Building a Web Thesaurus from Web Link Structure

Zheng Chen[1], Shengping Liu[2], Liu Wenyin[3], Geguang Pu[2], Wei-Ying Ma[1]

[1]Microsoft Research Asia
5F, Sigma Center, 49 Zhichun Rd
Beijing 100080, P.R.China
{zhengc, wyma}@microsoft.com

[2]Dept. of Information Science
Peking University
Beijing 100871, P.R.China
{lsp, pgg}@is.pku.edu.cn

[3]Dept. of Computer Science
City Univ. of Hong Kong
Kowloon, Hong Kong
csliuwy@cityu.edu.hk

ABSTRACT

Thesaurus has been widely used in many applications, including information retrieval, natural language processing, and question answering. In this paper, we propose a novel approach to automatically constructing a domain-specific thesaurus from the Web using link structure information. The proposed approach is able to identify new terms and reflect the latest relationship between terms as the Web evolves. First, a set of high quality and representative websites of a specific domain is selected. After filtering out navigational links, link analysis is applied to each website to obtain its content structure. Finally, the thesaurus is constructed by merging the content structures of the selected websites. The experimental results on automatic query expansion based on our constructed thesaurus show 20% improvement in search precision compared to the baseline.

Categories and Subject Descriptors

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

General Terms

Algorithms, Experimentation

Keywords

Thesaurus, Link Analysis, Content Structure, Query Expansion.

1. INTRODUCTION

The amount of information on the Web is increasing dramatically, which makes it even harder for Web search. Although existing search engines work well to a certain extent, they still face many challenging problems. One of the problems is word mismatch: the Web editors and the users often do not use the same vocabulary. Another problem is short query: the average length of queries by the user is less than two words [13]. Short queries are often ambiguous in expressing the user's intention. The technique such as query expansion has long been suggested as an effective way to address these two problems.

A query is expanded using words or phrases with similar meanings to increase the chance of retrieving more relevant documents [14]. The central problem of query expansion is how to select expansion terms. Global analysis methods construct a thesaurus to model the similar terms by corpus-wide statistics of co-occurrences of terms and select terms most similar to the query as expansion terms. Local analysis methods use only some initially retrieved documents for selecting expansion terms. Both methods work well for traditional documents, but the performance drops significant when applied to the Web. The main reason is that there is too much irrelevant information contained in a web page, e.g. banners, navigation bars, and hyperlinks that can distort the co-occurrence statistics of similar terms and degrade the query expansion performance. Hence, we need a better way to deal with the characteristics of web pages while building a thesaurus from the Web.

The discriminative characteristic between a web page and a pure text lies in hyperlinks. Besides the text, a web page also contains hyperlinks which connect it with other web pages to form a network. A hyperlink contains abundant information including topic locality and anchor description [19]. Topic locality means that the web pages connected by hyperlinks are more likely of the same topic than those unconnected. A recent study [2] shows that such topic locality is often true. Anchor description means that the anchor text of a hyperlink always describe its target page. Therefore, if all target pages are replaced by their corresponding anchor texts, these anchor texts are topic-related. Furthermore, the Web's link structure is a semantic network, in which words or phases appeared in the anchor text are nodes and semantic relations are edges. Hence, it is possible to construct a thesaurus by using this semantic network information.

In this paper, we refer to the link structure as the navigation structure, and the semantic network as the content structure. A website designer usually first conceives the information structure of the website in his mind. Then he compiles his thoughts into cross-linked web pages using HTML language, and adds some other information such as navigation bar, advertisement, and copyright information. Since HTML is a visual representation language, much useful information about the content organization is missed after the authoring step. So our goal is to extract the latent content structure from the website link structure, which in theory reflects the designer's view on the content structure.

The domain-specific thesaurus is constructed by utilizing the website content structure information in three steps. First, a set of high quality websites from a given domain is selected. Second, several link analysis techniques are used to remove noise links and convert the navigation structure of a website into the content structure. Third, a statistical method is applied to calculate the mutual information of the words or phases within the content structures to form the domain-specific thesaurus. The statistic step

SIGIR'03, July 28–August 1, 2003, Toronto, Canada.

attempts to keep the widely acknowledged information while removing irrelevant information.

Although there is much noise information in the link structure of a website, the experimental results have shown that our method is robust to the noise as the constructed thesaurus effectively represents the user's view on the relationships between words on the Web. Furthermore, the experiments on automatic query expansion also show a great improvement in search precision compared to the traditional association thesaurus built from pure full-text.

The rest of this paper is organized as follows. In Section 2, we review the recent works on the thesaurus construction and web link structure analysis. Then we present our statistical method for constructing a thesaurus from website content structures in Section 3. In Section 4, we show the experimental results of our proposed method, including the evaluation on the website content structure and the use of the constructed thesaurus by query expansion. We summarize our main contributions in Section 5.

