Abstract. Electronically available documents on the Web are exploding at an ever-increasing rate. Many Web documents, however, contain rich knowledge that describes the document's content. The Web can be viewed as a body of text containing two fundamentally different types of data: the contents and the tags. A tag is in HTML (Hyper-Text Markup Language) meta-data describing the layout and linking structure between the text. For these kinds of documents we can apply different approaches to extract and structure terms automatically. These approaches are based on a statistical model and syntactic analysis that describe the data of interest, including relationships, and context keywords. In this paper, we discuss an approach to extracting and structuring terms from documents posted on the Web to construct a thesaurus. The proposed tool, ThesWB is used to construct domain independent thesaurus from HTML pages. ThesWB is used to capture the internal structure of meta information embedded in HTML documents.

Keywords: Thesaurus construction, HTML Documents, Information Extraction, Knowledge Management and Discovery, Intelligent Information Retrieval, and Knowledge Extraction.
ThesWB: A Tool for Thesaurus Construction from HTML Documents

Yousef Abuzir¹ and Fernand Vandamme²

¹LAE, University Ghent, Blandijnberg 2, 9000 Gent, Belgium
yousef.abuzir@rug.ac.be

²LAE, University Ghent, Blandijnberg 2, 9000 Gent, Belgium
fernand.vandamme@rug.ac.be

Abstract. Electronically available documents on the Web are exploding at an ever-increasing rate. Many Web documents, however, contain rich knowledge that describes the document's content. The Web can be viewed as a body of text containing two fundamentally different types of data: the contents and the tags. A tag is in HTML (Hyper-Text Markup Language) meta-data describing the layout and linking structure between the text, graphics, audio and interactive components. For these kinds of documents we can apply different approaches to extract and structure terms automatically. These approaches are based on a statistical model and syntactic analysis that describe the data of interest, including relationships, and context keywords. In this paper, we discuss an approach to extracting and structuring terms from documents posted on the Web to construct a thesaurus.

1 Introduction

The World-Wide Web is potentially the world's largest knowledge base but only if new Web techniques are developed to take advantage of its unique characteristics, particularly the semi-structured information within pages, across pages, and in page names.

Users of the Web make use of both intra-document information, such as a page's logical structure or how it should be visually rendered, and inter-document information, such as what pages are connected by hyperlinks. Even the uniform resource locator "URL" identifying a page is structured: typically a host name followed by a location in the file hierarchy. All of these types of information are used by human readers but largely ignored by automated Web tools. This information is potentially valuable to automated tools and show the feasibility of making use of structural information through a system that allows easy creation of thesaurus.

There is information in the layout of a web page, and that by looking at the HTML formatting in addition to the text on a Web page; one can use this information in tasks such as constructing domain dependent thesaurus.
Because of these types of structure, a specialized tool is required to gather structural information and knowledge to build a thesaurus. We provide a tool called ThesWB, which uses these different types of structure in a uniform manner, allowing the user to extract terms and relationships from the Web.

A rich representation for web pages, the HTML Structure, is described. A parsing tool for extracting, creating the hierarchical relationships between terms are presented, as well as an experiment that use this tool to build a thesaurus from Web pages. Our goal is to extract terms and to find the relationships between them either entirely automatically, or with minimal human intervention "semi-automatically".

In this paper, we consider the problem of extracting terms as well as extracting a relation between these terms from Web. In section 2 we start with basics of thesaurus, thesaurus construction and the different approaches to thesaurus construction. Then we show how our ideas differ from the other approaches. Section 3 discusses what type of information about web pages can be used to construct thesaurus. We review related works in section 4. An overview of ThesWB tool and experimental is presented in Section 5. Evaluation is presented in section 6. Finally conclusion and future work are discussed in section 7. Future work is included in this section to present a progression ideas and resulting improvements.

2 Thesaurus

A thesaurus is a set of concepts in which each concept is represented with at least synonymous terms, broader concepts, narrower concepts, and related concepts. A term is a word or sequence of words that refers to a concept [1]. The relations or links broader, narrower, related, and synonymous have been defined for thesauri [2,3]. The relation ‘broader’ can mean: Class inclusion, whole-part. The relation ‘narrower’ includes the reverse relations of those listed for ‘broader’. Two terms are considered to be ‘related’, if they are related but neither is ‘broader’ than the other. ‘Related’ may be used to identify terms that are related to each other from a certain point of view, such as usage, action, or process. The ‘synonymous’ relation points to terms that mean the same thing. The standards for thesauri restrict the relations to little more than the above.

