [11].
- Word Frequency: This tool outputs word frequencies from the input texts.

3.5 Data Shrink Functions

In this subsection, data shrink methods by numerical and words conditions are described. Input data is given as transaction data that consists of pairs of items and those values as in Table 1. Data shrink means extraction of partial transaction data matched with numerical or words conditions. If a data set consists of review results and a review contains evaluation values and comments, only reviews including a specified evaluation value or including a specified word in comments are extracted by this data shrink.

3.5.1 Data Shrink by Numerical Conditions

Numerical conditions can be given for data shrink by using the data mining tool "DataMining." Figure 7 shows the procedures of data shrink by the numerical conditions.

First, a user sees the displayed numerical data and selects a column to investigate. Second, the user can give a specific number or a range of the value by the form of the tool as a condition. Third, data matched with the conditions are highlighted in the data table. If

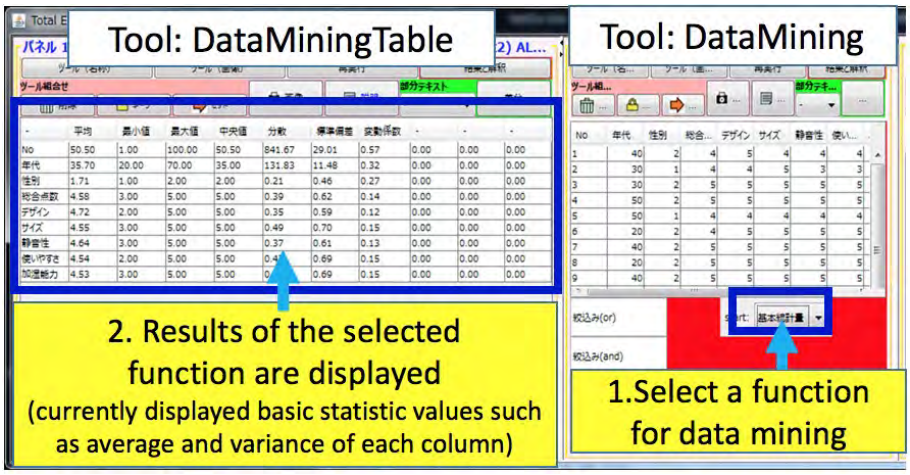

Figure 6: Data Mining Tools, “DataMining” is in the right panel and “DataMining Table” is in the left panel.

the user wants to investigate more characteristics of the highlighted data, the user can push the button at the bottom of the tool to create a partial data. Then, the user can continue to analyze by using data mining tools and text mining tools with the partial data.

By using this condition, users can investigate why some people give a specific score such as 5 points, or how about the case of women or middle age is, and so on.

3.5.2 Data Shrink by Words Conditions

Words conditions can be given for data shrink by using the text mining tool “Segment Extraction.” Figure 8 shows the procedures of data shrink by the words conditions.

First, a user sees the result of a text mining tool and finds a point to investigate. Second, the user can give words by the form of the tool as a condition. Third, segments including input words are checked in the list of segments. If the user wants to investigate more characteristics of the checked data, the user can push the button at the top of the tool to create a partial data. Then, the user can continue to analyze by using data mining tools and text mining tools with the partial data.

By using this condition, users can investigate why some people refer to a specific word such as “smell”, or how about people interested in the size of the products as “big” or “small” is, and so on.

3.6 Sample Environment for Combining Data Mining and Text Mining

In this subsection, created five sample combinations of data mining and text mining tools are described. Tool combinations are prepared by the framework shown as Figure 9. As for the data mining part, data mining tools output rules or statistical values and have a data shrink function with the numerical conditions. As for the text mining part, text mining tools output results including words information and have a data shrink function with the words conditions.

