Extracting Topic Maps from Web browsing histories

Motohiro Mase[1] and Seiji Yamada[2]
[1] CISS, IGSSE, Tokyo Institute of Technology
4259 Nagatsuta, Midori, Yokohama, Japan
[2] National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda, Tokyo, Japan
{m_mase, seiji}@nii.ac.jp

Abstract

In this paper, we propose a method of clustering to extract Topic Map from the Web browsing history. The our method is based on the traditional agglomerative clustering with the constraint of the Web structure and the weight of link relation. Topic Map shows 2D-visualized overview graph of the Web browsing history, and the relations between the topics that are gathered by a user and extracted from the pages around the history pages. Using the Web browsing history, we experimentally evaluate the extracted Topic Maps.

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 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 is a difficult task. In 10th GVU WWW User Survey[2], 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 information gathered as the biggest problems. Users typically have to look for a target page from bookmarks or favorites and browsing history manually. It is therefore necessary to support revisiting pages and organizing the information gathered from Web.

We propose the system that makes visualization of user's Web browsing history clustered by the topic of pages. Our system shows *Topic Map* that indicates 2D-visualized graph of Web browsing history, and relation between the information that a user acquired from Web browsing and the information of pages in the surroundings. Topic Map provides overview of a Web graph that includes history path, discovery of topic that has not been seen, and relations between the topics. In this paper, we propose a new method that uses traditional clustering method with constraint by web structure to construct Topic Map. Finally, we conduct preliminary experiments to evaluate our system.

2 Related Work

2.1 Web history visualization

Many researches have been reported in the area of Web history visualization. Domain Tree Browser[3] and Zooming Web Browser[4] provide 2D visualized history tree by Web browsing with thumbnail images of Web pages. VISVIP[5] visualizes user's path as 2D graph that node represents a Web page and an edge represents link between pages. WebPath[6] 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 gathered from Web by topics and provides relations between the topics in addition to visualization.

2.2 Clustering

Clustering is division of data set into groups of similar data. Hierarchical agglomerative clustering and k-means clustering[8] are some of basic clustering methods. In hierarchical clustering, data set is recursively merged into clusters in descending order of similarity until the number of clusters in one. In k-means clustering, the data set is divided into k clusters desired. These methods are based on similarity between two

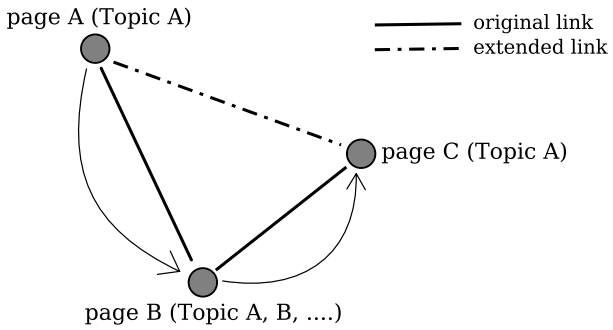

Figure 1: The relaxation of the constraint

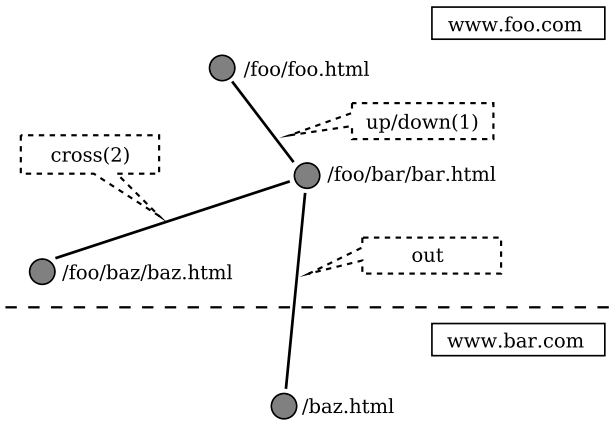

Figure 2: types of link relations

data. In a data set that linked as network graph, Girvan and Newman have reported the methods based on network graph theory[9][10]. Our method is based on the hierarchical clustering method.

3 Proposed Method

3.1 Topic Maps

Topic Maps provide clusters that classified by topics of pages and relation between topics in addition to user's Web browsing history. The methods categorize pages according to the topics of pages include clustering. Clustering methods that based on only similarity between Web pages exclude the relation between pages on Web. We focus link structure on Web that authors of pages set, and introduce following two ideas to clustering methods.