Originally intended for indexing and retrieving documents, thesauri have increasingly been seen as knowledge bases and used beyond the domain of librarianship. (Kosovac, 1998)

2.1 Thesaurus Construction

Building manual thesauri requires a lot of human labour from linguists or domain experts and they are expensive to build. Since it is difficult and expensive to build thesauri manually, many researchers attempted to construct thesauri automatically [3]. There are three approaches to construct a thesaurus. The first approach, on designing a thesaurus from document collection, is a standard one [4,5,6,7,8,9,10].
Here the idea is to use a collection of documents as the source for thesaurus construction. This assumes that a representative body of text is available. By applying statistical or linguistic procedures we can identify important terms as well as their significant relationships.

The second approach is merging existing thesauri \[11,12,13,14\]. This second approach is appropriate when two or more thesauri for a given subject exist, that need to be merged into a single unit. If a new thesaurus can indeed be served by merging two or more existing thesauri, then a merger perhaps is likely to be more efficient than producing the thesaurus from scratch.

The third automatic is based on tools from expert systems \[15\]. In this approach thesauri are built using information obtained from users. For example, if a retrieval system user combines two terms by OR, in his/her query these terms are probably synonyms. The majority of the term pairs found were either morphologically similar like “net” and “network” or translations of each other. Other synonyms were hardly found.

The original motivation behind automatic thesaurus construction is to find an economic alternative to manual thesauri. Therefore, many researchers attempted to construct thesauri automatically. The various method methods are:

− The simplest approach is to reuse existing online lexicographic databases, such as WordNet \[16,17\] or Longman's subject codes \[18\].

− Corpus-based methods perform a computation on the text of the documents in the corpus to induce thesauri. For example a hierarchical thesaurus can be constructed from computed list of complex noun phrases where subsumption \[19\] roughly corresponds to the subset relation defined on terms (i.e., "intelligence" subsumes “artificial intelligence”).

− Head-modifier relationships have been used to determine semantic closeness \[9,10\]. Two terms that modify the same words (or are modified by the same words) often belong to the same semantic category

− Another approach is semantic relatedness \[7\] by considering the occurrence of terms in documents. Documents are clustered into small groups based on a similarity measure that considers two documents similar if they share a significant number of terms. With medium frequency terms preferentially weighted. Terms are then grouped by their occurrence in these document clusters.

All the approaches to thesaurus construction mentioned so far perform what we might call construction by content, since information for constructing a thesaurus is extracted from the text of the document. Thesaurus construction by content does not exploit an essential aspect of a hypertext environment like the Web, namely the structure of documents and the link topology. In this paper we investigate a technique for automatic thesaurus construction, which we have called construction by context, since it exploits the context in the HTML document to extract useful information for building thesaurus. Thesaurus construction by context exploits relevance hints that are present in the structure and topology of the HTML documents published on the Web. Combining a large number of such hints, an adequate degree of accuracy can be achieved in constructing the thesaurus.
3 characteristic of the Web

A fundamental characteristic of the Web is the structural information it encodes, which has been obscured through there being a number of different ways to represent different kinds of structure. In this section, we discuss what type of information about Web pages can be used to construct a thesaurus.

3.1 Types of Information on the Web

The World-Wide Web consists of pages of data, addressed by uniform resource locators (URLs). We focus on pages written in hypertext markup language (HTML). HTML includes tags and attributes to specify intra-document structure "via annotations indicating headers, lists, and formatting directives", and inter-document structure "hyperlinks" such as what pages are connected by hyperlinks.

URLs, or Uniform Resource Locators, are the method by which documents or data are addressed in the World Wide Web. In the default configuration, hyperlinks are divided into three categories, internal links (within a page), local links (within a site), and global links.

Web document authors highlight the content with HTML tags, such as title, heading, list, table, italic, bold, underline, etc. Additionally, META tag allows users to add extra information for the documents such as “classification” and “keyword”. Apparently, HTML tags provide significant information to index and classification as well as extracting semantic and hierarchical knowledge.