2. RELATED WORKS

A simple way to construct thesaurus is to construct manually by human experts. WordNet [8] is an online lexical reference system manually constructed for general domain. Besides the general thesaurus, there are also some thesauri for special domains. wordHOARD [27] is a series of Web pages prepared by the Museum Documentation giving information about thesauri and controlled vocabularies. Although the manually made thesauri are quite precise, it is a time-consuming job to create a thesaurus for each domain and to keep track of the change of the domain. Hence, many automatic thesaurus construction methods have been proposed to supplement the shortcoming of the manual solution. MindNet [22] tries to extract the word relationship by analyzing the logic forms of the sentences by NLP technologies. Pereira et al. [7] proposed a statistical method to build a thesaurus of similar words by analyzing the mutual information [16] of the words. All these solutions build the thesauri from offline analysis of words in the documents.

Our thesaurus is different in that it is built based on web link structure information. Research based on web link structure has attracted much attention in recent years. Early works focus on aiding the user's Web navigation by dynamically generating site maps [3]. Recent hot spots are finding authorities and hubs from web link structure [1] and its application for search [15], and community detection [4]. Web link structure can also be used for page ranking [17] and web page classification [5]. These works stress on the navigational relationship among web pages, but actually there also exist semantic relationship between web pages [21]. However, as far as we know, there is no work yet that has formal definition of the semantic relationship between web pages and provides an efficient method to automatically extract the content structure from existing websites. Another interesting work is proposed by S. Chakrabarti [24] that also considers the content properties of nodes in the link structure and discuss about the structure of broad topic on the Web. Our works focus on discovering the latent content knowledge from the underlying link structure at the website level.

3. BUILDING THE WEB THESAURUS

To construct a domain-specific Web thesaurus, we firstly need some high quality and representative websites for the domain. We use the domain name, for example, "online shopping," as a query to Google Directory search (http://directory.google.com) to obtain a list of authority websites. These websites are believed to be popular in the domain based on the Google's ranking mechanism. After this step, a content structure for every selected website is built, and then all the obtained content structures are merged to construct the thesaurus for this particular domain. Figure 1 shows the entire process. We will discuss each of the steps in detail in the following.

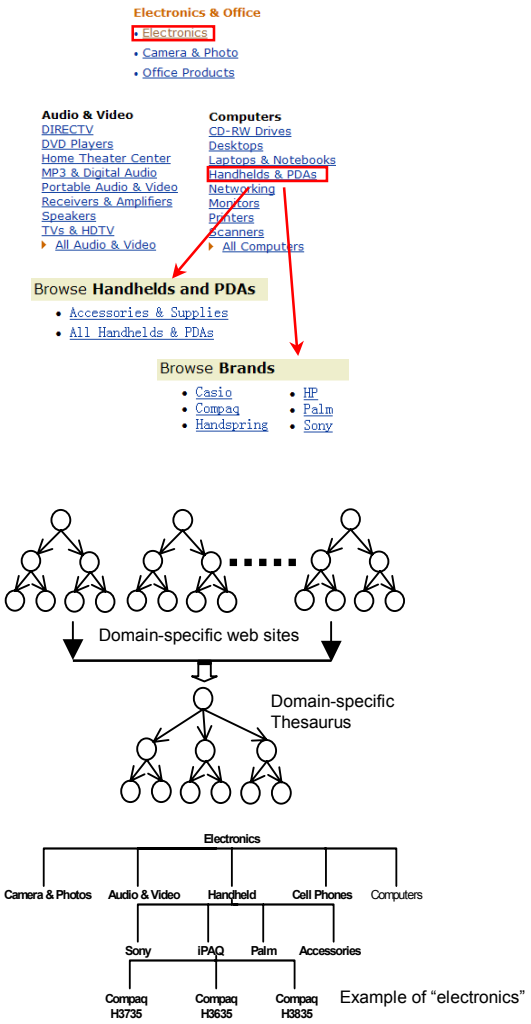

Figure 1. The overview of constructing the Web thesaurus

3.1 Website Content Structure

Website content structure can be represented as a directed graph, whose node is a web page assumed to represent a concept. The concept here stands for the generic meaning of the web page. Thus, the semantic relationship among web pages can be seen as the semantic relationship among the concepts of web pages.

There are two general semantic relationships for concepts: aggregation and association. Aggregation relationship is a kind of hierarchy relationship, in which the concept of a parent node is broader than that of a child node. The aggregation relationship is non-reflexive, non-symmetric, and transitive. The association relationship is a kind of horizontal relationships, in which concepts are semantically related to each other. The association relationship is reflexive, symmetric, and non-transitive. In

addition, two child nodes have association relationship if they share the same parent node.

