Extracting Topic Maps from Web histories by clustering with Web structure and contents

Motohiro Mase
CISS, IGSSE, Tokyo Institute of Technology
4259 Nagatsuta, Midori, Yokohama, Japan
m_mase@nii.ac.jp

Seiji Yamada
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda, Tokyo, Japan
seiji@nii.ac.jp

Abstract

In this paper, we propose a clustering method to extract Topic Maps from the Web browsing history. We improve the structure-based hierarchical clustering method using the contents similarity of the pages and the weight by the types of links and the hierarchical difference of the directories in which the pages are located. The topic maps show the topics that user has seen or not in Web browsing and the relationships between the topics. Using the Web browsing history, we experimentally extract the topic map and evaluate it.

1 Introduction

Information gathering using a huge amount of Web pages is very useful, and essential for users. The Web pages increase every year, and exceed at least 11.5 billion in January, 2005[1]. It is very difficult to look for information that agrees with user's purpose from the Web. In many cases, users find target information by using a search engine such as Google. On the one hand, users have the problems that revisiting a page visited once and organizing the information from the Web are difficult tasks. In 10th GVU WWW User Survey[4], 547 of 3,291 users regarded not being able to return to a page once visited as one of the biggest problems in using the Web, and 908 of 3,291 users regarded not being able to efficiently organize the gathered information as the biggest problems. Users typically have to look for a target page from bookmarks or favorites and browsing history manually, but it is not convenient enough. It is therefore necessary to support revisiting pages and organizing the information.

We propose a method that extracts the topics of the Web pages and the relationships between the topics from Web histories pages of users and vizualize them as Topic Maps. User can find the topics of Web histories pages and neighborhood pages that have not been read and the relationship between the topics. Then, user can organize the collected information automatically and easily access the Web pages using the topic as the query. In introducing weight by contents similarity and types of links between pages and hierarchical difference of directories in which the pages are located into structure-based clustering, we extract Topic Maps that provide the topic and the relationships between them from Web histories pages of users.

2 Related Works

Many researches have been reported in the area of Web history visualization. Domain Tree Browser[8] provides 2D visualized history tree by Web browsing with thumbnail images of Web pages. VISVIP[2] visualizes user's path as 2D graph that node represents a Web page and an edge represents link between pages. WebPath[3] visualizes browsing history path as 3D graph. Browsing Icons[7] shows a task and a session based 2D graph that dynamically draws Web history path. These researches basically visualize user's Web browsing history. Our research, however, classifies the information from Web by topics and provides the relations between the topics in addition to visualization.

3 Topic Maps

Topic Maps are ISO/IEC 13250 standard and a solution for organizing, accessing and retrieving information[5]. A topic map are composed of topics that represent any concepts and subjects in real world, associations that represent the relations of the topics and occurrences that represent the relations between the topics and related information resources. We extract the three elements from Web pages using our proposed method that is based on a hierarchical clustering method as follows. Topics are the extracted clusters, associations are the edges between the clusters and oc-

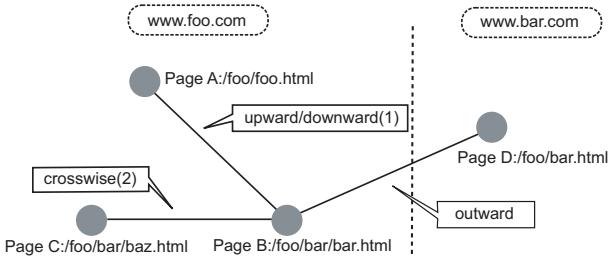

Figure 1. Types of links

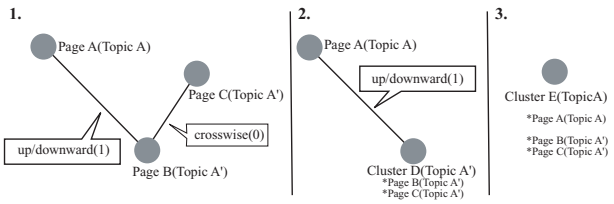

Figure 2. Weight by types of links

currences are the related Web pages that compose the clusters.

4 Proposed Method

In this paper, we propose the method that is based on structure-based hierarchical clustering method using constraint of Web structure and weights from contents similarity and types of links between Web pages.

Topic Maps need to represent not only topics that is from the extracted clusters but also relationships between the topics. The conventional contents-based hierarchical clustering can extract clusters from Web pages and strength of similarities of the clusters. It can not provide types of relationships between the clusters. Then, we focus on Web structure and types of links between Web pages to extract the relationships in addtion to the clusters.

4.1 Web structure

Much of links between Web pages are made by the authors considering these pages to be similar. The links probably include the relationships between the topics of the linked pages. Then, we extract the relationships by the clustering method with the constraint of the Web structure. The constraint is merging only the linked Web pages. The clustering with the constraint results in clusters and edges between them. We regard the clusters and the edges as topics and associations.

