Extracting Topic Maps from Web Pages by Web Link Structure and Content

Motohiro Mase, Seiji Yamada, and Katsumi Nitta

***Abstract*— We propose a framework to extract topic maps from a set of Web pages. We use the clustering method with the Web pages and extract the topic map prototypes. We introduced the following two points to the existing clustering method: The first is merging only the linked Web pages, thus extracting the underlying relationships between the topics. The second is introducing weighting based on the similarity from the contents of the Web pages and relevance between topics of pages. The relevance is based on the types of links with directories in the Web sites structure and the distance between the directories in which the pages are located. We generate the topic map prototypes by assuming that the clusters are the topics, the edges are the associations, and the Web pages related to the topics are the occurrences from the results of the clustering. Finally, users complete the prototype by labeling the topics and associations and removing the unnecessary items. We incrementally use a user's evaluation of the topic maps to judge whether a Web page is unnecessary or necessary and then reduce the number of unnecessary pages. We use the relevance feedback along with a Support Vector Machine (SVM) to judge the Web pages. For this paper, at the first step, we mounted the proposed clustering method and conducted experiments to evaluate the effectiveness of extracting topic map prototypes. We eventually discussed the effectiveness of our two additional points by evaluating the extracted topic map prototypes.**

I. Introduction

THE size of the Web continues to grow and exceeded 11.5 billion pages in January 2005 [7]. Information gathering that references a huge amount of Web pages is very useful and essential for Web users. However, finding the necessary information from the Web when users need it and organizing the gathered information are both big problems [8]. There are many works currently underway to solve these problems, and one of them is Topic Maps [9]. Topic Maps are an international standard for organizing and classifying information along with a user's knowledge and concepts. Topic maps are composed of three elements; topics, associations, and occurrences. This standard can connect the various information resources with the knowledge and concepts of the user, represent the relations between the concepts, and help users more easily access the information they need. It takes the information resources and the selection of target domains and topics to build topic maps. The user needs to manually do these tasks, although the user can reduce costs by using editors for the topic maps, such as Ontopoly [1]. By converting the existing metadata, such as XML and RDF, previous studies have been able to automatically generate topic maps [11][16]. Although the amount of structured metadata on the Web is growing, many of the Web pages that users utilize on a daily basis are semi-structured data and HTML files, and using the previous method with them is difficult. Therefore, this process requires a method to directly extract topic maps from semi-structured data. However, we aim to extract topic map prototypes from the Web pages by using the clustering method, because it is difficult to automatically extract complete topic maps. Then, users finally complete the topic map prototypes by evaluating the topics and associations, and by adding and removing them. We incrementally use the users' evaluations of the prototypes to extract a more appropriate topic map by reducing the unnecessary number of Web pages beforehand. We use relevance feedback with a Support Vector Machine [3][14][15] to judge whether Web pages are necessary or not for the users. In this paper, we at first propose a method for extracting a topic map prototype from a set of Web pages and conduct an experiment to evaluate the extracted topic maps. By utilizing the topic map that is extracted from the Web pages in the user's browsing history and neighboring pages, which the user has not visited, a user can easily manage the collected information and then access the information that they have never seen before.

The contents of Web pages and structures of Web graphs that consist of nodes as pages and edges as links contain the underlying topics of pages' contents and relationships between the topics. Our approach is to extract the involved information as topic maps by clustering the Web pages by the contents of their pages and the structures of the Web graphs. In this field, many previous works have studied on the extraction of information from Web communities from the structures of the Web graphs. Broder et al. extracted the communities by searching a complete bipartite graph which is a community signature [2]. Flake et al. found the communities from the Web by using a maximum flow algorithm [4]. Girvan and Newman extracted the structures of the communities within networks by using a clustering method based on the edge betweenness, which is the number of shortest paths between all pairs of vertices that run through it [6]. Newman proposed a hierarchical clustering method that maximizes the modularity[13], which is a function to quantify how good a particular division is: we call the method the Newman's method. These preliminary works focused only on the extraction of structures from the Web graphs. Although our method is based on Newman's method, the method uses the similarity between the contents of the Web

Motohiro Mase and Katsumi Nitta are with the Department of Computational Intelligence and Systems Science, The Tokyo Institute of Technology, Tokyo, Japan (email: m_mase@nii.ac.jp, nitta@ntt.dis.titech.ac.jp

Seiji Yamada is with Digital Content and Media Sciences Research Division, National Institute of Informatics, Tokyo, Japan (email: seiji@nii.ac.jp)

[1] http://www.ontopia.net/solutions/ontopoly.html

978-1-4244-1823-7/08/$25.00©2008 IEEE

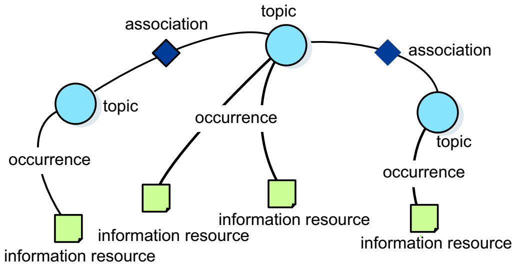


Fig. 1. Topic Maps Overview

pages and the relevance between the topics of pages which is based on the structures of the Web site's directories in addition to the structures of the Web graphs, to extract topic maps.

