4th WSEAS Int. Conf. on E-ACTIVITIES, Miami, Florida, USA, November 17-19, 2005 (pp21-26)

Visualization of Discussions in Comments of a Blog Entry using KeyGraph and Comment Scores

KOHEI TSUDA and RUCK THAWONMAS
Intelligent Computer Entertainment Laboratory
Graduate School of Science and Engineering,
Ritsumeikan University
1-1-1, Noji-higashi, Kusatsu, Shiga, 525-8577

http://www.ice.ci.ritsumei.ac.j/

Abstract: - KeyGraph is a visualization tool for discovery of relations among text-based data. This paper discusses a new application of KeyGraph for visualization of discussions in the comments of a blog entry in Slashdot. We propose an approach that applies KeyGraph successively to multiple chunks of comments, each chunk having a different range of moderation scores provided by Slashdot. This approach gives a higher number of scenarios with more specific meaning than a common approach that applies KeyGraph to the whole comments at once.

Key-Words: - weblogs, blogs, comments, discussions, KeyGraph, Slashdot, moderation

1 Introduction

Recently, weblogs (or blogs) whose number of users is growing rapidly have gained a lot of interests among researchers [1]. As far as data mining is concerned, most of the researchers target at the whole blog space and attempt to discover trends (such as key words, key phrases, or key persons) [2] [3] or to grasp information propagation and epidemics [4] [5].

In this paper, we are interested in visualization of discussions in the comments of a blog entry. Our approach uses a tool called KeyGraph [6]. The basic idea behind our approach is that, rather than being applied to the whole comments at once, KeyGraph is applied successively to multiple chunks of comments, each chunk having a different range of scores. In particular, we take as our research target a blog entry in Slashdot Japan [7], which is a site composed of story submissions and comments to them by a large number of users. We generate comment chunks according to the moderation score of comments provided by Slashdot Japan.

2 KeyGraph

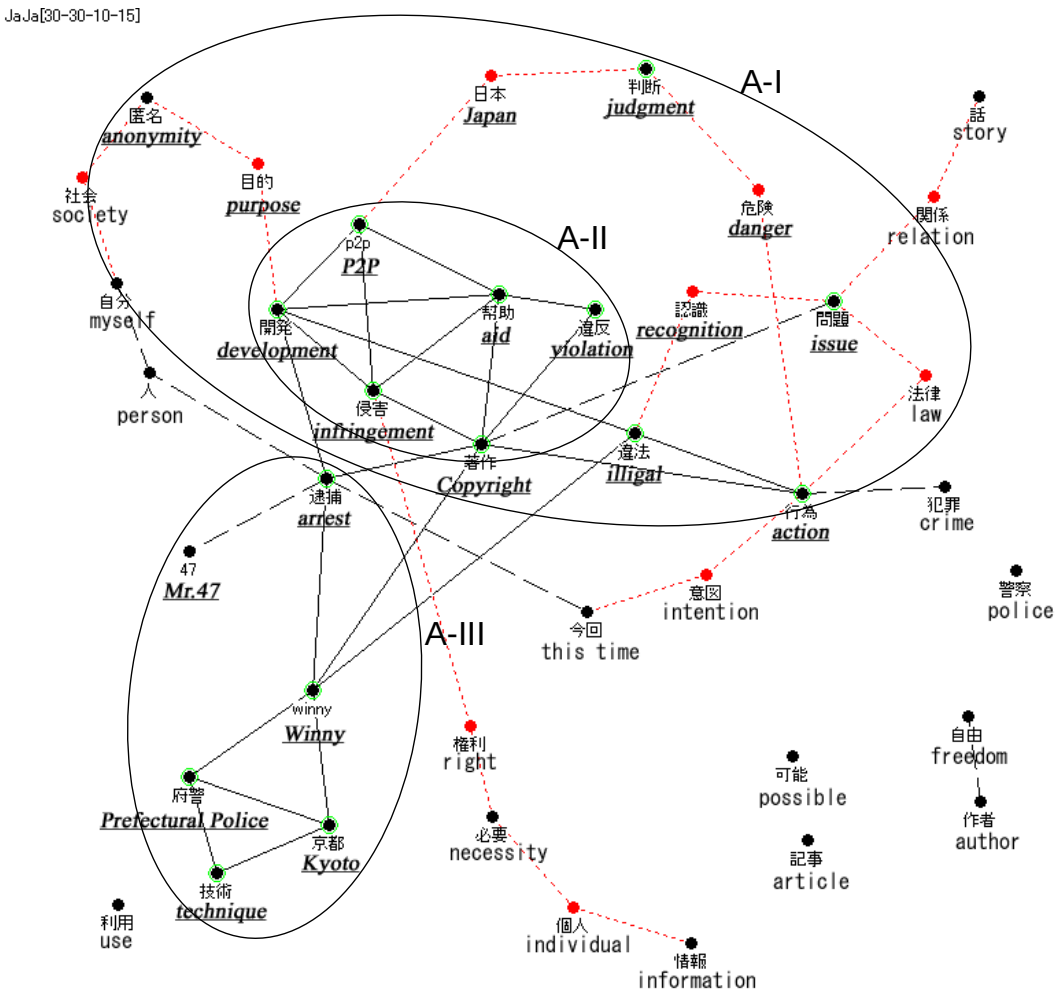
Keygraph was originally developed for extracting keywords in a document. It has been recently applied to many applications [8], recently including discovery of online-game player characteristics [9]. Here rather than giving a detailed explanation of it, we briefly describe an outline of KeyGraph. KeyGraph consists of three major components derived based on building construction metaphor. Each component is described as follows:

