A Link Structure Based Website Topic Hierarchy Extracting Approach *

Zhao Xu, Qingcai Chen, Hongzhi Guo
xuzhao1984@gmail.com qingcai.chen@hitsz.edu.cn
hongzhi.guo@gmail.com

Abstract

Visualizing hierarchy of a website is very helpful for both users' navigating and search engine efficiently presenting results. In this paper, treating webpages as nodes and hyperlinks as directed edges, the link structure is firstly modeled as weighted directed graph. Considering multiple website features, which include directory path, contents and anchor texts etc.,the weight is determined by semantic relevance between webpages. The single source shortest path algorithm is finally applied to extract the Topic hierarchy. Conducted experiment on real web to evaluate the proposed algorithm shows the proposed method gets an average precision gain of 11.67% than baseline method.

Keywords: Link Structure; Website Topic Hierarchy; Weighted Directed Graph

1. Introduction

Nowadays, the two predominant methods for seeking information on the WWW are navigation and search[1]. For the navigation method, to find the desired information, web users typically explore a website by follow the hyperlinks from the starting pages to the subsequent web pages that are thought as relevant with the starting pages. For the latter method, users firstly search the related content via a search engine and then follow the returned links to find out desired webpages. As web sites in Internet become more and more complicate, it has become more difficult for users to efficiently find out desired information on large Web sites. For the users who are directed to one webpage of a large website, it is even more difficult to quickly locate the position of directed webpage in the website. To overcome this problem, many websites provide sitemaps to facilitate navigation, which provide users an overview of the topic hierarchy of the web site. Since most of the site maps are constructed manually, usually just a little part of web pages among the entire website are covered. And the update of sitemap is also an exhausting and boring work. It is obvious that to conduct some automatic approaches for building sitemap is a very useful and important task for both the web site masters and search engines.

Since the hierarchical model is simplicity and clarity. Web sites organizing their web pages in a hierarchical way are more convenient for both web site de-

* This work is supported by Chinese National Programs for High Technology Research and Development (863) Program (2006AA01Z197).

signers to manage the information and users to navigate in the web site. Under this model, a large website is firstly divided into several topics, and which are recursively divided into more subtopics. This hierarchical content structure is called a topic hierarchy[2]. A partial link structure of "www.sina.com.cn" is illustrated in Fig.1.

In the topic hierarchy, hyperlinks between two pages have two functions, one is for carrying semantic related information, which is called aggregation links [3]. illustrated as solid arrows in Fig.1.

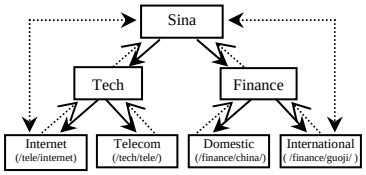

Fig.1 : Partial Link Structure of www.sina.com.cn

Unfortunately, Fig.1 also shows that, for the search engines that just contain web pages for one website, to reconstruct its topic hierarchy is not a trivial task. In fact, the hyperlinks and the web pages usually form a directed graph G (V, E) rather than a hierarchical tree. Here the vertex in V is a web page set and edges in E are corresponded to hyperlinks among web pages. So the main task for topic hierarchy extraction is trying to build a hierarchy tree from the graph. There had been many applications presented for this purpose. Durand and Kahn[10] proposed a heuristically-determined and user-assigned weight that is associated with each hyperlink between two pages to reflect the topic distance between these pages. The disadvantage of this method is that the determination of weight requires the users' participation. Zhu, J., Hong et al.[5] proposed another method to construct the link hierarchy of a web site, which is based on the users' usage data contained in a web log file. They treat the number of user traversals on a link as an important feedback of relevance. Since the web server's log file is usually not opened for a search engine, the application of this algorithm is greatly limited. Liu and Yang[2][6] suggested to analyze the semantic relationships among web pages by link structure, web page content and directory information. Because that just the features directly contained in the web pages are involved, Liu and Yang's method is more practicable for search engines. Our work is based on Liu and Yang's work, but it overcomes the shortcoming of previous method in the following aspects, i.e., it firstly introduces the link type analysis based on the directory structure and links pointing to enhance the URL based weighting. And then the identification of topic entry web page is applied. The content dissimilarity computing is also enhanced by using the vector space model rather than just calculating the word overlap. More discussions can be found in Sub-section 2 and 3.

