IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 4, April 2022

DOI: 10.17148/IJARCCE.2022.114207

Data Analysis Support by Combining Data Mining and Text Mining

Pooja J. Shirure

Department of Computer Science, Universal College of Engineering and Research, Pune.

Abstract: In recent years, data mining and text mining techniques have been frequently used for analyzing questionnaire and review data. Data mining techniques such as association analysis and cluster analysis are used for marketing analysis, because those can discover relationshipsand 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 textanalysis, 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 text part cannot be connected for interpretation.
In this paper, a mining framework that can treat both numerical and text data is proposed. We can iterate data shrink and data analysis with both numerical and text analysis tools in the unique framework. Based on experimental results, the proposed system was effectively used to data analysis for review texts.

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

INTRODUCTION:

have been frequently used for analyzing questionnaire and review data. 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. Onthe other hand, text mining 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 in text data. If we can use data mining and text mining analysis coincidentally, we can grasp both objective patterns or rules and subjective meanings that can be the reasons of extracted rules.
However, 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 R1 are embedded to a text mining system TETDM [1],
Total Environment for Text Data MIning2. We can iterate data shrink and data analysis with both numerical and text analysis tools in the unique framework.

II. TETDM

TETDM, Total Environment for Text Data Mining, is used as a basic environment for constructing the proposed framework. This interface consists of four panels and each panel has one mining tool and one visualization tool. TETDM has about 40 mining tools and 40 visualization tools, so users can assign one of mining tools and one of visualization tools to each panel.
Currently, though TETDM has only tools for text mining, the environment can accept anykind of tools if the tools meet the TETDM specification. Therefore, we incorporate data mining tools into TETDM to realize the proposed framework.

III. ANALYSIS FRAMEWORK FOR COMBINING DATA MINING AND TEXT MINING

A. Target Data

In this framework, target data contains both numerical/ categorical data and text data. Onerecord consists of values of items, and some of values can be numerical/categorical data
and text data written in natural language. That is, what we called transaction data such as inTABLE. I is used as a target.

B. Framework of Data Analysis

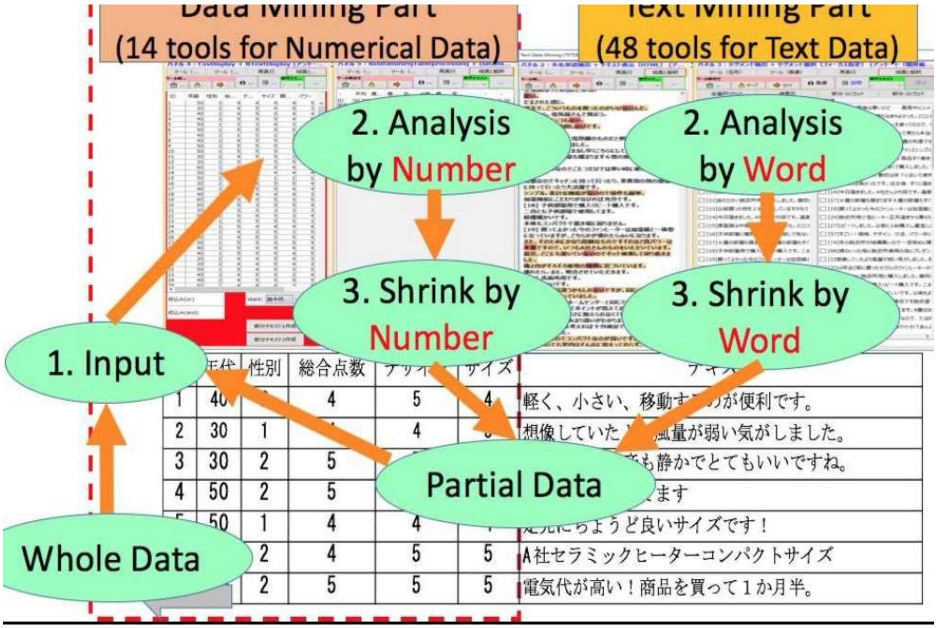


Fig.2 shows the framework for combining data mining and text Mining.

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 areavailable in this framework.

D. Tools for Analysis

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

- 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 byJAVA. Two tools for data mining, “DataMining” and “DataMining Table” are implemented and embedded into TETDM.