Each combination has the same DM (Data Mining) tools described in 3.4.1 for data mining analysis and data shrink with the numerical conditions, the same TM tool, “Segment

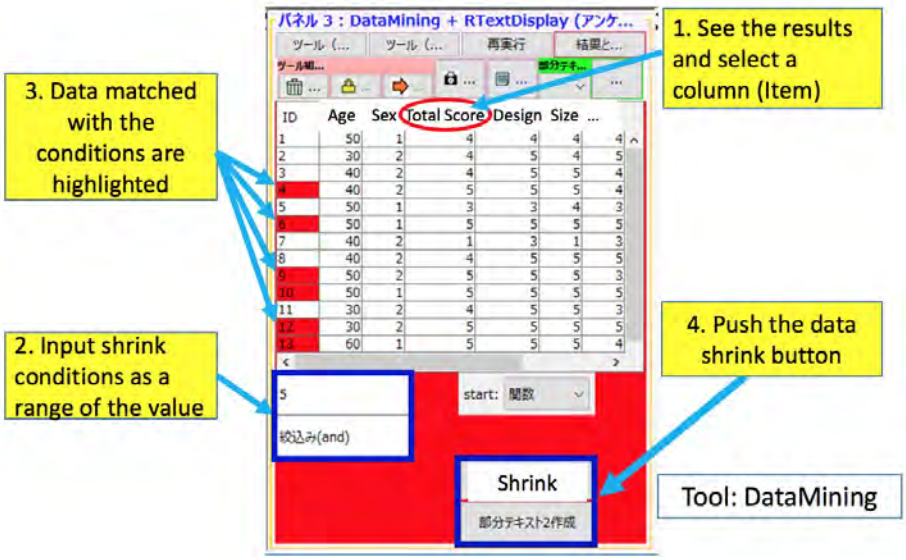

Figure 7: Data Shrink by Numerical Conditions

Extraction," described in 3.5.2 for data shrink with the words conditions, and has one of text mining tools described in 3.4.2 for text mining analysis.

In the following subsections, details of each combination are described.

3.6.1 DM and Reference of Original Document

This combination has "Data Mining," "Segment Extraction" and "Word Extraction." As described in the above, "Data Mining," and "Segment Extraction" are common tools for DM and TM for data shrink. The tool "Word Extraction" can highlight some specific words that a user inputs. Therefore, users can see results of association rules for specific segments and can find words included in segments that have a specific numerical value.

3.6.2 DM and Summarization

This combination has "Data Mining," "Segment Extraction" and "Text Summarization." The tool "Text Summarization" can extract important sentences with keywords of a document. Therefore, users can see results of association rules for specific segments and can find words included in topic sentences that have a specific numerical value.

3.6.3 DM and Impolite Words Extraction

This combination has "Data Mining," "Segment Extraction" and "Impolite Words Extraction." The tool "Impolite Words Extraction" can extract words that may be impolite. Therefore, users can see results of association rules for specific segments and can find impolite expressions in segments that have a specific numerical value.