The rest of the paper is organized as follows. A brief introduction of Liu and Yang's Algorithm is described in section 2. Main processes of our approach to extracting topic hierarchy are given in section 3. In Section 4, we show the experimental results of our method, and results of Liu and Yang's method will be given as a contrast. Section 5 we summarize our method and point out the directions of future work.

2. Baseline Extracting Algorithm

As we discussed above, the main challenge in our work is to distinguish aggregation links from navigational links. We can assign the aggregation links smaller weight, otherwise the navigational links will get bigger weight. After that, topic hierarchy can be obtained by single source shortest path algorithm.

Here is the Extracting Algorithm proposed by Liu and Yang[2][6]:

First, The directory dissimilarity between two pages is calculated based on the number of common folders they are both under:

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

Then they propose measuring two page's content dissimilarity based on their vocabulary overlap:

$$d_{content}(u,v) = 1 - \frac{|S_u \cap S_v|}{|S_u \cup S_v|} \quad (2)$$

Su and Sv are the sets of terms on u and v. two page's content similarity can be presented by intersection of Su with Sv divide by the union of Su with Sv.Thus, The total weight of hyperlink (u, v) can be represented as:

$$cost(u \to v) = \alpha \times d_{path}(u,v) + (1-\alpha) \times d_{content}(u,v) \quad (3)$$

Where α and 1-α are weights controlling the relative importance of the two types of dissimilarity.After assigning costs to edges in G(V,E). Using the single source shortest path algorithm can build a shortest path tree which is satisfies all the properties of a topic hierarchy.

Since formula(1) calculate the dissimilarity of the two URLs to weight the link, which ignores the difference of link's direction information.And calculate the vocabulary overlap between the two pages to measure their content dissimilarity.We propose an Improved Topic Hierarchy Extracting Algorithm in the following section.

3. Improved Topic Hierarchy Extracting Algorithm

Recently research indicates that trying to understand the designer's intention can help us identifying the navigational links[3]. So we introduces the link type analysis based on the directory structure and links pointing to enhance the URL based weighting.

3.1. URL feature for weighting

The URL syntax contents structural information, which we can use to attempt to understand the designer's intention[7]. In URL, a sequence of domain labels separated by ".", components of hierarchical schemes are separated by "/". Using the following URL as an example http://class.edu.sina.com.cn/org/xinhangdao/school.html. Its hostname field is "sina.com.cn", and its directory information can be denoted as(edu, class, org, xinhuangdao, school.html).

Based on the directory information and the Link direction, we can categorize the links of the web sites as follows[3]:

1) Upward link: the target page is in a parent directory.

2) Downward link: the target page is in a subdirectory.

3) Forward link: a specific downward link that the target page is in a sub-subdirectory.

4) Crosswise link: the target page is in other directory other than the above cases.

Based on the result of the above link analysis, if a link is Upward link then it is a navigational links too. So we give the biggest weight to this kind of link, for the

other 3 kinds of links we use directory dissimilarity to measure the weight of this hyperlink.

Given a link from page u to page v, with page u the directory information can be presented as (u_1,u_2,…u_n), similarly page v's directory can be illustrated as (v_1,v_2,…v_m). dpath is the URL feature part weight of hyperlink(u,v):

Assume:

$$k = \min\{i \mid 1 \le i \le \min(m,n) \wedge u_i \ne v_i\}$$

$$d'_{path}(u,v) = \begin{cases} 1 & \text{if } k=m, m<n \\ 1-2\times k/(m+n+2) & \text{otherwise} \end{cases} \quad (4)$$

3.2. Content and anchor text feature weighing

In this section, we describe how to weight hyperlinks using page's content and anchor text feature. We use Vector Space Model to model each page into a vector, then the similarity between two pages can be measured by the similarity of the vector representation of the two pages.

When users navigating a particular topic, the first page which meant to be visited is called Topic entry page[8]. Usually the content in these entry pages is so limited that it cannot represent to whole idea the page had expressed. Since these pages have rich anchor texts which are relevant to the topic, we can add the anchor texts into the content to represent the topic of this page. Meanwhile pages which are not topic entry pages, their contents are enough to represent the idea they want to express. We just model the content into the feature vector of the page.