When authoring a website, the designer usually organizes web pages into a structure with hyperlinks. Generally speaking, hyperlinks have two functions: one for navigation convenience and the other for connecting semantic related web pages together. For the latter one, we further distinguish explicit and implicit semantic relationship: an explicit semantic relationship must be represented by a hyperlink while an implicit semantic relationship can be inferred from explicit semantic relationships and thus does not necessarily correspond to a hyperlink. Accordingly, in the navigation structure, a hyperlink is a semantic link if the connected web pages have explicit semantic relationship; otherwise it is a navigational link. For example, in Figure 2, each box represents a web page in http://eshop.msn.com. The text in the box is the anchor text over the hyperlink which targeted to the web page. The arrow with solid line is a semantic link and the arrow with dashed line is a navigational link.

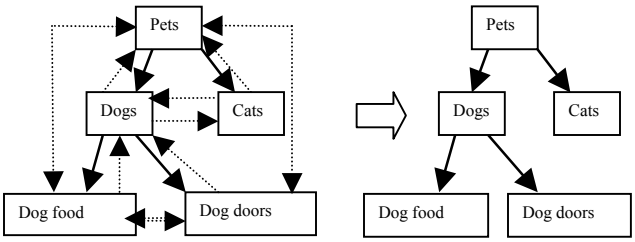

Figure 2. A navigation structure vs. a content structure that excludes navigational links

A website content structure can be represented as a directed graph *G= (V, E)*, where *V* is a collection of nodes and *E* is a collection of edges in the website, respectively. Each node is a 3-tuple (*ID*, *Concept*, *Description*) where

- *ID* is the identifier of the node;
- *Concept* is a keyword or phrase that represents the semantic category of a web page;
- *Description* is a list of name-value pairs to describe the attributes of the node, such as the corresponding web page URL and whether the page is an index or content page, etc.

The root node is the home page of the website. An edge is a 4-tuple (*SourceNode*, *TargetNode*, *Type*, *Description*) where

- *SourceNode* is the node defined previously which contains a semantic hyperlink;
- *TargetNode* is the node pointed by the link of the *SourceNode*;
- *Type* is either aggregation or association;
- *Description* is a list of name-value pairs to describe the attributes of the edge, such as the anchor text of the corresponding link, file name of the images, etc.

3.2 Constructing Website Content Structure

Given a website navigation structure, the construction of the website content structure includes three tasks:

1. Distinguishing semantic links from navigational links;
2. Discovering the semantic relationship between web pages;
3. Summarizing a web page to a concept category.

Since the website content structure is a direct reflection of the designer's point of view of the content, some heuristic rules according to canonical website design rationale [12] are used to help us extract the content structure.

3.2.1 Distinguishing semantic links from navigational links

To distinguish semantic links from navigational links, the Hub/Authority analysis [10] is not very helpful because hyperlinks within a web site do not necessarily correspond to recommendation or citation between pages. Instead, we use the Function-based Object Model (FOM) analysis, which attempts to understand the designer's intention by identifying the functions of objects on a page [11], such as the navigation bar and navigation list objects in a page. In addition, FOM can decide whether a web page is an index page or content page.

To understand the designer's intention, the structural information encoded in URL [6] can also be used. In URL, the directories information is always separated by a slash. Based on the directory structure, links pointing within a web site can be categorized into the following five types:

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 where the target page is in a sub-subdirectory.
4. Sibling link: the target page is in the same directory.
5. Crosswise link: the target page is in other directory other than the above cases.

Based on the result of the above link analysis, a link is classified as a navigational link if it is one of the following:

1. Upward link: because the website is generally hierarchically organized and the upward links always function as a return to the previous page.
2. Link within a high-level navigation bar: High-level means that the link in the navigation bar is not downward link.
3. Link within a navigation list which exists in many web pages: because the link is not specific to a page, so it is not semantically related to the page containing the link. Such link is usually highlight news or recommended product grouped by a navigation list and appeared in many web pages.

Although the proposed approach is very simple, the experiments have proved that it is efficient to recognize most of the navigational links in a website.

3.2.2 Discovering the semantic relationship between web pages

After recognizing and removing navigational links, the remaining are considered semantic link. We then analyze the semantic relationship between web pages based on these semantic links according to the following heuristic rules.

1) A link in a content page conveys association relationship: because a content page always represents a concrete concept and is assumed to be the minimal information unit that has no aggregation relationship with other concepts.

2) A link in an index page usually conveys aggregation relationship: Because index page always functions as a hub to access the content pages and then represents a more generic concept.

3) If two web pages have aggregation relationship in both directions, the relationship is changed to association.

