Observing Evolution of Web Communities

Masashi Toyoda and Masaru Kitsuregawa
Institute of Industrial Science, University of Tokyo
4-6-1 Komaba Meguro-ku, Tokyo, JAPAN
{toyoda, kitsure}@tkl.iis.u-tokyo.ac.jp

ABSTRACT

We propose a method for observing evolution of web communities. A web community is a set of web pages created by individuals or associations with a common interest on a topic. So far various techniques have been developed to extract web communities by link analysis. Since a community represents a certain topic, we can understand when and how the topic emerged and evolved in the Web. We have developed a system for observing evolution of web communities by comparing three Japanese web archives periodically crawled in 1999, 2000, and 2001. It first extracts whole web communities, and their relevances from each archive. Then it provides a web community evolution viewer. The user can observe the evolution, using the relevances and evolution metrics, such as growth rate and novelty. Several evolution examples are shown using our system.

Keywords

Link analysis; Web community; Web community chart; Evolution

1 INTRODUCTION

We propose a method for observing evolution of *web communities*. A web community is a collection of web pages created by individuals or associations with a common interest on a topic, such as fan pages of a baseball team, and official pages of computer vendors. Recent research on *link analysis* [2, 5, 4, 1] shows that we can identify a web community on a topic from densely connected structure of the web graph, in which nodes are web pages and edges are hyperlinks. Although these techniques can automatically identify communities, they have not considered evolution of communities yet.

Since a web community represents a certain topic, we can understand when and how topics emerged and evolved in the Web. For example, when mad-cow disease became a serious problem, and what kinds of pages have been created about the disease. Such information is important in the following situations: (1) answering historical queries about topics on the Web; (2) observing the emergence of quality web communities on a specific topic; (3) understanding sociology of web community creation related to real social activities.

For extracting such information, we have developed a system for observing evolution of web communities by comparing three Japanese web archives periodically crawled in 1999, 2000, and 2001. This system first extracts whole web communities and their relevances from each archive. This is based on our previous work of *web community chart* [6] that is a graph of communities, in which relevant communities are connected by edges. Compared with prior work such as [2, 5, 4, 1], the main advantage of our community chart is existence of relevances between communities. We can navigate through related communities, and can locate evolution around a particular community. Finally, our system provides an evolution viewer that allows the user to extract when and how communities evolved, using the relevances and evolution metrics, such as growth rate, and novelty.

Year	Period	#URLs	#links	#seeds	#communities
1999	Jul. to Sep.	16.8M	126M	627K	70K
2000	Jun. to Aug.	14.1M	100M	600K	68K
2001	Early Oct.	40.5M	343M	1135K	131K

Table 1: Details of web archives

In the following, we first briefly describe our technique for building the web community chart. Then show our web archives and community charts built from them. Finally, we demonstrate the evolution viewer with some examples.

2 WEB COMMUNITY CHART

Our web community chart is a graph that consists of communities as nodes, and weighted edges between relevant communities. The weight of each edge represents the relevance of communities at both ends. The main idea of our technique is applying a related page algorithm, Companion–[6], to a number of seed URLs, then investigate how each seed derives other seeds as related pages. Companion– takes a seed URL as an input, and calculates related pages to the seed based on the concept of hubs and authorities [3]. It returns authorities near the seed as related pages.

To identify communities and to find their relationships, we focus on symmetric derivation relationship (SDR) between s and t, that is s and t derive each other by the related page algorithm. This often means that the both pages s and t are pointed to by similar sets of hubs, and have strong similarity. Using this relationship, we define that a community is a set of pages that are connected by the symmetric relationships, and that two communities are related, if a member of one community derives a member of an another community.

The community chart is built from a given seed set. We first apply Companion– to each seed in the seed set. Then, we put seeds into a community, when they are connected by symmetric relationships. Every two communities are connected by a edge, when there exists a derivation from a seed in one community to an another seed in the other community. A weight of a edge is the number of such derivations. Refer to [6], for more detailed descriptions.

3 WEB ARCHIVES AND CHART EXTRACTION

Our system first builds the web community chart from each web archive as described in Section 2. We used three web archives of Japanese web pages (in jp domain) that were periodically crawled in 1999, 2000, and 2001 (Details are shown in Table 1.) The 2000 archive is smaller than one in 1999, because we randomly lost about 3 million pages due to disk crash. The size of the 2001 archive is more than twice of other archives, because we improved the crawling rate significantly in 2001.

From each archive, the system selects a seed set including popular URLs (with 3 or more inlinks from different servers), then builds a web community chart using the seed set. The number of seeds and extracted communities are also shown in Table 1.