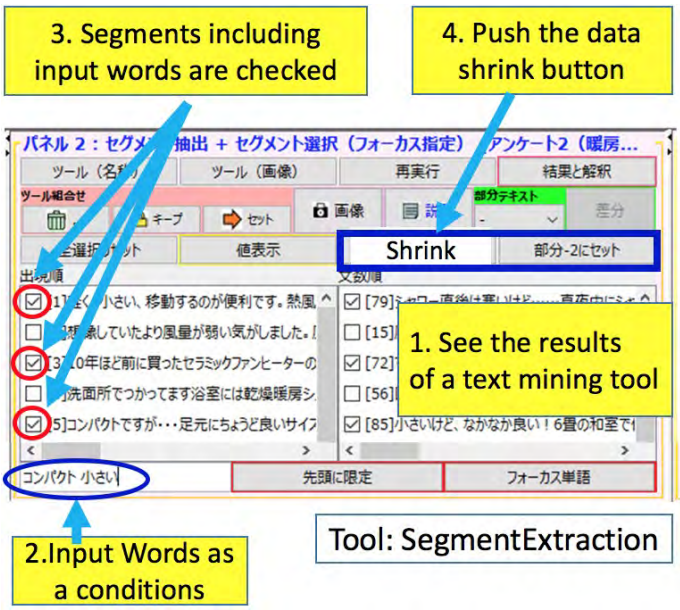

Figure 8: Data Shrink by Words Conditions

3.6.4 DM and Text Clustering

This combination has "Data Mining," "Segment Extraction" and "Text Clustering." The tool "Text Clustering" can classify segments into groups hierarchically. Therefore, users can see results of association rules for specific segments and can find relationships among segments that have a specific numerical value. Figure 10 shows the interface of this tool combination.

3.6.5 DM and Word Frequency

This combination has "Data Mining," "Segment Extraction" and "Word Frequency." The tool "Word Frequency" shows frequency of words in the table. Therefore, users can see results of association rules for specific segments and can find frequently appearing words in segments that have a specific numerical value.

3.7 Sample Procedures of DM and Summarization Combination

In this subsection, sample procedures of the proposed environment is described along with Figure 11. This environment of the figure is set as the combination of DM and Summarization described in 3.6.2. We suppose that a review data that contains numerical five-stage evaluations and a reviewer's comment. The numbers of the following procedures from four to ten are correspond to the numbers in Figure 11.

1. A review data set that contains both numerical and text data is prepared.
2. Numerical part of data is placed in the folder "csvfile" as csv format.
3. Text part of data is placed in the folder "text" as text format. Text file is input using the "File" button in the left top of the interface.

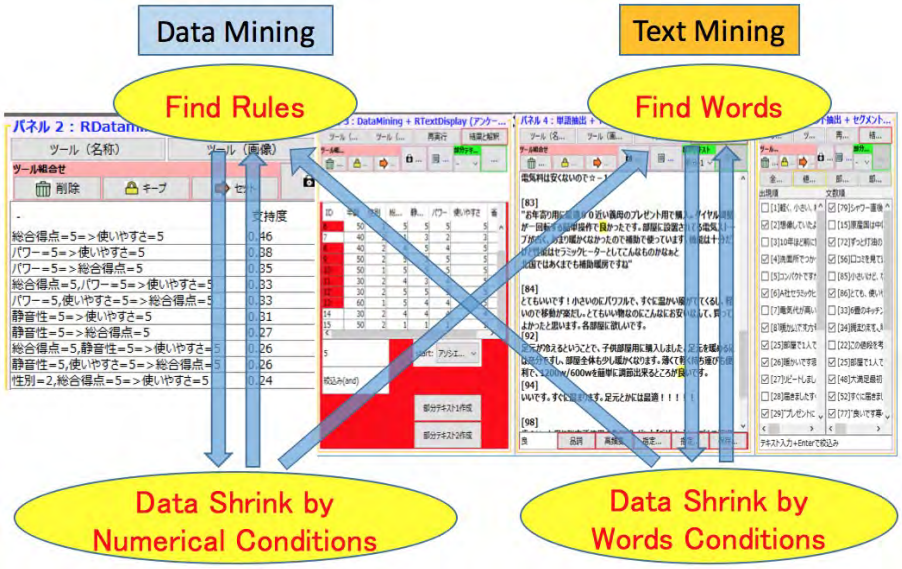

Figure 9: Framework for Preparing Tool Combinations

4. Push the “DM <-> Text Summarization” button to use the combination of DM and Summarization.

5. Output rule sets for data whose total score is 5 by association analysis.

6. Find a rule such as “the point of easy to use is 5 -> the point of total score is 5”.

7. Shrink data whose score of easy to use is 5 by data shrink using numerical conditions.

8. Refresh the display of results by the shrunk data.

9. Find words such as “convenience” and “function” that relate easy to use from the listed keywords, and summarize the texts.

10. Give an interpretation by seeing the summarization results such that this product is evaluated in its simple function.