II. Topic Maps

Topic Maps is an ISO/IEC 13250 standard and a solution for representing a concept and connecting the concept to its related information resources [9]. Fig. 1 gives an overview of Topic Maps. A topic map is composed of topics, which represent any concept or subject in the real world: associations, which represent the relation between topics; and occurrences which represent the connection between topics and the information resources related to them. Topic maps have a lot of flexibility, and allow their creators to define the types of topics, associations and occurrences, and good for representing various topics and relationships concerning them on the Web.

One of the syntaxes of Topic Maps is XML Topic Maps (XTM) [20]. Some information items and named properties are essential for representing topic maps using XTM. We extract only the items concerning the topics, associations, and occurrences, because it is hard to extract all the items and properties from Web pages. We were able to get three types of items from the results of using our clustering method on a set of Web pages, by assuming the clusters were topics, the edges were associations, the pages related to the clusters were occurrences, and then generate prototypes of the topic maps with the extracted items. Finally, users complete the prototype by labeling the topics and associations and removing the unnecessary items.

III. Extraction of a topic map from a set of Web pages

A. Overview

In Fig. 2 shows the overview of the framework to extract a topic map. We extract a topic map from a set of Web pages using the following procedure.

1) Collect a set of Web pages from the user's Web history.
2) Use the clustering method on the set of pages.
3) Extract the items of the topic map from the result of clustering.
4) Build a topic map prototype using the extracted items.

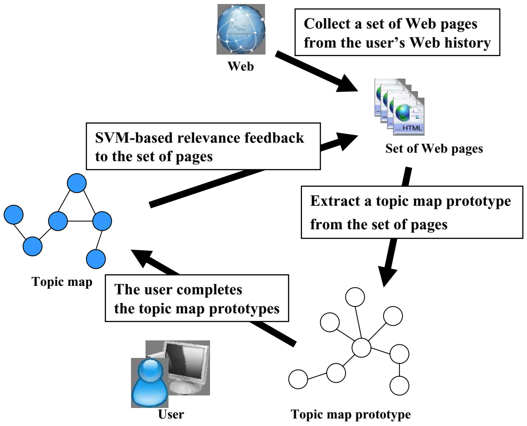


Fig. 2. Framework for extraction of topic maps

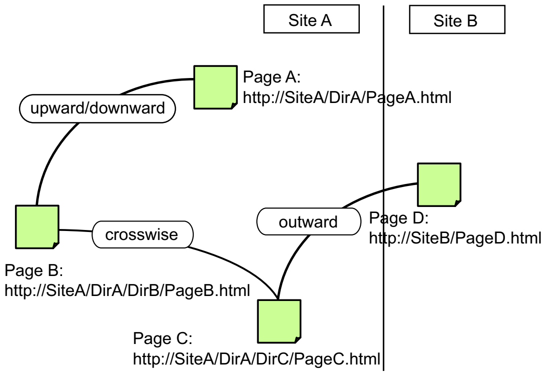
5) A user completes the extracted prototypes by evaluating, labeling, and modifying the items.
6) Apply the SVM-based relevant feedback from the user's evaluation to make a selection of the appropriate Web pages.

In this paper, we propose a method to extract the topic map prototypes. Topic maps have to show not only the topics that are found in the Web pages, but also the relationships between them. However, existing contents-based clustering methods [10] can measure the similarities between the topics that the clusters represent by using the contents of the Web pages, but not extract what the relationship between the topics is. In contrast, structured-based clustering methods focus only on the structures of networks without any reference to the contents of the pages. We proposed a clustering method based on Newman's method, structure-based clustering methods, regarding both the similarities between the contents of Web pages and the relevance between the topics of pages based on the graph structures of links on Web.

1) Structure of links between Web pages: In general, site creators manually generate links on the Web, and the linked pages cover the relevant topics [12]. These links have underlying relations between the topics. However, even if a creator link pages that have relevant topics and the links represent some relation to the topics, the contents-based methods can't find the relation for lower value similarities for the pages' contents. We cluster pages by merging only the linked pages and extract the relations from the remaining links, which are represented as the edges between clusters, at the end of clustering.

