

Web Link-Analysis for Automated Hot Web Directory (Position Paper)

Hiroshi TSUDA
Fujitsu Laboratories Ltd./INTAP
4-1-1 Kamikodanaka, Nakahara-Kawasaki 211-8588, JAPAN
e-mail: htsuda@jp.fujitsu.com

(Keyword: hyperlink analysis, Web directory, automated metadata creation)

1. INTRODUCTION

To cope with the information overload stemming from with the growth of the Internet and intranets, domain-specific portals such as vertical portals (Vortals) and enterprise information portals (EIP) are becoming increasingly important. Yahoo-like Web directories are still popular as an interface for organizing selected Web pages. Compared to robot-based search engines, directories have a small numbers of irrelevant documents and are provide an easy user interface.

As determined by Yahoo Web surfers and the thousands of volunteers working on the Open Directory Project (ODP¹), however, most directories are maintained manually, and at a very high cost.

2. AUTOMATED METADATA CREATION

To reduce the maintenance const, we have been doing researches on automated metadata creation especially for semi-automated Web directories [1, 2].

Web contains varieties of information in various page types. Link-based popularity information such as Google's PageRank is also useful to eliminate garbage information. As the Web is rapidly changing, such "trend" information is also required.

We developed a hyperlink-based approach to automatically derive the following type of metadata. [1, 2]:

- navigation type: menu, links, contents, multi media files, ...
- popularity: (link-based) page ranking
- relatedness: link topological distance from pages in the target domains (communities).

¹<http://www.dmoz.org/>

*Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission by the authors.
Semantic Web Workshop 2002 Hawaii, USA Copyright by the authors..*

- trend: movement of Web site popularities as the gradient of the linear regression of temporal sequence of page popularities.

Figure 1 is a snapshot of our automated Web directory (compound category : "Middle East" (area) and "Terrorism" (genre)). Figure 2 shows the movement of page popularity ranking of the RAWA page in Afganistan. After the terrorism in 9/11/2002, you can see the page gained the popularity.

3. DISCUSSION: CREATION AND MAINTENANCE OF SEMANTIC WEB METADATA

As a member of a Semantic Web committee of Interoperability Technology Association for Information Processing, Japan (INTAP²), I am doing research and survey on Web Metadata, Ontology, and their processing techniques.

To realize the Semantic Web, one of the serious problems is to create and maintain metadata as the management of Web directories previously mentioned. For this WWW2002 Semantic Web Workshop, I would like to have discussions about the maintenance process of the Semantic Web metadata:

- How automated approaches are possible for SW Metadata?
- How to maintain metadata and ontologies in the rapidly changing Web world.
- How to find authoritative metadata? META tags in HTML were used as "word spam"s for search engines. How SW avoid the "metadata spam" problem?

4. REFERENCES

- [1] H. Tsuda, T. Ugai, and K. Misue. Link-based Acquisition of Web Metadata for Domain-specific Directories. In *PKAW2000*, pages 317–324, 2000.
- [2] H. Tsuda, T. Ugai, and K. Misue. An Approach to Automated Web Metadata Creation for Web Directories. In *Proc. 19th Symposium on Informatics*, pages 17–24. Information Processing Society of Japan, 2002. (in Japanese).

²<http://www.intap.or.jp/>



Figure 1: A compound category of our automated Web directory: “Middle Eeast” and “Terrorism”

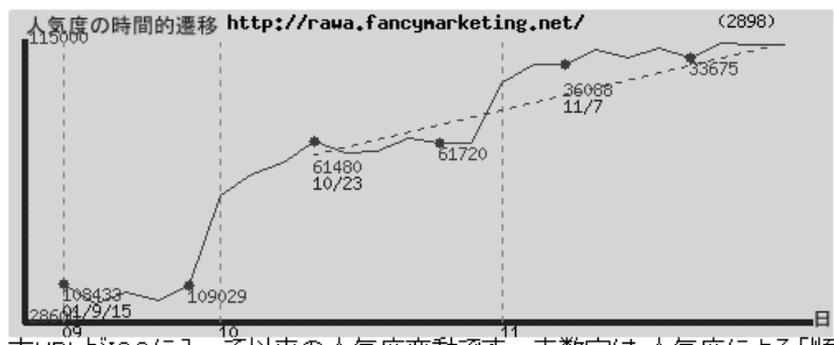


Figure 2: Movement of popularity ranking of RAWA (Revolutionary Association of the Women in Afghanistan) home page from Sept. to Nov. in 2001