Semiautomatic Extraction of Topic Maps from Web Pages using Clustering with Web Contents and Structure

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

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

Katsumi Nitta
Tokyo Institute of Technology
4259, Nagatsuta, Midori, Yokohama, Japan
nitta@dis.titech.ac.jp

Abstract

In this paper, we describe a method to semi-automatically extract Topic Maps from a set of Web pages. We introduce the following two points to the existing clustering method: The first is merging only the linked Web pages, to extract the underlying relationship of the topics. The second is introducing the similarity by contents of Web pages and the types of links, and the distance between the directories in which the pages are located, to generate dense clusters. We generate the topic map by assuming the clusters as topics, the edges as associations, the Web pages related to the topic as occurrences from the result of clustering. We experimentally extracted the topic map and evaluated it.

1 Introduction

The size of Web has augmented and exceeds at least 11.5 billions in January 2005[3]. Information gathering utilizing a huge amount of Web pages is very useful and essential for Web users. However, finding out the necessary information from the Web when users need and organizing the gathered information are big problems[4]. There is many works for these problems, and one of them is Topic Maps. Topic Maps is international standard to organize and classify the information along with user's knowledge and concept. This standard can connect the various information resources with knowledge and concept of user, represent relations of the concepts, and help user easier access to the information one needs. It takes the information resources and selection of target domain and topics to build topic maps. The user basically needs to do these tasks by hand power, though user can reduce the cost of them by using editors of topic maps such as Ontopoly. By converting existing metadata such as XML and RDF, the previous study can automatically generate topic maps[8]. Although the amount of structured metadata on the Web is growing, many Web pages which user utilizes on a daily basis are semistructured data, HTML files, applying the previous method to them is difficult. Therefore it requires a method to directly extract topic maps from the semistructured data.

We aim to extract a skeleton of topic map and completely build topic map through interaction with user, because it is difficult to automatically extract complete topic maps from Web pages. We propose a method to automatically extract a skeleton of topic map from a set of Web pages. Contents of Web pages and structures of Web graph which consists of nodes as pages and edges as links have the underlying topics of pages' contents and relationships between the topics. Our approach is extracting the involved information as topic maps by clustering the Web pages on the ground that the contents of pages and the structure of Web graph. One of the studies in the field of information extraction from Web is Web structure mining.

In this field, many previous works studied the extraction of the Web community from the structure of Web graph. Broder and others extracted the communities by searching a complete bipartite graph which is a community signature[1]. Girvan and Newman extracted the structure of community in networks by using a clustering method based on edge betweenness which is the number of shortest paths between all pairs of vertices that run through it[2]. Newman proposed a hierarchical clustering method that maximizes the modularity[7], which is function to quantify how good a particular division is: we call the method as Newman's

0-7695-3028-1/07 $25.00 © 2007 IEEE
DOI 10.1109/WI-IATW.2007.85

method. These preliminary works focused only the extraction of structure from Web graph, our method is based on Newman's method, utilizing the similarity of contents for Web pages, weighting by the structure of Web site's directories in addition to the structure of Web graph, to extract topic maps.

2 Topic Maps

Topic Maps is ISO/IEC 13250 standard and a solution for representing concept and connecting the concept to related information resources[5]. A topic map is composed of topics, which represent any concepts and subjects in real world; associations, which represent the relation between topics; and occurrences which represent connection between topics and information resources related to them. Topic maps have a lot of flexibility, and allow creators of topic maps to define the types of topics, associations and occurrences, and good for representing various topics and relationships of them on the Web.

One of the syntaxes of Topic Maps is XML Topic Maps (XTM)[9]. Some information items and named properties are essential to represent topic maps using XTM. We extract only items of topics, associations and occurrence, because it is hard to extract all the items and properties from Web pages. We get three types of items from results of applying our clustering method to the set of Web pages, by assuming the clusters as topics, the edges as associations, the pages related to the clusters as occurrences, and generate skeletons of topic maps with extracted items.

3 Extraction of a topic map from a set of Web pages

3.1 Overview

We extract a skeleton of topic map from a set of Web pages, along the following procedure, by clustering the Web pages and extracting the items of topic maps from the result of clustering. Topic maps have to show not only topics that are found in Web pages, but also the relationship between them. However, existing contents-based clustering methods can measure similarities of topics that clusters represent by contents of Web pages, but not extract what is the relation between the topics. In contrast, structured-based clustering methods focus only the structures of networks without any reference to contents of pages. We proposed a clustering method based on Newman's method, structure-based clustering methods, regarding both the similarities between contents of Web pages and graph structures of links on Web.

3.1.1 Structure of links between Web pages

In general, site creators manually generate links on Web, linked pages cover relevant topics[6]. These links have underling relations between the topics. But even if creators link pages which have relevant topics and the links represent some relations of the topics, the contents-based methods can't find the relation at lower values of the similarities for pages' contents. We cluster pages by merging only linked pages and extract the relations from remaining links, which are represented as edges between clusters, at the end of clustering.

