

Constructing Knowledge Using Exploratory Text Mining

Naoya Otsuka
Graduate School of Informatics
Kansai University
Osaka, Japan
Email: k589149@kansai-u.ac.jp

Mitsunori Matsushita
Faculty of Informatics
Kansai University
Osaka, Japan
Email: mat@res.kutc.kansai-u.ac.jp

Abstract—Our goal is to support users who want to discover or create knowledge from a large amount of text data. Text mining is a process that extracts novel knowledge from unstructured data. A variety of applications for text mining, such as Total Environment for Text Data Mining, have been proposed. However, to the best of our knowledge, these methods are not effective at conducting text mining tasks aimed at finding novel knowledge. In this paper, we discuss characteristics of the text mining process and propose a design principle for building text mining applications based on two concepts: (1) text mining is an exploratory search task, and (2) text mining is a process for creative knowledge work.

I. INTRODUCTION

Vast amounts of text data are accessible on the World Wide Web and can be used not only for browsing information, but also as resources to obtain useful *knowledge* which can contribute to solve their own problems. In order to obtain useful and novel knowledge from the resources, text mining techniques are often employed.

Text mining is a technique for finding useful knowledge from unstructured text data [1], and it employs a complex and difficult method that includes natural language processing, data mining, and information visualization, which are all adaptable to the process of finding new knowledge. To make analyzing text easier and more efficient, many techniques have been developed such as keyword extraction, automatic document summarization, document clustering, sentiment analysis, and topic detection and/or tracking. In addition, other methods for finding useful knowledge for text mining have also been proposed.

When most users begin to search for novel knowledge from given text, their information needs are not often clear because such searching must be accomplished by means of a complex exploration process. In other words, text mining is a method for obtaining access to explore information rather than an isolated process, such as that of goal-oriented tasks. To obtain useful knowledge from a large collection of text, users are required to perform an exploratory text analysis by means of a repeated trial-and-error processes by viewing a text collection from a variety of perspectives. From the perspective of such users seeking information, text mining can be considered an exploratory search [2].

When users perform a text mining operation, a combination of text analysis techniques must be employed to analyze

text. In addition, to view these results from a variety of perspectives, users must alternate among several information visualization techniques that have been proposed in previous studies. However, because these techniques are designed for individual purposes, users cannot combine them. In addition, environments do not exist that allow users to be flexible with and thus take advantage of these techniques.

To address this problem, Total Environment for Text Data Mining (TETDM) has recently been proposed [3]. TETDM provides an environment in which users can combine several text analysis tools. TETDM aims to support users who want to mine useful information in a large volume of text. However, the TETDM interface is ineffective for performing this type of task, because users cannot explore the text intuitively and users exploration processes are hindered. In our previous study, we solved this problem by redesigning the interface in order to facilitate users' trial and error during text mining [4], and in this paper, we proposed a design principle to improve the interface.

Our goal is to support users searching for useful information and novel knowledge from an extensive amount of text. To this purpose, we first consider the main features of the text mining operation. Second, we examine relationships between the nature of the text mining operation and certain user perspectives, such as exploratory search and creative knowledge work. Finally, based on these relationships, we propose a design principle for developing an environment in which a text mining operation can be performed based on user interaction behavior with the text mining process.

Our design principle contains the following three features: (1) users can understand their current stage of exploration, which is how did users progress advance in their exploration, and which determine the tools does users to use to analyze texts, (2) users can reflect their exploration history, and (3) users can express their exploration progress, results or partial solutions (during exploration) without restriction.

II. RELATED WORKS

In this section, we first clarify the nature of the text mining operation from the perspective of its processes. Second, we describe Total Environment for Text Data Mining (TETDM) and our previous work. Third, we describe creative knowledge work because we consider a particular task for finding novel

knowledge as creative knowledge work. In order to using high-functionality applications such as text mining applications, such applications should need to lead users to be able to use work well. Finally, we describe an approach for such an applications.

