Development and Case Study of Trend Analysis Software Based on FACT-Graph

Ryosuke Saga[1], Hiroshi Tsuji[2], Takao Miyamoto[2], and Kuniaki Tabata[1].

[1]Kanagawa Institute of Technology, 1030 Shimo-ogino, Atsugi, Kanagawa, Japan
[2]Osaka Prefecture University, 1-1 Gakuencho, Nakaku, Sakai, Osaka, Japan
(Tel : 81-46-291-3235; Fax : 81-46-242-8490)
({saga,tabata}@ic.kanagawa-it.ac.jp, tsuji@cs.osakafu-u.ac.jp, aki@center.osakafu-u.ac.jp)

Abstract: This paper proposes text mining software to analyze FACT-Graph and describes case study by using the software. FACT-Graph is a trend graph which visualizes what kinds of topics exist and shows the change of trend in time-series text data. However, FACT-Graph itself does not have enough environments to analyze trend although it provides the clue for trend. In order to resolve the problem, we develop the software called Loopo. This software provides the functions of adding the consideration of analyst as the keywords and operating FACT-Graph itself such as moving, adding, and clearing nodes. Also, the system allows analysts to refer information source, keyword information, and network information in order to analyze and consider FACT-Graph. In a case study about criminal trend using the title of articles of newspaper between 1987 and 2007, we confirm the usability of this software.

Keywords: Trend Analysis, FACT-Graph, Text Mining,, Text Mining System

I. INTRODUCTION

Recently, the usage of text data on web page has been promoted on the back of fast networks and large volume storage. Text mining has been developed to use this text data. Text mining uses unstructured and voluminous text data and discovers new usage knowledge. Text mining methods are used for keyword extractions, summarizations, and visualizations [1]-[4].

This paper focuses on the trend analysis that uses time-series text data and several text mining methods [5]. In many cases of trend analysis, terms and topics in text data (i.e. *bag-of-words*) are used. Consequentially, analysts who analyze trends need to find out what are the terms that receive a lot of attention in a certain time period. To identify and summarize trends, Saga et al. have developed a trend visualization method called *FACT-Graph* [6]. FACT-Graph classifies terms on the basis of their importance and the relationships between them. FACT-Graph visualizes how these two elements change as time elapses. Fig. 1 shows the visualized information as a co-occurrence graph, and an analyst speculates about the trends from this graph.

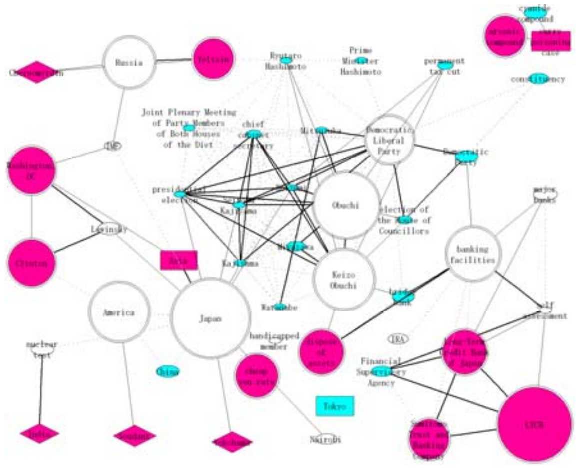


Fig.1. FACT-Graph

However, text mining based on human-centric actions is currently receiving much attention [7]. This text mining emphasizes the centrality of analyst interactivity in knowledge discovery processes. It repeats the processes in which an analyst considers the text mining results, reflects the considerations in the result, and analyzes again. For human-centric text mining operations, text mining necessitates the system-based support that aims to show the evidences and reflects the analyst's intentions.

This paper describes text mining software called *Loopo* that extend FACT-Graph as human-centric text mining. Loopo provides information on the processes of analyzing terms, networks, and evidences for trends analysis to identify trends.

The rest of this paper is organized as follows; Section 2 introduces FACT-Graph. Section 3 describes Loopo in detail. Section 4 details a case study about crime trends in Japan using newspaper articles. Finally, Section 5 concludes our papers.

II. FACT-Graph

FACT-Graph visualizes the change of keywords trend and relationships between terms over two time periods (Fig. 1). The usefulness of FACT-Graph has been discussed in several previous studies [6][8].

