Comparison Analysis for Text Data by Integrating Two FACT-Graphs

Ryosuke Saga and Hiroshi Tsuji

Abstract. This paper describes a method to visualize contrast information about two targets the Frequency and Co-occurrence Trend (FACT)-Graph. FACT-Graph is a method to visualize the changes in keyword trends and relationships between terms over two time periods. We have used FACT-Graphs as comparison method between two targets in previous research; however, the method cannot compare them as equals. To visualize contrast information, we combine two FACT-Graphs generated from different viewpoints and express the features in one graph. In case study by using 132 articles from two newspapers, we compare topics such as politics and events in them.

Keywords: Contrast Mining, Visualization, FACT-Graph, Text Mining, Knowledge Management.

1 Introduction

With the increasing availability of data on Web, several business organizations have to create business value and sustain competitive advantage by using data in data-warehouses [1][2]. To make these data-warehouses work to their advantage, they have to recognize their strong points, develop a strategy, and make effective investments.

To recognize advantages, comparison analysis is often done by using not only traditional methods such as cross-tabulation and visualization analysis but also data/text mining methods. Comparison analysis is relatively easy when the comparative data are expressed quantitatively. However, most significant data often occur in text data and are difficult to obtain from pre-defined attributes. Therefore, text data in questionnaires, reports, and so on must be analyzed.

For the target of text data, text mining is used to obtain new knowledge [3]. In text mining, the applicable areas are wide-ranging such as visualization, keyword

Ryosuke Saga · Hiroshi Tsuji
Osaka Prefecture University, Graduate School of Engineering,
1-1 Gakuen-cho, Nakaku, Sakai, 559-8531, Japan
e-mail: {saga,tsuji}@cs.osakafu-u.ac.jp

T. Watanabe et al. (Eds.): Intelligent Interactive Multimedia: Systems & Services, SIST 14, pp. 143–151.
springerlink.com

extraction, summarization of text, and so on. As previous researches, we have developed the Frequency and Co-occurrence Trend (FACT)-Graph for trend visualization of time-series text data [4] and try to visualize comparison information by using FACT-Graph. However, the comparison information is shown from the viewpoints of one side target and FACT-Graph cannot show the features of both targets as equal.

Therefore, this paper describes a method to compare two targets by using FACT-Graph. The rest of this paper is organized as follows: Section 2 describes the overview and underlying technologies of FACT-Graph. Next, Section 3 describes how to apply FACT-Graph for comparison analysis. After that, Section 4 performs a case study of two Japanese newspapers. Finally, we conclude this paper.

2 FACT-Graph

A FACT-Graph is a method to create visualized graphs of large-scale trends [4]. It is shown as a graph embedding a co-occurrence graph and the information of keyword class transition. A FACT-Graph enables us to see the hints of trends, which have been used for analyzing trends in different fields such as politics and crime, by using analysis tools [5][6]. In addition, the FACT-Graph has been used for web access log data and we have acquired useful knowledge such as the result shown in Fig. 1 [7].

A FACT-Graph uses nodes and links. It embeds the changes in a keyword's class transition and co-occurrence in nodes and edges. In addition, a FACT-Graph allocates the last keyword class attribute to nodes because we assume that recent information is important to carry out trend analysis.

There are two essential technologies in order to compile a FACT-Graph: class transition analysis and co-occurrence transition. Class transition analysis shows the transition of a keyword class between two periods [8]. This analysis separates keywords into four classes (Class A–Class D) on the basis of term frequency (TF) and document frequency (DF) [8]. The four classes are classified by the status of high/low TF and DF separated by two thresholds. The results of the analysis detail the transition of keywords between two time-periods (before and after) as 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." A FACT-Graph identifies these trends by the node's color. For example, red means fashionable, blue refers to unfashionable, and white stands for unchanged.

In addition, a FACT-Graph visualizes relationships between keywords by using co-occurrence information to show and analyze the topics that consist of multiple terms. As a result, useful keywords can be obtained from their relationship with other keywords, even though that keyword seems to be unimportant at a glance, and the analyst can extract such keywords by using a FACT-Graph. Moreover, from the results of the class-transition analysis, the analyst can comprehend trends in keywords and topics (consisting of several keywords) by using the FACT-Graph. In addition, a FACT-Graph considers the transition of the co-occurrence relationship between the keywords. This transition is classified into the following types.