A. Text Mining

Text mining is a technique for obtaining new knowledge from text data [1]. It is similar to data mining with respect to extracting useful information from a large amount of data. The difference is that text mining aims to extract information from unstructured text data. Natural language processing techniques are critical for extracting structured information from unstructured text data. In addition, to help users understand the results of these techniques, information visualization techniques are often used [5]. Therefore text mining is a complex combination of techniques that include natural language processing, data mining, and information visualization.

Several text mining tools have been proposed. Fan et al. briefly described the text mining tools developed by major commercial vendors [5]. Fan et al. implemented the following techniques in their tools: information extraction, topic tracking, summarization, categorization, concept linkage, clustering, information visualization, and question answering. In their paper, Fan et al. outlined a generic process model for a text mining application and argued that users sometimes iterate the text mining process until targeted information is obtained. In addition, they explained that users can employ a combination of text analysis techniques depending on their goals.

Nasukawa and Nagano proposed a text mining system known as Text Analysis and Knowledge Mining (TAKMI) to take advantage of textual databases in personal computer help centers [6]. They designed an interactive feature that allows users to easily confirm analysis results of analysis with an original document. In addition, a statistical analysis tends to ignore minor patterns. Thus, by using this interactive feature, users can find not only major patterns treated as noise but also minor patterns. In order to find novel knowledge, using a computer's ability to manage vast quantities of data and human's ability to notice subtle differences in patterns is essential.

Hearst defined data mining, information access (such as information retrieval), and corpus-based computational linguistics, and discussed their relationship to text data mining (Table I) [7]. She argues that finding novel information is a critical part of text mining, but states that information retrieval and data mining do not contribute to the acquisition of new knowledge. This is clear for the following reasons: (1) Documents returned by information retrieval systems do not contain new information because the such information is already known by the document authors., (2) data mining applications help to find trends and patterns automatically from extensive datasets.

However, users who view their results might find unexpected trends or patterns. This is critical for discovering new knowledge because it suggests that such a search requires cognitive ability. These techniques do not enable to find novel information. However, by analyzing their results from different perspectives, users can acquire novel knowledge. In addition,

TABLE I. A CLASSIFICATION OF DATA MINING AND TEXT DATA MINING APPLICATIONS ACCORDING TO HEARST (REPRINTED FROM [7]).

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

Hearst also states that text mining is a form of exploratory data analysis [8] because users can find unknown information in order to test a hypothesis. In exploratory data analysis with a computer, the following processes occur: (1) users submit a query to the computer based on an uncertain idea of the information they want to acquire, (2) the computer obtains a result based on the query, and (3) based on this returned result, users increase their understanding of the data and formulate a new query. In exploratory data analysis, users collect useful information related to problem solving and decision making by means of these processes until the targeted information or knowledge is obtained [9]. In addition, to support effective exploratory data analysis using a computer, users' exploratory and reflection actions must be supported [10].

Text mining can not only extract well-known information and finds patterns but also finds novel information in a large volume of text.

B. Total Environment for Text Data Mining

The tools necessary for the proposed text mining methods and techniques are often experimental or simply unavailable. Because users must apply a combination of different text analysis tools during text mining, building each tool separately, formatting the data for each tool, and implementing an interface to compare the analytical results generated by each tool might prove demanding. Therefore, performing text mining with a variety of text analysis tools is difficult. To address this problem, TETDM has been proposed as a text mining system [3]. Using TETDM, users can employ multiple text analysis tools in a parallel arrangement onscreen, and can combine these tools flexibly (Figure 1).

However, the TETDM interface is ineffective for text mining tasks. In our previous study, we attempted to facilitate the users' trial-and-error process during text mining and proposed a redesigned interface for TETDM [4]. The proposed interface includes a graph in which nodes indicate tools and links denote the process flow between nodes (Figure 2). Our interface accommodates the trial-and-error process during text mining by allowing users: (1) to combine/alternate among text analysis tools by manipulating nodes directly, and (2) to understand the current state of each tool. This paper updates our previous study with additional research.

