Mining Web Site's Topic Hierarchy

Nan Liu
Department of Systems Engineering and
Engineering Management
The Chinese University of Hong Kong
nliu@se.cuhk.edu.hk

Christopher C. Yang
Department of Systems Engineering and
Engineering Management
The Chinese University of Hong Kong
+852 2609 8239
yang@se.cuhk.edu.hk

ABSTRACT

Searching and navigating a Web site is a tedious task and the hierarchical models, such as site maps, are frequently used for organizing the Web site's content. In this work, we propose to model a Web site's content structure using the topic hierarchy, a directed tree rooted at a Web site's homepage in which the vertices and edges correspond to Web pages and hyperlinks. Our algorithm for mining a Web site's topic hierarchy utilizes three types of information associated with a Web site: link structure, directory structure and Web pages' content.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Search and Retrieval – search process, retrieval models

General Terms: Algorithms, Experimentation

Keywords: Content Structure, Web Mining, Topic Hierarchy

1. INTRODUCTION

Recent years have seen an increasing interest in the modeling of various Web structures. Such models have been shown to effectively enhance the performance of Web search tasks. However, almost all the existing works focus on the modeling of the entire Web. Recently, there have been several interesting applications [1, 2, 3], which only deal with a subset of the Web, in particular, pages within a particular Web site. Such applications call for the effective modeling of a complete Web site instead of the entire Web. In this work, we propose to use a topic hierarchy to model a Web site's content structure.

2. WEB SITE CONTENT STRUCTURE

Unlike the whole Web, Web sites are much more structured. Organizing content is a necessary step when one creates a Web site. The most common devices for content organization include hyperlinks among related pages and folders dividing large number of pages into smaller groups. In addition, similar to other types of publications such as book, a hierarchical structure is typically used to present information in a Web site for its simplicity and clarity. A large Web site includes very broad topics For example, the information of an organization Web site (e.g. IBM) includes several subtopics, such as products, services & solutions, support & downloads, etc. Usually, sub topics can be further divided into sub topics. Such a process resembles the division of a book into chapters, sections and subsections. However, a Web site doesn't provide the table-of-content to display its hierarchical content structure as a book does. Although a site map may be provided at some Web sites, existing site maps only covers a very small proportion of Web pages within a site partially due to the tedious process of constructing it manually. Constructing site maps become even more difficult when parts of a site are created independently by different authors. For example, within a departmental site in an academic institution, different professors' personal sites are designed by themselves and the linked to the departmental site.

In this work, we propose to model Web site's content structure using a **topic hierarchy**. A topic hierarchy for a Web site is a directed tree structure whose vertices and edges correspond to Web pages and hyperlinks from the Web site. A topic hierarchy must have the following properties:

- It must be rooted at the homepage of the Web site.
- It covers all the pages in a Web site that can be reached from the Web site's homepage by following a series of hyperlinks.
- All hyperlinks included in the topic hierarchy must be downward links, i.e., pointing from ancestors to their direct descendants.

Each sub tree ***t*** within the topic hierarchy captures a particular topic ***T***, which is elaborated by the pages in the sub tree. The root page of ***t*** connects directly to a set of pages ***P*** in the topic hierarchy. Those pages in ***P*** may themselves be root pages of sub trees contained by ***t***, which correspond to sub topics of ***T***. Figure 1 below shows part of the topic hierarchy for the Stanford database group's Web site: www-db.stanford.edu.

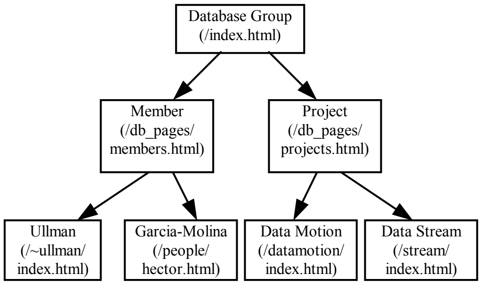

Figure 1: Partial Topic Hierarchy for www-db.stanford.edu

3. TOPIC HIERARCHY MINING

As a content representation model, topic hierarchy aims at capturing the semantic relationships among Web pages. Chen et al. [3] suggests using the presence of hyperlinks as an indication of semantic relationships among pages. They classify the semantic

WWW 2005, May 10-14, 2005, Chiba, Japan.
ACM 1-59593-051-5/05/0005.

relationships among pages into two broad classes: association and aggregation. Aggregation is kind of hierarchical relationship, in which the parent node represents a broader concept than the child node. It is non-reflexive, non-symmetric and transitive. An example of aggregation relationship is the link from the *staff* page of an organization to an individual staff's personal page. Association is non-hierarchical, reflexive and symmetric. Examples of association include links from a project page to its group members' pages, where a project doesn't conceptually subsume its members.

Following these notions, topic hierarchy mining can be viewed as the process of selecting a sub set of hyperlinks indicating aggregation relationship to connect the web pages under the constraint that resulted structure must have the properties of a topic hierarchy.