3.2 Structure within and across web pages

HyperText Markup Language (HTML) [20], which consists of ordinary text augmented with tags and attributes. Tag names are followed by zero or more attribute assignments, all enclosed in angle brackets. In some cases, subsequent text is affected until the appearance of a concluding tag, prefixed with a forward slash (/). This enclosed text is referred to as anchor text. For example, this piece of hypertext:

<font face="Times New Roman" color="#ff0000">Thesaurus</font>

contains the tag /font", which indicates that the anchor text "Thesaurus" should appear in the font specified by the attributes, i.e., in the face " Times New Roman" and the color "#ff0000" is red. HTML is used to express information about a document's internal structure, hyperlinks, and meta-information. We will discuss these in the following separate sections, although a single tag can serve more than one of these purposes.
3.2.1 Internal structure
An HTML document can contain internal structure in the form of headers and lists. The tag \H1" is used to designate a top-level header, \H2" a second-level header, and so forth through \H6", a sixth-level header. Four tags are used to designate different types of simple lists: \OL" (ordered list), \UL" (unordered list), \MENU", and \DIR" (directory list). The \LI" (list item) tag appears before each item.

3.2.2 Hyperlinks
Tags are used to designate hyperlinks from the current page to another location on the Web, typically another page. A link is designated with the \A" (\anchor") tag, and the \HREF" attribute is used to express the destination.

3.2.3 Meta-information
The \META" tag is used to express meta-information meant for Web tools rather than for browsers. The \NAME" and \CONTENT" attributes are used to specify the type of information being given and its value, respectively. Common values of the \NAME" attribute are \description" and \keywords", both used by search engines to categorize, rank, and display URLs.

In summary, we categorize HTML tag into the following types:

- Informative: Contents in these tags are regarded most important, such as \META Classification Keyword>, <TITLE>, <H1-H3>, <P>, <B>, <I>, etc. Thus, the content enclosed by these tags is extracted and stored in a database.
- Uninformative: These tags or the contents enclosed by these tags have no effects, such as <BR>, <AREA>, <COL>, <COMMENT>, <!DOCTYPE>, <APPLET>, <OBJECT>, <SCRIPT> etc., are omitted during parsing.
- Hyperlink: These tags are extracted.

By utilizing the structure of HTML documents, the terms in the title, h1, h2 and h3 tags, meta-tag keywords and anchor text are most indicative of a document’s content. Thus our approach to construct thesaurus is based on those tags and the hyperlinks.

4 Related Works

We describe previous related researches to the problem of our work, finding information on the web, taking into account the structure and the additional meta-data provided by web authors. The solutions are presented in a chronological order and it is interesting to note that, through time, solutions rely more and more on the information provided by the authors of the web pages [17] (e.g. Met tag Keywords, anchor text, link structure, etc.). Since this paper suggests using the information embedded
in the anchors, the link structure, tags, etc., the studies described below were chosen in order to show past trends in using such information.

Minsky M.[22], treating the Web as a knowledge base is reminiscent of semantic networks, graphs whose nodes represent concepts or entities and whose arcs represent the relations among them.

Frei and Stieger [23] describe a way for using the semantic content of hypertext links for retrieval. They present an indexing algorithm that makes use of the document's text and link content. The content of the link is marked as being "referential" or "semantic". Semantic links are further marked for textual content and pointing/node relations. They have discussed how knowledge of the adjacency of nodes via hyperlinks can be used to help a user navigate or find the answer to a query.

McBryan [24] suggests that searches can be performed on titles, reference hypertext, or within components of URL name strings. In his system he indexes each URL with its anchor and title of page plus the title of the target page "to maximize the available contextual information".

Weiss et al. [25] combine link structure and textual information search techniques, by Performing TF-IDF on single documents and then on a higher level collection and so on, generating "content labels" for link clusters. The search is performed on the labels of the clusters and then - gradually - on a document level. Synonymous information is added to the labeling or, when needed, some of the more frequent terms in the cluster labeling are filtered. In this implementation documents are considered similar if they share similar links or are linked by similar documents.

Chakrabarti et al. and Brain D. [26,27] introduce a notion that the text around href links pointing to a page is descriptive of the content of this page.