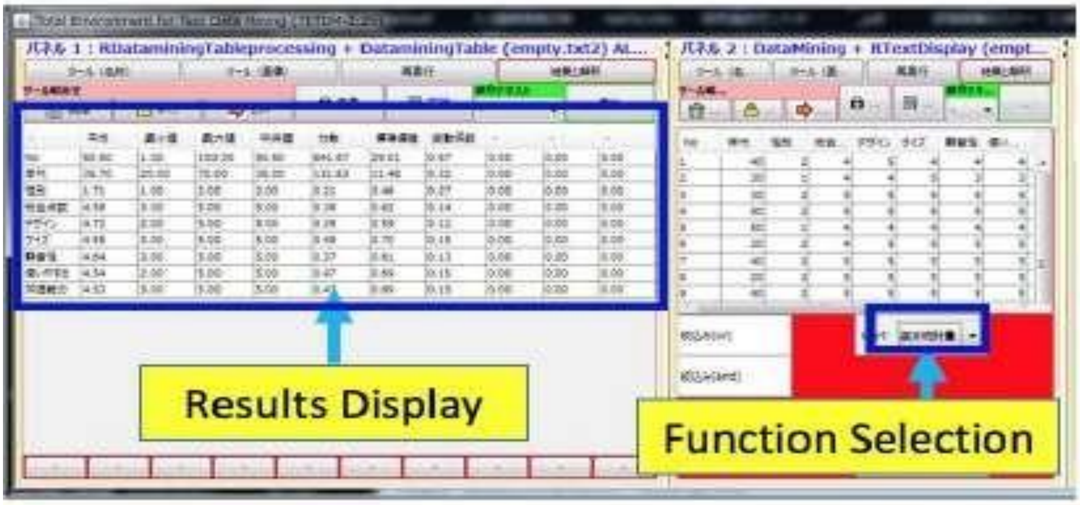

Total Environment for Text Data Mining (TETDM)
パネル 1 : RDataminingTableprocessing + DataminingTable (empty.txt2) All...
パネル 2 : DataMining + RTextDisplay (empt...
Results Display
Function Selection

Fig. 3 shows the display of two data mining tools

DOI: 10.17148/IJARCCE.2022.114207

“DataMining” and “DataMining Table.” In the right part of the interface, input data isdisplayed, and users can select a function to use. 14 functions are available such as average,
minimum, maximum, median, variance, standard deviation, correlation, association analysisand 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. Correlationcalculates 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 “DataMiningTable.” That is, users select a function in the right panel, then the results are displayed in
the left panel.

- **Tools for Text Mining:**

TETDM contains about 40 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 thisframework, five text mining tools that can be used with text mining tools are selected for
the prototype system. By using these text mining tools, users can grasp the tendencies ofthe whole or the part of input data.

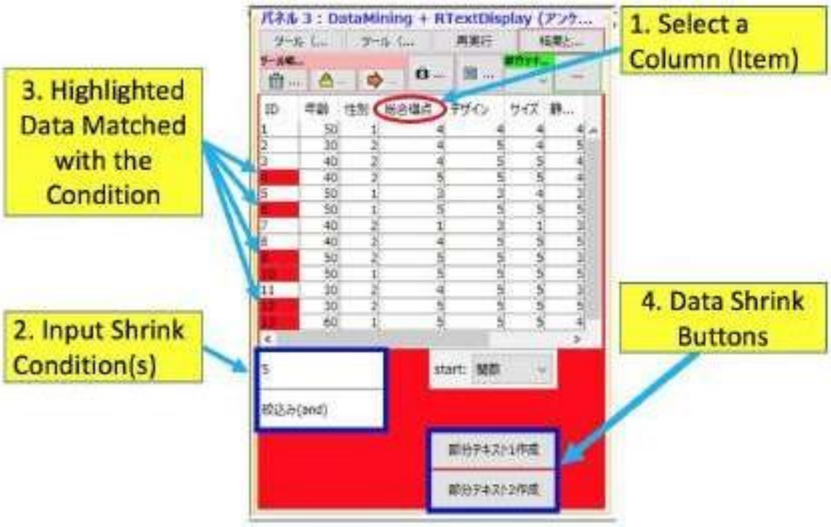
- Word Extraction: This tool extracts input word from input texts and highlight the word.
- Text Summarization: This tool summarizes input texts by extracting important sentences.
- Impolite Words Extraction: This tool extracts impolite words from input texts and highlightthe words.
- Text Clustering: This tool classifies input texts hierarchically by the word relationship among texts (segments).

E. Data Shrink Functions

In this section, data shrink methods by numerical andwords conditions are described.

1) Data Shrink by Numerical Conditions: Numerical conditions

can be given for data shrink by using the data mining tool “DataMining.”