In this way, users can iterate data shrink and interpretation by using the proposed environment.

4 Evaluation Experiments: Review Text Analysis

In this chapter, experiments for verifying the effectiveness of the proposed environment are described.

4.1 Settings

In order to verify the effectiveness of the proposed environment, we conducted experiments that the test subjects analyze review data and create ideas for new products. We supposed

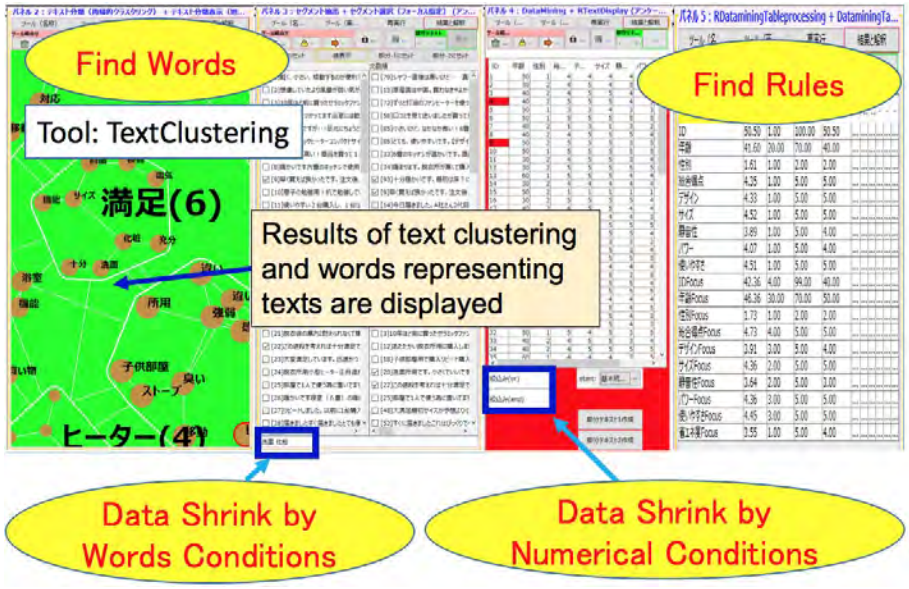

Figure 10: Interface for DM and Text Clustering

that the test subjects could create concrete ideas by using the proposed system rather than to use an ordinary one.

Then, two systems were prepared, the proposed system and a comparative system. The proposed system consists of five tool sets that contain both DM and TM described in 3.6. The comparative system consists of six tool sets, DM and five TM tools described in 3.4.2. That is, the test subjects who used the comparative system could only use DM tools and TM tools separately.

Test subjects were 16 university or graduate school students who major engineering but not so familiar to information analysis. 8 subjects were assigned to the proposed system, and the others were assigned to the comparative system.

The data sets for analysis were review texts downloaded from Rakuten Market[4]. The products were “Humidifiers” and “Fan Heaters,” and downloaded 100 reviews for each. Each review consists of numerical values for functions, design and price, and a text comment of the product.

Test subjects were instructed to analyze review texts by using the assigned system and create ideas for new products that can acquire better review results.

The following basic procedures were given to the test subjects.

1) Select a tool set for analysis.

2) Shrink data by numerical or words conditions.

3) Examine the results. Go to 4) or go back to 2).

4) Register an interpretation by using the registration window as in Figure 3. Go back to 1).