2) Types of links with directories in Web sites structure: We build denser clusters, taking into consideration the weights of links between Web pages. Utilizing the similarities between the contents of linked pages is one of the calculation methods for the weights. In contrast, we use the relevance between the topics of the Web pages based on the types of links.

The relevance between the topics based on the types of



Site A
Site B
upward/downward
Page A:
http://SiteA/DirA/PageA.html
crosswise
outward
Page D:
http://SiteB/PageD.html
Page B:
http://SiteA/DirA/DirB/PageB.html
Page C:
http://SiteA/DirA/DirC/PageC.html

Fig. 3. Types of links between Web pages

links are computed by grouping the links with directories in which linked pages exist on a Web site and using the distance between the directories, without focusing on the contents of the pages. We presume that the creators of Web pages probably make directories on Web sites and locate pages in the directories as follows: 1) Web pages are classified into directories along with the topics of pages. 2) The topics of pages in child directories are a specialization of pages' topics in their parent directories. 3) The topics of pages are similar to them in closer directories than distant directories. We compute the relevance from the previously mentioned estimation. As for related research, Spertus categorized the types of links with hierarchical relationships of directories in which linked pages exist, and introduced heuristics to estimate the meanings of the links [19].

To estimate the relevance between the topics of the Web pages using the types of links, we sorted all the links between pages to three types listed by the directories in which the pages are located, and introduced a measure, the distance of the directories. The measure is the number of directories on the shortest path between any two pages, which we put in a tree structure that consists of pages and directories used as nodes. Fig. 3 shows the types of links between Web pages.

1) **upward/downward** is a link between a page and it in a higher or lower directory on the same Web site. In Fig.3, the link between Pages A and B is an upward/downward link.
2) **crosswise** is a link between the pages on the same Web site, except for the upward/downward links. In Fig.3, the link between Page B and Page C is a crosswise link.
3) **outward** is a link between pages on two different Web sites. In Fig. 3, the link between Page C and Page D is an outward link.

We calculated the relevance by giving priority to the types of links as follows.

- We prioritize by assigning crosswise links over upward/downward ones. Since the creators of pages sort them into directories by content, the topics of pages reached through upward/downward links are specialized or generalized topics and are uncorrelated with the topics of pages in the same directory.
- We assign priority to the outward links over the upward/downward ones, because the outward links represent pointers to related information resources on other sites made by the creators of them, and linked pages with outward links have mutual topics.
- We prioritize links that have lower values for the distances of their directories. Pages in the same directory have similar topics and the topics of the pages are poorly correlated as the distances between the directories in which the pages are involved increase.

With this in mind, we compute the relevance by the types of links. We generate concentrated clusters by weighting based on the similarities between the contents of the Web pages and the relevance between the topics of the pages by the types of links. As for the details, we note them in the next section.

B. Weighting links between Web pages

For the weight between Web pages p and q, $w(p,q)$ is computed as a weighted liner sum of the similarity between contents of the pages, $s(p,q)$, and the relevance between the topics of pages by the types of links, $r(p,q)$. $w(p,q)$ is defined as follows:

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

Where, at this time, the value of α is 0.5. We aim to generate dense clusters by using the relevance between the Web pages' topics which is unable to calculate with the similarity between the contents of the pages in addition to the similarity. Then, we evenly use the weighting based on values of the similarities and the relevance.

1) Similarity between contents of pages: We construct a document vector of a page by applying the TF-IDF method [18] to all the 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 pages p and q, $s(p,q)$ is defined as follows:

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

2) Relevance between the topics of pages by types of links: We calculate the relevance between the topics of pages based on the types of links by using the relationships between the directories of the linked pages and the distances of the directories, along with the line of weighting shown in Section III-A.2. The relevance between topics of pages p and q,

$r(p,q)$ is defined as follows:

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

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

C. Clustering method

To generate topic maps, we use the clustering method for the Web pages and extract the items for the topic maps from the results of the clustering. Our clustering method is based on Newman's method, although Newman's method only focuses on the structure of the network and we introduce the weighting for the links that were introduced in Section III-A.2 to Newman's method. We calculate the value of the weight in the way hereinafter prescribed. Newman's method [13] is the agglomerative hierarchical clustering method and has the modularity Q, which shows the quality of a cluster division. In the clustering in Newman's method, a pair of clusters that provide the maximal value of increment in Q are joined at each step. The definition of Q is as follows:

$$Q = \sum_i (e_{ii} - a_i^2) \quad (4)$$

$$\Delta Q_{ij} = 2(e_{ij} - a_i a_j) \quad (5)$$

where e_{ij} represents the fraction of edges between clusters i and j in the network, and is calculated as the value for the number of edges between clusters i and j divided by the total number of the edges in the network. a_i is the fraction of edges belonging to cluster i and is denoted as $a_i = \sum_i e_{ij}$. Q is the difference between the fraction of the number of edges within all the clusters in the network and the expected value for the same cluster division in a network whose edges are randomly connected. If the number of edges within the clusters is lower than the random connected edges, we will obtain $Q = 0$. The maximal value of Q is 1; a high value represents a good partitioning of the network.

