1A3-IOS-3a-2

Graphical Interface that Supports Users' Trial-and-Error Process of Text Mining

Naoya Otsuka [*1] Mitsunori Matsushita [*2]

[*1] Graduate School of Informatics, Kansai University
[*2] Faculty of Informatics, Kansai University

This paper proposes an interface for supporting a user's activity of mining useful information from a large amount of text. Finding useful knowledge from texts requires a trial-and-error process; thus, a user's information request becomes clearer with the execution of the process. Our proposed interface attempts to expedite this process. When a user performs a text mining task, it is necessary for him or her to use combinations of different text analysis tools to analyze the text. In order to combine different text mining tools, the Total Environment for Text Data Mining (TETDM) was recently proposed. However, currently, the interface of TETDM is not sufficiently good to perform such a task; it does not provide a means to combine/switch text analysis tools smoothly. In this study, we redesigned the interface of TETDM such that it enables smooth switching/combining of text analysis tools. The proposed interface includes a graph in which nodes denote the tools and links denote the processing flow between them. The interface facilitates (1) combining/switching text analysis tools by directly manipulating nodes, and (2) understanding the current state of each tool.

1. Introduction

Large amounts of text data can be found via the World Wide Web and stored on computers. Such a text is used for browsing and as information resources for finding useful knowledge based on its analysis. Text mining is a technique for finding useful knowledge from unstructured text data. This is a complex technique that includes natural language processing, data mining, and information visualization, which are used adaptively to find useful knowledge. In addition many techniques have been developed, such as keyword extraction, important sentence extraction, automatic document summarization, and document clustering. Other examples of the extraction of useful knowledge and novel information by text mining have been proposed.

In general, text mining is not a goal-oriented task because the user's informational requirements are clear when they start their exploration. To extract useful knowledge from a large text collection, a user is required to perform exploratory information analysis, such as repeated trial-and-error processes and viewing a text collection from a variety of perspectives. When a user performs a text mining task, it is necessary to use combinations of different text analysis techniques to analyze the text. However, few methods of information access can satisfy the variety of requests made by people who want to acquire useful knowledge from a large volume of text data. In addition, a variety of information visualization techniques have been proposed but an environment is not available that allows users to take advantage of these techniques.

To address this problem, the Total Environment for Text Data Mining (TETDM) was proposed recently [Sunayama 11a]. TETDM provides an environment in which users can combine a variety of text analysis tools. TETDM aims to support users who want to mine for useful information in a large volume of text. This type of task is known as exploratory data analysis.

However, the TETDM interface is not adequate for performing these types of tasks. In this paper, we propose a redesigned interface for TETDM. The proposed interface includes a graph where the nodes denote the tools while links denote the process flow between nodes. This interface facilitates: (1) combining/switching text analysis tools by manipulating nodes directly, and (2) understanding the current state of each tool.

2. Related Work

In this section, we describe the characteristics of text mining and how TETDM was designed to perform this task.

2.1 Text Mining

Text mining is a complex technique that combines various techniques such as natural language processing, data mining, and information visualization. Text mining aims to extract useful information and novel knowledge from unstructured text data using a combination of these techniques. To improve the performance of information extraction from a text, natural language processing is an indispensable preprocessing step for text mining, because text mining aims to analyze unstructured data, i.e., raw text data [Rajman 97].

Table 1 shows the relationships between text mining, data mining, and information retrieval [Hearst 99]. Importantly, text mining differs from information retrieval. An information retrieval system does not facilitate the discovery or derivation of new information from data. Hearst stated that: "The fact that an information retrieval system can return a document that contains the information a user requested implies that no new discovery is being made: the

Contact: Mitsunori Matsushita, Faculty of Informatics, Kansai University, 2-1-1 Ryozenji-cho, Takatsuki-shi, Osaka, 569-1095, JAPAN, Tel: +81-72-690-2437, Fax: +81-72-690-2491, E-mail: mat@res.kutc.kansai-u.ac.jp

Table 1: A classification of data mining and text data mining applications [Hearst 99]

	Finding Patterns	Finding Nuggets	
		Novel	Non-Novel
Non-textual data	standard data mining	?	database queries
Textual data	computational linguistics	real TDM	infomation retrieval

information had to have already been known to the author of the text; otherwise the author could not have written it down." In addition, data mining is a technique used to find patterns in large volumes of data. Consequently, information retrieval and data mining do not contribute to finding novel information. Text mining can not only extract well-known information and finds patterns but also find novel information in a large volume of text.