[4]Rakuten Market: (URL) http://www.rakuten.co.jp

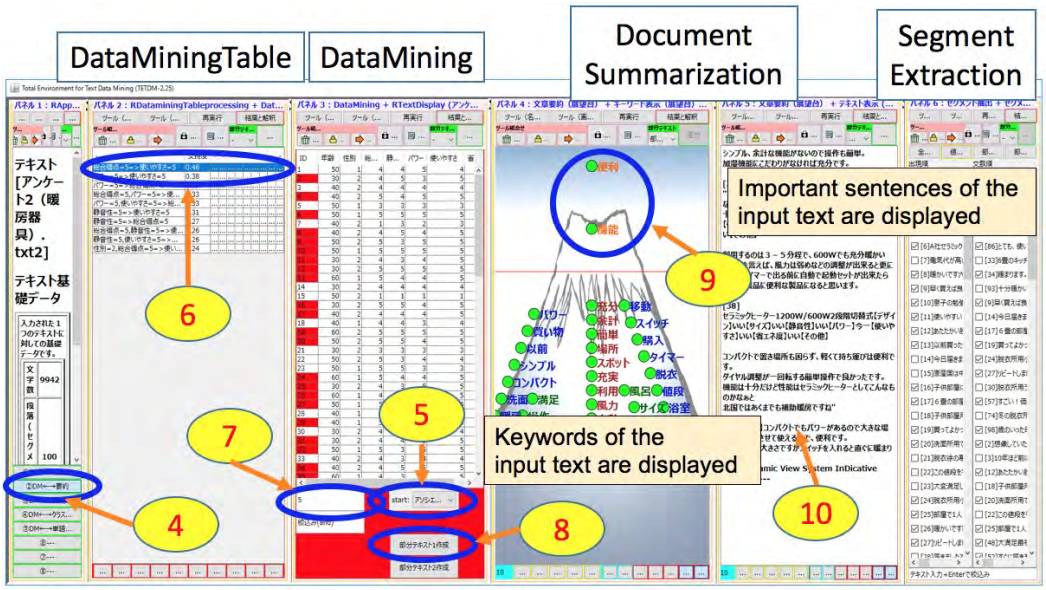

Figure 11: Interface for DM and Summarization ("Document Summarization" consists of two tools, Keyword Extraction and Summarization.)

5) After the iterations of the above procedures from 1) to 4), converge the collected interpretations by using the interface for knowledge emergence as in Figure 4 in order to create ideas for a new product.

4.2 Results

Created ideas were evaluated by a viewpoint score and a concreteness score. The viewpoint score, one point, was given to an idea that includes target customers such as gender or age, or an important viewpoint to improve. The concreteness score, one point, was given to an idea that includes concrete direction or degree of an improvement topic.

Figure 12 shows the scoring rate of the created ideas for "Humidifiers" and "Fan Heaters." Both means the rate of ideas that have acquired both the viewpoint and the concreteness scores. The rates of the proposed system were greater than those of the comparative system. Therefore, test subjects who used the proposed system could investigate and analyze the reviews deeply.

A test subject created an idea, "The product was purchased by women, and women think design is important. So the design must be cute or elaborated." This idea includes viewpoints such as "women" and "design is important," and includes a concrete idea "design must be cute or elaborated." In this case, "women" was one of the numerical data and the mention to "design" was included in the text data. Therefore, combination of data mining and text mining could lead the user to a concrete idea.

That is, data mining gives a fact that can be a ground and text mining gives an interpretation of the fact. Both a fact and an interpretation are required to create significant ideas.

Table 2 shows the rates of used tools where "DM only" means the rate of used tool combinations that consist of data mining tools only. Rates are calculated as the number of