FACT-Graph is shown using nodes and links. FACT-Graph embeds the change in a keyword's class transition and co-occurrence in nodes and edges. The keyword's class transition, which is one of the essential elements of FACT-Graph, is based on class transition analysis that separates keywords into four classes based on term frequency (TF) and document frequency (DF) [9]. The result of the analysis shows the transition of keywords between two time-periods shown in Table 1. For example, if a term belongs to Class A in a certain time period and moves into Class D in the next time period, then the trend regarding that term is referred to as "fadeout". FACT-Graph identifies these trends by the node's color. For example, red color shows fashionable, blue color shows unfashionable and white shows unchanged in FACT-Graph

Additionally, a FACT-Graph visualizes keywords and relationships between keywords by using co-occurrence information. As a result, useful keywords can be obtained from their relationship with other keywords, even though that keyword does not seem to be important at a glance, and the analyst can extract such keywords by using FACT-Graph. Moreover, from the result of the class-transition analysis, the analyst can comprehend trends in keywords and in topics (consisting of several keywords) by FACT-Graph.

Table 1. Class transition and keyword trends

		After			
		A	B	C	D
Before	A	Hot	Cooling	Bipolar	Fade
	B	Common	Universal	-	Fade
	C	Broaden	-	Locally Active	Fade
	D	New	Widely New	Locally New	Negligible

The steps for generating FACT-Graph are as follows:

1. Separate time-series text data in accordance with the analysis periods
2. Extract keywords in each period by morphological analysis and TF-IDF algorithm
3. Carry out class transition analysis and extract co-occurrence relations.
4. Visualize keywords and relations.

A FACT-Graph is composed of three factors: time, keywords, and a co-occurrence network. The analyst of a FACT-Graph configures several parameters related to these factors such as analysis period, keyword filtering, and threshold of co-occurrence.

III. Loopo

Loopo is software made to extend FACT-Graph as a human-centric text mining system. Loopo enable users to analyze FACT-Graph efficiently by improving the analysis process of FACT-Graph. In order to improve the analysis process, Loopo aims to repeat user input such as parameters' setting and reflection onto FACT-Graph and output such as smoothly generating FACT-Graph and information references.

Fig. 2 is a screenshot of how Loopo draws FACT-

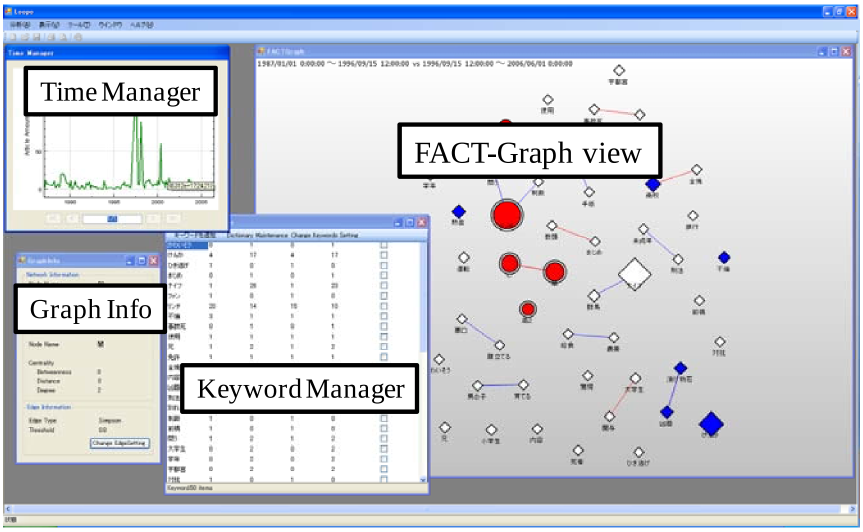


Fig.2. Loopo

Graph from certain text data. Loopo consists of four windows: "FACT-Graph View," which shows and operates the FACT-Graph itself; "Keyword Manager," which manages keywords; "Time Manager," which manages information and parameters concerning analysis periods; and "GraphInfo," which shows and manages parameters concerning the network of the FACT-Graph. User can configure the parameters to generate FACT-Graph via these windows.

The analysis by Loopo starts with the import of time-series text data. From the data, Loopo makes visualization at the same time as generating FACT-Graph. Loopo renders FACT-Graph dynamic so a user can moves keywords via FACT-Graph View to facilitate visualization. Also, the user can add terms that are deselected by the parameter to FACT-Graph and removes unnecessary keywords from the graph.