Foundations -- sub-graphs of highly associated and frequent terms that represent basic concepts in the data,
Roofs -- terms that are highly associated with foundations,
Columns -- associations between foundations and roofs that are used for extracting keywords, i.e., main concepts in the data.

In KeyGraph, associations between terms are the co-occurrence among them in same sentences, and keywords are the terms in either foundations or roofs that are connected to strong columns. In addition, foundations are depicted by solid lines and their touching black nodes, columns by dotted lines, roofs by red nodes, and keywords by double circles.

3 KeyGraph for visualization of discussions in comments

Sub-graphs in a given KeyGraph are used for deriving scenarios, i.e., textual explanations of the data. If the targeted data are large, the resulting KeyGraph and its sub-graphs will become complicated, from which only scenarios having broad meaning can be derived. To solve this problem, rather than applying KeyGraph to the whole data only once, we apply Keygraph to the whole data first,



Scenario A-I: Issue on recognition and dangerous judgment for illegal action in P2P development with anonymous purpose in Japan
Scenario A-II: P2P development aid for violation and infringement of copyright
Scenario A-III: Arrest of Mr. 47 by Kyoto Prefectural Police due to Winny and the technique

Figure 1: KeyGraph and scenarios of the whole comments (1018 comments)

and then successively to a smaller-size chunk of comments but with a higher score. We anticipate that, through this procedure, scenarios having broad meaning are derived first and followed successively by those with more specific meaning. This should lead to a better understanding of the data.

Existing methods that can be applied to scoring comments in blogs or bulletin board services include RI (Reply-Index, i.e., number of replies), IDM (Influence Diffusion Model [10] or MIR (Measuring Influence Rates) [11]. In this work, however, we directly use the results from the Slashdot moderation system [12], where comments are reviewed and scored from -1 (lowest quality) to 5 (highest quality) by a selected group of users called moderators.

From a given KeyGraph, we derive scenarios by focusing mainly on terms in sub-graphs that include keywords, then with less priority on terms in sub-graphs that include roofs and have clear cluster patterns. These terms are then combined into phrases or sentences exploiting the knowledge on the content of the original article and, when necessary, on the content of its comments where the selected terms reside. Though scenario derivation in our approach is done manually, one can summarize discussions in the comments without the need to thoroughly read all comments.

4 Experiment and Evaluation

We applied KeyGraph to comments on a blog entry titled "Winny developer, Mr. 47, was arrested." [13]. This blog entry is the most active story, having the largest number of comments (1018 comments), in the Hall of Frame of Slashdot Japan. First, Keygraph was applied to all 1018 comments (with the score

range of [-1, 5]), then to the chunk of comments with the score range of [1, 5] (355 comments), and finally to the chunk of comments with the score range of [3, 5] (21 comments). Henceforth, the first, the second, and the third comment chunks are called Data Set A, Data Set B, and Data Set C, respectively. We note here that Data Set A includes Data Sets B, while Data Set B also includes Data Set C.