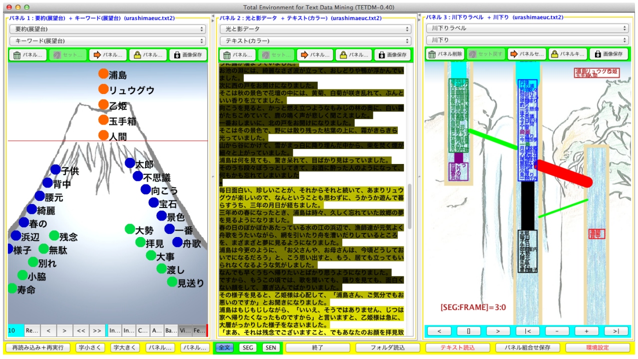
According to Nasukawa, text mining provides knowledge candidates that contribute to users awareness [Nasukawa 09]. Considering a variety of information other than the data for analysis, users will determine whether the acquired awareness is useful. Text mining is not a technique that finds knowledge automatically, but one that enables users to find knowledge. They find knowledge by manipulating the system interactively. Nasukawa also reported a case where automobile failure information was analyzed using text mining, which demonstrated the importance of trial-and-error during text mining.

2.2 Exploratory Data Analysis

When a user begins a text mining task to extract useful knowledge from text, their information request might not be clear. Thus, a user's information request is clarified by repeating a trial-and-error process. The request may also change. This is also the case when a user has a predetermined request. This type of task belongs to the category of exploratory data analysis [Hartwig 79]. Hearst described text mining as quoted in the previous section. He also stated that text mining is a form of exploratory data analysis from the view point of finding novel information and information to solve a problem [Hearst 99]. When performing exploratory data analysis using a computer, a user iterates the following processes: (1) a user has an ambiguous idea of what they want to find in the data so they submit a query to the computer; (2) the computer retrieves a result based on the query; (3) based on the result returned by the computer, the user obtains a better understanding of the data and formulates a new query. The user gradually collects useful information related to problem-solving and decision-making via these processes [Matsushita 05a]. Hence the facilitation of these processes is necessary to support exploratory data analysis.

2.3 Total Environment for Text Data Mining (TETDM)

Various text mining methods and basic techniques have been proposed. However, the tools proposed based on these methods are frequently tentative or unavailable. During text mining, a user has to apply a combination of different text analysis tools. This may be a burden for the user who has to build each tool separately, format the data for each tool, and implement an interface to compare the analytical results generate by each tool. Therefore, it is not easy for the user to perform text mining with a variety of text analysis tools.



お池の淵には、綺麗なさざ波が立って、おしどりや鴨が浮かんでいました。
次に西の戸をお開けになりました。
そこは秋の景色で花壇の中には、黄菊、白菊が咲き乱れて、ぷんといい香りを立てました。
向こうを見ると、かっと燃え立つようなもみじの林の奥に、白い霧がたちこめていて、鹿の鳴く声が悲しく聞こえました。
一番おしまいに、北の戸をお開けになりました。
そこは冬の景色で、野には散り残った枯葉の上に、霜がきらきら光っていました。
山から谷にかけて、雪がまっ白に降り埋んだ中から、柴を焚く煙が細々と上がっていました。
浦島は何を見ても、驚き呆れて、目ばかり見はっていました。
そのうち段々ぼうっとしてきて、お酒に酔った人のようになって、何もかも忘れてしまいました
■
毎日面白い、珍しいことが、それからそれと続いて、あまりリュウグウが楽しいので、なんということも思わずに、うかうか遊んで暮らすうち、三年の月日が経ちました。
三年めの春になったとき、浦島は時々、久しく忘れていた故郷の夢を見るようになりました。
春の日のぽかぽかあたっている水の江の浜辺で、漁師達が元気よく舟歌をうたいながら、網を引いたり舟を漕いだりしているところを、まざまざと夢に見るようになりました。
浦島は今更のように、「お父さんや、お母さんは、今頃どうしておいでになるだろう」と、こう思い出すと、もう、居ても立ってもいられなくなるような気がしました。
なんでも早くうちへ帰りたいとばかり思うようになりました。
ですから、もうこの頃では、歌を聞いても、踊りを見ても、面白くない顔をして、塞ぎ込んでばかりいました。
その様子を見ると、乙姫様は心配して、「浦島さん、ご気分でもお悪いのですか」とお聞きになりました。
浦島はもじもじしながら、「いいえ、そうではありません、じつは家へ帰りたくなったものですから」と言いますと、乙姫様は急に、大層がっかりした様子をなさいました。
「まあ、それは残念でございますこと、でもあなたのお顔を拝見致