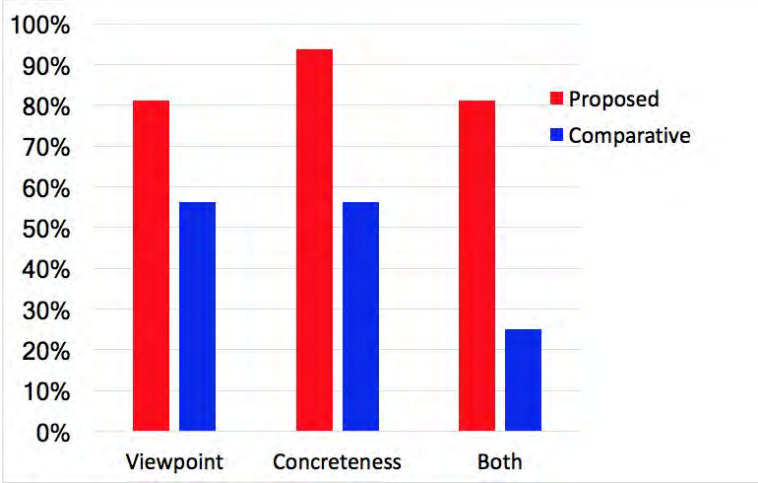

Figure 12: Scoring Rate of Created Ideas for “Humidifiers” and “Fan Heaters”

specified tool combinations out of the total number of used tool combinations. Difference means the tendency that the test subjects have not used DM and TM tools equally, and are calculated by the difference between “DM only” and “TM only.” Table 2 also shows the final score (four points maximum), the total of the viewpoint and the concreteness points for humidifiers and fan heaters, of each test subject.

As a result, the difference of the proposed group A was less than that of the comparative group B. Difference values of the test subjects whose final score was 0 or 1 were 0.76, 0.88 and 0.89. Therefore, we can conclude that a good interpretation requires balanced use of data mining and text mining tools. Most of the test subjects in the proposed group could use both type of mining tools and could have achieved reasonable ideas.

Two questions were given to the test subjects after the experiments as a questionnaire, “Q1. As for the data shrink, which data shrink function did you frequently used, numerical or words conditions?” and “Q2. As for the data analysis, which type of tools did you frequently used, data mining or text mining?” The answers were collected as five stages such that 1 and 5 were the mostly biased to one side, 2 and 4 were some biased to one side, and 3 was the similar extent. Table 3 shows the results of the questionnaire as deviation scores calculated by the absolute values of $(answered\text{-}score - 3)$.

As a result, only test subjects whose total scores were less than 2 could have achieved final score more than 3. This means that users must be conscious to use both data mining and text mining tools both in shrinking and analyzing data. In addition to this, balanced use of data mining and text mining tools in analyzing data is superior to those in shrinking data because the test subject B8 in Table 3 had achieved 3 in the final score in spite of he/she scored 2 in data shrink deviation and no one whose score was 2 in data analysis deviation could have achieved high final scores.

Table 2: Deviation Rates of Used Tool (Ascending order in Difference, A: Proposed Group, B: Comparative Group)

Test Subjects	DM only	TM only	Difference	Final Score
A1	-	0.15	0.15	4
A2	-	0.17	0.17	4
A3	-	0.21	0.21	4
A4	-	0.36	0.36	4
A5	-	0.42	0.42	4
A6	-	0.53	0.53	3
A7	-	0.73	0.73	4
A8	-	0.76	0.76	1
Average	-	-	0.42	3.5
B1	0.71	0.29	0.43	4
B2	0.14	0.86	0.71	2
B3	0.14	0.86	0.72	3
B4	0.13	0.88	0.75	2
B5	0.94	0.06	0.88	1
B6	0.95	0.05	0.89	0
B7	0.04	0.96	0.92	3
B8	1.00	0.00	1.00	3
Average	-	-	0.79	2.3

5 Related Works

Data mining tools such as R and Weka [5] exist. Also, text mining tools such as KH Coder[6] and UserLocal[7] exist. Those systems can basically treat numerical or text data only.

Though VidaMine[12] developed as an environment including all process of knowledge discovery, this framework does not include text mining tools. On the other hand, UIMA[13] is proposed as an platform to deal with various unstructured data such as text, voice and movies. However, this platform is only for text data.

Currently, though R project develops some of text mining functions[8], text mining functions are independently supplied from data mining functions.

Text mining is placed at the important term in data mining, and text mining extracts quality information [14]. That is, data mining extracts quantity information and requires quality information by words.

