Visualization of Topic Distribution Based on Immune Network Model

Yasufumi Takama

Tokyo Metropolitan Institute of Technology
6-6 Asahigaoka, Hino, Tokyo 191-0065 JAPAN
PREST, Japan Science and Technology Corporation, JAPAN
ytakama@cc.tmit.ac.jp

Abstract. A method is presented that applies the properties of immune system for visualizing the topic distribution over a document set, as well as the topic stream through a sequence of document sets.

1 Introduction

As the Web provides us with new topics constantly, it is difficult for us to grasp the trends or change of topics. In particular, there exists the information, which cannot be found from a single document but from a sequence of document sets. We are developing a Web information visualization method that can find the topic distribution over the document set(s)[1,2]. As the fundamental method, a **plastic clustering method**[2] has been proposed for generating a topic-sensitive keyword map. One of the characteristic features of the plastic clustering method is the generation of **keyword map** as well as document clustering. The keyword's activation value is calculated based on the **immune network model**, which is also useful as the visualization metaphor to improve the understandability of the keyword map. The model of **memory cell** is also incorporated, so that it can find the topical relation among different document sets.

2 Examples of Extracted Topic Stream

The experiment shown here is performed on the sequences of Yahoo! Japan online news article sets that were issued from 17 to 21 September 2001. Fig. 1 shows the landmarks extracted from the sequence. The landmarks extracted by both methods (with/without memory cell) are indicated with dotted texture. In Fig. 1, only one landmark is reused more than once, among 18 landmarks (5.56%) when no memory cell is used. On the contrary, 3 landmarks are reused among 17 landmarks (17.6%) with using memory cell.

As the news articles within the sequence was issued just after the tragedy in N.Y., many articles contain the related topics. Two landmarks, 'Performance', 'Multiple', which are reused more than once, concern the topics related to this N.Y. tragedy. That is, 'Multiple' literally concerns the simultaneous multiple terrorist attacks occurred in New York. After that disaster, many concerts, events,

E. Cantú-Paz et al. (Eds.): GECCO 2003, LNCS 2723, pp. 246–247, 2003.

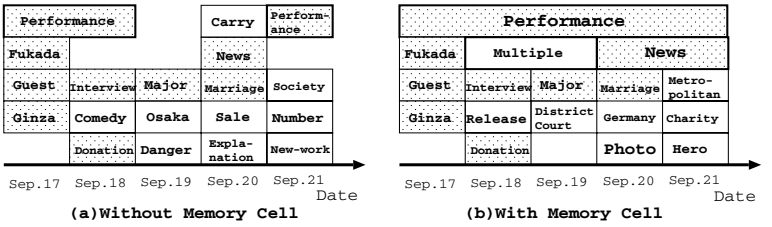

Fig. 1. Landmarks Extracted from Online News Sequence

etc., were performed/held for the purpose of charity, donation, or encouragement for the victims. The landmark 'Performance' mainly refer to this kind of topics.

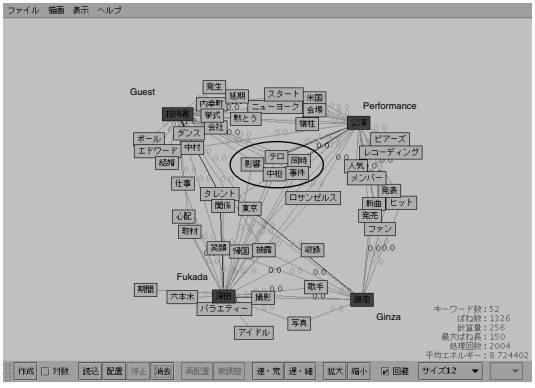

Fig. 2. Keyword Map Generated From Document Set of 17 Sep. 2001.

3 Conclusion

A method for extracting topic distribution over document space based on immune network model is presented. We are now developing the keyword map-based visualization tool that can visualize the extracted topic distribution as shown in Fig.2.

References

1. Takama, Y. and Hirota, K., "Web Information Visualization Method Employing Immune Network Model for Finding Topic Stream from Document-Set Sequence," *J. of New Generation Computing*, Vol. 21, No. 1, pp. 49–59, 2003.
2. Takama, Y., Hirota, K., "Web Information Visualization Based on Immune Network Metaphor," *J. of Japan Society for Fuzzy Theory and Systems*, 14, 5, pp. 472–481, 2002.