A Link Classification Based Approach to Website Topic Hierarchy Generation

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
yang@se.cuhk.edu.hk

ABSTRACT
Hierarchical models are commonly used to organize a Website's content. A Website's content structure can be represented by a topic hierarchy, a directed tree rooted at a Website's homepage in which the vertices and edges correspond to Web pages and hyperlinks. In this work, we propose a new method for constructing the topic hierarchy of a Website. We model the Website's link structure using weighted directed graph, in which the edge weights are computed using a classifier that predicts if an edge connects a pair of nodes representing a topic and a sub-topic. We then pose the problem of building the topic hierarchy as finding the shortest-path tree and directed minimum spanning tree in the weighted graph. We've done extensive experiments using real Websites and obtained very promising results.

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

General Terms: Algorithms, Experimentation

Keywords: Content Structure, Website Mining, Topic, Hierarchy

1. INTRODUCTION
Users looking for information on the web frequently need to explore particular websites carefully to locate individual pages with interesting information. Many websites provide sitemaps to facilitate navigation. Sitemaps usually list the major topics of a website. By taking a look at the site map, one may quickly settle on one or more topics that are of interest. Despite the usefulness of sitemaps, websites with sitemaps only account for a small portion of the entire web since sitemaps are usually manually constructed. Besides, such manual process is very time consuming.

Individual websites are more organized compared with the entire web. Hierarchical model is most commonly used for the organization of complex bodies of information on websites because of its simplicity and clarity. Using this model, a large website is first divided into a number of broad topics, which are recursively divided into subtopics. We define such hierarchical content structure of a website as its **topic hierarchy**. More formally, a topic hierarchy is a directed tree which is rooted at the homepage of the website and provides a path formed by hyperlinks from the root to every page in the website. In this paper, we study the problem of constructing a website's topic hierarchy, in particular, how to extract it from the link structure of the website, which is a general directed graph.

WWW 2007, May 8–12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.

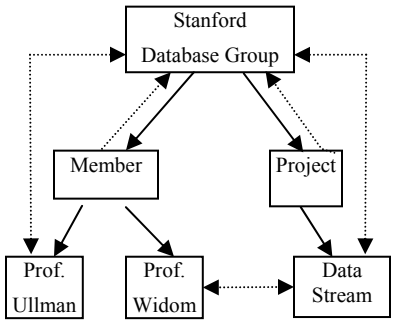

Figure 1: Partial Topic Hierarchy of www.db.stanford.edu

2. TOPIC HIERARCHY MINING

2.1 Overview:
Our approach for generating a Website's topic hierarchy is based on analyzing its link structure, which is a general directed graph as shown in Figure 1. We define two types of hyperlinks based on the different functions they serve. When designing a Website, there are always hyperlinks pointing from a topic to its subtopics, as these are the essential links for browsing a Website. We refer to the hyperlinks connecting a topic to its subtopics as ***aggregation links***, which are represented by the solid arrows in Figure 1. In the mean time, there exist a large number of hyperlinks that are not used to connect topics to subtopics, but are created simply to facilitate navigation between pages, which we refer to as ***short-cut links***. These are indicated by the dashed arrows in Figure 1. A key challenge in extracting the topic hierarchy from the link structure is therefore to distinguish aggregation links from short-cut links.

There exists several graph algorithms that have be adapted to extract a tree structure from a directed graph including breadth-first search, shortest-path search and directed-minimum spanning tree [3]. Breadth first search deals with unweighted graph and minimizes the number of edges between root and other pages. Shortest path search and directed minimum spanning tree work on weighted directed graph and minimizes the total weight of all tree paths and the total weight of all edges respectively.

2.2 Edge Weight Function
Both the shortest path search and the directed minimum spanning tree algorithm work on weighted directed graph. The edge weight function therefore plays an important role in the selection of edges during topic hierarchy construction. Clearly, the edges in a topic hierarchy should be aggregation links. It should be noted that both shortest path search and directed minimum spanning tree favor edges with smaller weight. Therefore, the edge weight function should be designed in such

a way that aggregation links are assigned smaller weights than short-cut links. Our idea is to treat the edge weighting problem as a standard classification problem, in which we attempt to classify a hyperlink into one of two types: aggregation vs. short-cut.