パネル3：DataMining + RTextDisplay (アンケ...
ツール (...
ツール (...
再実行
結果と...
1. Select a Column (Item)
3. Highlighted Data Matched with the Condition
2. Input Shrink Condition(s)
4. Data Shrink Buttons
5
絞込み(and)
start: 開始
部分テキスト1作成
部分テキスト2作成

ID	年齢	性別	総合得点	デザイン	サイズ	静...
1	50	1	4	4	4	4
2	30	2	4	5	4	5
3	40	2	4	5	5	4
4	40	2	5	5	5	4
5	50	1	3	3	4	3
6	50	1	5	5	5	5
7	40	2	1	3	1	3
8	40	2	4	5	5	5
9	50	2	5	5	5	3
10	50	1	5	5	5	5
11	30	2	4	5	5	3
12	30	2	5	5	5	5
13	60	1	5	5	5	4

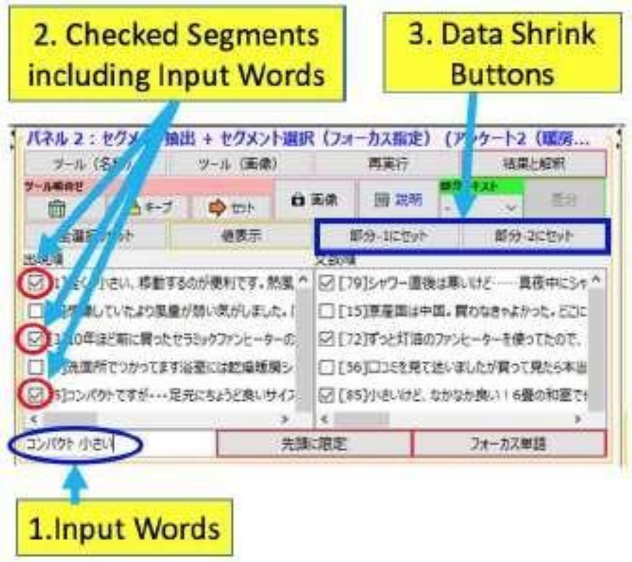
Fig. 3 shows the procedures of data shrink by the numerical conditions.

First, a user sees the numerical data and notices a point to investigate. Second, a user can give a specific number or a range of number by the form of the tool. After that, shrunk data are highlighted in the data table. If a user wants to investigate more characteristics ofshrunk data, the user can push a button at the bottom of the tool to create partial data.
Then, the user can continue to analyze by data mining or text mining tools with the partial
data. By using this condition, users can investigate why some people give a specific scoresuch as 5 points, or how about the case of women or middle age, and so on.

2) Data Shrink by Words Conditions:

Words conditions can be given for data shrink by using the text mining tool “SegmentExtraction.”

DOI: 10.17148/IJARCCE.2022.114207



☑ [1]とても小さい。移動するのが便利です。熱風
☐ [2]想像していたより風量が弱い気がしました。
☑ [3]10年ほど前に買ったセラミックファンヒーターの
☐ [4]洗面所でつかってます浴室には乾燥暖房シ
☑ [5]コンパクトですが・・・足元にちょうど良いサイズ

☑ [79]シャワー直後は寒いけど……真夜中にシャ
☐ [15]原産国は中国。買わなきゃよかった。どこに
☑ [72]ずっと灯油のファンヒーターを使ってたので、
☐ [56]口コミを見て迷いましたが買って見たら本当
☑ [85]小さいけど、なかなか良い！6畳の和室で

Fig. 4 shows the procedures of data shrink by the words conditions.

First, a user sees the results of text mining and notices a point to investigate. Second, a usercan give words by the form of the tool. After that, shrunk data are checked in the list of segments. If a user wants to investigate more characteristics of shrunk data, the user can push a button at the top of the tool to create partial data. Then, the user can continue to analyze by data mining or text mining tools with the partial data.

By using this condition, users can investigate why some people refers to a specific wordsuch as “smell”, or how about people who are interested

F. RELATED WORKS

Data mining tools such as R and Weka 4 exist. Also, text mining tools such as KH Coder5 andUserLocal6 exist. Those systems can basically treat numerical or text data only.

Currently, though R project develops some of text mining In another case, if we have noticed a significant opinion that includes some specific words in text data, we hope toknow 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 forthe effective data analysis and comprehension. A study uses both text mining and data mining for examining students’ online interaction.

CONCLUSION:

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 operations and visualization for combining data mining and text mining. Intelligence of computers and humans will be merged and utilized by activating such an integrated system.

REFERENCES:

1. 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 &4. ISIS2014), Kitakyushu, TP6-2-7-(3), (2014)
2. Yogapreethi.N, Maheswari.S: A Review On Text Mining in Data Mining, International Journal on Soft Computing (IJSC), Vol.7, No.2,6. pp.1–8 (2016)
3. 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)
4. 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, pp.150–176 (2016)