Web structure The first one is constraint by the Web structure. This means the cluster can be made only from the pages linked mutually. It is necessary that Topic Maps show not only relation between topics but also overview of user's Web browsing history path. Classifying Web pages by topics leaving the link structure of Web pages is suitable to purpose of Topic Maps. Since original Web link structure is a strong constraint, we relaxed the constraint by putting a link between the pages that are reachable within two steps. This intends to prevent the clusters that are dissimilar in topics from merging into a cluster. Figure 1 shows the relaxation of the constraint. Extending the Web structure allows page A and page C that have the same topic to merge into a cluster without making the cluster which has variation with page C such as the index page. In that case, we treat a Web graph as an undirected graph for simplification.

Link relations The second one is weighting by kind of link between Web pages. The links are due to the intention of authors of Web pages. Therefore, the links have some meanings. Parasite[11] categorized links based on hierarchical relations of directory including linked pages, and introduced heuristics to estimate meanings of the links. In this research, we use the characterization of link relations based on hierarchical relations of linked pages, and hierarchical difference.

Figure 2 shows relations of linked pages. We classified link relation into the following three types.

- *up/down*

  This type is relation between a page in the upper directory and a page in the lower directory in the same Web site. The number in parentheses is depth that shows hierarchical difference between directories of the two pages. Topic of page in lower directory means specialization of topic of page in upper. The similarity of topics that the linked pages have is inversely proportional to hierarchical difference.

- *cross*

  This type is relation between pages in different directories in same Web sites. The number in parentheses is *depth* that shows hierarchical difference between directories of pages. (For Figure 3, depth from '/foo/bar/' to '/foo/' and from '/foo/' to '/foo/baz/' are (1) and (2).) The similarity of topics that linked pages have is inversely proportional to hierarchical difference.

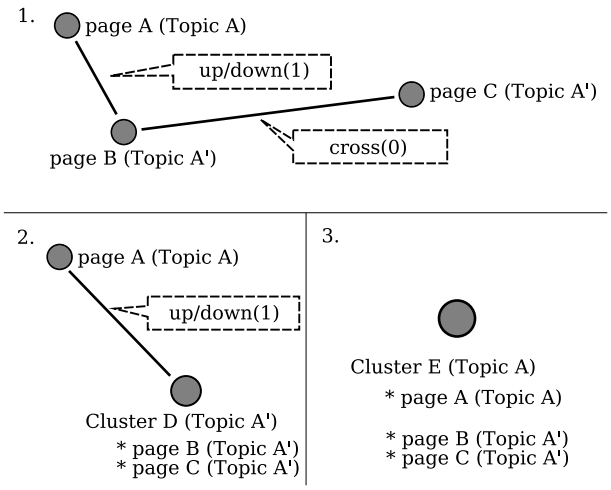

Figure 3: usage of link relations

- *out*

  This type is relation between pages in other Web sites. Both pages have similar topics, basically.

Our method is not simply clustering method based on only similarity but characterized by constraint of link structure and weight of link relation in addition to similarity. Figure 3 shows the example of usage of link relation. Topics of page B and page C are subtopic of the topic on page A. In this case, page B and page C are first merged into cluster D that has topic A'. Next, page A and cluster E are merged into cluster E with topic A. Such accumulation order of pages provides the pseudo hierarchical relation of topics. When there is little difference of similarities between pages, the usage of correction with weight that gives priority to relation between page B and page C more than relation between page A and page B result in the order.

3.2 Clustering algorithm

3.2.1 Preprocessing

We extract Topic Maps from a Web page set that is user's Web browsing history pages and pages around the history pages. The page set is history pages and pages that are accessible by expanding outbound links and inbound links from history pages within n steps.

As shown in section 3.1, pages that are reachable within two steps on Web undirected graph are linked on top of the pages linked on original Web graph. Only these linked pages merge into clusters.

Each page of the page set is an initial cluster. The following two evaluations are calculated for all pair of clusters that have link relations. The first is similarity of clusters, and the second is cluster cohesiveness.

similarity of clusters Similarity of clusters is weighted liner sum of these two values. The first is similarity of contents between clusters, and the second is cohesiveness of cluster. Here is description of the values.

