Automatic Generation of Link Collections and their Visualization

Osamu Segawa
Chubu Electric Power Co., Inc.
Nagoya, 459-8522 Japan
Segawa.Osamu@chuden.co.jp

Jun Kawai
TIS Inc.
Osaka, 564-0051 Japan

Kazuyuki Sakauchi
TIS Inc.
Osaka, 564-0051 Japan
sakauchi@karl.tis.co.jp

ABSTRACT

In this paper, we describe a method of generating link collections in a user-specified category by comprehensively collecting existing link collections and analyzing their hyperlink references. Moreover, we propose a visualization method for a bird's-eye view of the generated link collections. Our methods are effective in grasping intuitively the trend of significant sites and keywords in a category.

Categories and Subject Descriptors

H.5.4 [**Information Systems**]: Hypertext/Hypermedia

General Terms

Algorithms

Keywords

Link collection, Hyperlink analysis, Visualization

1. INTRODUCTION

Link collections are useful resources in information retrieval on the Web. However, their creation and maintenance require much cost. Moreover, even if we observe a link collection page in a category, both coverage and usefulness of the sites referred from the page are not always guaranteed. In a popular or topic field, it is expected that a number of well-classified link collections exist. Therefore, by comprehensively collecting existing link collections in various categories and analyzing their hyperlink references, we will be able to identify useful sites and grasp the trend of the Internet. Based on this concept, in this paper, we describe a method of generating and visualizing link collections in a user-specified category.

2. AUTOMATIC GENERATION OF LINK COLLECTIONS

The input to the system is only a user-specified category word. The algorithm is shown below.

1. The acquisition of existing link collections is performed using a Web crawler. Gathering pages that are determined to be link collection pages are saved. The rules for this determination are as follows: 1) more than a certain number of external links exist; 2) the title tag includes the string "link collection" or "link"; and 3) the file name includes the string "link".

WWW 2005, May 10–14, 2005, Chiba, Japan.
ACM 1-59593-051-5/05/0005.

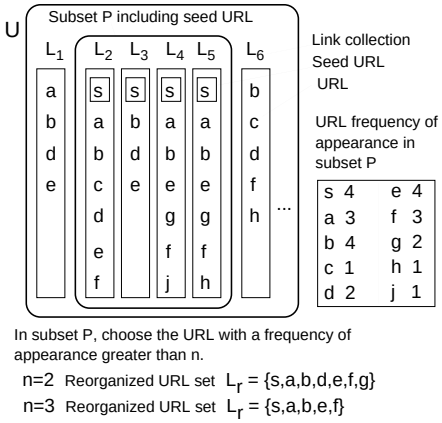

Figure 1: Method of generating link collections.

2. As a preprocess, a content page is acquired from all the external links included in the link collection set, and feature words are extracted for every URL. Here, feature words are selected by evaluating the tf*idf value of the words included in the title tag and text of each content page.
3. Let U be the list of external URLs included in the link collection set in the user-specified category. Among the elements of U, the URL for which the frequency of appearance is greater than a certain number, or the URL that includes a category word in its feature words is selected as a seed URL (*i.e.,* a typical site in the category).
4. From among the elements of subset P that includes the seed URL (see Fig.1), choose URLs with a frequency of appearance greater than n ($n > 1$) and reorganize a new URL set. Then, determine the theme label words for each reorganized URL set based on the feature words of the set.
5. Finally, the degree of semantic relevance between the category and each reorganized URL set is evaluated, and the sets that have high relevance scores are selected as final link collections. Here, semantic relevance score is defined as the cosine similarity measure of the vector space model between the feature vector of the category and the feature vector of a reorganized URL set.

3. VISUALIZATION OF LINK COLLECTIONS

Using the method described above, a number of link collections are generated by the seed URL's difference. If a

large number of link collections are displayed, an overview of the results is not always easy.

Therefore, we propose a method of visualizing generated link collections on a plane. In our method, sites and feature words that are elements of a link collection are displayed as a node on a plane, and a graph structure in which an arc expresses the connection to a link collection is generated. In the visualization, nodes connected with the arc receive tension mutually by the dynamics system similar to the spring model. At the same time, each node is distributed in the direction of the perimeter by a weak expansion pressure from the center of the plane. According to the layout procedure, a node belonging to many link collections remains near the center of a visualization map, while a node belonging to specific link collections localizes near the perimeter and forms a cluster. If a node has a high arc density, it is a worth-visit site or a significant keyword in a category.

4. IMPLEMENTATION

4.1 Search directory

We have developed a search directory system using the proposed method described in section 2. The system generated link collections in 607 categories automatically (in January 2005). An example of the generated link collection (category: RDF) is shown in Fig.2.