3.1.2 Types of links with structure of directories in Web sites

We build denser clusters, regarding weights of links between Web pages. Utilizing the similarities between linked pages by contents of them is one of calculation methods for the weights. In contrast, we calculate the weight by using both liked pages' similarities based on types of links and contents of pages. The similarities with types of links are computed by grouping links with directories in which linked pages exist on Web site and using distance between the directories, without focusing the contents of pages.

We presume that the creators of Web pages probably make directories on Web sites and locate pages in the directories. First, Web pages are classified into directories along topics of pages. Next, topics of pages in child directories are specialization of pages' topics in their parent directories. Finally, topics of pages are similar to them in close directories than distant directories.

To estimate the similarities between linked pages by utilizing only the types of links, we assort all links between pages to three types by directories in which the pages are located, and introduce a measure, distance of directories. The measure is the number of directories on shortest path between any two pages, in tree structure which consists of pages and directories as nodes. The three links is as follows. 1) **upward/downward** is link between a page and it in higher or lower directory in same Web site. 2) **crosswise** is link between pages in same Web site, except upward/downward links. 3) **outward** is link between pages in two different Web sites.

We calculate the similarity with giving priority to the types of links. First, we prioritize crosswise over upward/downward. Because the creators of pages sort into directories by each content, and topics of pages reached through upward/downward links are specialized or generalized topics and are uncorrelated than topics of pages in a same directory. Next, we assign priority to outward over upward/downward, because outward links means pointer to related information resources on other site by the creators of them, linked pages with outward link have mutual topics. Finally, we prioritize links which have lower values for distance of directories. The Pages in a same directory have similar topics and topics of pages are correlate poorly as distance between directories in which pages are involved increases.

3.2 Weighting links between Web pages

The weight between Web pages p and q, $w(p, q)$ is computed as weighted liner sum of the similarity with contents of pages, $s_c(p, q)$ and the similarity with the types of links, $s_l(p, q)$. $w(p, q)$ is defined as follows:

$$w(p, q) = \alpha s_c(p, q) + (1 - \alpha) s_l(p, q) \qquad (1)$$

3.2.1 Similarity with contents of pages

We construct a document vector of a page, applying TF-IDF method to all words extracted from the page's HTML file. The document vector v_i is $v_i = (w_{i1}, w_{i2}, ..., w_{in})$, where w_{ij} is the tfidf value of word j in page i. The similarity between page p and page q, $s_c(p, q)$ is defined as follows:

$$s_c(p, q) = \frac{v_p \cdot v_q}{\| v_p \| \| v_q \|} \qquad (2)$$

3.2.2 Similarity with types of links

We calculate the similarity with types of links, using the relationships between the directories of the linked pages and the distance of the directories, along the line of weighting shown at section 3.1.2. This similarity between pages p and q, $s_l(p, q)$ is defined as follows:

$$s_l(p, q) = \begin{cases} \frac{0.25}{d} C_l & \text{(upward/downward)} \\ \frac{0.5}{d} C_l & \text{(crosswise)} \\ 0.4\ C_l & \text{(outward,} \quad \tau_s \leq s(p, q)) \\ 0 & \text{(outward,} \quad s(p, q) < \tau_s) \end{cases} \qquad (3)$$

where C_l is decided by the direction of link. When the link between pages is cross-linked, C_l = 2. The value of C_l for the single link is 1, C_l = 1. d is the distance between directories of pages, as has been previously described. τ is average of all values of $s_c(p, q)$, similarities between pages which are connected by "outward" link.

3.3 Clustering method

To generate topic maps, we apply the clustering method to Web pages and extract the items for topic maps from the result of clustering. Newman's method only focuses only the structure of network, so we introduce the weighting for the links that is shown in section 3.1.2 to Newman's method. We calculate the value of the weight in the way hereinafter prescribed. Newman's method[7] is the agglomerative hierarchical clustering method and has the modularity Q which shows the quality of cluster division. In clustering of Newman's method, a pair of clusters that provides the maximal value of increment in Q are joined at each step. The definition of Q is as follows:

$$\begin{aligned} Q &= \sum_i (e_{ii} - a_i^2) && (4) \\ \Delta Q_{ij} &= 2(e_{ij} - a_i a_j) && (5) \end{aligned}$$

where e_{ij} represents the fraction of edges between cluster i and cluster j in the network and is calculated as the value for the number of edges between cluster i and j divided by the total number of edges in the network. a_i is the fraction of edges belonging to cluster i and denoted as $a_i = \sum_i e_{ij}$. We cluster Web pages using our method which is based on Newman's method with the mentioned weights, according to the following procedure.