- similarity of contents

  Web pages are characterized by terms and weight of terms in vector space.The terms and weight are extracted by Chasen and TermExtract from HTML files that removed tags, and corrected by inverted document frequency. Then, the document vector of page s is defined as: $d_s = (w_0, w_1, ..., w_N)$, where w_i is weight of term i. Therefore, similarity between page p and page q is defined as the following equation.

$$sim_c(p,q) = \frac{d_p \cdot d_q}{\|d_p\| \times \|d_q\|}$$

  Similarity between clusters are similarly calculated by above-mentioned equation.

- weight of link relation

  Weight value of link relation between pages are defined by kind of the relation. The value between page p and page q is defined as the following three equations.

$$weight_{lr}(p,q) = \frac{0.25}{(depth+1)} \times C_{lt} \qquad (up/down)$$

$$weight_{lr}(p,q) = \frac{0.5}{(depth+1)} \times C_{lt} \qquad (cross)$$

$$weight_{lr}(p,q) = \begin{cases} 0.4 \times C_{lt} & (out, \tau_{sim} \leq sim_c(p,q)) \\ 0 & (out, sim_c(p,q) \leq \tau_{sim}) \end{cases}$$

  C_{lt} shows type of link. C_{lt} is 2 if the type is two-way on the Web graph, and if the type is one-way on the Web graph, C_{lt} is 1, and if type is the extended link, C_{lt} is 0.5. τ_{sim} is average of all sim_c of links that are 'out' type.

  The weight between cluster s and cluster t is average of all $weight_{lr}$ of pages that are a page belong to cluster s and a page belong to cluster t which have link relation.

The similarity of clusters is defined as the following equation.

$$f(p,q) = \alpha \times sim_c(p,q) + (1-\alpha) \times weight_{lr} \qquad (0 \leq \alpha \leq 1)$$

Cohesiveness of cluster Cohesiveness of a cluster is average of similarities between the cluster and pages belong to the cluster. The cohesiveness is defined as the following equation.

$$c = \frac{\sum^N sim_c(s,i)}{N}$$

i is a page belong to a cluster s. N is number of pages that are attached to cluster s.

3.3 Clustering algorithm

Our method is based on the agglomerative clustering with the constraint of the Web structure and the weight of link relation. The following is the procedure of our clustering method. τ_c is threshold of cluster cohesiveness and running from 0 to 1.

1. **Initialize**

   Calculate all similarities between two clusters having link relation. A threshold of cluster cohesiveness τ_c is set to β.

2. **Select a pair of clusters**

   Select a pair of clusters that have the highest similarity.

3. **Merge the clusters**

   Merge the selected clusters merge into a new cluster. The document vector of the new cluster is made by composition of the vectors of the clusters and normalization. The link relations of the cluster are total of the link relations of the clusters.

4. **Calculate Cohesiveness**

   Calculate cohesiveness of the new cluster. If the value of cohesiveness is lower than the threshold τ_c, exclude the cluster from a candidate for merging and go to step 6

5. **Recalculate new similarities**

   Recalculate all new similarities between the new cluster and clusters linking to the cluster.

6. **Check state**

   Check whether pairs of the clusters remain. If the pairs remain, return to step 2, if not so clustering is done.

The procedure for clustering provides Topic Map with the parameter that τ_c is β. The procedure can provide Topic Map of other τ_c recursively with the clusters remaining and the new value of cohesiveness.

4 Extracted Topic Maps

We conduct preliminary experiments to evaluate our proposed method. A Web page set is obtained by expanding outbound and inbound links from the Web history pages by n steps. The number of the links expanded in each step is b. The links are selected at random. In the following experiments, n and b are set to 4 and 3. The method extracts Topic Maps from the Web browsing histories. Topic Maps show the graph generated from clusters and edges by spring model method. The weight of edge is based on similarity between clusters. A node represents a cluster. The scale of a node is based on the number of pages belong to a cluster. A node with ring has the history pages. A label of a node shows upper three words of the vector.

4.1 Test 1 : Searching the font settings of Meadow

This browsing history is searching the font settings of Meadow (Emacs editor on Windows OS). The total of history pages is 5. Figure 4 shows Topic Map which is extracted from the history by clustering in twice. τ_c is 0.5 and 0.4. The nodes which have relevant topics is basically around the ringed nodes. The node in the center is interesting. The node of topic is Cygwin (Linux-like environment for Windows). Although the similarity of the contents between Cygwin and Meadow is not apparent, the graph shows their relevance.