We cluster Web pages using our method which is based on Newman's method, using above mentioned weights, and according to the following procedure.

1) Set a cluster as a page, and an edge as a link in a given set of Web pages.
2) Calculate all the values of the weight for the links, and apply them to the edges. Each weighted value of an edge is normalized by the total value of the edges.
3) Select a pair of linked clusters that have a maximal value of ΔQ. Terminate if the ΔQ of the selected pair has a 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 back to Step 3.

D. Extracting and visualizing the topic maps prototypes

Our process of extracting three types of items for topic maps is as follows.

- **topic:** A topic is a concept represented by the Web pages within a cluster. A cluster presents a topic. When the topic is extracted, it has no name. A user gives a proper name to the topic because it's hard to automatically name it.
- **association:** An association is represented as an edge between clusters, which remains at the end of the clustering. As is for the topics, it's difficult to automatically annotate the association. A user adequately gives the annotation to the association.
- **occurrence:** Occurrences are pointers to information resources, as in a set of Web pages, which are related to the cluster.

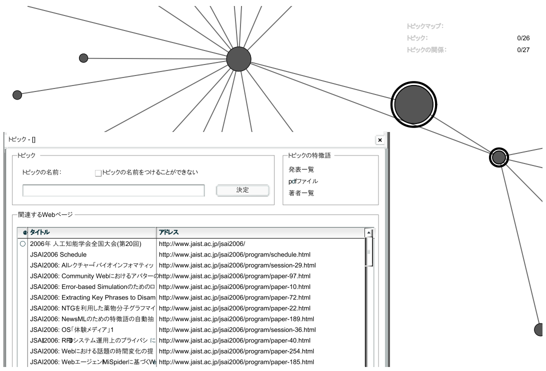
We generate the topic map prototypes with extracted items and visualize the topic maps on a graph using graphviz [5], which is one of the graph visualization tools. On the graph, the topics are represented as nodes and the associations are represented as edges. The area of a node is dependent on the number of pages related to the topic.

IV. Experiments

We conducted experiments to evaluate our proposed method by comparing the topic map prototypes that were generated using our method to those using Newman's method. The comparison with the forms generated by the two methods explains the effect of the introduction of the weighting. Sixteen participants took part in this experiment. Each participant supplied two sets of browsing histories, which were used to collect the sets of Web pages, and also evaluated four topic maps generated from the sets of pages using both methods. The order for evaluating the topic maps was random.

A. Extracting topic map prototypes

In this experiment, we used the sets of Web pages that were collected using the browsing histories of the participants as seeds. As these sets of pages consisted of pages related to the browsed pages, they easily evaluated the extracted topic map prototypes. The histories are restricted to a sequence of five pages searching for some sort of issue. The pages in the histories have to be connected to each other, because our method is based on the links between pages. We obtained a set of Web pages by following the links four times from the pages of the histories. There were three selected links in each of the pages and the links were selected at random. The sets of pages contained around 600 pages. We extracted the prototypes from these sets of pages using both methods, and



トピックマップ:
トピック: 0/26
トピックの関係: 0/27

トピック - []

トピック

トピックの名前: □トピックの名前をつけることができない

決定

トピックの特徴語

発表一覧
pdfファイル
著者一覧

関連するWebページ

@	タイトル	アドレス
○	2006年 人工知能学会全国大会(第20回)	http://www.jaist.ac.jp/jsai2006/
	JSAI2006 Schedule	http://www.jaist.ac.jp/jsai2006/program/schedule.html
	JSAI2006: AIレクチャー「バイオインフォマティッ	http://www.jaist.ac.jp/jsai2006/program/session-29.html
	JSAI2006: Community Webにおけるアバターの	http://www.jaist.ac.jp/jsai2006/program/paper-97.html
	JSAI2006: Error-based Simulationのためのロ	http://www.jaist.ac.jp/jsai2006/program/paper-10.html
	JSAI2006: Extracting Key Phrases to Disam	http://www.jaist.ac.jp/jsai2006/program/paper-72.html
	JSAI2006: NTGを利用した薬物分子グラフマイ	http://www.jaist.ac.jp/jsai2006/program/paper-22.html
	JSAI2006: NewsMLのための特徴語の自動抽	http://www.jaist.ac.jp/jsai2006/program/paper-189.html
	JSAI2006: OS「体験メディア」1	http://www.jaist.ac.jp/jsai2006/program/session-36.html
	JSAI2006: RFIDシステム運用上のプライバシ に	http://www.jaist.ac.jp/jsai2006/program/paper-40.html
	JSAI2006: Webにおける話題の時間変化の提	http://www.jaist.ac.jp/jsai2006/program/paper-254.html
	JSAI2006: WebエージェントMiSpiderに基づくWe	http://www.jaist.ac.jp/jsai2006/program/paper-185.html