C. Exploratory Search

Search activities occur in a variety of situations and for multiple purposes. Exploratory search is a model illustrates the interaction between the search process and the problem context [2], and it is designed for users who are (1) unfamiliar with the domain in which they are performing a search, (2) unsure about the ways to achieve their searching goals, and (3) unsure about their goals. In exploratory search,

problem contexts are open-ended, persistent, and multifaceted, whereas search processes are opportunistic, interactive, and multi-tactical. Exploratory search behavior consists of two major aspects: exploratory browsing and focused searching. In exploratory search, users generally have vague or unclear information needs. However, by iterating exploratory browsing and focused searching, users gradually clarify their information needs. In addition, exploratory search is related to exploratory data analysis. Exploratory browsing is similar to exploratory data analysis, where the goal is to generate hypotheses from data.

- 1) Supports querying and rapid query refinement.
- 2) Offers faceted category-based and metadata-based result filtering.
- 3) Leverages search contexts.
- 4) Offers visualizations to support insight and decision making.
- 5) Supports learning and understanding.
- 6) Facilitates collaboration.
- 7) Provides exploration histories, virtual workspaces, and progress updates.
- 8) Supports task management.

Ye and Fischer discussed the necessities and benefits of information delivery in supporting learning on demand in high-functionality applications [12]. They indicated that users have different levels of knowledge about such applications. In Figure 3, rectangle L4 represents an entire information space, and the ovals, L1, L2, and L3, represent a different levels of knowledge. L1 represents are well-known elements, L2 represents vaguely known elements, and L3 represents elements anticipated to exist. In addition, they also describe four modes of information use: (1) use by memory, (2) use by recall, (3) use by anticipation, and (4) use by delivery. Each mode requires a different information acquisition approach. In information-use modes (3) and (4), information delivery is necessary because it is difficult for users to find information concerning levels of knowledge such as L3 and (L4-L3) in Figure 3 without receiving system support.

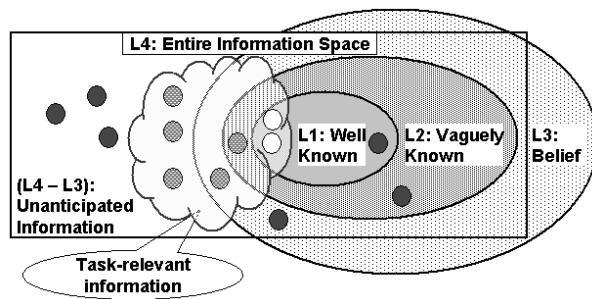


Fig. 3. Different levels of knowledge about a high-functionality application (reprinted from [12]).

III. DESIGN PRINCIPLES

During the exploration process, it requires two actions: exploration action and reflection action. An exploration action is to externalize beneficial findings obtained from the repeating the trial-and-error process, and a reflection action organizes and interprets the externalized findings to utilize them in the users' confronting task. These two actions are significantly related.

In order to extract useful information and obtain novel knowledge from extensive text data during a text mining task, an interactive feature is required. Furthermore, appropriate interaction style for each action is different. Thus, in order to enable users to execute these two actions seamlessly, we should consider the type of information and objects in the system that are shared between these two actions [10].

In this section, we discuss the requirements for a text mining application from the perspective of users. We examine three points of view: (1) facilitation of the user exploration process, (2) reflection of the user exploration process, and (3) facilitation in learning how to use the system.

We examine three main user concerns: (1) how can the exploration process be facilitated? (2) how can the exploration process be reflected in the application? and (3) how can learning the system's features and functionality be enhanced?

A. Facilitation of Users' Exploration Process

During a text mining task, users analyze texts by iterating a variety of text analysis techniques during a trial-and-error process.

To execute these analytical methods effectively, a text mining application interface should be intuitive and possess the requisite text analysis tools that allow users to combine and alternate among such tools easily and effectively. Because users typically combine text analysis tools given the extremely variable nature of the analytical process, users must be able to review the analytical process in order to understand the current analysis stage, and if necessary, choose a more suitable analytical approach.

B. Reflection of User Exploration Process

To support effective exploratory data analysis when using a computer, users' exploration actions and their reflection actions must be considered [10]. In exploration process, a user towards

a his/her goal, which is often vague, and the goal may be changed. In the exploration process, a user's progress towards a his/her goal, which is often vague, and could the goal may be changed. That is, various exploration branches might exist.