4.2 Test 2 : Searching the simple backup system

This browsing history is searching the backup system. The total of history pages is 5. Figure5 shows Topic Maps which is extracted from the history by clustering in twice. τ_c is 0.5 and 0.4. The backup system pdumpfs is based on dumpfs on Plan9 OS and mounted by Ruby scripting language. Then, the nodes including these topics are adjoining. The nodes at the right of the graph are closely linked. The topics of these nodes include OpenSSH and OpenBSD roughly. The fact that OpenSSH is primarily developed by the

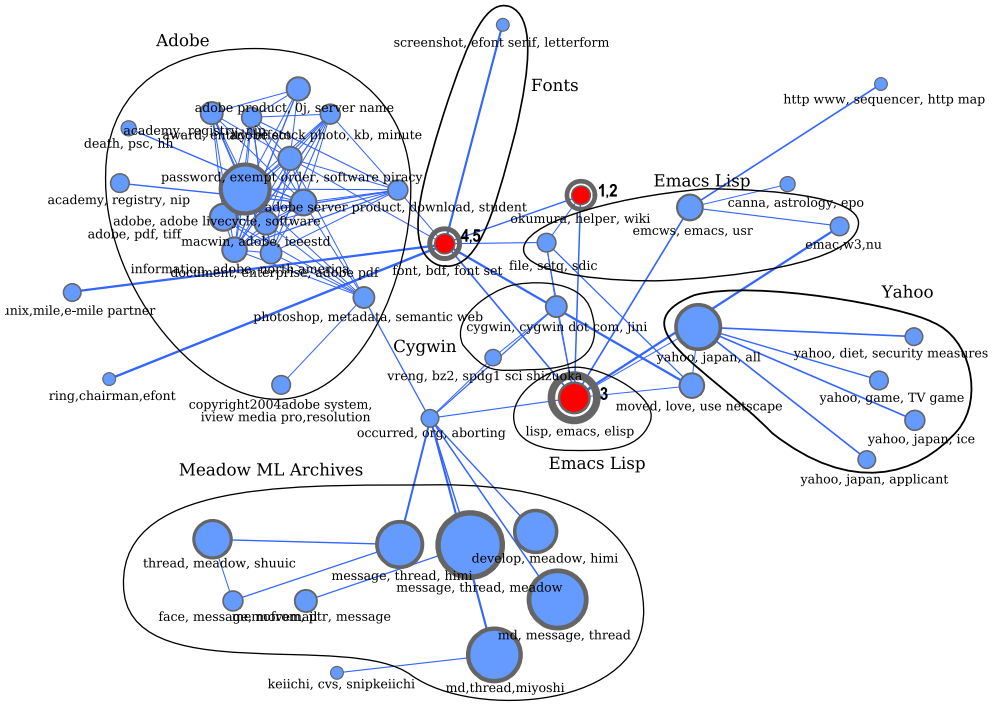

Figure 4: Test1. Searching the font settings of Meadow

OpenBSD Project causes the close link relation on the Web.

5 Discussion

As shown by the experiments, use of the link relations on the Web enlarges the relations of the topics extracted by the only similarities. Basically, the link relations on the Web are set by the authors of the pages. Then, the expansion by the relations is adoption of a new viewpoint. In section 4.1, Linking Meadow and Cygwin is the viewpoint that they are UNIX tools on Windows. This viewpoint in the Web structure is concern to users. Topic Map helps users with discovering the viewpoint. In our method, the calculation of the similarities is limited to the linked clusters. Then, the method is capable of reduction in the calculation cost based on the number of links, although the method of the computational complexity is $O(n^2)$ as well as the traditional hierarchical clustering. On the one hand, the constraint of the Web structure allows only the linked clusters to merge into a cluster in our method. Therefore, the cluster that has various topics is merged, if the pages with several independent topics or multiple topics are adjacent. It causes the noise though the relaxation of the constraint reduces it. In addition, our method employs the weights of the link relations based on the hierarchical difference of the directories with pages. Then, the method cannot be very effective in clustering the pages that are dynamically generated such as Wiki pages; the pages are basically in the same directory and distinguished by query words.

6 Conclusion

We proposed a new clustering method with a constraint of the Web structure and a weight based on a link relation. The constraint of the extended structure prevent clusters that are dissimilar in topics from merging into a new cluster, leaving the Web structure. Applying the weight to a traditional clustering provides a cluster that has few differences. The method extracts Topic Map from a Web page set based on a browsing history. We conducted preliminary experiments to evaluate the method and verified that Topic Maps show useful relations between topics.