5 Architecture

The subject of this paper is the construction and evaluation of an automatic tool for thesaurus construction from HTML pages. ThesWB [28] is a prototype system to construct thesaurus from web sites. ThesWB is a tool which, given a specific URL address of the web page that is broad and well-represented of the domain, will find out and return a list of terms that it considers the most authoritative for that topic. ThesWB is built to perform a local analysis of text, meta-tags and links to arrive at a "global consensus" of the best terms and relationships between them for specific domain. Our study suggests that, although our result lists are compiled wholly automatically, they fare relatively well reviewed by human.

5.1 System Overview

Figure 1 shows an overview of the construction process, which is made up of several steps. First the application accepts URL address as input and the HTML page is fetch.
From this HTML document plain text, meta-data like meta tags, titles, headings, list, etc and URL are extracted. The outcome is stored in a database and used in later stages. Each of these steps is described more fully in the following section.

For each HTML document, the tool parses the text and classifying it into a maximum of three different groups. The groups give a local analysis on text, met-data and links found in the HTML page. These groups are (see figure 1):

- **Metadata**: as present in HTML META-tags and important text, like document title and all HTML headings.
- **Plain text**: all other text.
- **URL list**: the list of the URL in the HTML page.

![Figure 1 Architecture of the ThesWB System](image-url)

Figure 1 Architecture of the ThesWB System
5.2 URL Extraction Model

Any URL found during the parsing is passed to processing if it points to a document within the current site, and stored for later analysis if it points to an external site, currently limited to depth of 3. This allows performing a depth-first visit of a site, collecting any categorization information it contains about itself and other sites.

Quite often the HTML pages of a site have a characteristic structure represented by links across pages. For instance there can be references to the main page, or links to the general services available in the site, like searching within the site, help or information pages. Finally, there can be advertisement banners in precise positions in each page. We want to avoid processing such pages. In order to identify these structural links, we used some heuristics about the structure of URL and an initial analysis of pages reachable from the starting page. These links placed in a stop list of URLs and discarded in the subsequent analysis of the site.

This task starts from a list of URLs, retrieving the documents referred by each of them and analyzing the structure of the document expressed in terms of its HTML tags.

5.3 Meta-Tags Analysis

The tags considered are currently <TITLE>, <Hn>, <UL>, <OL>, <A>, etc. and a meta-tags <META Classification Keywords>. Whenever one of these tags is found, a context phrase is recorded, which consists of the title within a pair <Hn> </Hn>, or the first portion of text after a <UL> or <LI> tag, or the phrase within a <A> tag.

In the analysis, tags related to layout or emphasis (<EM>, <B> <I>, <CENTER>, <FONT> etc.) are discarded here. Such tags can be effectively used in plain text analysis to extract thesaurus terms.

An HTML document can be viewed as a structure of different nodes. This document can be parsed into a tree according to its definition in SGML/DTD [29]. Tree structure provides an effective way to capture the common logical structure of the document[30].

The tree structure is constructed from parsing the tags and the corresponding content. The set of tags such as <H1>, <H2>, <P>, <UL>, etc. can be used to create the layout structure of the HTML page. For example, the structure shown in Figure 2 consists of Head node and Body node. The Head is further divided into TITLE, META NAME="KEYWORD "…, etc. The Body node has other low sub-levels like <H1>, <P>, <UL>, etc. The layout structure of an HTML document can be approximates by a subset of its tags.

In general the system will build the structure of the web pages as Follows;

1. Accept a URL, i.e. HTML document.
2. The system defines a target node using a start tag and an end tag, start tag is used to label each node.
3. Define a target low-level node within the node.
4. Repeat step 3 if more low-level nodes to be defined.
5. Repeat steps 2-4 if more nodes to be defined.

In order to extract the relationship between terms, ThesWB exploits the hierarchical structure and sets of extraction rules. For each node in the structure, it is associated with rules that extract that particular information or element.

During the parsing process, ThesWB apply text extraction rules for each type of tags. There are rules to extract and build the tree structure of the tags <Hn>. Such a tree has the tag <TITLE> as a root. The extraction rule for each of these tags is applied until all tags have been extracted.

If a tag is a list, the extraction rule for the list will be applied iteratively to extract the list elements for that node in the structure. After that, other extraction rules will be applied to extract individual items like terms, compound terms, lexico-syntactic patterns, looking for subsumption, etc. As a final step, we use the structure to group together the individual items to assemble the thesaurus.