A. FACT-Graph View

FACT-Graph View shows the analysis results for the text data which is imported to Loopo as FACT-Graph. The analyst can move, clear and fix keywords for trend analysis easily via the window. For example, the "fixing keyword" function is used to fix the locations of noteworthy keywords between multiple analysis periods. The analyst can therefore browse through remarkable keywords and their related keywords over the periods at ease. FACT-Graph View also allows the analyst to refer to original text data from remarkable keywords and helps them to comprehend macro/micro trends.

B. Time Manager

It is important to configure time period for time-series analysis. Usually, the number of articles is shown as a clue of setting of time periods. By displaying the trend in article volume as a chart, Time Manager helps the analyst configure the parameter concerning analysis period. The window indicates how the time periods are divided up for analyzing a FACT-Graph (which is output according to the time periods). Time Manager also has a function for setting time periods forward or backward. With this function, the analyst can view a series of FACT-Graphs via FACT-Graph View along with the change of time period.

C. Keyword Manager

Keyword Manager is the window for listing and managing the keywords currently shown in a FACT-Graph. The analyst can add and delete keywords, and refer to the original text data from a keyword via Keyword Manager or FACT-Graph View. As a result, the window can reflect the analyst's awareness in FACT-Graph. The analyst can also configure the parameter, such as thresholds, concerning keyword extraction.

D. Graph Info

One of the measures for identifying whether a FACT-Graph is meaningful is network information such as network size and density. GraphInfo shows the network information about FACT-Graph. GraphInfo shows network size, density, and the type of links as an overview of a FACT-Graph. It also shows several centralities such as betweenness centrality and closeness centrality when analyst selects a node of interest via FACT-Graph View [11][12]. Moreover, GraphInfo allows the analyst to change co-occurrence type and thresholds of co-occurrence.

IV. CASE STUDY

A. Environment

This section describes a case study to confirm the usability of Loopo. In this case study, we use article titles from Japanese newspaper between 1987 and 2007, and we try to analyze crime trends. The detailed data is shown in Table 2. In this analysis, we use the keywords the TF-IDF weight of which contains the top 30 in each time period, and we adopt Simpson coefficient for co-occurrence. We also adopt more than 0 as thresholds of keyword and co-occurrence because articles titles are generally abbreviated sentences.

Table 2. Basic information of analysis data

The number of articles		2971
Analysis Span		1987-2007
Co-occurrence Type		Simpson
Parameters	Keywords	30
	Thresholds(TF, DF)	0
	Thresholds (Co-occurence)	0