4.2 Visualization tool

We have developed a visualization tool based on the method described in section 3. An example of the visualization result is shown in Fig.3. This figure shows the visualization result of 8 link collections (category: RDF). On the visualization map, the nodes of sites and feature words are displayed with no duplication. A red oval node indicates a site, and a green oval node indicates a feature word. A yellow circle expresses the organization as a link collection. The more arcs there are in a site or a feature word node, the deeper the color of the node is. In this example, the W3C RDF page (the significant site in this category) is located in the center of the map, and the other related sites in RDF field are located around the perimeter of the map.

The map provides a kind of search interface. If users double-click a site node, they can access the corresponding page via a browser. Moreover, if users click a word node, the site that includes it as a feature word is highlighted.

5. RELATED WORK

In the field of Web community analysis, Kleinberg proposed the *HITS* [1]. The aim of the *HITS* is to find *authorities* and *hubs* in a category. The focus of our method is to generate various new *hubs* (link collections) using existing link collections. Kumar *et al.* proposed the *Web Trawling* [2]. Our method is similar to Kumar's in that the co-occurrence of references is analyzed. In our method, on the other hand, there is no premise of a complete bipartite graph structure in the discovery of a site set.

Moreover, in the field of Web visualization, several methods concernig the relevance between pages (sites) or between keywords have been proposed. The former example is TouchGraph's *Google Browser* [3], a visualization tool for search results. The latter example is *Keyword Map* [4]. Our method is different from these approaches in two points. One is that our method does not depend on commercial search engines, and the other is that our method visualizes the relevance of a site and a keyword simultaneously.

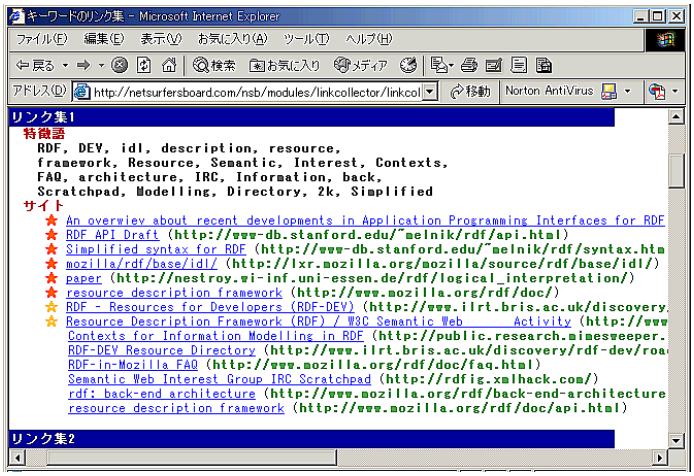

Figure 2: Example of the generated link collection (category: RDF).

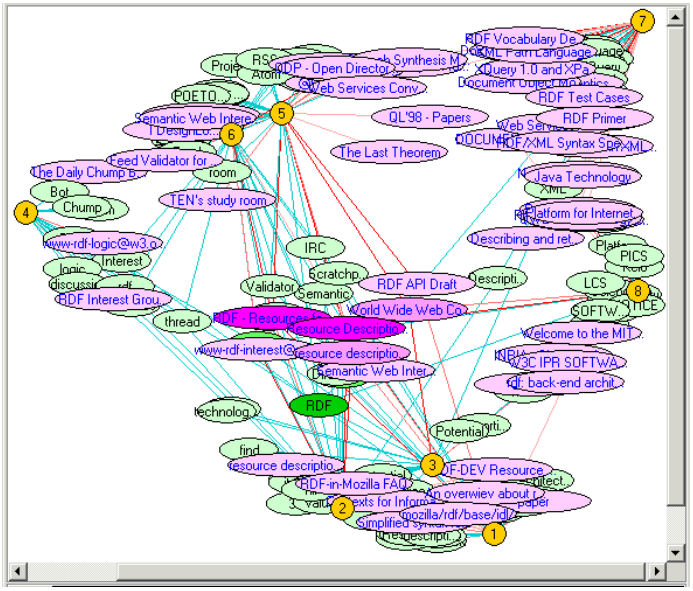

Figure 3: Example of visualizing link collections (category: RDF).

6. SUMMARY

In this paper, we have proposed a method of automatically generating link collections in a user-specified category, and a method of visualizing the generated link collections. Our methods are effective in grasping intuitively the trend of significant sites and keywords in a category.

7. REFERENCES

[1] J. Kleinberg, "Authoritative sources in a hyperlinked environment", in *Proc. 9th ACM-SIAM Symposium on Discrete Algorithms*, 1998.

[2] R. Kumar *et al.*, "Trawling the web for emerging cyber-communities", in *Proc. of 8th WWW Conference*, 1999.

[3] TouchGraph http://www.touchgraph.com

[4] Y. Takama *et al.*, "Finding landmarks in keyword map based on immune network", 2nd International Symposium on Advanced Intelligent Systems, 2001.