Figure 1: Total Environment for Text Data Mining.

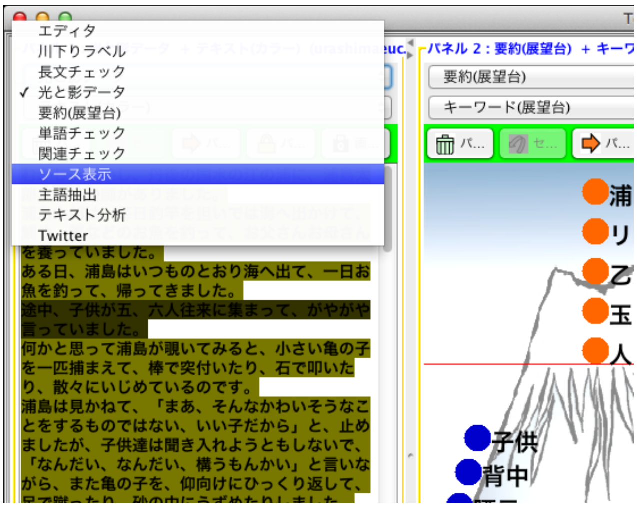
To address this problem, TETDM was proposed as a system for performing text mining [Sunayama 11a]. In TETDM, multiple text analysis tools can be displayed horizontally (Figure 1). In addition, each text analysis tool performs separate mining and visualization processes, which are implemented as modules in the system. Thus, users can perform a variety of text analyses by combining/switching modules depending on their requirements.

The following section describes examples of the tools that are available in TETDM.

2.4 Available Tools on TETDM

Nishihara et al. presented a method for visualizing text based on topic relevance [Nishihara 11] where the background of each sentence was colored from black to yellow depending on the relevance of each sentence, which supported the user's understanding of the text. In TETDM, this method was implemented by combining *MakeLight* (Item 1 in Table 2) with *ScoreDist* (Item 1 in Table 3) or *TextDisplayColor* (Item 3 in Table 3). The central area in Figure 1 shows a visualization of the results generated by this method.

A system that evaluates the subject relevance of each sentence has been proposed [Sunayama 11b]. This system requires two types of input data: text and a set of terms that it contains. In TETDM, this method is implemented by combining *LabelData* (Item 2 Table 2) and *FlowPanel* (Item 4 in Table 3). It also uses the output data from *Panoramic* (Item 3 in Table 2) as input data, which is a set



を養っていました。
ある日、浦島はいつものとおり海へ出て、一日お魚を釣って、帰ってきました。
途中、子供が五、六人往来に集まって、がやがや言っていました。
何かと思って浦島が覗いてみると、小さい亀の子を一匹捕まえて、棒で突付いたり、石で叩いたり、散々にいじめているのです。
浦島は見かねて、「まあ、そんなかわいそうなことをするものではない、いい子だから」と、止めましたが、子供達は聞き入れようともしないで、「なんだい、なんだい、構うもんかい」と言いながら、また亀の子を、仰向けにひっくり返して、足で蹴ったり、砂の中にうずめたりしました。

Figure 2: Interface for switching/combining a module.

of terms. The right area in Figure 1 shows a visualization of the results generated using this method.