If we are in the occasion that we have to analyze a questionnaire data set including both numerical and text data, first we need to separate the data in numerical part and text part. Then, we input numerical data to a data mining tool, and input text data to a text mining tool. After that, we have to compare the both results for finding corresponding data.

In the analysis process using both numerical and text data, when a rule has extracted from a data mining system, we should find the grounds of the rule from the text data.

[5]Weka (URL)http://www.cs.waikato.ac.nz/ml/weka/downloading.html
[6]KH Coder: (URL)http://khc.sourceforge.net/
[7]UserLocal: (URL)http://textmining.userlocal.jp
[8]Text Mining tools for R: (URL)http://www.rdatamining.com/examples/text-mining

Table 3: Deviation Scores from Questionnaire Results (Ascending order in total deviation scores, A: Proposed Group, B: Comparative Group)

Test Subjects	Q1:Data Shrink	Q2:Data Analysis	Total(Q1+Q2)	Final Score
A4	0	0	0	4
A3	1	0	1	4
A5	1	0	1	4
A7	1	0	1	4
A1	1	1	2	4
A2	1	1	2	4
A6	1	1	2	3
A8	2	2	4	1
B7	0	0	0	3
B1	0	1	1	4
B3	1	1	2	3
B8	2	0	2	3
B6	1	2	3	0
B2	2	2	4	2
B4	2	2	4	2
B5	2	2	4	1

Therefore, we have to identify a part of text data that matches to the extracted rule.

In another case, if we have noticed a significant opinion that includes some specific words in text data, we hope to know the personal information such as gender, age, and salary expressed as numerical/categorical data.

In general, we need to shrink and analysis in both way from text mining and data mining for the effective data analysis and comprehension. In these days, deep learning methods are frequently used in various systems. Though those systems output answers with high accuracy, no grounds of the answer will be output. Some studies for interpretation of deep learning in image processing are proposed [15, 16]. However, no interpretation support for text mining is proposed and the meanings lead to our decision making must be attached by human beings no matter how much support information is supplied to us.

A study uses both text mining and data mining for examining students' online interaction [17], and an another study classifies bug reports by combining text mining and data mining [18]. Though these studies use both numerical and text data, no interactive interface for iterative data analysis is proposed. A study also combines data mining and text mining methods to detect early stage dementia [19]. This study focuses on the specific field to analyze but our study supplies a generic framework for the combination.

6 Conclusions

In this paper, a system that can treat both numerical and text data for data analysis is proposed. Based on the experimental results, users of the proposed system could have created concrete ideas.

In future works, we continue to develop a new framework that includes intuitive opera-

tions and visualization for combining data mining and text mining. In addition, we consider the method to combine numerical and text data that have no common identification numbers by calculating correlations between those data.

Though development of machine learning methods are remarkable in the artificial intelligence field, roles of human beings will never vanished. That is, we have to clarify computer's and human's roles and tasks that each can take an advantage. Thus, intelligence of computers and humans will be merged and utilized by activating such an integrated system.

References

[1] Pekka Paakkonen and Daniel Pakkala: Reference Architecture and Classification of Technologies, Products and Services for Big Data Systems, Big Data Research, Vol.2, No.4, pp.166 – 186 (2015)

[2] Savi Gupta and Roopal Mamtora: A Survey on Association Rule Mining in Market Basket Analysis, International Journal of Information and Computation Technology, Vol.4, No.4, pp.409 – 414 (2014)

[3] Pavel Turcinek and jana Turcinkova: Exploring Consumer Behavior: Use of Association Rules, Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis, Vol.63, No.3, pp.1031 – 1042 (2015)

[4] Hearst, M.A.: Untangling text mining, Proc Annual Meeting of the Association for Computational Linguistics ACL99, (1999)

[5] Masao Kakihara and Carsten Sorensen: Exploring Knowledge Emergence: From Chaos to Organizational Knowledge, Journal of Global Information Technology Management, Vol.5, No.3, pp.48 – 66 (2002)