3.2.3 Summarizing a web page to a concept

After the previous two steps, we summarize each web page into a concept. Since the anchor text over the hyperlink has been proved to be a pretty good description for the target web page [23], we simply choose anchor text as the semantic summarization of a web page. While there maybe multiple hyperlinks pointing to the same page, the best anchor text is selected by evaluating the discriminative power of the anchor text by the TFIDF [9] weighting algorithm. That is, the anchor text over the hyperlink is regarded as a term, and all anchor texts appeared in a same web page is regarded as a document. We can estimate the weight for each term (anchor text). The highest one will be chosen as the final concept representing the target web page.

3.3 Merging Website Content Structure

After the building of content structure for the selected websites, we merge these content structures to construct the domain-specific thesaurus. Since the method for building individual website content structure is a reverse engineering process with no deterministic result, incorrect recognition may occur. So we proposed a statistical approach to extracting the common knowledge and eliminating the effect of wrong links from a large amount of website content structures. The underlying assumption is that the useful information exists in most websites and the irrelevant information seldom occurs in the large dataset.

In the "traditional automatic thesaurus" method, some relevant documents are selected as the training corpus, from which a gliding window is used to move over the documents to divide each document into small overlapped pieces. Then, a statistical approach is used to count the terms, including nouns and noun phases, co-occurred in the gliding window. The term pairs with higher mutual information will be formed as a relationship in the constructed term thesaurus. We apply a similar algorithm to find the relationship of terms in the content structures of web sites. Since the content structures of web sites are different because of different views of website designers on the same concept. The content structures of similar websites can be considered different documents in the automatic thesaurus method. The sub-tree of a node with constrained depth performs the function of the gliding window on the content structure. Then, the mutual information of the terms within the gliding window can be counted to construct the relationship of different terms. The process is described in detail as follows.

Since the anchor text over hyperlinks are chosen to represent the semantic meaning of each concept node, the format of anchor text is different in may ways, e.g. words, phrases, short sentence, etc. In order to simplify our calculation, anchor text is segmented by NLPWin [20], which is a natural language processing system that includes a broad coverage of lexicon, morphology, and parser developed at Microsoft Research, and then formalized into a set of terms as follows

$$n_i = [w_{i1}, w_{i2}, \cdots, w_{im}] \tag{1}$$

where n_i is the i^{th} anchor text in the content structure; w_{ij},(j=1,…,m) is the j^{th} term for n_i. Delimiters, e.g. space, hyphen, comma, etc., can be identified to segment the anchor texts. Furthermore, stop-words should be removed in practice and the remaining words should be stemmed into the same format.

The term relationship extracted from the content structure may be more complex than traditional documents due to the structural information in the content structure. That is, we should consider the sequence of the words while calculating their mutual information. In our implementation, we restrict the extracted relationship into three formats: ancestor, offspring, and sibling. For each node n_i in the content structure, we generate the corresponding three sub-trees ST_i with the depth restriction for the three relationships, as shown in Equation (2).

$$\begin{aligned} ST_i(offspring) &= (n_i, sons_1(n_i), \cdots, sons_d(n_i)) \\ ST_i(ancestor) &= (n_i, parents_1(n_i), \cdots, parents_d(n_i)) \\ ST_i(sibling) &= (n_i, sibs_1(n_i), \cdots, sibs_d(n_i)) \end{aligned} \tag{2}$$

where ST_i(*offspring*), ST_i(*ancestor*), and ST_i(*sibling*) are the sub-trees for calculating the offspring, ancestor, and sibling relationship respectively. $sons_d$, $parents_d$, and $sibs_d$ stand for the d^{th} level's son, parent, and sibling nodes for node n_i, respectively.

While it is easy to generate the ancestor and offspring sub-trees by adding the children's nodes and the parent's nodes, generating the sibling sub-tree is difficult because sibling of sibling does not necessarily stands for a sibling relationship. Let us first calculate the first two relationships.

For each generated sub-tree, the mutual information of a term-pair is counted as Equation (3).

$$\begin{aligned} MI(w_i, w_j) &= \Pr(w_i, w_j) \log \frac{\Pr(w_i, w_j)}{\Pr(w_i)\Pr(w_j)} \\ \Pr(w_i, w_j) &= \frac{C(w_i, w_j)}{\sum_k \sum_l C(w_k, w_l)}, \Pr(w_i) = \frac{C(w_i)}{\sum_k C(w_k)} \end{aligned} \tag{3}$$

where $MI(w_i,w_j)$ is the mutual information of term w_i and w_j; $\Pr(w_i,w_j)$ stands for the probability that term w_i and w_j appear together in the sub-tree; $\Pr(x)$ (x can be w_i or w_j) stands for the probability that term x appears in the sub-tree; $C(w_i,w_j)$ stands for the counts that term w_i and w_j appear together in the sub-tree;$C(x)$ stands for the counts that term x appears in the sub-tree.