Users can combine the same mining modules or different mining modules, depending on their requirements. TETDM aims to build an environment that can use a variety of text analysis tools by integrating text analysis techniques from existing and future studies into a single environment. Thus, users will not have to perform extra work so they can focus on their analysis.

3. Problems of the TETDM interface

As mentioned earlier, text mining is a form of exploratory data analysis based on a trial-and-error process. The TETDM system aims to perform this task. The interface should allow the user to perform the trial-and-error process smoothly. However, the current interface of TETDM is not adequate for performing this task. In this section, we describe the problems of the TETDM interface.

3.1 Module Selection

During text mining using the TETDM system, the user must be able to select or switch a module that suits their needs. The system uses drop-down lists as an interface to select or switch each text analysis tool (Figure 2). However, the user cannot access all of the available modules until a list has been opened. This prevents the smooth performance of a task by the user. A user cannot execute text mining smoothly if they cannot access all of the available analysis tools and the results generated using each tool.

3.2 Input/Output Data for Each Module

In the TETDM system, a user can apply combinations of text mining techniques in a flexible manner. To achieve this, each text analysis technique is separated into two processes, i.e., the mining process and visualization process, and each process is implemented as a module in the system. By combining mining modules and visualization modules, the user can use them as text analysis tools. Furthermore, it is possible to view the analysis results from various perspectives and to compare the analytical results by switching the visualization module for a mining module or the mining module for a visualization module. Current modules in TETDM system, however, are quite restricted in terms of their combination. These modules are not available to use in flexible combinations with other modules for various reasons. Thus, modules can be combined only with a specified module where a mining module has two mining processes and a specified module.

3.3 Data Flow between Modules

In the TETDM system, the text data input into the system is processed using three steps: (1) preprocessing by natural language processing such as morphological analysis; (2) processing using a mining module; (3) processing using a combined visualization module with a mining module. In addition, the system has a function that allows interactions with the input/output data between a module and other uncombined modules. Concequently, the data flow required to input the text data from the displayed analytical result is highly complex. Furthermore, with the exception of the flow to/from modules in current use, the data flow is not indicated explicitly to a user. For this reason, it is difficult for a user to understand how their current analysis will affect other analytical results. Hence, an unintended operation performed by the user may interfere with their trial-and-error process.

4. Redesign

In the previous section, we described the problems of TETDM from the perspective of the user's trial-and-error process during text mining. To address these problems, we considered the requirements for the TETDM interface where users need the following features: (1) to switch/combine modules smoothly, (2) to access each module in use, and (3) to access the data flows between modules. In this section, we describe the design guidelines from two perspectives: (1) the user and (2) the data, to ensure the development of a satisfactory interface.

4.1 User's Behavior

During a text mining task in the TETDM system, a user analyzes text by combining/switching a variety of module in their trial-and-error process. To execute these analytical processes effectively, the TETDM interface should not interfere with the user's trial-and-error process. To facilitate the user's trial-and-error process, an interface containing the requisite switching/combining modules should be intuitive. It is also assumed that a user employs a combination of text analysis tools to analyze texts because the contents of the analytical procedure may be highly variable. Consequently, a user must be able to access their analytical process to appreciate their current state and to choose their requisite analysis.

To respond to requests, our proposed interface includes a graph where the nodes denote the modules and the links denote the process flow between nodes (see Figure 3). The proposed interface allows a user to switch/combine text analysis tools by manipulating the nodes directly. This is

Table 2: Available Mining Modules (TETDM version 0.34)

	Module Name	Description
1	*MakeLight*	evaluate the subject relevance of each sentence
2	*LabelData*	evaluate the coherence of a text
3	*Panoramic*	extract characteristic terms and sentences
4	*SubjectExtraction*	extract subject terms
5	*PaperCheck*	extract designated phrases
6	*LongSentenceCheck*	extract long sentences from a document
7	*TextInfo*	summarize the results processed by other modules
8	*Twitter*	twitter search results as input text data
9	*RelevancesKM*	output the relevance of each term
10	*ToTagData*	output data for *TagCloudView* (Item 7 in Table 3) based on the term frequency
11	*AnnotationMining*	extract sentences to be annotated based on the importance of each sentence

Table 3: Available Visualization Modules (TETDM version 0.34)