[6] Wataru Sunayama: Knowledge Emergence using Total Environment for Text Data Mining, In Proceedings of the Joint 7th International Conference on Soft Computing and Intelligent Systems and 15th International Symposium on Advanced Intelligent Systems (SCIS & ISIS2014), Kitakyushu, TP6-2-7-(3), (2014)

[7] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth: From Data Mining to Knowledge Discovery in Databases, AI Magazine, Vol.17, No.3, pp.37–54 (1996)

[8] Ronald J. Brachman, Tom Khabaza, Willi Kloesgen, Gregory Piatetsky-Shapiro, and Evangelos Simoudis: Mining Business Databases, Communications of the ACM, Vol.39, Nol.11, pp.42 – 48 (1996)

[9] Amruta Kulkarni, Jyoti Nighot, Ashish Ramdasi: Text Mining Methodology to Build Dependency Matrix from Unstructured Text to Perform Fault Diagnosis, Proceedings of The First International Conference on Smart Trends in Information Technology and Computer Communications, pp.534–540 (2016)

[10] Wataru Sunayama and Masahiko Yachida: Panoramic View System for Extracting Key Sentences Based on Viewpoints and an Application to a Search Engine, Journal of Network and Computer Applications, Elsevier Science, Netherlands, Vol.28, No.2, pp.115–127 (2005)

[11] Wataru Sunayama, Shuhei Hamaoka and Kiyoshi Okuda: Map Interface for a Text Data Set by Recursive Clustering, In Workshop Proceedings of The 6th International Workshop on Chance Discovery(IWCD6), held with the Twenty-second International Joint Conference on Artificial Intelligence(IJCAI2011), pp.63–68 (2011)

[12] S. Kimani, S. Lodi, T. Catarci, G. Santucci and C. Sar- tori: VidaMine:A Visual Data Mining Environment, Journal of Visual Languages and Computing, Vol.15, No.1, pp.37–67 (2004)

[13] Ferrucci, D. and Lally, A. : UIMA: an architectural approach to unstructured information processing in the corporate research environment, Natural Language Engineering, Vol.10, No.3-4, pp.327–348 (2004)

[14] Yogapreethi.N, Maheswari.S: A Review On Text Mining in Data Mining, International Journal on Soft Computing (IJSC), Vol.7, No.2, pp.1–8 (2016)

[15] Scott M. Lundberg and Su-In Lee: A Unified Approach to Interpreting Model Predictions, Proceedings of Neural Information Processing Systems 30 (NIPS 2017), (2017)

[16] Raymond A. Yeh, Jinjun Xiong †, Wen-mei W. Hwu, Minh N. Do, and Alexander G. Schwing, Interpretable and Globally Optimal Prediction for Textual Grounding using Image Concepts, Proceedings of Neural Information Processing Systems 30 (NIPS 2017), (2017)

[17] Wu He: Examining students' online interaction in a live video streaming environment using data mining and text mining, Computers in Human Behavior, Vol.29, No.1, pp.90–102 (2013)

[18] Yu Zhou, Yanxiang Tong, Ruihang Gu and Harald Gall: Combining text mining and data mining for bug report classification, Journal of Software, Evolusion and Process, Vol.28, No.3, pp.150–176 (2016)

[19] Christopher Bull, Dommy Asfiandy, Ann Gledson, Joseph Mellor, Samuel Couth, Gemma Stringer, Paul Rayson, Alistair Sutcliffe, John Keane, Xiaojun Zeng, Alistair Burns, Iracema Leroi, Clive Ballard, Pete Sawyer: Combining data mining and text mining for detection of early stage dementia: the SAMS framework, Proceedings of LREC 2016 Workshop, Resources and Processing of Linguistic and Extra-Linguistic Data from People with Various Forms of Cognitive/Psychiatric Impairments (RaPID-2016), pp. 35 - 40 (2016)