The relevance of a pair of terms can be determined by several factors. One is the mutual information, which shows the strength of the relationship of two terms. The higher the value is, the more similar they are. Another factor is the distribution of the term-pair. The more sub-trees contain the term-pair, the more similar the two terms are. In our implementation, entropy [26] is used to measure the distribution of the term pair, as shown in Equation (4):

$$entropy(w_i, w_j) = -\sum_{k=1}^{N} p_k(w_i, w_j) \log p_k(w_i, w_j)$$

$$p_k(w_i, w_j) = \frac{C(w_i, w_j \mid ST_k)}{\sum_{l=1}^{N} C(w_i, w_j \mid ST_l)} \quad (4)$$

where $p_k(w_i,w_j)$ stands for the probability that term w_i and w_j co-occur in the sub-tree ST_k; $C(w_i,w_j|ST_k)$ is the number of times that term w_i and w_j co-occur in the sub-tree ST_k; N is the number of sub-trees. This information can be combined with the mutual information to measure the similarity of two terms, as defined in Equation (5).

$$Sim(w_i, w_j) = MI(w_i, w_j) \times \frac{entropy(w_i, w_j) + 1}{\alpha \log(N)} \frac{1}{2} \quad (5)$$

where α is the tuning parameter to adjust the importance of the mutual information factor vs. the entropy factor. In our experiment, α=1.

The term pairs with similarity exceeding a pre-defined threshold will be selected as candidates for constructing the thesaurus for "ancestor relationship" and "offspring relationship."

Then, we calculate the term thesaurus for "sibling relationship." For a term w, we first find the possible sibling nodes in the candidate set $ST_i(sibling)$. The set is composed of three components. The first is the terms which share the same parent node with term w, the second is the terms which share same child node with term w, and the third is the terms that have association relationship with the term w. For every term in the candidate set, we apply the algorithm in Equation (5) to calculate the similarity, and choose the terms with similarity higher than the threshold as the sibling nodes.

In summary, our Web thesaurus construction is similar to the traditional automatic thesaurus generation. In order to calculate the proposed three relationships for each term pair, a gliding window moves over the website content structure to form the training corpus, then the similarity of each term pair is calculated, finally the term pairs with higher similarity value are used to form the final Web thesaurus.

4. EXPERIMENTAL RESULTS

In order to test the effectiveness of our proposed algorithm, several experiments are conducted. First, since distinguishing semantic links from navigational links is an important step for ensuring the quality of constructed thesaurus, a performance evaluation for the selection of semantic links is conducted. Second, the obtained Web thesaurus is used for query expansion to measure the search improvement in comparison with the use of traditional hand-coded thesaurus.

Because the web pages in the standard TREC Web track do not have the same quality of link structure as those from a collection of websites in the real world, we can not build Web thesaurus from the TREC corpus. Therefore we perform our experiments on the web pages downloaded by ourselves.

4.1 Data Collection

Our experiments were conducted on three selected domains, i.e. "online shopping," "photography" and "personal digital assistant (PDA)." For each domain, the domain name was used as a query to Google's search engine to retrieve top ranked websites. The 13 top websites are selected except those with robot exclusion policy which prohibits crawling. These websites are of high quality and representative in the corresponding domain. They are used to extract the website content structure information. Table 1 illustrates the detailed information for the obtained data.

Table 1. Statistics on our text corpus

Domains	Shopping	Photography	PDA
# of websites	13	13	13
Size of raw text (MB)	428	443	144
# of web pages	56480	55868	17823

4.2 Evaluating the Link Selection Algorithm

In order to evaluate the quality of obtained website content structure, we randomly selected 25 web pages from every website based on the sampling method described in [18]. We asked four users to manually label the link as either semantic link or navigational link in the selected web pages. The classic IR performance metrics, i.e., precision and recall, are used to measure the effectiveness of our link classification scheme. However, because the anchor text on a semantic link is not necessary a good concept in the content structure, e.g. anchor texts with numbers and letters, a high recall is usually accompanied by a high noise ratio in the website content structure. Therefore we only show the precision in this paper. Due to the space constraint, this paper only shows the result of the online shopping domain in Table 2.