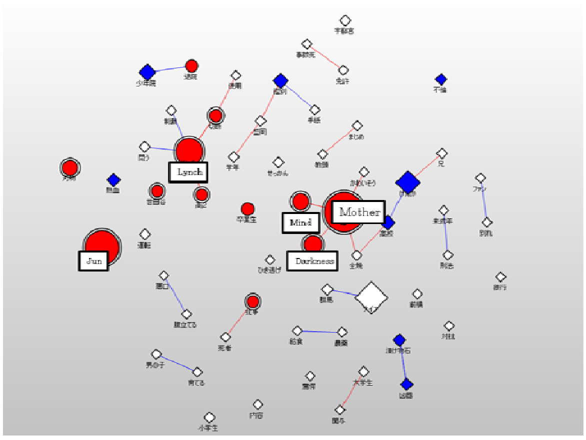


Fig.3. FACT-Graph between 1987 and 2007

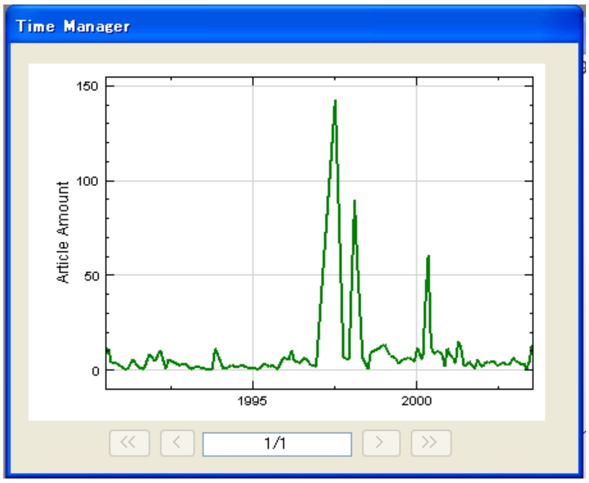


Fig.5. Time Manager in Case Study

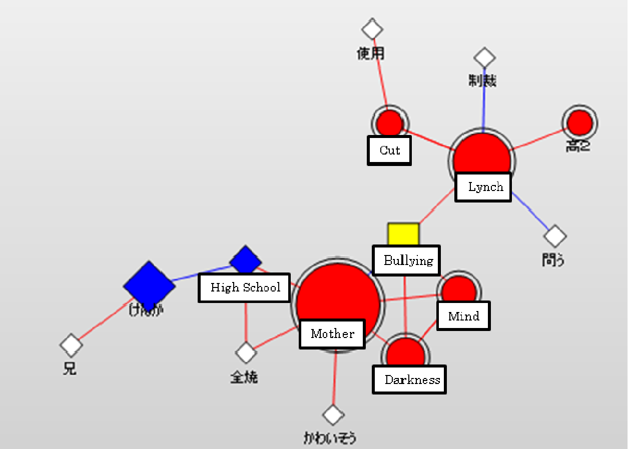


Fig.4. FACT-Graph added the keyword “Bullying”

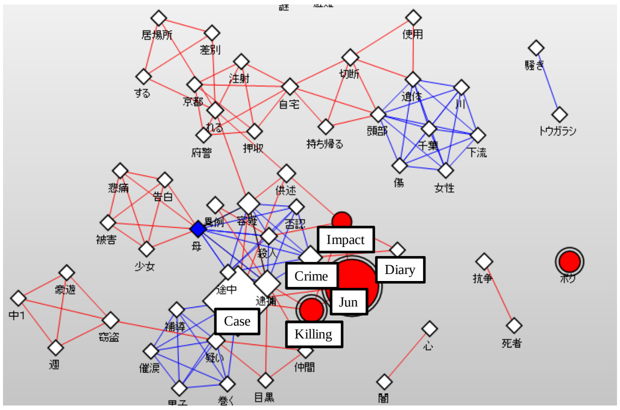


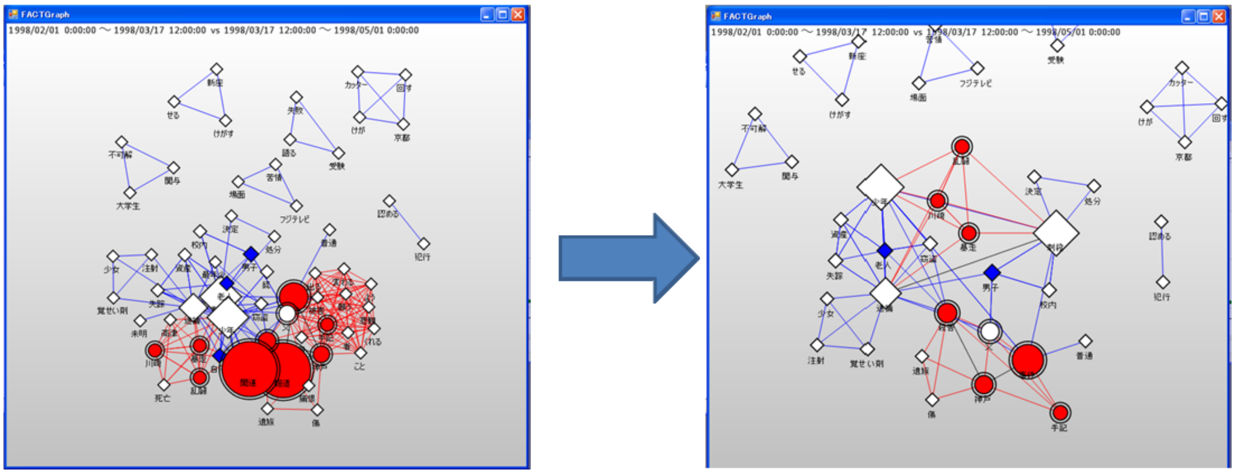
Fig.6. FACT-Graph
between November 1996 and October 1997

B. Analysis Result

At first, we separated entire text data into two parts, that is, the period from 1987 to 1997 and the period from 1997 to 2007, and then generated FACT-Graph. The result is shown in Fig. 3. FACT-Graph visualizes a fashionable keyword as a red node. Therefore we could obtain the several fashionable keywords such as "Mother", "Lynch", "Mind", and "Darkness" from the figure. Then By using the function of Loopo, we could found that "Mind" and "Darkness" were used frequently as one phrase "Darkness of Mind" and the phrase did not appear in before period.