4 A WEB COMMUNITY EVOLUTION VIEWER

By using our evolution viewer, the user can observe various behaviors of communities, such as growing, emerging, merging, and splitting. Our viewer visualizes the evolution of given communities by comparing web community charts through two years as shown in Figure 1. Each communities are represented as a rectangle including the list of its URLs, and are arranged in columns that represent years. Each URL in communities can be browsed with a web browser by clicking it with the mouse. Labels on communities are automatically attached by selecting frequent keywords from anchor texts that point to URLs in the community.

When the user selects communities by specifying keywords or a URL, it finds and displays the corresponding community in each year, which share the most URLs with the given community in that year. As shown in Figure 1, These communities are arranged horizontally, and lines are drawn between them, so that the user can easily compare their differences. If the user needs more detailed view (including merging and splitting), the viewer can show all communities that share URLs with the given community.

Then the user can sort or filter communities by several evolution metrics, such as growth rate, and novelty. These metrics are calculated by counting such as newly appeared URLs, and preserved URLs in communities. For example, the novelty is a ratio of newly appeared URLs in the community. The user can find emerged communities, by sorting communities by their novelty. In the following, we show some evolution examples observed with the evolution viewer.

Figure 1 is also an example of a stably growing community. That is, the number of URLs in the community grows, preserving most of its URLs. We can find such communities by sorting all communities by their growth rate, and filtering out unstable communities. Figure 1 shows a fan community of Major League Baseball stably grew in Japan. Especially, it rapidly grew from 2000 to 2001, since a Japanese star player Ichiro Suzuki is transferred to Seattle Mariners, and he became an outstanding player in Major League (You can see his name, "ichiro," in some URLs in the community in 2001).

We can find emerged communities related to a particular community. In Figure 2, we first select a community of Islam and muslim information in 2001, then sort the related communities by their novelty from 2000 to 2001. There emerged two communities about Afghanistan. We guess that such hubs are rapidly created after the attack on America by terrorists on 11 September, 2001. (Note that our web archive in 2001 was crawled in early October.) From this example, we can see that communities grow very quickly when their topic has a great impact to the real society.

References

[1] G. W. Flake, S. Lawrence, and C. L. Giles. Efficient Identification of Web Communities. In *Proceedings of KDD 2000*, 2000.

[2] D. Gibson, J. Kleinberg, and P. Raghavan. Inferring Web Communities from Link Topology. In *Proceedings of HyperText98*, 1998.

[3] J. M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, 1998.

[4] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Extracting large-scale knowledge bases from the web. In *Proceedings of the 25th VLDB Conference*, 1999.

[5] S. R. Ravi Kumar, Prabhakar Raghavan and A. Tomkins. Trawling the Web for emerging cyber-communities. In *Proceedings of the 8th World-Wide Web Conference*, 1999.

[6] M. Toyoda and M. Kitsuregawa. Creating a Web Community Chart for Navigating Related Communities. In *Conference Proceedings of Hypertext 2001*, pages 103–112, 2001.

