A Web Community Chart for Navigating Related Communities

Masashi Toyoda and Masaru Kitsuregawa
Institute of Industrial Science, University of Tokyo
7-22-1 Roppongi Minato-ku, Tokyo, JAPAN
+81-3-3402-6231 (ext. 2358)
toyoda, kitsure@tkl.iis.u-tokyo.ac.jp

ABSTRACT

Recent research on hyperlink analysis has shown the existence of numerous communities on the Web. A community is a collection of web pages created by individuals or any kind of associations that have a common interest on a specific topic. We have developed a technique to create a community chart for navigating from one community to other related communities. The community chart allows the user to perform new type of navigation through the Web. It provides additional paths not only to related pages, but also to related communities. The community chart can be also used for a 'What's Related Community' service, which provides not only a community including a given URL, but also shows other related communities. In this paper, we briefly describe the technique and experiments for creating a community chart of companies, organizations, and associations, from thousands of seed pages. The result shows that our technique clearly identifies communities on various categories of business, and correctly connect related communities for navigation.

Keywords

Link analysis; Community; Related communities

1. INTRODUCTION

Recent research on hyperlink analysis has shown that numerous web communities can be automatically identified from the Web [1, 2, 3]. A community is a collection of web pages created by individuals or any kind of associations that have a common interest on a specific topic, such as fan pages of a baseball team, and official pages of PC vendors. However, those techniques have not concerned the relationship between communities. Our goal is not only to identify communities, but also extract relationships between communities, and create a global community chart for navigating from one community to other related communities. For example, one can navigate from fan pages of a baseball team to official pages of baseball teams, then to fan pages of an another baseball team. The community chart allows the user to perform new type of navigation through the Web. It provides additional paths not only to related pages, but also to related communities. The community chart can be also used for a community version of 'What's Related' service, which provides not only a community (a set of related pages to a given URL), but also shows other related communities.

As the first step to our goal, we developed a technique that creates a subset of the global community chart from thousands of seed URLs on a broad topic. We created a community chart using around 5000 URLs of companies, organizations, and associations. The result shows that our technique clearly identify communities on various categories of business, and related communities are correctly connected for navigation.

2. METHOD FOR CREATING A COMMUNITY CHART

The main idea of our method is applying a related page algorithm, Companion [5], to a number of URLs, then investigate how each URL derives other URLs as related pages. Companion takes a seed URL as an input, then calculates related pages to the seed. It is based on the concept of hubs and authorities [4], and it finds authorities near the seed as related pages.

To identify communities and to find their relationships, we investigate the relationship between a seed URL and related pages derived by the algorithm. Consider that a page s derives a page t as a related page, and t also derives s as a related page. This often means that the both pages s and t are pointed to by similar sets of hubs. For example, a fan page of a baseball team derives other fan pages as related pages. When we apply the related page algorithm to one of the other fans, the page derives the original fan, because those fan pages are mutually linked by each other, that is, pointed to by similar sets of hubs. If each fan derives other fans as related pages, we can consider that these fans form a fan community. Under these observations, we say that s and t *respect each other*, when s and t derive each other using the related page algorithm. Using this Respect-Each-Other (REO) relationship, we refine the definition of communities and their relationships. We define that a community is a set of pages that are connected by the REO relationships, and that two communities are related, if a member of one community derives a member of an another community.

Based on these definitions, we create the community chart, which is a graph that includes communities as nodes, and weighted edges between related communities. A weight of a edge represents the strength of the relationship. We first extend a given seed set by applying Companion to each seed and gathering results. Then, we apply Companion to each seed in the extended seed set. Finally, we put seeds connected by REO relationships into a community, and create edges between related communities. A weight of a edge is the number of derivations between members of two communities. The details of our method will be published later.