Fig. 4. Interface for evaluation of topic maps

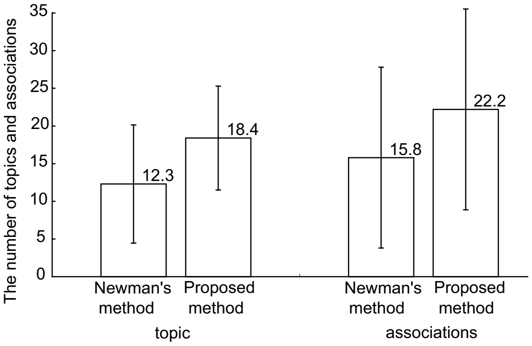


Fig. 5. No. of valid topics and associations

visualized them on an interface for evaluation. The interface shows the clusters that have pages in the histories as ringed nodes.

B. Evaluation

The topic maps had no labels for the topics and associations when they were extracted. The participants gave appropriate names and annotations to the topics and associations. Fig. 4 shows the interface. The interface shows the extracted topic map, representing the topics as nodes and the associations as edges between the topics. By clicking the nodes and edges, the interface shows the information window for the topics and associations that correspond to the nodes and edges. The participants get the information concerning the topics and associations through the window.

1) Evaluating topics: As the topic of the cluster is a concept represented by a set of Web pages within it, the name of the topic is the name of the concept. The participants judge what the concept of the set of pages is, and appropriately name the topic, referring to the following information:

- Title, URL, and content of the pages related to the topic.
- Three domain-specific terms.

When the cluster has a number of topics of the Web pages, the participants evaluate the cohesiveness of the topics of the Web pages and select a major one and name it. If the topics in the cluster are vary widely, the participants don't give the proper name to the cluster. The participants also evaluate the granularity of the cluster's topic in terms of results. Then, the granularity of the individual topics in the topic maps varies. We call the topics named properly "valid topics". On other hand, the topics that have no name are called "invalid topics".

2) Evaluating associations: The associations are described as edges between clusters that present topics, and represent some kind of relationship between the topics. The participants judge what relation between the topics is. If feasible, they give proper annotation to the associations, referring to following information:

- The information concerning the two topics that have an association.
- The information concerning the links between the pages that consist of the edges.

As in the case of the topics, we call the associations that are provided with appropriate annotations, "valid associations". The associations having no annotations are called "invalid associations".

C. Experimental Results

We evaluate whether the topics and the associations are valid and whether the individual ones are appropriately extracted. We don't evaluate whether they are useful for the participants. Because the problems of evaluating the utility of the topic and the associations depend on the knowledge and information of the participants, it's hard to evaluate it. Thus, in this experiment, we evaluate the validity of the topic map prototypes by the participants' evaluation for the cohesiveness and the granularity of the individual topics and associations in the prototypes.

We extract topic map prototypes from the set of pages, show them to the users. The users finally have to modify and fix the topic map prototypes, so they can be used as the topic maps. It's preferable that the prototype contains valid topics and associations that are extracted broadly. The tasks for adding insufficient topics and associations are more costly than the tasks for reducing the unnecessary items.

To evaluate the fitness of the prototypes, we evaluate the number of valid topics and associations, especially those regarding the topic maps extracted by both methods. In this case, recall [1], which is often used to evaluate the performance in the field of IR, is unusable, because the complete topic maps in the sets of pages is not obvious and the topics and associations that are required to be extracted are unclear.

D. Analyzing Results

Fig. 5 shows the average number of valid topics and associations. The results showed that the average number of valid topics was 12.3 (standard deviation: 7.84) when using the Newman's method, 18.4 (6.89) for the proposed method, and the same value of valid associations was 15.8 (12) for the Newman's method, but 22.2 (13.22) for our method.