Table 2. The precision for recognizing the navigational links in the "online shopping" domain

Websites ("www" are omitted in some sites)	#Sem. Links labeled by user	#Nav. links labeled by user	#Nav. Links recog-nized	Precision for nav. links recognition
eshop.msn.com	394	646	638	98.76%
galaxymall.com	308	428	422	98.60%
samintl.com	149	787	737	96.82%
govinda.nu	160	112	80	71.43%
lamarketplace.com	124	416	392	94.23%
www.dealtime.com	198	438	412	94.06%
www.sotn.com	400	1056	1032	97.73%
stores.ebay.com	324	1098	918	83.61%
storesearch.com	308	286	276	96.50%
mothermall.com	54	168	162	96.43%
celticlinks.com	80	230	210	91.30%
internetmall.com	260	696	686	98.56%
lahego.com	86	140	124	88.57%
Average precision				92.82%

From Table 2, we see that the precision of recognizing the navigational links is 92.82%, which shows the simple method presented in Section 3 is effective.

4.3 Evaluating the Constructed Thesaurus

Thesaurus is in general evaluated by the performance of using it for query expansion, so we conducted an experiment to compare the performance using our constructed thesaurus with full-text

automatic thesaurus. Here, the full-text automatic thesaurus is constructed from the downloaded web pages by counting the co-occurrences of term pairs in a gliding window. Terms which have relations with other terms are sorted by the weights of the term pairs in the full-text automatic thesaurus.

4.3.1 Full-text search: the Okapi system

We chose the Okapi system Windows2000 version [25] as our baseline full-text search engine to evaluate the query expansion. In our experiment, the term weight function is BM2500 as shown below:

$$\sum_{T \in Q} w^{1} \frac{(k_1+1)tf\,(k_3+1)qtf}{(K+tf)(k_3+qtf)} \tag{6}$$

where Q is a query containing key terms w^1, tf is the frequency of occurrence of the term within a specific document, qtf is the frequency of the term within the topic from which Q was derived, and w^1 is the Robertson/Spark Jones weight of T in Q and is calculated as follows

$$w^{1} = \log \frac{(r+0.5)/(R-r+0.5)}{(n-r+0.5)/(N-n-R+r+0.5)} \tag{7}$$

where N is the number of documents in the collection, n is the number of documents containing the term, R is the number of the documents relevant to a specific topic, and r is the number of relevant documents containing the term. In Equation (6), K is calculated using Equation (8):

$$K = K_1((1-b)+b*dl/avdl) \tag{8}$$

where dl and $avdl$ denote the document length and the average document length measured in word unit. Parameters k_1, k_3 and b are tuned to optimize the performance. In our experiment, the values for k_1, k_3 and b are 1.2, 1000, 0.75, respectively.

4.3.2 The Experimental Results

For each domain, we used 10 queries to retrieve the documents and the top 30 ranked documents from the Okapi system were evaluated by 4 users. The queries were the following.

1. Online shopping domain: women shoes, mother day gift, children's clothes, antivirus software, listening jazz, wedding dress, palm, movie about love, Canon camera, cartoon products.
2. Photography domain: Kodak products, digital camera, color film, light control, camera battery, Nikon lenses, accessories of Canon, photo about animal, photo knowledge, adapter.
3. PDA domain: PDA history, PDA game, price, top sellers, software, PDA OS, size of PDA, Linux, Sony, and java.

Then, the automatic thesaurus built from pure full-text was applied to expand the initial queries. The terms in the thesaurus were extracted according to their similarity with original query words and added into the initial query. Since the average length of query we used was less than three words, we chose six relevant terms with highest similarities to expand the query. The weight ratio between original terms in the initial query and the expanded terms from the thesaurus was 2.0. After query expansion, the Okapi system was used to retrieve the top 30 relevant documents for evaluation.

Similarly, our constructed thesaurus is also used to expand the query. Note that there are three relationships in our constructed thesaurus. In the experiments, we only used two of them, i.e. offspring and sibling relationship. The ancestor relationship was not used because it would make the query border and thus decrease the search precision. Six terms with the highest similarities to the query words from the obtained thesaurus were chosen to expand initial queries. The weight ratio and the number of documents to be evaluated was the same as full-text thesaurus.

After obtaining the retrieval, we asked four users to provide their subjective evaluation on the results. The evaluation was conducted on a query basis based on the following four system configuration: (1) the baseline with no query expansion (QE), (2) QE with full-text thesaurus, (3) QE with sibling relationship, and (4) QE with offspring relationship. In order to evaluate the results fairly, each user did not know what kind of query results to be evaluated in advance. The search precision for online shopping and photography domain is shown in Table 3 and Table 4. The result for PDA domain is similar to photography domain.