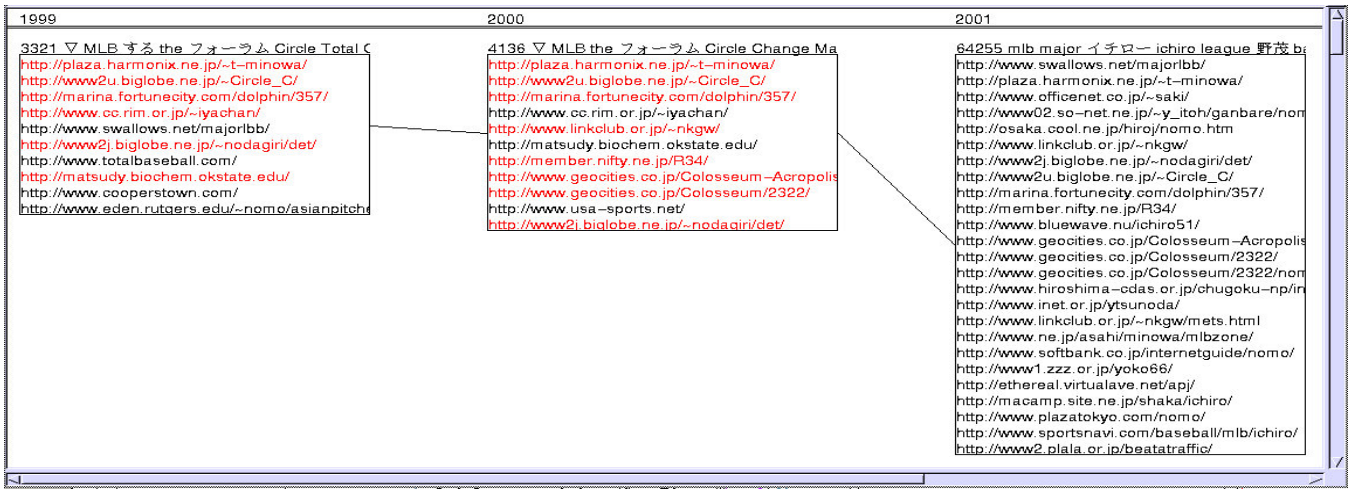

Figure 1: Evolution of a MLB fan community

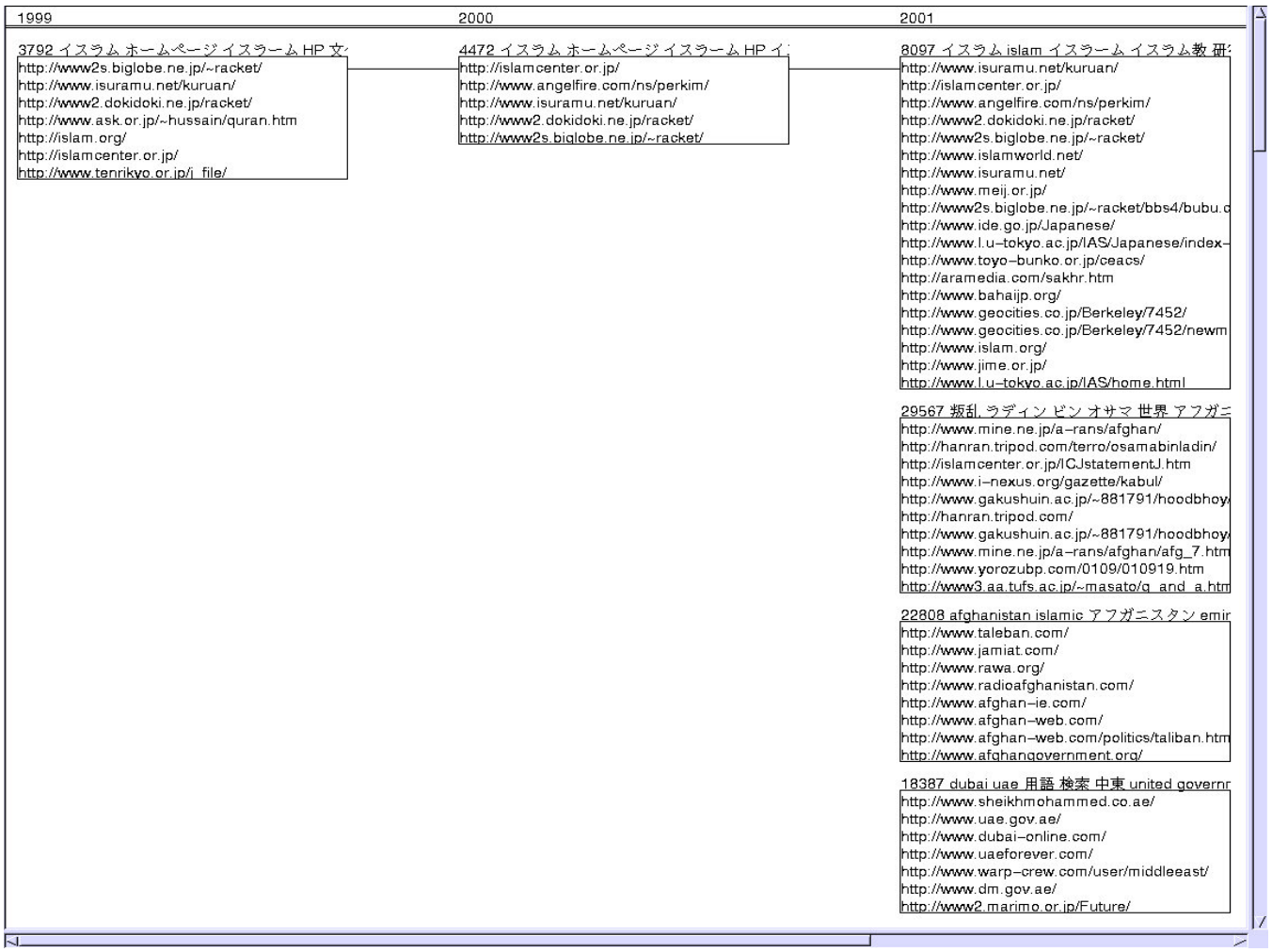

Figure 2: Communities about Afghanistan emerged around an Islam information community