1. Set a cluster as each page, and an edge as each link in a given set of Web pages.
2. Calculate all values of the weight for links, and apply them to edges. Each weighted value of edge is normalized by total value of edges.
3. Select a pair of linked clusters which has maximal value of ΔQ. Terminate, if ΔQ of the selected pair is negative value.
4. Merge the pair of clusters into a new cluster.
5. Recalculate the values of e_{ij} and a_i which are related to the new cluster and Go to step 3.

3.4 Extraction and Visualization of topic maps

We extract three types of items for topic maps from the result of clustering. A topic is a concept represented by Web pages within a cluster. However, it's hard to automatically name the topic, user has to give the proper name to the topic. An association is represented as an edge between clusters, which is remained at the end of the clustering. As in the case of the topics, it's difficult to automatically annotate the association. User adequately gives the annotation to the association. Occurrences are pointers to information resources, as the set of Web pages, which are related to the cluster. We generate the skeleton of topic maps with extracted items and visualize topic maps on graph. On the graph, topics are represented as nodes and associations are represented as edges. The area of node is dependent on the number of pages related to the topic.

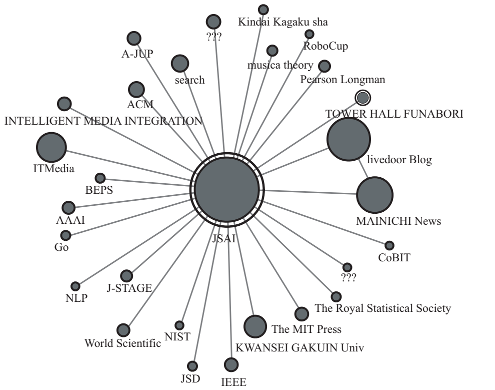

Figure 1. Topic map (Newman's method)

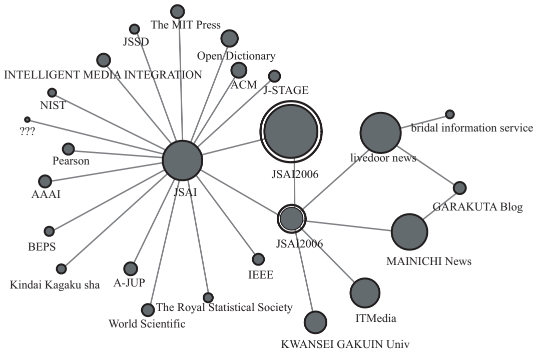

Figure 2. Topic maps (proposed method)

4 Extracted topic map

We experimentally extract topic maps with both methods. We utilize the sets of Web pages which are collected using browsing histories of the Web user as seeds. We obtain the set of Web pages by following links four times from the pages of histories. The number of selected links in each of pages is 3 and the links are selected at random. The sets of pages have around 600 pages. Fig.1 shows the topic map by Newman's method. Fig.2 shows one by the proposed method.

In this case, the participant search information about "JSAI2006", the annual conference of JSAI in 2006. In both figure, we find some topics, "AAAI", "ACM", "IEEE", "NIST" around "JSAI". These topics are societies and research institute, which are relevant to "JSAI". In the topic maps of Newman's method, the topic, "JSAI2006", which is the participant's target information, is involved in "JSAI". In contrast, in the topic map of our methods, we can find these two topics separately and associations that is hidden in the topic map of Newman's method. Topic "JSAI2006" is connected to "ITmedia", "livedoor News", "MAINICHI News" at right in Fig.2. These associations shows that each Japan-based news site runs an article about "JSAI2006". It is difficult for the participant to find the associations from the topic maps by Newman's method, because topic "JSAI2006" and associations between the topics is hidden.

5 Conclusion

In this paper, we proposed the method to extract the skeleton of the topic maps. Our method was based on Newman's method with the weighting based on the similarities by contents of pages and types of links. 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] Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J.: Graph structure in the web: experiments and models, in *5th WWW Conference* (2000)

[2] Girvan, M. and Newman, M. E. J.: Community structure in social and biological networks, http://arxiv.org/abs/cond-mat/0112110/ (2001)

[3] Gulli, A. and Signorini, A.: The indexable web is more than 11.5 billion pages, in *Special interest tracks and posters of the 14th WWW Conference*, pp. 902–903, New York, NY, USA (2005)

[4] GVU's WWW Surveying Team: GVU's 10th WWW User Survey:Problem Using the Web, http://www.gvu.gatech.edu/user_surveys/ (1998)

[5] International Standard Organization: ISO/IEC 13250 Topic Maps: Information Tecknology Document Description and Markup Language (2000)

[6] Menczer, F.: Lexical and semantic clustering by web links, *Journal of American Society Information Science and Technology*, Vol. 55, No. 14, pp. 1261–1269 (2004)

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

[8] Reynolds, J. and Kimber, W. E.: Topic Map Authoring With Reusable Ontologies and Autoated Knowledge Mining, in *XML 2002 Conference* (2002)

[9] TopicMaps.Org: XML Topic Maps (XTM) 1.0, http://www.topicmaps.org/xtm/1.0/ (2001)