1

4.2 Types of links

The authors of Web pages probably make a directory in which the pages are located with following intention. 1)Web pages in a same directory have a mutual topic. 2)A topic of a page in the lower directory is specialization of a topic of a page in upper. 3)The similarity between topics of Web pages depends on difference of directories in which the pages are located. Therefore, we presume types of the links with the directories. Parasite[9] categorized the types of links with the hierarchical relationships of the directories in which the linked pages and introduced heuristics to estimate meanings of the links. In this paper, we consider the hierarchical difference of the directories in which the linked pages in addition to the relationship of the directories and categorize the links as the following three types.

- *upward/downward*

  This is a relationship between a page in an upper directory and a page in a lower directory in a same Web site. For Figure1, the relationship between page A and page B is this type. The number in the parentheses is d that shows hierarchical difference between the directories. The topic of the page in the lower directory means specialization of the topic of the page in the upper. The similarity of the topics is inversely proportional to hierarchical difference.

- *crosswise*

  This is a relationship between pages that are in different directories in a same Web site. For Figure1, the relationship between page B and page C is this type. The d is 2(From '/foo/bar/' to '/foo/' (1), from '/foo/' to '/foo/baz/' (2)) . The similarity of the topics is inversely proportional to hierarchical difference.

- *outward*

  This is a relationship between pages in other Web sites. For Figure1, the relationship between page B and page D is this type. Both pages have similar topics, basically.

We weight links based on these types as follows.

- Weighting in inverse proportion to the value of d. It results in prioritizing the link that has high value of similarity between the topics.

- Weighting by giving preference to the link of $crosswise$ type over the link of $upward/downward$. For Figure2, merging the link between the pages B and C is given priority over merging the link between the pages A and B. It results in extracting the relationships

between the topics A and A' from the links between the page A and cluster D.

4.3 Clustering

Our proposed method is based on the Newman's method[6]. The Newman's method is structure-based hierarchical clustering method and forms clusters to maximize modularity Q that is evaluation function to check the validity of network division.

We introduce the weight by contents similarity and the types of the links between the pages and the hierarchical difference of the directories in which the pages are located into the modularity Q as follows.

$$\begin{aligned} Q &= \sum_i (e_{ii} - a_i^2) \\ \Delta Q_{ij} &= 2(e_{ij} - a_i a_j) \end{aligned}$$

e_{ij} is the fraction of the summation of the weights of the edges between the pages in the cluster i to the summation of the weights of all edges. a_i is the fraction of the summation of the weights of the edges between the page in cluster i and the page in other clusters to the summation of the weights of all edges. This gives the results that are based on the weight of contents similarity and the types of the links.

Weights of edges between the pages The weight of the edge between the pages p, q is defined as follows.

$$W(p,q) = \alpha s(p,q) + (1-\alpha) w(p,q)$$

$s(p,q)$ is the contents similarity between the pages p, q. $w(p,q)$ is the weight that is based on the types of the links and the hierarchical difference of the directories. α is 0.5.

Contents similarity Contents similarity is defined as the cosine value of the document vectors of the pages. The document vector of the page is formed by the TFIDF value of the terms in the page.

Weight of link Weight of link is calculated with the types of the links and the hierarchical difference of the directories. The value between pages p, q is defined as follows.

$$w(p,q) = \begin{cases} \dfrac{0.25}{d+1} C_l & (upward/downward) \\ \dfrac{0.5}{d+1} C_l & (crosswise) \\ 0.4\ C_l & (outward,\ \ \tau_s \leq s(p,q)) \\ 0 & (outward,\ \ s(p,q) \leq \tau_s) \end{cases}$$

C_l is the value that is based on the direction of link. C_l is 2 when the direction is two-way on the Web. And when the direction is one-way, C_l is 1. τ_s is average of all $s(p,q)$ of the pages of which the link is 'outward' type.

4.4 Extraction of Topic Maps

We extract the three elements of topic maps from the result of the forementioned clustering method.

- topic

  The topic is the concept that is represented by the Web pages in the cluster. User gives the name to the topic manually at this time, because it is difficult to name the topic automatically.

- association

  The association is the one that is represented as the edges between the clusters. The association is characterized by the ratio of the types of the link between the pages in the cluster and the pages in other cluster.

- occurrence

  The occurrence is the set of Web pages that are located in the cluster and are related to the topic of the cluster.