Simultaneously, we noticed that "Bullying" which was not in FACT-Graph appeared with "Darkness of Mind" frequently. Here, we added the "Bullying" via the Keyword Manager in Loopo. The result is shown in Fig. 4. From the result, "Bullying" appeared as a bridge between two topics. Also the betweenness from GraphInfo is highest among other keywords, so "Bullying" was possibly important over 20 years of crimes.

Next, we focused on the two peaks of the number of articles between November 1996 and May 1998 because Time Manager indicated that the number of articles increased twice suddenly in this period (Fig. 5). Fig. 6 visualizes the FACT-Graph for the first peak around 1997. In this figure, "Jun", a male name that also appears in Fig. 3, is most fashionable keyword. On the other hand, the FACT-Graph in the second peak around 1998 shows Fig. 7(a). However, there are so many unnecessary keywords such as "Related" and "Reports" that it is difficult to analyze trends. Therefore, by using Loopo, we removed these keywords by referring to the original information source, so we can retain the important keywords in Fig. 7(b). From the processed figure, we found that teenage crimes and information on particular criminals are focused on in these periods.



(a) Output FACT-Graph

(b) After removing unnecessary keyword

Fig.7. FACT-Graph between October 1997 and October 1998

V. CONCLUSION

This paper proposes text mining software to analyze FACT-Graph and describes case study by using the software. FACT-Graph is a trend graph which visualizes what kinds of topics exist and shows the change of trend in time-series text data. However, FACT-Graph itself does not have enough environments to analyze trend.

In order to resolve the problem, we develop Loopo which provides the functions of adding the consideration of analyst as the keywords and operating FACT-Graph itself such as moving, adding, and removing nodes. Also, the system allows analysts to refer information source, keyword information, and network information in order to analyze and consider FACT-Graph. In a case study about criminal trend using the title of articles of newspaper between 1987 and 2007, we confirm the usability of this software.

REFERENCES

[1] Salton G (1989), Automatic Text Processing. Addison-Wesley Publishing Company

[2] Harman D (1992), Ranking algorithms, in Information Retrieval, chapter 14. Prentice Hall

[3] Yamanishi K, Li H (2002), Mining Open Answers in Questionnaire Data. IEEE Intelligent Systems 17(5):58-64

[4] Feldman R, Aumann Y, Zilberstein A and Ben-Yamada Y (1998), Trend graphs: Visualizing the evolution of concept relationships in large document collections. Second European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD 1998):38-46

[5] Havre S, Hetzler B and Nowell L (2002), ThemeRiver(TM): In search of trends, patterns, relationships. IEEE Trans Visualization Computer Graphics8:9-20

[6] Saga R, Terachi M and Tsuji H (2009), FACT-Graph: Trend Visualization by Frequency and Co-occurrence. IEEJ transactions on electronics, information and systems 129(3):545-552

[7] Brachman R, Anand T (1996), The Process of Knowledge Discovery in Databases: A Human CenteredApproach. A KDDM, AAAI/MIT Press:37-58

[8] Saga R, Terachi M, Sheng Z and Tsuji H(2008), FACT-Graph: Trend Visualization by Frequency and Co-occurrence. in Lecture Notes on Artificial Intelligence (LNAI 5243: Ed by A. Dengel et al.(Eds)). Springer-Verlag Berlin Heidelberg:308-315

[9] Terachi M, Saga R and Tsuji H (2006), Trends Recognition in Journal Papers by Text Mining. Systems. Man & Cybernetics (IEEE/SMC 2006):4784-4789

[10] Ozaki N and Ohsawa Y (2003), Polaris: An Integrated Data Miner for Chance Discovery. Proceedings of Workshop of Chance Discovery and Its Human Computer Interaction Conference(HCI2003), Crete, Greece

[11] Matsuo Y, Ohsawa Y and Ishizuka M (2002), Keyword Exraction using Small World Structure in a Document. Transactions of Information Processing Society of Japan 43(6):1825-1833

[12] Freeman L C (1979), Centrality in social networks: Conceptual clarification. Social Networks 1(3):215-239