For generation of KeyGraphs, we used Polaris [14], a data-mining tool with the KeyGraph function, and selected the Jaccard coefficient for computation of associations between terms. With this set of parameters, KeyGraph was applied to Data Sets A, B, and C. Scenarios were then derived based on the procedure given at the end of Section 3. For illustration purpose, in each KeyGraph, the sub-graph corresponding to a derived scenario is superimposed by an oval, and terms used for that scenario are underlined.

4.1 KeyGraph of Data Set A

We applied KeyGraph to the whole comments (Data Set A). Figure 1 shows the resulting KeyGraph and scenarios derived from it. As one can see from this figure, the KeyGraph is complicated. As a result, the derived scenarios are quite general though they give big pictures of discussions in the comments.

4.2 KeyGraphs of Data Sets B and C

Figures 2 and 3 show the resulting KeyGraphs and scenarios of Data Sets B and C, respectively. One can see that sub-graphs are more separated and the derived scenarios become more specific, compared to those in Figure 1. In addition, one can see that scenarios in the three KeyGraphs are related, namely,
Scenarios A-I, B-IV, B-V
Scenarios A-II, B-II
Scenarios A-III, B-III, C-III.

4.3 Scenario Summaries

Table 1 shows a summary of scenarios derived when KeyGraph is applied to the whole comments and a summary of scenarios derived when KeyGraph is applied to multiple comment chunks (our approach). For the latter, scenarios with a lower score are not placed in the summary if they have related scenarios with a higher score. One can see that the latter approach can give not only a higher number of scenarios but also more specific ones than the former approach.

5 Conclusions and future work

We have shown that the proposed approach that applies successively KeyGraph to multiple chunks of dada, each with a different range of scores, is superior to a common approach that applies KeyGraph to the whole data at once. Namely, for comments of the targeted Slashdot Japan's blog entry, the proposed approach gives a higher number of scenarios with more specific meaning than the common approach. With the proposed approach, a user can effectively grasp discussions in the comments of a blog entry without having to read all comments. In our current work, scenario derivation is done manually according to the procedure given in the paper. Our future work is that of automating this procedure.

Acknowledgements

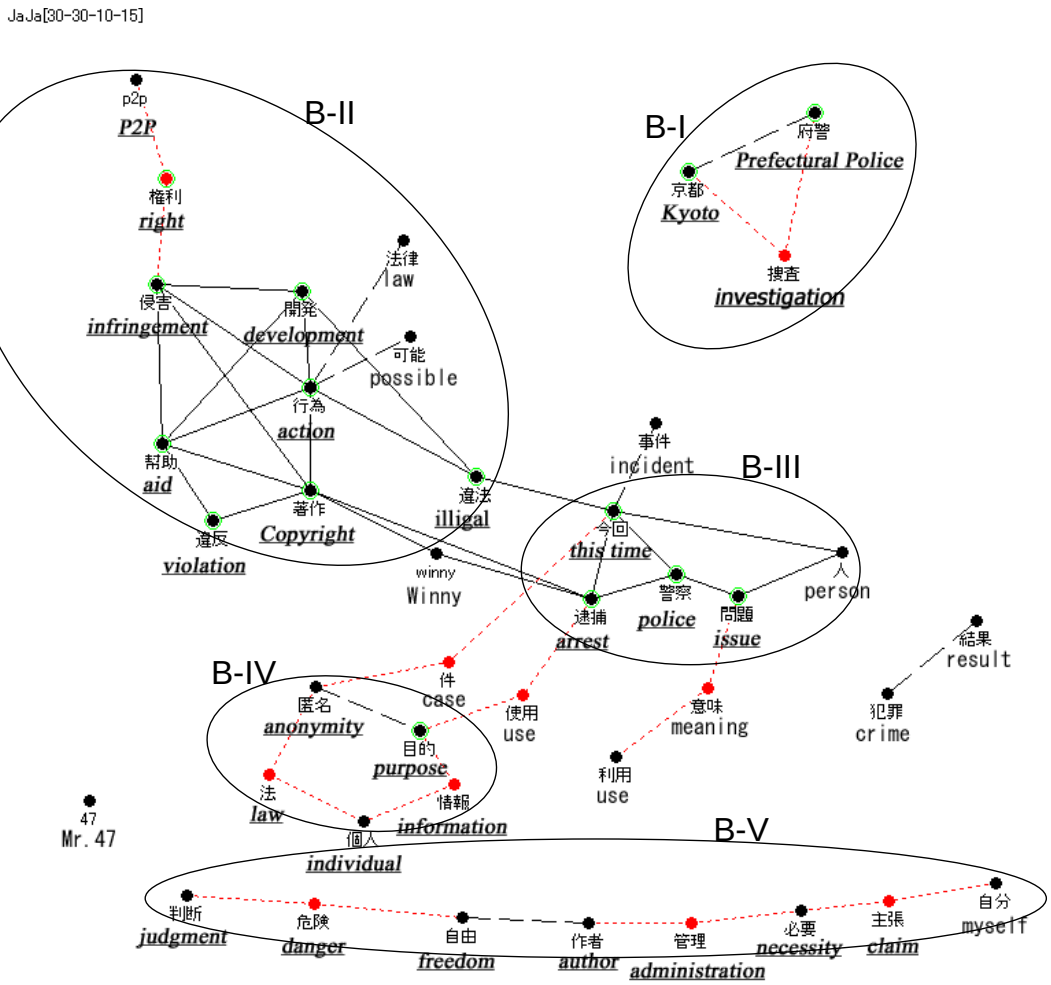
This work has been supported in part by the Ritsumeikan University's **Kyoto Art and Entertainment Innovation Research**, a project of the 21st Century Center of Excellence Program funded by the Japanese Ministry of Education, Culture, Science and Technology; and by Grant-in-Aid for Scientific Research (C), Number 16500091, the Japan Society for Promotion of Science.