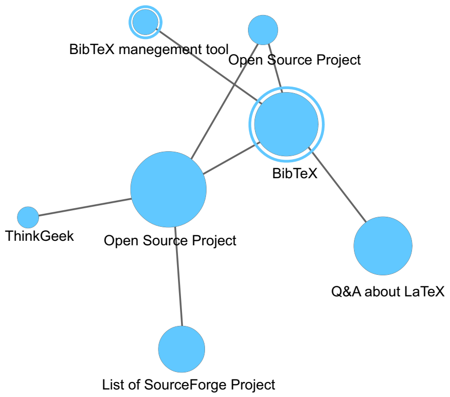


Fig. 6. Topic map 1 (Newman's method)

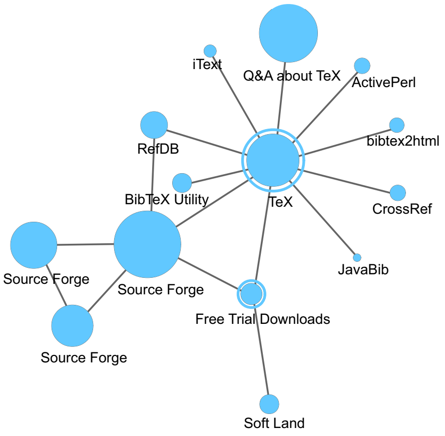


Fig. 7. Topic map 1 (proposed method)

Regarding the number of topics, a comparison made using Wilcoxon signed-ranks test showed a significant difference between both methods (p= 0.000012, α= 0.05). The comparison in the number of associations using a paired t-test also showed a significant difference (p= 0.002, α= 0.05). These results show that our proposed method can extract topic maps that have more valid topics and associations than that by the Newman's method.

V. Discussion

The results of the experiments represented that our method can extract enough topics and associations to be suitable for prototyping topic maps. When a user modifies a prototype, the tasks for generating scarce topics and appending them are more costly than those for reducing the noise from already extracted items. When a user generates topic maps that involve topics and associations that have never been seen before, the process of finding and inserting the unknown topics is very costly. The topic maps extracted by the proposed method have a sufficient enough number of topics and associations, and potentially show the useful items. We represent interesting examples from extracted topic maps in the following way.

A. Discovery of topics

In this case, we see the participant search tools related to "BibTeX". Fig. 6 shows the topic map generated using Newman's method. Fig. 7 shows the one made by our method. These figures show that the number of topics by the proposed method is larger than those by Newman's method, and our method's topics are segmentalized. In Fig. 7, we can find the topics "Utilities of BibTeX", "bibtex2html", "JavaBib", and others around the topic "TeX". Topic "BibTeX" involves these topics in Fig. 6. These divided topics have occurrences connected to Web pages on several independent Web sites. As the proposed method regards the types of links and the distance of the directories, it can extract these topics. Then, the participant is able to find relevant tools, such as "bibtex2html", which is unknown to him/her.

B. Discovery of associations

In this case, the participant search information concerned "JSAI2006", the annual conference of JSAI in 2006. Fig. 8 shows the topic map generated using Newman's method, and Fig. 9 shows the one generated by our proposed method. In both figures, we find some similar topics, such as "AAAI", "ACM", "IEEE", and "NIST" around "JSAI". These topics are societies and research institutes that are relevant to "JSAI". In the topic maps generated using Newman's method (Fig. 8), the topic, "JSAI2006", which is the participant's target information, is involved in "JSAI". In contrast, in the topic map generated by our method (Fig. 9), we can find these two topics separately and some associations that are hidden in the topic map made by Newman's method. The topic "JSAI2006" is connected to "ITmedia", "Livedoor News", and "MAINICHI News" at the right in Fig. 9. These associations show that each Japan-based news site runs an article about "JSAI2006". It is difficult for the participant to find these associations from the topic maps made by Newman's method, because the topic "JSAI2006" and the associations between the topics is hidden.

C. Open problems

Inadequate topic maps are extracted from sets of Web pages, and this involves a number of dynamic pages like Wiki pages. The Wiki pages swap links with each other. Our proposed method does not carry out the preliminary processing that selects and reduces the number of links, because we have assumed that some links have some kind of relationship between the topics of pages. The Wiki pages have URLs that are dynamically generated on flat directories in the Web site structure, so it's hard to accommodate these pages, because our proposed method weights the links depending on the directories in the Web site structure. Some news sites have the problem in the same way. The Web pages

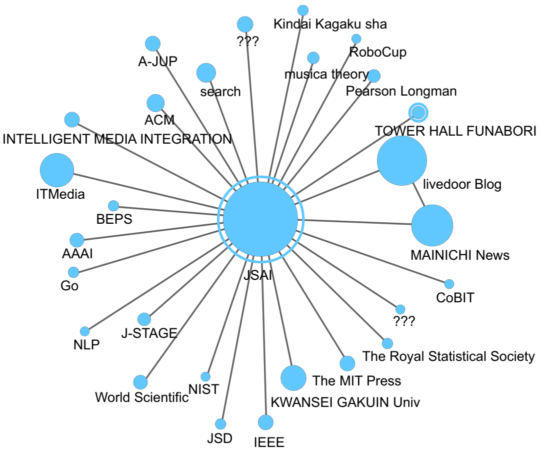


Fig. 8. Topic map 2 (Newman's method)