Table 3. Query expansion results for online shopping domain

Online Shopping domain	Avg. Precision (% change) for 10 queries		
# of ranked documents	Top-10	Top-20	Top-30
Baseline	47.0	44.5	44.0
Full-text thesaurus	50.0 (+6.4)	47.5 (+6.7)	46.7 (+6.1)
Our Web thesaurus (Sibling)	52.0 (+10.6)	48.0 (+10.1)	38.3 (-12.9)
Our Web thesaurus (Offspring)	67.0 (+42.6)	66.5 (+49.4)	61.3 (+39.3)

Table 4. Query expansion results for photography domain

Photography domain	Avg. Precision (% change) for 10 queries		
# of ranked documents	Top-10	Top-20	Top-30
Baseline	51.0	48.0	45.0
Full-text thesaurus	42.0 (-17.6)	39.5 (-17.8)	41.0 (-8.9)
Our Web thesaurus (Sibling)	40.0 (-21.6)	31.5 (-34.4)	26.7 (-40.7)
Our Web thesaurus (Offspring)	59.0 (+15.7)	56.0 (+16.7)	47.7 (+6.0)

From Table 3 and Table 4, we find that query expansion by offspring relationship can improve the search precision significantly. And we also find that query expansion by full-text thesaurus or sibling relationship almost does not improve the search precision or even makes it worse. Furthermore, we find that the contribution of offspring relationship varies from domain to domain. For online shopping domain, the improvement is the highest, which is about 40%; while for PDA domain, the improvement is much lower, which is about 16%. We know that different websites may contain different link structures; some are easy to extract the content structure while others are difficult.

Figure 3 illustrates the average precision of all domains. We find that the average search precision for baseline system (full-text search) is quit high, which is about 55%. And the query expansion with full-text thesaurus and sibling relationship can not help the search precision at all. The average improvement for QE with

offspring relationship compared to the baseline is 22.8%, 24.2%, 9.6% on top 10, top 20, and top 30 web pages, respectively. From the results, we can make the following conclusions.

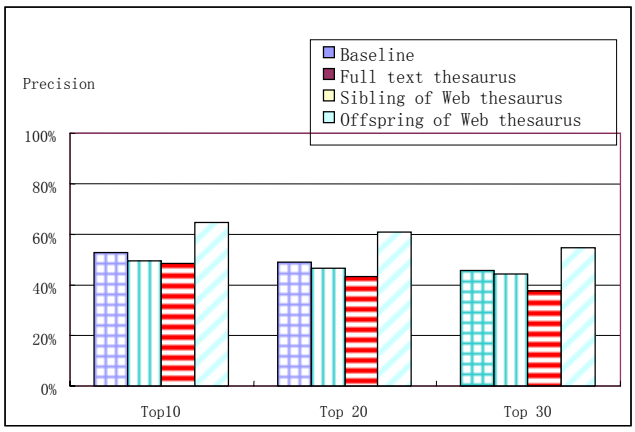

Figure 3. Average performance of all domains

1) The precision of baseline retrieval of a specific domain is quite high. Figure 3 shows that the average retrieval precision of top 30 ranked documents is above 42%. The reason is that the corpus in a specific domain is less divergent than the Web.

2) From Table 3 and Table 4 we can find the query expansion based on the full-text thesaurus decreases the precision of retrieval in most cases. A reason is that that we did not try to tune the parameters to optimize the result. It also seems that the naïve automatic text thesaurus for query expansion is not good. For example, to the initial query "children's clothes" of the online shopping domain, the most six relevant terms of the query in the thesaurus are "book," "clothing," "toy," "accessory," "fashion" and "vintage." When these terms are added to the initial query, they may decrease the retrieval performance.

3) The query expansion based on sibling relationship is bad. It decreases the precision of retrieval in each domain. The reason is that the sibling relationship is more likely to be the words that are relevant to some extent but not similar. For example, when querying the "children's clothes" in the online shopping domain, the most six relevant terms according to sibling relationship in the constructed thesaurus are "book," "toy," "video," "women," "accessories," and "design." Even though these words are related to the initial query, the precision of search result is apparently declined due to the topic divergence.

4) The retrieval precision is improved if the query expansion is based on the offspring relationship. The reason is that the terms of offspring relationship are semantic narrower and can refine the query. For example, when the user submit a query "children's clothes," the system will automatically expand the most relevant terms, "baby," "boy," "girl," "shirt" and "sweater," which are the narrower terms appeared in the offspring sets. Thus, the returned documents will be more likely related to children's clothes.

In summary, query expansion based on the offspring relationship can significantly improve the retrieval performance. The other two relationships are not suitable for query expansion.

5. CONCLUSIONS AND FUTURE WORKS

Although much effort has been devoted to hand-coded thesaurus, to keep up with the speed of growth for new terms and concepts on the Web, automatic thesaurus construction continues to be an important research area for information management. In this paper, we proposed a new automatic thesaurus construction method which extracts term relationships from the link structure of websites. The proposed scheme is able to identify new terms and reflect the latest relationship between terms as the Web evolves. Experimental results have shown that the constructed thesaurus, when applied to query expansion, outperforms traditional association thesaurus.