	Module Name	Description
1	*ScoreDist*	display each sentence weight as a bar chart
2	*TextDisplay*	display text
3	*TextDisplayColor*	display text with terms and sentences with a background color
4	*FlowPanel*	for *LabelData* (Item 2 in Table 2)
5	*XDrawDisplay*	display keywords based on their importance
6	*TextDisplayHtml*	display HTML text
7	*SimpleKeywordMap*	display a keyword map
8	*TagCloudView*	display a tag cloud

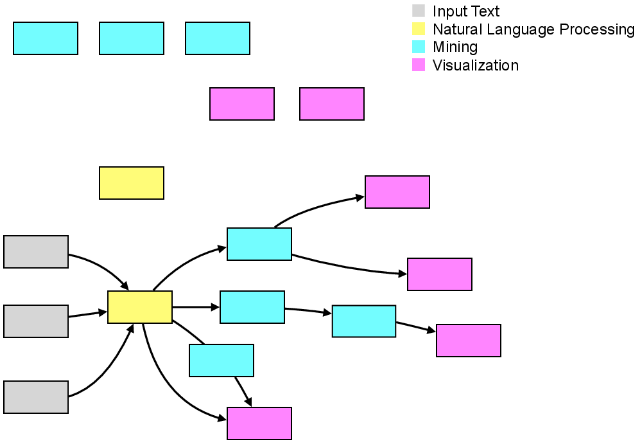


Figure 3: Concept of Proposal Interface.

achieved intuitively by manipulating the nodes directly as objects.

4.2 Data Format Limitation Between Modules

In the TETDM system, the text mining process has three stages: natural language processing, data mining, and information visualization. Each text analysis tool is implemented by combining mining modules and visualization modules. It imposes the implementation of a single module to perform multiple processing on module developers if they implement a text mining tool that requires complex processing. This reduced the availability of a module in a wider usage.

In our proposed interface, we reorganize each module to perform single processes by setting constraints on their input/output data format. Consequently, a user can combine or switch each module in a flexible manner. As shown in Figure 3, it can also connect a series of many mining modules. We can also consider the natural language processing as a module like a mining module and a visualization module. Each module has input and output formats and the pairing of combinable modules is determined by them. For example, *Panoramic* (Item 3 in Table 2) is a mining module for automatic document summarization where the input format is a text document while its morphological analysis results and output formats are keywords and a summarized text, respectively. The modules that can be combined with *Panoramic* are determined by *XDrawDisplay* and *TextDisplay* in each input format (shown at the top of Figure 4). However, *Panoramic* includes multiple processes and has multiple data output formats for various processes. We propose that one module performs one process. For example, we can split the module into three processes, as shown at the bottom of Figure 4: (1) keyword extraction, (2) important sentence extraction, and (3) automatic document summarization. By organizing the input/output data format of modules, users can combine more modules in a flexible manner. As shown in Figure 3, the user can access all

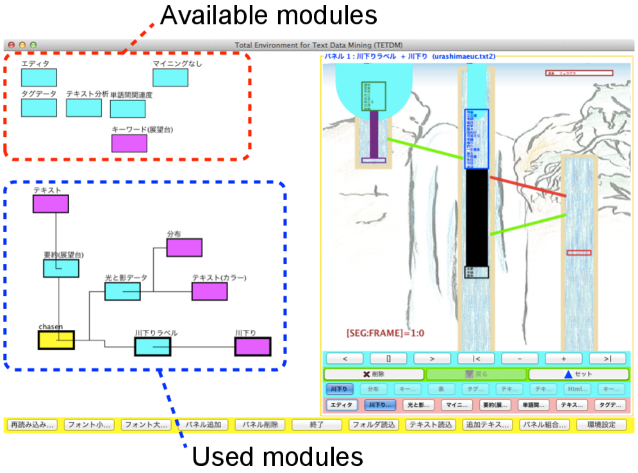


Figure 5: The Prototype System.