References:

[1] *Annual Workshop on Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics.*
http://www.blogpulse.com/www2004-workshop.html
http://www.blogpulse.com/www2005-workshop.html

[2] Glance, N., Hurst, M., Tomokiyo, T.: BlogPulse: Automated Trend Discovery for Weblogs. *Proc. WWW2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, New York (2004)

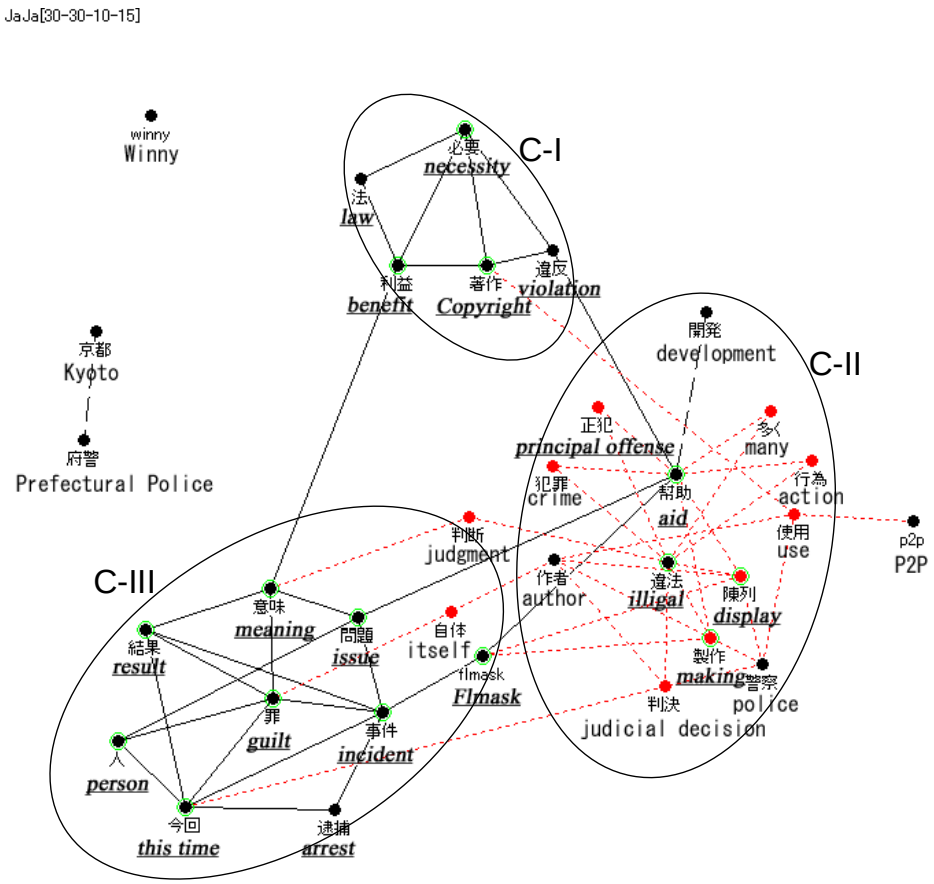
[3] Fujiki, T., Nanno, T., Okumura, M.: Differences between Blogs and Web Diaries. *Proc. WWW2005 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, Chiba (2005)