For this reason, we must examine the following.

1) Users can understand their current exploration stage

During a text mining task, exploration could become complicated and various branches. If users do not fully understand their current exploration stage, they cannot progress to the next analytical step. Therefore, users must easily comprehend the current phase of exploration in which they find themselves.

For this purpose, the system also requires functions for users to understand their exploration branches, such functions can be as zooming in/out of branches and pruning unnecessary branches.

2) Users can reflect on their exploration history

Historical information is critical to the exploratory search process. An exploration history allows users to review their past exploratory behavior. By providing an exploration history, the system can help users find insight into the process of locating new knowledge.

3) Users can review their exploration progress, results, or partial solutions without restriction

It is necessary to review solutions without restrictions because users are facilitated knowledge creation by expressing externalized objects as interpretation-rich representations. Externalized objects are, for example, visualized analysis results, annotations to help with user saving their memories, insights and findings derived from the exploration process, and partial solutions to a final problem result. In addition, in creative knowledge work, it is important for users to interact with externalized objects. To address this issue, users are allowed to (1) move or resize objects, (2) add or delete objects, and (3) zoom in and out of a workspace that manages objects. Through such actions, users can reflect on their previous exploration, remember what they thought originally, or notice new insights.

C. Learning How to Use the System

Text mining applications often possess too many functions for users to master easily. Most users cannot comprehend all functions when beginning to use a text mining application. Therefore, users must gradually learn how to use such systems based on task relevant information while using the systems. These systems should provide users with a way to learn such knowledge on demand.

In addition, when performing text mining tasks, users should know the following: (1) basic functions for manipulating the system, such as how to combine or switch between text analysis tools and how to view user exploration history, (2) the nature of each analytical technique so that users can select the most appropriate one based on their information needs, and (3) the nature of or domain of the target data. To address (1), information delivery of learning on demand is effective. However, regarding (2) and (3), other approaches are required. Chen et al.

suggested that knowledge-assisted visualization is necessary in the context of developing information visualization technology [13]. Knowledge-assisted visualization, which is an advanced form of information visualization, takes advantage of user knowledge regarding visualization techniques, such as how to use the techniques and how to control parameters for such techniques, stored in a database. Such knowledge corresponds to (2). Thus, using this knowledge, users can more easily learn to employ text analysis tools.

In addition, users can gradually learn (3) by conducting exploratory searches, which are iterations of exploratory browsing and focused searching.

IV. FUTURE WORK

We currently implement an interface on the TETDM system based on our design principles, especially in order to support users' reflection actions. Figure 4 shows our interface prototype. Our prototype interface based on TETDM. Our interface contains the module (text analysis tool) selection panel and history tree panel. The module selection panel is used to combine and alternate text analysis tools and to understand users' current exploration process by visualizing the tools that are combined or used. The history tree panel is used to reflect on users' past exploration. We continue to develop our proposed interface.

Shneiderman provides visual information seeking mantra and taxonomy of tasks for information visualization [14]. The visual information seeking mantra is known as "Overview first, zoom and filter, then details-on-demand." In addition, Shneiderman classifies seven tasks for information visualization at a high level of abstraction: overview, zoom, filtering out uninteresting items, details-on-demand, viewing relationships among items, keeping history of actions, extracting sub-collections and query parameters. In information visualization literature, researchers have discussed a methodology for facilitating user understanding using visual representations and for interacting with such representations. In text mining literature, researchers often described techniques for the text mining tasks discussed in the related work [5]. However, tasks for text mining at a high level of abstraction are not often discussed. In other words, the technology required for text mining is discussed, but the tasks required in text mining are not discussed sufficiently. We need to discuss tasks required in text mining for browsing texts and finding new knowledge from the perspective of interaction behavior in performing text mining tasks. Information visualization is significantly related to text mining. Thus, we consider that text mining can employ such a concept as information visualization literature. By classifying tasks for text mining at high level of abstraction, the interaction model for text mining tasks can be built based on such task classification.

In the future, we consider the tasks required for text mining. In addition, we build a interaction model and design principles for text mining applications for discovering knowledge.