Generally speaking, topic entry pages have the following characteristics:

1) The filename of URL contains field like index.html or default.html etc.

2) The page has a large number of outdegree links.

3) There are little content in the page, while has rich anchor text. This can be represented by text-link ratio(length of content/length of anchor text).

As discussed above, we represent the content (for topic entry page is content add anchor text) as a vector. Assume we model page u, v as (u_1, u_2, …, u_n)and(v_1, v_2, …, v_n), by Content and anchor text feature the weight for hyperlink from u to v is measured by the distance of the two vectors:

$$d'_{content}(u,v) = 1 - \frac{\sum_{i=1}^{n}(u_i \times v_i)}{\left[\sum_{i=1}^{n}(u_i^2) \times \sum_{i=1}^{n}(v_i^2)\right]^{1/2}} \quad (5)$$

To sum up, the total weight of hyperlink (u->v) can be represented as:

$$cost'(u \to v) = \alpha \times d'_{path}(u,v) + (1-\alpha) \times d'_{content}(u,v) \quad (6)$$

Base on the formula (6), we get the weight for each edge in G(V, E). Then the whole website is donated as a weighted graph. Using the homepage of the website as the source node, we can find a shortest path for each page by single source shortest path algorithm. The result of the algorithm gives us a shortest path tree, which is the topic hierarchy of the website.

4. Experimental Results

We crawled three different type of websites(Portal website: ww.sina.com.cn, Specialized website in finance and economics: www.hexun.com and Specialized website in hardware information: www.zol.com.cn) to test our algorithm. For each of the three websites, we first use a Word Splitter to choose feature words as the site's feature space. Then based on topic entry page identification, we model each page into a vector. Using the weighting algorithm propose above, we get a weighted graph for each website.

By running the single source shortest path algorithm, we can extract the website's topic hierarchy form the weighted graph.

We picked up pages whose in-degree in more than 2, to from our evaluation set. Which means there are at least two different paths from homepage to a particular page in the evaluation set. By adjust the α value in formula 6, we got several path(or only one path)from homepage to a particular page, we manually chose a reasonable one for each page in evaluation to make up our evaluation benchmark.

By checking if the path in the topic hierarchy matches with the one in our evaluation benchmark, we can get the precision of different methods. To find out the effect of directory dissimilarity and content dissimilarity in formula 6, we assigned α five values 0, 0.25, 0.5, 0.75, 1.0, and the precision of each algorithm are shows in Table 1(Ⅰ: www.sina.com.cn; Ⅱ: www.hexun.com; Ⅲ: www.zol.com.cn).

Table 1. Precision comparison for Improved algorithm with different α

	0	0.25	0.5	0.75	1.0
Ⅰ	82.93%	88.31%	93.16%	96.80%	87.90%
Ⅱ	57.35%	67.98%	83.06%	90.21%	74.40%
Ⅲ	45.20%	73.39%	85.95%	94.70%	84.25%

For Baseline algorithm in formula 3, assigning α with five different values 0, 0.25, 0.5, 0.75 and 1.0, we calculate the precision which are showed in Table 2 (Ⅰ: www.sina.com.cn; Ⅱ: www.hexun.com; Ⅲ: www.zol.com.cn).

Table 2. Precision comparison for Baseline algorithm with different α

	0	0.25	0.5	0.75	1.0
Ⅰ	78.17%	81.75%	83.23%	87.80%	81.26%
Ⅱ	60.46%	66.76%	72.71%	76.57%	72.79%
Ⅲ	61.26%	72.70%	77.71%	82.33%	84.06%

Compare the result between table 1 and table 2 for each web site, we get three figures Fig.2 Fig.3 and Fig.4.