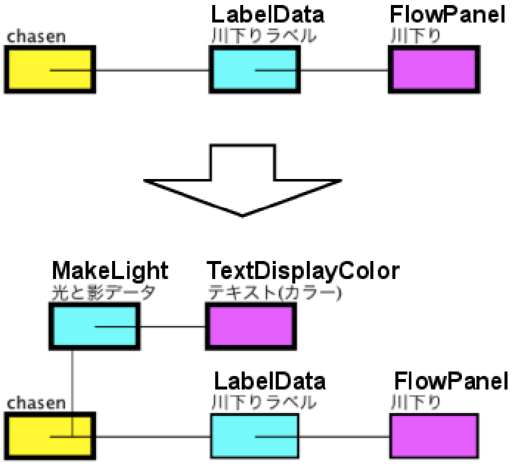


Figure 6: Add modules.

of the analytical processes involving the input text and the results as simple paths.

5. Implementation

In this section, we describe a prototype system, which we implemented based on our design guidelines. The prototype system was implemented using TETDM system version 0.34.

5.1 Overview

In the proposed interface, we use a graph where the nodes denote each module and the links denote the process flow. This allows users to understand how each module is used. In addition, we made the switching module more intuitive by manipulating a node directly, which denotes a module. In this implementation, the system cannot display multiple text analysis tools at the same time. Because we focused on facilitating the smooth switching and combining of modules to help the user understand how the modules are used.

Figure 5 shows the prototype system with the proposed interface. The system comprises a tool panel, which is displayed as a visualization result in the basic system (Figure 5, right side) with the proposed interface (Figure 5, left side). The rectangles on the interface indicate the currently available modules, which are nodes in the graph. The module names are associated with each node. The TETDM system uses two types of modules: mining modules and visualization modules. We also added another type of module as a node, which shows the preprocessing that occurs such as the processes between text input and natural language processing. When nodes are connected by a link, it shows that the nodes are combined. When a large number of nodes are displayed, it is difficult for a user to determine the module type. Thus, each module is highlighted with different colors according to the module's process.

5.2 Manipulation

A user can manipulate nodes directly, where each module is operated using a mouse to place a node in the neighborhood of a node so they can be combined by dragging and dropping, and these nodes are linked automatically.

In the left of Figure 6, we use the mining module, *LabelData*, and the visualization module, *FlowPanel*. In the right of Figure 6, we use the mining module, *MakeLight*, and the visualization module, *TextDisplayColor*. After combining a pair of modules, a user can switch on the visualization module by clicking the node that denotes it.

6. Discussion

In this section, we discuss the user exploration process during text mining and our future work.

6.1 User's Exploration Process on Text Mining

Text mining requires exploratory information analysis based on trial-and-error processes. This task comprises two of exploration processes: (1) exploratory browsing and (2) focused searching [White 09]. The process that corresponds to exploratory browsing requires that a user understands the text collection by considering it from various perspectives, while the process that corresponds to focused search requires that a user perform a detailed analysis of a portion of a text collection because the user's information request is clarified. Thus, a user can clarify their information request by growing and contracting the exploration space repeatedly.

The user performs a trial-and-error process during the former exploration procedure so the proposed interface is beneficial for this process. During the latter process, a user might not need to switch the text analysis tool frequently because their information request has been clarified. During text mining, however, the user's information request may be changed. Therefore, the user moves from a focused search to exploratory browsing so our proposed interface indicates how each tool is used and the user can determine the tool that they require.

6.2 Future Work

In the prototype system, multiple text analysis tools cannot be displayed horizontally. In addition, we implemented our proposed interface based on TETDM and we did not