Three types of information are used in the mining of topic hierarchy: link information, content information and directory information. Link information about a Web site can be represented by a directed graph *G(V, E)*, where *V* and *E* are vertex and edge sets. Content information can be obtained by extracting the textual content from Web pages. Directory information of Web pages is revealed by the path description within their URLs. The content and directory information are incorporated into the directed graph *G(V, E)* by assigning a cost to each link ($u \rightarrow v$) based on the **content dissimilarity** and **path dissimilarity** between *u* and *v*.

- Content Dissimilarity:

The set of terms used on a web page can roughly tell us about its content. Semantically related pages often mention many common terms. We represent a page's textual content by the simple bag-of-terms model, where each page is associated with a set of terms contained in its text. The dissimilarity between two pages *u* and *v* is then computed as follows:

$$d_{content}(u,v) = 1 - \frac{|S_u \cap S_v|}{|S_u \cup S_v|} \qquad (1)$$

where S_u and S_v are the sets of terms on *u* and *v* respectively.

- Path Dissimilarity:

Directory is a commonly used for organizing large number of documents. A web designer creates folders and subfolders to group similar pages together. Therefore, the locations of pages in the web site's directory can also be used to measure their semantic relationships. A Web pages's URL is of the format *protocol::/hostname[:port number][/path]*. By splitting a URL's *path* component using "/" as delimiter, we can represent it by a list of tokens corresponding to the folder names and the file name. For instance, a path such as "$n_1/n_2/n_3$" is represented by (n_1, n_2, n_3). Given two pages *u* and *v*, whose paths are represented as (u_1, u_2,.., u_n) and (v_1, v_2,.., v_m), the path dissimilarity between them is :

$$d_{path}(u,v) = 1 - 2 \times \frac{\min\{i \mid 1 \le i \le \min(m,n) \wedge u_i \ne v_i\} - 1}{m+n} \qquad (2)$$

For example, the two paths "people/widom/research.html" and "people/ullman/teaching.html" have a dissimilarity of 2/3 for they start to differ from each other at the second folder.

The total cost of a hyperlink ($u \rightarrow v$) is a weighted sum of the two types of dissimilarities, i.e.,

$$cost(u \rightarrow v) = w_{content} \cdot d_{content}(u,v) + w_{path} \cdot d_{path}(u,v) \qquad (3)$$

where $w_{content}$ and w_{path} are weights controlling the relative importance of the two types of dissimilarity.

After assigning costs to edges in *G(V,E)*, we use the single source shortest path algorithm to build the topic hierarchy, which has a time complexity of $O(|V|^2)$. The homepage of the Web site is used as the source node. After each iteration, we can find the shortest path to one particular node as well as determining its parent in the topic hierarchy. The resulted shortest path tree is satisfies all properties of a topic hierarchy.

4. EXPERIMENTS

We have tested our algorithm on two Web sites: www-db.stanford.edu and www.cs.cmu.edu. For each of the two sites, we have crawled all pages on these two hosts that are reachable from their homepages, which include about 4600 pages from www-db.stanford.edu and 1300 pages from www.cs.cmu.edu. From these two sets of pages, we picked out pages which are linked from more than one pages within the same site to form the evaluation set.

For each page in the evaluation set, we manually examined the pages linking to it and selected the one having an aggregation relationship with the linked page as its parent. To evaluate a topic hierarchy, we check if a page's parent in the topic hierarchy matches with the one assigned by human judge. This allows us to measure the accuracy on the evaluation set.

We used the breadth first traversal algorithm (I) as the baseline method, as it can be viewed as a simplistic version of shortest path algorithm which weighs all edges uniformly. To study the effect of content and path dissimilarity, we tested 3 schemes of weighing the edges for the shortest path algorithm: (II) w_{path}=1,$w_{content}$=0, (III) w_{path}=0,$w_{content}$=1, (IV) w_{path}=1,$w_{content}$=1. Table 1 below shows the accuracies of I, II, III and IV on the two data sets.

Table 1. Accuracy Comparison of Four Algorithms

	I	II	III	IV
www-db.stanford.edu	74.3%	81.6%	81.8%	83.1%
www.cs.cmu.edu	81.8%	82.5%	83.1%	83.1%

5. Conclusion

In this work, we have proposed the topic hierarchy to model a Web site content structure in order to support the searching process. In addition to the hyperlink relationship, we also adopt the content and path dissimilarities. Best first search is utilized to construct the topic hierarchy. In the experiment, we find that our proposed technique outperforms the baseline method that uses hyperlink relationship only.

6. REFERENCES

[1] W.S. Li, O. Kolak, Q. Vu and H. Takano. Defining Logical Domains in a Web Site. Proc. of ACM Hypertext, San Antonio, 2000

[2] M. Ester, H.P. Kriegel and M. Schubert. Web Site Mining: A new way to spot Competitors, Customers and Suppliers in the World Wide Web. In Proc. of ACM SIGKDD 2002

[3] Z. Chen, S. Liu, W. Liu, G. Pu and W.Y. Ma. Building a Web Thesaurus from Web Link Structure. In Proc. of ACM SIGIR, Toronto, Canada, 2003