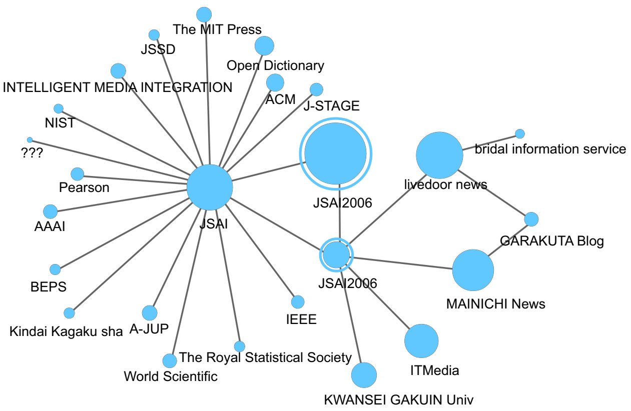


Fig. 9. Topic map 2 (proposed method)

of the news sites are sorted to the directories by date, our proposed method is unconformable for these Web sites. So, we need to adjust our method to accommodate the increasing number of these types of pages.

To solve the problem, we used tags or categories given to Web pages, in place of the directories sorted by date in the Web sites. When URL of a Web page has the flat directory or the directories sorted by date and the page has the tags or the categories, we introduced a virtual directory to the URL of the page by using given tags and convert the URL as shown by Fig. 10. We similarly calculated the weights as shown in Section III-A.2 and extracted the topic map prototypes. We experimentally extracted a topic map from a set of Web pages that are collected from PCUSER, which is a Japanese news site about PC, by our proposed method with converted URLs. Fig. 11 shows the extracted topic map from the Web pages of PCUSER. The pages of PCUSER[2] are sorted into the directories by date, and have URLs as shown by Fig. 10. The pages of PCUSER have some tags given from 293 tags according to the topics of the pages and are sorted into the categories as follows: "notebook PC", "desktop PC", "Mac", "printer", "peripheral", "PC parts", "Akihabara", "mother board", "Software", "event", "joke". The classification by the tags is too detailed and complicated. We can find the topics presented by the categories in the Web pages of PCUSER, but which the topics are related is unclear. In contrast, the extracted topic map shown as Fig. 11 shows some topics related to PC and relationships between the topics. We find the topic "pc parts" and the related topics, such as "graphic card", "mother board", "power supply", "TV tuner" around "PC parts". Also, We can find the relationships between "PC parts" and the related topics. The extracted topic map supports a survey of the Web sites.

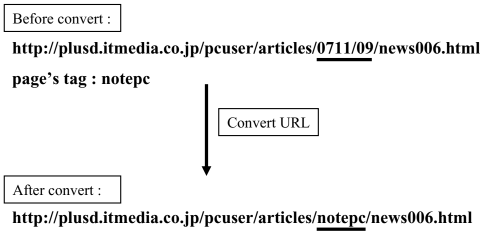

Before convert :
http://plusd.itmedia.co.jp/pcuser/articles/0711/09/news006.html
page's tag : notepc
Convert URL
After convert :
http://plusd.itmedia.co.jp/pcuser/articles/notepc/news006.html

Fig. 10. Convert URL

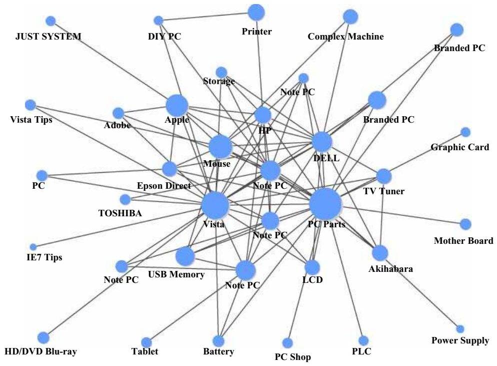


Fig. 11. Topic map from PCUSER

The user needs to complete the topic map prototypes by labeling, adding and removing the topics and the associations, to use the prototypes as topic maps. In future work, we will give Support Vector Machine-based relevant feedback [14][15] by using the data of user's modification to the prototypes, to reduce the process of removing the unnecessary topics and associations. Relevance feedback [17] is used in Information Retrieval and a method of modifying the query by a user's evaluation about whether a document is relevant or not to improve retrieval performance. In general,

[2] http://plusd.itmedia.co.jp/pcuser/

relevance feedback modifies the query by adjusting the document vector of query to them of the relevant documents evaluated by user. In contrast, SVM-based relevant feedback method regards the relevant and non-relevant documents as positive and negative examples and applies SVMs as the classifier to the relevance feedback method. The SVMs are learning machines which can perform the binary classification problems and have great performance for the problems. The SVM-based relevant feedback method generates a hyper-plane for classifying relevant and non-relevant documents by using SVM learned by user's judgments of the documents. Next, the method classifies the unchecked documents, which are not evaluated by the user, as relevant or non-relevant by using the hyper-plane as the discriminant function. Finally, the method maps the all documents into the feature space that has the hyper-plane as the discriminant function, and ranks the documents according to the distance between the documents and the hyper-plane, and shows the documents ranked by the order according to the rank.