V. CONCLUSION

In this paper, we proposed a design principle for developing an exploratory text mining environment in order to find new and creative knowledge. Text mining task requires iterative processes and combines a variety of text analysis techniques.

The process of text mining is classified as both an exploratory search task and creative knowledge work. Thus, our design principle possesses three main requirements: (1) the system must facilitate the user exploration process, (2) it must allow users to reflect on the exploration process, and (3) it must facilitate learning to use the system.

Regarding requirement (1), the proposed interface in our previous study included a graph in which nodes indicate the modules and links denote the process flow between nodes. In addition, the interface allows users to combine and alternate between text analysis tools by manipulating the nodes directly. This is achieved by manipulating the nodes directly as objects.

For building text mining applications, the interaction between a users and the system needs to become intuitive. Therefore, the combinations of the human ability for finding trends or patterns and the computer's ability for managing large amounts of data, are important points for hastening to find novel knowledge.

In a future study, we hope to address requirements (2) and (3). We will further refine our design principle and consider a means to create an interface that achieves such design principle. Furthermore, we will implement the created interface in a text mining system and evaluate the system's usability and practicality.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number 22300048.

REFERENCES

- [1] M. Rajman and R. Besançon, "Text mining: Natural language techniques and text mining applications," in *In Proceedings of the 7th IFIP Working Conference on Database Semantics (DS-7)*. Chapam. Hall, 1997, pp. 7–10.
- [2] R. W. White and R. A. Roth, *Exploratory Search: Beyond the Query-Response Paradigm*. Morgan & Claypool Publishers, 2009.
- [3] W. Sunayama, Y. Takama, Y. Nishihara, T. Kajinami, M. Kushima, and H. Tokunaga, "Practical application in development and use of mining tools with total environment for text data mining," *Journal of The Japanese Society for Artificial Intelligence*, vol. 29, no. 1, pp. 100–112, 2014, in Japanese.
- [4] N. Otsuka and M. Matsushita, "Graphical interface that supports users' trial-and-error process of text mining," in *Proceedings of the 27th Annual Conference of the Japanese Society for Artificial Intelligence (JSAI 2013)*, 2013.
- [5] W. Fan, L. Wallace, S. Rich, and Z. Zhang, "Tapping the power of text mining," *Communications of the ACM*, vol. 49, no. 9, pp. 76–82, 2006.
- [6] T. Nasukawa and T. Nagano, "Text analysis and knowledge mining system," *IBM systems journal*, vol. 40, no. 4, pp. 967–984, 2001.
- [7] M. A. Hearst, "Untangling text data mining," in *In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, 1999, pp. 3–10.
- [8] F. Hartwig and B. Dearing, *Exploratory data analysis*. SAGE Publications, 1979.
- [9] M. Matsushita, "Supporting exploratory data analysis by preserving contexts," in *Proc. 9th International Conference on Knowledge-Based Intelligent Information and Engineering Systems*, 9 2005, pp. 540–546.
- [10] M. Matsushita and Y. Shirai, "Supporting exploration and reflection in exploratory data analysis," in *Proc. 4th International Workshop on Chance Discovery*, 8 2005, pp. 3–8.
- [11] Y. Yamamoto and K. Nakakoji, "Interaction design of tools for fostering creativity in the early stages of information design," *Int. J. Hum.-Comput. Stud.*, vol. 63, no. 4-5, pp. 513–535, Oct. 2005.