5 Extracted Topic Map

We conduct preliminary experiments to evaluate our proposed method. We extracted topic maps from the set of Web pages that is collected using user's Web history pages as seed with Newman's method and our method and compare the topic maps. This browsing history is searching the font settings of Meadow (Emacs editor on Windows OS). The total of history pages is 5. The set of Web pages is obtained by expanding outbound and inbound links from the Web history pages by 4 steps. The number of the links expanded in each step is 3. The links are selected at random. The extracted topic maps are shown in Figure 3 and Figure 4. Figure 3 shows the topic maps by Newman's method. Figure 4 shows the topic maps by our method. The node represents the topic. The scale of the node is based on the number of

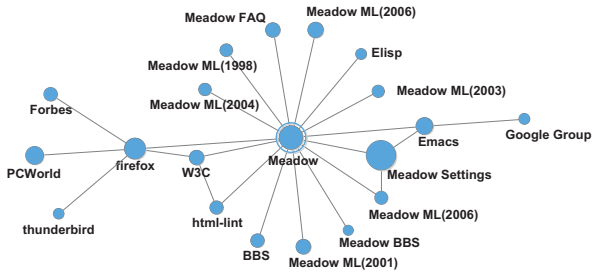

Figure 3. The topic maps by Newman's method

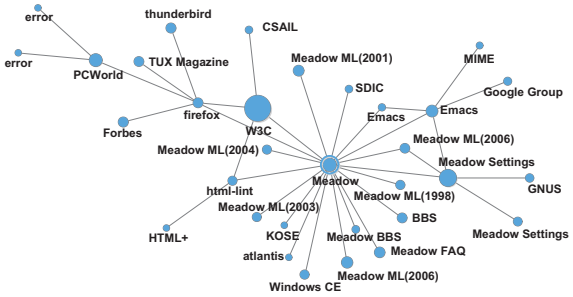

Figure 4. The topic maps by proposed method

occurrences that is related to the topic. The ringed node includes the history pages. The label of the node shows the topic name that is named manually.

From Figure3 and Figure4, we can see as follows. The both topic maps have the ringed node which includes all Web browsing history pages. Then, the topics which the other nodes present have not been seen in Web browsing. The topic 'Meadow' is at the center and the related topics such as 'Emacs', 'Meadow FAQ' and 'Meadow ML' are around it. Each 'Meadow ML' topics are formed with respect to year, because the pages are closely linked according to year. This seems to be the characteristics of structure-based clustering. Though 'Meadow' and 'firefox' are not directly related, the topic maps show it. In this case, the author of pages about 'Meadow' advises 'firefox' as the browser and links the page to the official pages of 'firefox'. In this way the topic maps show the relationship by the author's interest and preference that does not depend on content similarity. In addition, we can find the relation between 'firefox' and 'thunderbird' from the maps.

We can find the topics only from the topic maps by proposed method as follows. The topic 'SDIC' is an elisp to look up in a dictionary. The pages which are related to this topic are in the topic 'Elisp' of the topic maps by Newman's method. As these pages is located in the other directories in the same Web site, our proposed method can make these pages into on cluster by clustering with weight that is based on the types of links and the hierarchical difference of directories and content similarity. The topic 'GNUS', 'WindowsCE' and 'MIME' are similar cases. These cases show that clustering method using the forementioned weight can extract the interesting topics and relationships.

6 Conclusion

We proposed the improved clustering method with the constraint of the Web structure and the weight based on the types of links and the hierarchical difference of the directories in which the pages are located.

The method extracts Topic Map from the set of Web pages that is based on the Web browsing history of user. We conducted preliminary experiments to evaluate the method and verified that the extracted topic map show the useful topics and relationships between them.

References

[1] A. Gulli and A. Signorini. The Indexable Web is More than 11.5 Billion Pages. In *The 14th International World Wide Web Conference 2005*, 2005.

[2] J. Cugini and J. Scholtz. VISVIP: 3D Visualization of Paths through Web Sites. In *DEXA '99: Proceedings of the 10th International Workshop on Database & Expert Systems Applications*, page 259, Washington, DC, USA, 1999. IEEE Computer Society.

[3] Emmanuel Frecon and Gareth Smith. WebPath - A Three-Dimensional Web History. In *INFOVIS '98: Proceedings of the 1998 IEEE Symposium on Information Visualization*, pages 3–10, Washington, DC, USA, 1998. IEEE Computer Society.

[4] GVU's WWW Surveying Team. GVU's 10th WWW User Survey: Problem Using the Web, 1998.

[5] International Standard Organization. ISO/IEC 13250 Information Technology - SGML Applications - Topic Maps. *ISO/IEC 13250*, 2000.

[6] M. E. J. Newman. Fast algorithm for detecting community structure in networks. *Physical Review E*, 69:066133, 2004.

[7] M. Mayer and B. Bederson. Browsing Icons: A TaskBased Approach for a Visual Web History. *HCIL-200119, CS-TR-4308, UMIACS-TR-2001-85*, 2001.

[8] R. Gandhi and Benjamin B. Bederson and G. Kumar and Ben Shneiderman. Domain Name Based Visualization of Web Histories in a Zoomable User Interface. In *DEXA Workshop*, pages 591–600, 2000.

[9] E. Spertus. ParaSite: Mining Structural Information on the Web. *Computer Networks*, 29(8-13):1205–1215, 1997.