We intend to apply iteratively SVM-based relevance feedback to the set of Web pages with user's evaluation for the topic map prototypes and reduce the unnecessary Web pages from the next set of Web pages beforehand. We regard the Web pages, which are contained in the unnecessary topics evaluated by the user, as non-relevant documents and classify the next set of Web pages by using a hyper-plane generated by a SVM that is learned by the evaluation of the documents. We previously remove unnecessary Web pages which are classified as the non-relevant documents and reduce the procedure for removing the unnecessary topics and associations from the user's modification of the topic map prototypes.

In addition to this, we have to build the interface that enables interactive improvement of the prototypes by user. The interface needs the functions as follows: 1) Editing of topics and associations. User can name the topics and the associations, remove the unnecessary items, add the missing items, and merge the items which have the same name. 2) Display of information about the topics and associations. User can name and add the items by referring to the information. 3) Support for editing of items. User can spread the changes to the prototypes by an edit once. For example, when user remove an association between topic "A" and "B", the other associations between topic "A" and "B" are automatically deleted.

VI. Conclusion

We proposed a framework in this paper to extract topic maps from a set of Web pages. First, we proposed a method to extract the topic map prototypes. Our method was based on Newman's method using a weighting scheme based on the similarities between the contents of pages and the relevance between the topics of the Web pages by the types of links. We conducted experiments to evaluate the proposed method by comparing it to Newman's method. The results of the experiments showed that our method can extract the topic maps that have enough topics and associations with the topic map prototypes.

In our future work, we will apply SVM-based relevant feedback by using the data from user's evaluations for the topic map prototypes to Web pages, to extract more appropriate topic maps by reducing the number of unnecessary Web pages. And we will build the interface which has the function to assist user to complete the topic map prototypes.

References

[1] Baeza-Yates, R. and Ribeiro-Neto, B.: *Modern Information Retrieval*, Addison Wesley, 1999.

[2] 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 *Proceedings of the 9th International World Wide Web Conference*, pp. 309–320, 2000.

[3] H. Drucker, B. Shahrary and D.C. Gibbon, "Relevance Feedback using Support Vector Machines", in *Proceedings of the Eighteenth International Conference on Machine Learning*, pp. 122–129, 2001.

[4] Flake, G. W., S., L., and Giles, C. L.: Efficient identification of Web communities, in *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 150–160, 2000.

[5] Gansner, E. R. and North, S. C.: An open graph visualization system and its applications to software engineering, *Software – Practice and Experience*, Vol. 30, No. 11, pp. 1203–1233, 2000.

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

[7] A. Gulli and A. Signorini, "The indexable web is more than 11.5 billion pages", in *Special interest tracks and posters of the 14th international World Wide Web Conference*, pp. 902–903, New York, NY, US A, 2005.

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

[9] International Standard Organization: ISO/IEC 13250 Topic Maps: Information Tecknology Document Description and Markup Language, 2000.

[10] Jain, A. K. and Dubes, R. C.: *Algorithms for clustering data*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1988.

[11] R. Kerk and S. Groschupf, "How to Create Topic Maps", http://www.media-style.com/gfx/assets/Howto CreateTopicMaps.pdf, 2003.

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

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

[14] T. Onoda, H. Murata and S. Yamada, "SVM-based Interactive document Retrieval with Active Learning", *New Generation Computing*, Vol. 25, No. 3 (to be published, 2007).

[15] T. Onoda, H. Murata and S. Yamada, "Comparison of Learning Performance and Retrieval Performance for Support Vector Machines based Relevance Feedback Document Retrieval", in *Proceedings of Intelligent Web Interaction Workshop 2007*, pp.249-252, Silicon Valley, USA, 2007.

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

[17] G. Salton, editor., "Relevance feedback in information retrieval", pages 313–323. Englewood Cliffs, N.J.: Prentice Hall, 1971.

[18] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval", *Information Processing & Management*, Vol. 24, No. 5, pp. 513–523, 1988.

[19] E. Spertus, "ParaSite: mining structural information on the Web", in *Selected papers from the 6th international World Wide Web Conference*, pp. 1205–1215, 1997.

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