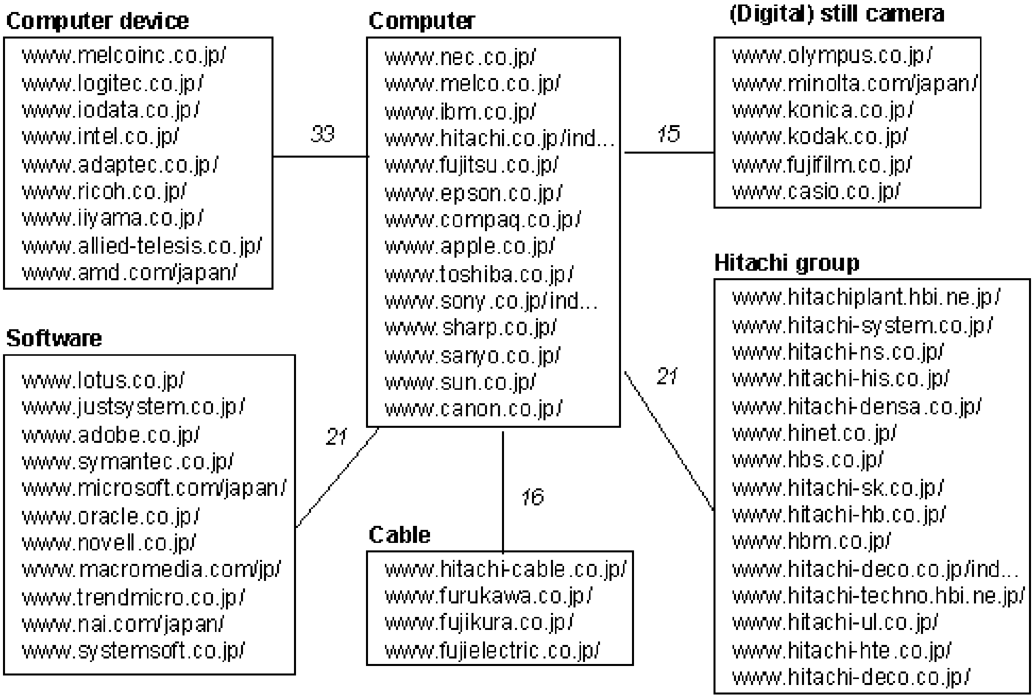

Figure 1: A part of community chart

3. EXPERIMENTS

Our data set for experiments is an archive of Japanese web pages that includes about 17 million pages (90GB). As a seed set, we use a manually maintained URL list that includes about 5,000 unique URLs of companies, associations, organizations, and schools. Note that the seed set is a small subset of all the company URLs in the archive.

Our community chart includes about 1,800 communities. Figure 1 shows a part of the community chart that consists of communities connected by highly weighted edges. Each box represents a community that includes list of URLs. Note that the category label on each box is attached manually. The number attached to each edge denotes the weight. We select the 'Computer' community as a center, since it has most edges in the chart. Each community, around the 'Computer', has edges to the 'Computer', and these weights are more than 15. Actually, there are more communities around the 'Computer' connected by lower weighted edges, that are not shown in Figure 1.

As shown in Figure 1, these communities are clearly classified and actually related to the 'Computer' community. The 'Software' community includes Lotus, Microsoft, etc., and obviously related to the 'Computer'. The companies in the 'Cable' community provides cables and optical fibers. The 'Hitachi group' community is slightly different from other communities. Although Hitachi is famous as a computer company, it is also one of the largest conglomerate in Japan. Since, all the companies in the 'Hitachi group' derive 'www.hitachi.co.jp' as one of authorities, the community has a highly weighted edge to the 'Computer'.

4. SUMMARY

We have proposed the technique to create the community chart from thousands of seed URLs, and applied the technique on the seed set of companies, organizations, and associations. The result chart consisted of many clearly classified communities by their categories of business, and navigation paths between related communities. The community chart provides additional paths not only to related pages, but also to related communities. The chart can also be used for a 'What's Related Communities' service, which provides not only related pages to a given URL, but also other related communities.

5. REFERENCES

[1] David Gibson, Jon Kleinberg, and Prabhakar Raghavan. Inferring Web Communities from Link Topology. *Proceedings of HyperText 98*, 1998.

[2] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan and Andrew Tomkins. Trawling the Web for emerging cyber-communities. *Proceedings of the 8th International World Wide Web Conference*, 1999.

[3] Gary W. Flake, Steve Lawrence, and C. Lee Giles. Efficient Identification of Web Communities, *Proceedings of KDD 2000*, 2000.

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

[5] Jeffrey Dean and Monika R. Henzinger. Finding related pages in the World Wide Web. *Proceedings of the 8th International World Wide Web Conference*, 1999.