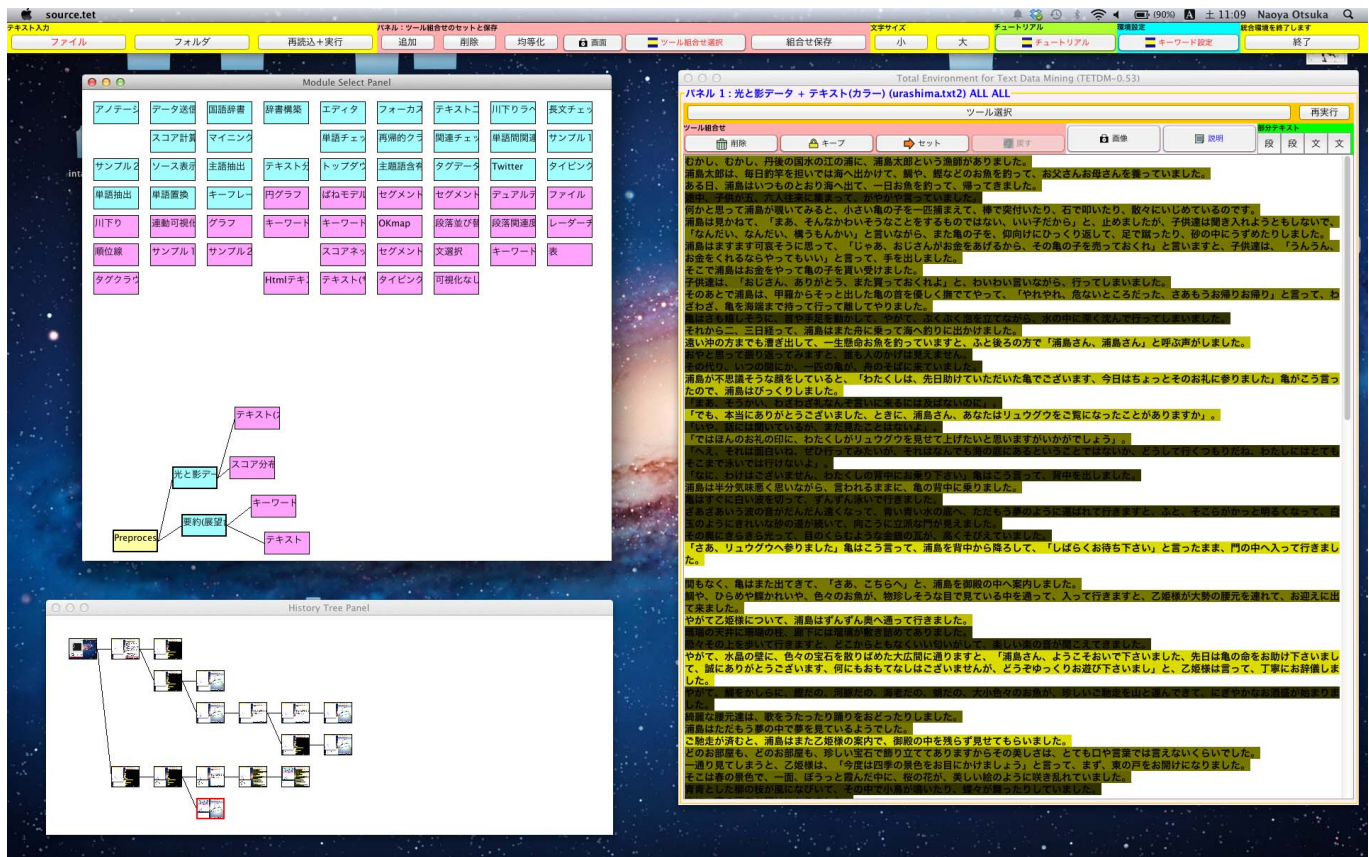


Fig. 4. Our prototype interface based on TETDM. Our interface contains the module (text analysis tool) selection panel and history tree panel. The module selection panel is used to combine and alternate text analysis tools and to understand users' current exploration process by visualizing the tools that are combined or used. The history tree panel is used to reflect on users' past exploration.

- [12] Y. Ye and G. Fischer, "Information delivery in support of learning reusable software components on demand," in *Proceedings of the 7th International Conference on Intelligent User Interfaces*, ser. IUI '02, 2002, pp. 159–166.
- [13] M. Chen, D. Ebert, H. Hagen, R. S. Laramée, R. Van Liere, K.-L. Ma, W. Ribarsky, G. Scheuermann, and D. Silver, "Data, information, and knowledge in visualization," *Computer Graphics and Applications*, IEEE, vol. 29, no. 1, pp. 12–19, 2009.
- [14] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," in *Proceedings of the 1996 IEEE Symposium on Visual Languages*, ser. VL '96. Washington, DC, USA: IEEE Computer Society, 1996, pp. 336–343.