The limitation of our work is that the current experiment is small in scale. To make the resulting thesaurus more useful, a large collection of data is needed to perform the analysis and testing. For our future work, we plan to extend our algorithm to construct a personalized thesaurus based on the user's navigation history and accessed documents on the Web. This personalized thesaurus will find many interesting applications such as making the Web search more personal.

ACKNOWLEDGMENTS

We are thankful to Jinlin Chen for many valuable suggestions and Xin Liu for developing the system.

6. REFERENCES

[1] A. Borodin, G. O. Roberts, J. S. Rosenthal, and P. Tsaparas. *Finding authorities and hubs from link structures on the World Wide Web*. In Proc. of WWW10, pp. 415-429, 2001.

[2] B. D. Davison. *Topical locality in the Web*. In Proc. of SIGIR'00, pp. 272--279, 2000.

[3] D. Durand and P. Kahn. *MAPA: a system for inducing and visualizing hierarchy in websites*. In Proc. of HyperText'98.

[4] D. Gibson, J. M. Kleinberg, and P. Paghavan. *Inferring Web Communities from Link Topology*. In Proc. of Hypertext'98, pp.225-234, 1998.

[5] E. Glover, K. Tsioutsiouliklis, S. Lawrence, D.Pennock, G. Flake. *Using Web Structure for Classifying and Describing Web Pages*. In Proc. of WWW2002, Hawaii, May 2002.

[6] E. Spertus. *ParaSite: Mining Structural Information on the web*. In Proc. of WWW6, pp. 587-595, 1997.

[7] F. Pereira, N. Tishby, and L. Lee. *Distributional clustering of English words*. In Proc. of ACL-93, pp. 183-190, 1993.

[8] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller. *WordNet: An On-line Lexical Database*, International Journal of Lexicography, Vol. 3, No. 4, 1990.

[9] G. Salton and C. Buckley. *Term-Weighting Approaches in Automatic Text Retrieval*. Information Processing & Management, 24(5), pp. 513-523, 1988.

[10] J. Kleinberg. *Authoritative sources in a hyperlinked environment*. Journal of the ACM, 46(5): 604-632, 1999.

[11] 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 WWW10, pp. 587-596, May 2001.

[12] J. L. Patrick, H. Sarah. *Web style guide - Basic design principles for creating web sites*. Yale University Press, 1999.

[13] J.R.Wen, J.Y.Nie and H.J. Zhang. *Clustering User Queries of a Search Engine*. In Proc. of WWW10, pp. 587-596, 2001.

[14] J. Xu and W. B. Croft. *Query Expansion Using Local and Global Analysis*. In Proc. of SIGIR'96, pp. 4-11, 1996.

[15] K. Efe, V. V. Raghavan, C. H. Chu, A. L. Broadwater, L. Bolelli, and S. Ertekin. *The shape of the web and its implications for searching the web*. In Proc. of SSGRR, 2000.

[16] K. W. Church and P. Hanks. *Word association norms, mutual information and lexicography*. Computational Linguistics, 16(1), 1990.

[17] L. Page, S. Brin, R. Motwani, and T. Winograd. *The PageRank Citation Ranking: Bring Order to the Web*. Technical Report, Stanford University, 1998.

[18] M. Henzinger, A. Heydon, M. Mitzenmacher and M. Najork. *On near-uniform URL sampling*. In Proc. of WWW9. pp. 295-308, May 2000.

[19] N. Craswell, D. Hawking, and S. Robertson. *Effective site finding using link anchor information.* In Proc. of SIGIR'01, pp.250-257, New Orleans, 2001.

[20] Natural Language Processing Group, Microsoft Research. *Tools for Large-Scale Parser Development.* In Proc. of COLING 2000.

[21] O. Liechti, M. Sifer, and T. Ichikawa. *Structured graph format: XML metadata for describing web site structure*. Computer Networks and ISDN Systems, 30:11-21, 1998.

[22] S. D. Richardson, W. B. Dolan, and L. Vanderwende. *MindNet: acquiring and structuring semantic information from text*. In Proc. of COLING'98, 1998.

[23] S. Chakrabarti, B. Dom, and P. Indyk. *Enhanced hypertext categorization using hyperlinks*. In Proc. of SIGMOD, 1998.

[24] S. Chakrabarti, M. Joshi, K Punera, and D. Pennock. *The structure of broad topics on the Web*. In Proc. of WWW11, 2002.

[25] S. E. Robertson and S. Walker. *Microsoft Cambridge at TREC-9: Filtering track*. In TREC-9, 2000.

[26] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, New York, 1991.

[27] wordHOARD, http://www.mda.org.uk/wrdhrd1.htm