Figure 3, which consists of a fragment of HTML page and the HTML source for this page, is an example. This HTML page is about thesaurus construction. The heading tags <Hn> were used to present the hierarchical structure. There are three ways to display the structure of thesaurus: alphabetical, systematic and graphic. Different methods may be present for each. Tree structure and Arrowgraph are two different methods for graphic display.
Displays

Alphabetical display: most common type

Systematic display
  - relates terms, often hierarchically
  - alphabetical index

Graphic displays: varied
  - Tree structure
  - Arrowgraph

Admin info etc
  - Administrative information:
    - term’s history in thesaurus eg
    - source of term

Figure 3  A Fragment of HTML page and the source of this fragment
The following fragment shows (Figure 4), how low level processing of the text can be used to build other relation. These rules are related to subsumption and lexicosyntactic.

- Hierarchical - Relationships subsumes Relationships Synonyms.
- Equivalence relation - using the symbol "::", an example is "Related terms" is UF Permanent Associations.

Figure 4 A Fragment of HTML page and the source of this fragment

6 Evaluation

HTML documents pose new research challenges because of the rich representation of a document and the connectivity between documents. Hyperlinks, content of linked documents, and meta-data (such as meta-tags, titles, etc.) about related web sites, all provide rich sources for thesaurus construction. Given this, the question of how to effectively use such information becomes important. More specifically, how useful or reliable are hyperlinks when they are used in building hierarchical relations from web pages or web sites?

From our experiment, we tentatively conclude that thesaurus construction from HTML documents would involve multiple phases or ranking. At one level, metadata like <TITLE> or other HTML tags should be given high consideration. At another
level, more typical Information Retrieval (IR) techniques should be employed in order to identify potentially useful terms that do not exist in the <TITLE> or other targeted tags. Then, other technique need to be developed to assess which terms from the two sets of candidates are best for thesaurus.

Other techniques need to be developed to predict which type of relationships can be find between these candidate thesaurus terms.

We described a system for automatically building a thesaurus from HTML pages and present performance results based on different combinations of term sets. ThesWB can serve as basis for a generalized framework for Automated thesaurus construction tool.

In our experiment we did not consider the hierarchical relations between the HTML pages in finding the relationships between terms.

We find that HTML meta-tags are good source for thesaurus construction, but are not in wide use despite their role in search engine.

The results of our experiments with ThesWB tool are quite encouraging. By exploiting different information, the tool achieves an effective and accurate result. The tool may evolve by incorporating further linguistic knowledge and techniques for manipulating stop list of URL that are general services, searching, help or information and advertisement banners.

7 Conclusions and Future Work

In this paper we discussed several issues related to automated thesaurus construction from web sites. We analyzed the nature of the web content and metadata in relation to requirements for thesaurus construction.

We addressed the problem how to effectively use the rich information, which is typically available in HTML to build thesaurus. We specified relevant hints about the hyperlinks and metadata about web sites, which we hope, will aid in the task of thesaurus construction and aim at exploiting the richness of HTML. We examined these hints with different collection of web pages. Our major findings include:

Using words in web pages alone often yielded sub-optimal performance, compared to exploiting additional sources of information.

Careful examination of hyperlink is crucial for the task of thesaurus construction.

Metadata about web pages or web sites can be extremely useful for improving the accuracy of the task of thesaurus construction. This suggests the importance of examining the availability of metadata in real world applications, and using Information Extraction techniques for automated acquisition of metadata.

ThesWB has the power for extracting the different sets of candidate terms and discovering the relationships between these terms. The results achieved with the current prototype are quite encouraging. In most cases, ThesWB achieves an effective and accurate result.

ThesWB, on the other hand, has the power for discovering relational regularities that cannot be explicitly identified using the other algorithms.
While the scope of the experimental results presented is necessarily confined to our HTML sets, we hope that our analysis will help future research into hypertext by providing some ideas about various types of information that may be present in a hypertext corpus and how one should take advantage of each.

We find that titles and meta-tag contain useful information for thesaurus construction task and result in performance that is good. In particular, titles words, anchor and meta-tags perform better than using all the words on the web page.

References