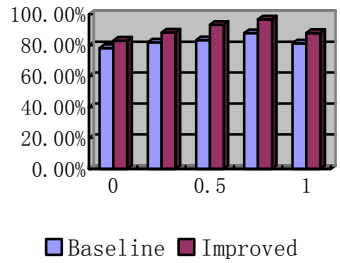

Fig.2 Precision for www.sina.com.cn

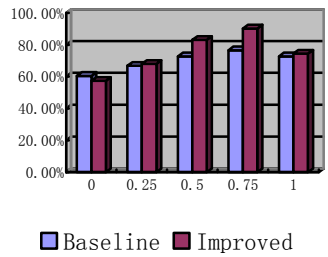

Fig.3 Precision for www.hexun.com

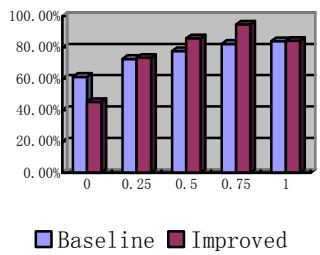

Fig.4 Precision for www.zol.com.cn

We can see that the maximum precision of improved algorithm is high than baseline algorithm on 9.0%, 13.64% and 12.37%. According to what has been discussed above, we can arrive at the conclusion that using vector space model to measure the pages' similarity and identifying the direction of the hyperlink had a significant improvement compared with using word overlap the measure the pages' similarity and ignoring the direction of the hyperlink. When the percentage of weighting factor are directory information 75% with semantic information 25%, our algorithm get the best result in all the three websites

5. Conclusions and future work

It is a tedious work for users to find desired information on large Web sites effectively and efficiently. Visualizing hierarchy of a website is very helpful for both users' navigating and search engine efficiently presenting results. And visualizing the hierarchy of the website, such as site map, can assist users in navigating a website. Since website's topic hierarchy are usually constructed manually by the web designer, it cost much time, and can only cover a limited number pages of the entire website. In this work, we propose a method to extracting the website's topic hierarchy from link structure.

First link type analysis based on the URL directory structure is imported to enhance the URL feature weighting, Then by applying entry page identification and using vector space model to determine semantic relevance between webpages. The single source shortest path algorithm is finally applied to extract the optimized topic hierarchy tree. Experiments on real web data illustrate that our algorithm gets 11.67% improvement than previous method.

The future work focus on optimizing feature word selection for vector space model. Another interesting work is applying the website's topic hierarchy to page ranking or other information retrieval issues.

6. Reference

[1] Olston, C. and Chi, E. H. (2003) ScentTrails: Integrating Browsing and Searching on the Web. ACM Transactions on Computer-Human Interaction (TOCHI), Vol. 10, No. 3, pp. 177-197..

[2] N. Liu and C. C. Yang. Automatic Extraction of Website's Content Structure from Link Structure. In Proc. Of ACM CIKM, 2005.

[3] Z. Chen, S. Liu, W. Liu, G. Pu and W.Y. Ma. Building a Web Thesaurus from Web Link Structure. In Proc. of the 25th ACM SIGIR Conference, Finland,2002.

[4] J. L. Chen, B. Y. Zhou, J. Shi, H. J. Zhang, and Q. F. Wu. Function-based Object Model Towards Website Adaptation, In Proc. of the 10th International World Wide Web Conference, Hong Kong, China, pp. 587-596, May 2001.

[5] Zhu, J., Hong, J. and Hughes, J. G. (2003) PageCluster: Mining Conceptual Link Hierarchies from Web Log Files for Adaptive Web Site Navigation. ACM Transactions on Internet Technology (ACM TOIT), in press, Pages: 185 - 208.

[6] N. Liu and C. C. Yang. Mining Web Site's Topic Hierarchy. In Proc. of International World Wide Web Conference, Tokyo, Japan, 2005.

[7] E. Spertus. ParaSite: Mining Structural Information on the web. In Proc. of 6th International World Wide Web Conference, 1997.

[8] Wen-Syan Li, Necip Fazil, Ayan Okan and Kolak Quoc Vu Constructing Multi-Granular and Topic-Focused Web Site Maps WWW10, May 1-5, 2001, Hong Kong.ACM 1-58113-348-0/01/0005..

[9] M. Maron. Automatic Indexing: an Experimental Inquiry. J. of the Association for Computing Machinery. 1961,8(3):404~417.

[10] Durand, D. G. and Kahn, P. (1998) MAPA: A System for Inducing and Visualizing Hierarchy in Websites. In Proc. of ACM Hypertext 1998, pp. 66-76.