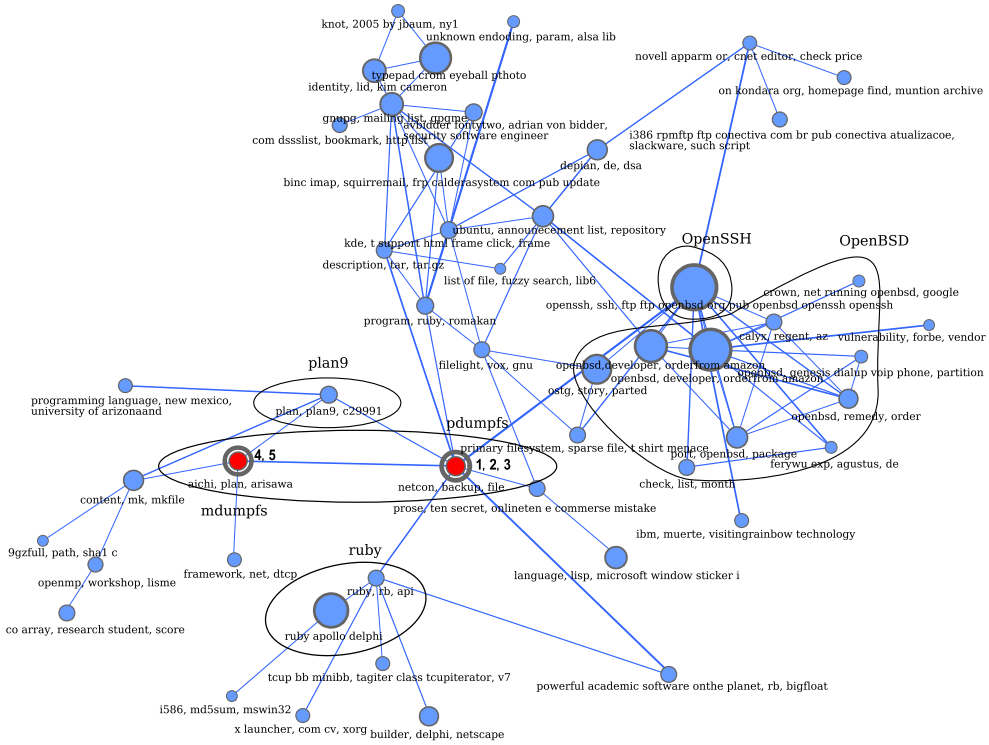

Figure 5: Test2. Searching the simple backup system

References

[1] Gulli, A.,Signorini, A. The Indexable Web is More than 11.5 Billion Pages, Proceedings of the 15th WWW conference, 2005.

[2] GVU's WWW Surveying Team, GVU's 10th WWW User Survey: Problem Using the Web, Gerogia Tech, Atlanta, GA, 1998.

[3] Gandhi, R., Kumar, G., Bederson, B. B., Shneiderman, B., Domain Name Based Visualiazation of Web Histories in a Zoomable User Interface, Proceedings of WebVis2000, pp. 591-598, 2000

[4] Benjamin B. Bederson, James D. Hollan, Jason Stewart, David Rogers, Allison Druin, David Vick, A Zooming Web Browser, Proceedings of 9th Annual ACM Symposium on User-Interface Software and Technology, 1996.

[5] Cugini and J. Scholtz, VISVIP: 3D Visualization of Paths through Web Sites, Proceedings of the International Workshop on Web-Based Information Visualization (WebVis'99), pp. 259-263, 1999.

[6] Frecon, E., Smith, G., WebPath - A Three Dimensional Web History,Proceedings of Information Visualization Symposium (InfoVis98),pp 3-10, 1998.

[7] Matthias, M., Benjamin, B, B., Browsing Icons: A Task-Based Approach for a Visual Web History, HCIL-200119, CS-TR-4308, UMIACS-TR-2001-85, HCI Lab, University of Maryland, Maryland, USA, 2001.

[8] A. K. Jain., R. C. Dubes., Algorithms for Clustering Data, Prentice Hall, 1998.

[9] Michelle, G., M. E. J. Newman, Community structure in social and biological networks, Proc. Natl. Acad. Sci. USA, Vol. 99, No. 12, pp. 7821-7826, 2002.

[10] M. E. J. Newman, Fast algorithm for detecting community structure in networks, Phys. Rev. E 69, 066133, 2004.

[11] Spertus, E., ParaSite: Mining Structural Information on the Web, Proceedings of the 6th WWW Conference, 1997.