Scenario B-I: Kyoto Prefectural Police investigation
Scenario B-II: P2P aid and illegal development action for violation and infringement of copyright (right)
Scenario B-III: Issue on this time arrest of the person by police arrest
Scenario B-IV: Anonymous purpose and individual information law
Scenario B-V: Dangerous judgment for a claim on the necessity of author freedom administration

Figure 2: KeyGraph and scenarios of the comment chunk with the moderation score range of [1, 5] (355 comments)

[4] Gruhl, D., Guha, R., Liben-Nowell, D., Tomkins, A.: Information Diffusion through Blogspace. *Proc. WWW2004 Conference*, New York (2004).
[5] Adar, E, Zhang, L., Adamic, L., Lukose, R.: Implicit Structure and the Dynamic of Blogspace. *Proc. WWW2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, New York (2004)
[6] Ohsawa, Y., Benson, N.E., Yachida, M.: KeyGraph: automatic indexing by co-occurrence graph based on building construction metaphor, *Proc. Advanced Digital Library Conference* (IEEE ADL'98) (1998) 12-18
[7] *Slashdot Japan*. http://slashdot.jp/
[8] Ohsawa, Y., McBurney, P. (eds.): *Chance Discovery - Foundation and Its Applications*, Springer Verlag (2003)



Scenario C-I: Necessity and benefit of copyright violation law
Scenario C-II: Principal offence for add of illegal display and making
Scenario C-III: Meaning of the issue on the result of the guilt Flmask incident and this time person arrest

Figure 3: KeyGraph and scenarios of the comment chunk with the moderation score range of [3, 5] (21 comments)

[9] Thawonmas, R., Hata, K.: Aggregation of Action Symbol Subsequences for Discovery of Online-Game Player Characteristics Using KeyGraph. Proc. IFIP 4th International Conference on Entertainment Computing (ICEC 2005), September, 2005, Sanda, Japan, published in *Lecture Notes in Computer Science*, Fumio Kishino et al. (Eds.), vol. 3711, pp. 126-135 (2005).

[10] Matsumura, N., Ohsawa, Y., and Ishizuka, M.: Influence Diffusion Model in Text-based Communication. *Proc. 11th International World Wide Web Conference* (WWW02) (2002)

[11] Tsuda, K., Thawonmas, R.: Keyword Discovery by Measuring Influence Rates on Bulletin Board Services. Proc. IFIP 4th International Conference on Entertainment Computing (ICEC 2005), September, 2005, Sanda, Japan, published in *Lecture Notes in Computer Science*, Fumio Kishino et al. (Eds.), vol. 3711, pp. 148-154 (2005).

[12] *Slashdot Moderation.* http://slashdot.org/moderation.shtml

[13] *The blog entry "Winny developer, Mr. 47, was arrested."* http://slashdot.jp/articles/04/05/10/0017250.shtml?topic=|

[14] *Polaris.* http://www.chokkan.org/research/chance_discovery/polaris/

Table 1: KeyGraph scenario summaries

Approach	Summary
KeyGraph for the whole comments	Issue on recognition and dangerous judgment for illegal action in P2P development with anonymous purpose in Japan
	P2P development aid for violation and infringement of copyright
	Arrest of Mr. 47 by Kyoto Prefectural Police due to Winny and the technique
KeyGraph for multiple comment chunks (our approach)	Kyoto Prefectural Police investigation
	P2P aid and illegal development action for violation and infringement of copyright (right)
	Anonymous purpose and individual information
	Dangerous judgment for a claim on the necessity of author freedom administration
	Necessity and benefit of copyright violation law
	Principal offence for add of illegal display and making
	Meaning of the issue on the result of the guilt Flmask incident and this time person arrest