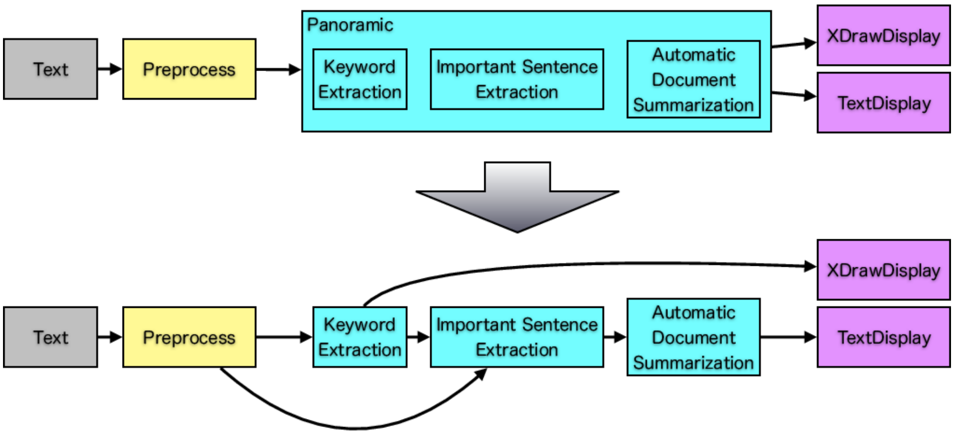

Text
Preprocess
Panoramic
Keyword Extraction
Important Sentence Extraction
Automatic Document Summarization
XDrawDisplay
TextDisplay
Text
Preprocess
Keyword Extraction
Important Sentence Extraction
Automatic Document Summarization
XDrawDisplay
TextDisplay

Figure 4: Example of module segmentation.

change the module specifications. We need to consider the best way to implement our system.

To support effective exploratory data analysis using a computer, we need to consider the user's exploratory actions and their reflective actions [Matsushita 05b]. In the proposed interface, we use a graph to indicate how each module is used. Thus, we consider that our proposed interface could support user reflection during the exploration process by adding functions that save the history of the graph states and annotations based on the user's awareness.

7. Conclusion

In this study, we redesigned the TETDM interface and implemented a prototype interface based on the existing TETDM system to support the trial-and-error process during text mining.

Text mining is a form of exploratory data analysis and the system applied to this type of task needs to support the analytical process. However, the TETDM interface used to perform text mining tasks is not adequate for this type of task. Therefore, we considered the user requirements for performing trial-and-error processes smoothly and we redesigned the TETDM interface by including a graph where the nodes denote the modules while the links denote the process flow between the nodes. In future, we will improve the interface further and investigate its utility.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Number 22300048.

References

[Hartwig 79] Hartwig, F. and Dearing, B.: *Exploratory data analysis*, SAGE Publications (1979)

[Hearst 99] Hearst, M. A.: Untangling text data mining, in *In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, pp. 3–10 (1999)

[Matsushita 05a] Matsushita, M.: Supporting Exploratory Data Analysis by Preserving Contexts, in *Proc. 9th International Conference on Knowledge-Based Intelligent Information and Engineering Systems*, pp. 540–546 (2005)

[Matsushita 05b] Matsushita, M. and Shirai, Y.: Supporting Exploration and Reflection in Exploratory Data Analysis, in *Proc. 4th International Workshop on Chance Discovery*, pp. 3–8 (2005)

[Nasukawa 09] Nasukawa, T.: Implementation of Text Mining in Practice, *Journal of The Japanese Society for Artificial Intelligence*, Vol. 24, No. 2, pp. 275–282 (2009), in Japanese

[Nishihara 11] Nishihara, Y. and Sunayama, W.: Text Visualization using Light and Shadow based on Topic Relevance, *International Journal of Intelligent Information Processing*, Vol. 2, No. 2, pp. 1–8 (2011)

[Rajman 97] Rajman, M., BESANON, R., and Besancon, R.: Text Mining: Natural Language techniques and Text Mining applications, in *In Proceedings of the 7th IFIP Working Conference on Database Semantics*, pp. 7–10 (1997)

[Sunayama 11a] Sunayama, W., Takama, Y., Bollegala, D., Nishihara, Y., Tokunaga, H., Kushima, M., and Matsushita, M.: Total Environment for Text Data Mining, *Journal of The Japanese Society for Artificial Intelligence*, Vol. 26, No. 4, pp. 483–493 (2011), in Japanese

[Sunayama 11b] Sunayama, W. and Tanikawa, N.: Text Coherence Evaluation related to Topics and its Application to Conclusion Extraction, *Jornal of Japan Society for Fuzzy Theory and Intelligent Informatics*, Vol. 23, No. 5, pp. 727–738 (2011), in Japanese

[White 09] White, R. W. and Roth, R. A.: *Exploratory Search: Beyond the Query–Response Paradigm*, Morgan & Claypool Publishers (2009)