Based on the content, layout and link information, we derived the following set of features for a hyperlink pointing from *u* to *v*:

- **Path Relation**: through examining the path information in each webpage's URL, we may identify if *u* is in the same, the parent, the child or the sibling folder as *v*.
- **Explicit Entry Page**: whether the file name of *u* is `index.html`
- **Content Similarity**: the textual similarity between u and v measured by the cosine between the *tfidf* vector representation of the two web pages.
- **Within Navigation Bar**: whether the link is within a navigation bar, which is identified as a node in the DOM tree containing a set of links.
- **Position in Text**: whether a link is at the beginning, middle or end of a paragraph.
- **Anchor Text Length**: number of words in the anchor text
- **Anchor Text Font Size**: font size of the anchor text

Based on these features, we could build hyperlink classifiers using a set of labeled links. In this work, we used three types of classifiers: naïve Bayes, decision tree and logistic regression. Instead of using the classifiers to produce a binary output indicating the type of the hyperlink, we adapt the classifiers to produce a real value output between 0 and 1 indicating the confidence that the link is an aggregation link. The resulted weighted graph would therefore be more fine-grained than one with edge weights are either 0 or 1.

3. EXPERIMENTS

We tested the algorithm using 5 different web sites. A human judge was asked to manually construct a topic hierarchy for each test website for use as benchmark. Given the benchmark, we evaluate the accuracy of a generated topic hierarchy based on the proportion of web pages which are connected to the same parent as in the benchmark. For each page in the benchmark, the link to it included in the topic hierarchy is an aggregation link while the other incoming links to it represent short-cut link. Table 1 shows some statistics about the benchmark data.

Table 1 Benchmark Dataset Statistics

Dataset	# aggregation links	# short-cut links
www.db.stanford.edu	260	658
www.cs.cmu.edu	124	1023
www.research.ibm.com	441	2057
www.whitehouse.gov	179	1601
www.palmsource.com	123	756

The breadth first search (BFS) algorithm is the simplest approach and represents the baseline in our experiments. Table 1 shows the performance of BFS. For the shortest path search (SPS) and directed minimum spanning tree (DMST) algorithms, we first build the hyperlink classifiers following the leave-one-site-out procedure: to classify the hyperlinks in one Website, we train the classifier using the hyperlinks from the other 4 Websites. This allows us to test if a general link classification model could be obtained to apply to unseen Websites. We build different versions of weighted directed graphs using each of the three types of classifiers and compare the performances of the algorithms on the different graphs Table 3 and 4.

As can be seen from the results, the SPS and DMST algorithms using weighted graph models outperform the simple BFS significantly. The result also shows decision tree as the most effective classifier for generating the edge weight. For 4 of 5 websites, the DMST algorithm with decision tree for edge weighting achieved the optimal performance.

Table 2 Performance of Breadth First Search

Dataset	Accuracy
www.db.stanford.edu	74.3%
www.cs.cmu.edu	77.2%
www.research.ibm.com	71.9%
www.whitehouse.gov	79.4%
www.palmsource.com	71.3%

Table 3 Performance of Shortest Path Search

Dataset	Decision Tree	Naïve Bayes	Logistic Regression
www.db.stanford.edu	90.9%	*92.5%	89.8%
www.cs.cmu.edu	*97.5%	91.5%	81.4%
www.research.ibm.com	*79.4%	75.3%	76.6%
www.whitehouse.gov	*87.8%	81.0%	81.6%
www.palmsource.com	*88.9%	69.2%	70.1%

Table 4 Performance of Directed Minimum Spanning Tree

Dataset	Decision Tree	Naïve Bayes	Logistic Regression
www.db.stanford.edu	88.2%	*88.6%	85.5%
www.cs.cmu.edu	*98.3%	92.4%	89.0%
www.research.ibm.com	*85.9%	80.2%	80.2%
www.whitehouse.gov	*94.6%	89.1%	79.6%
www.palmsource.com	*92.3%	91.5%	91.5%

4. CONCLUSION

We have developed a novel technique to build a Website topic hierarchy by analyzing the link structure, directory structure and content similarity. We have compared with previous techniques and showed that our proposed technique outperforms.

5. REFERENCES

[1] W.S. Li, O Kolak, Q. Vu and H. Takano. Defining Logical Domains in a Website. Proc. of 11th ACM Conf. on Hypertext and Hypermedia, San Antonio, 2000

[2] 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

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