On Intra-page and Inter-page Semantic Analysis of Web Pages

Jun Wang
Fujitsu R&D Center Co., Ltd.
Room B1003, Eagle Run Plaza
No.26 XiaoYun Road
Beijing, CHINA. 100016
jwang@frdc.fujitsu.com

Jicheng Wang, Gangshan Wu
Institute of Multimedia Computing
Nanjing Univ.
Nanjing, CHINA. 210093
{wjc,gswu}@graphics.nju.edu.cn

Hiroshi Tsuda
Fujitsu Laboratories, Ltd.
4-1-1 Kami-kodanaka,
Nakahara-Kawasaki,
Kanagawa 211-8588, Japan
htsuda@jp.fujitsu.com

Abstract

To make real Web information more machine processable, this paper presents a new approach to intra-page and inter-page semantic analysis of Web pages. Our approach consists of Web pages structure analysis and semantic clustering for intra-page semantic analysis, and machine learning based link semantic analysis for inter-page analysis. Based on the automatic repetitive patterns discovery in structure level and clustering in semantic level, we explore the intra-page semantic structure of Web pages and extend the processing unit from the whole page to a finer granularity, i.e., semantic information blocks within pages. After observing the various hyperlinks, we synthesize the Web inter-page semantic and define an information organizing oriented hyperlink semantic category. Considering the presentation of the hyperlink carrier and intra-page semantic structure, we propose corresponding feature selection and quantification methods, and then exploit the C4.5 decision-tree method to classify hyperlink semantic type and analyze the inter-page semantic structure. In our experiments, the results suggest that our approch is feasible for machine processing.

1 Introduction

In the recent years, the content and structure of Web pages become much more complex for business purposes with easy access and user-friendly. However, it brings a big challenge for automatic Web information processing system since most of the Web's content today is designed for humans to read, not for machines to manipulate meaningfully (Tim Berners-Lee 2001), and many problems occur in current Web machine processing due to lack of semantic analysis. According to these problems in practical application, two opposite kinds of semantic information of Web pages are analyzed in this paper.

Web pages often represent a collection of different topics and functionalities that are loosely knit together to form a single entity. People can easily identify the information areas with different meaning and function in a Web page, but it's very difficult for automatic processing systems since HTML is designed for presentation instead of for content description. Now most of current Web IR (Information Retrieval) and DM (Data Mining) systems regard a Web page as an atomic unit and do not give due consideration to the intra-page semantic structure of a Web page, and in this case many problems occur. (Xiaoli Li 2002) illustrate the precision decline occurred when the query words scatter at different semantic information areas in a Web page. The use of templates has grown with the recent developments in Web site engineering, because the template is very valuable from a user's point of view since it provides context for browsing, however, (Ziv Bar-Yossef 2002 and Soumen Chakrabarti 2001) reveal that the template of Web pages skew ranking, IR and DM algorithms and consequently, reduce precision, and (Shian-Hua Lin 2002) describe the intra-page redundancy caused by the common template. Obviously, it is important to analyze the intra-page semantic of Web page for improvement of current Web application.

On the other hand, Web is a hypertext environment, and hyperlink analysis has been widely used in the Web information processing system, such as Google (Sergey Brin 1998) and HITS (Jon M. Kleinberg 1999). However, these systems just regard the hyperlink as the simple reference and endorsement, actually from the point view of the author of the Web page, the hyperlink tells the diverse inter-page semantic information of Web pages, such as reference, related topic and document organizing structure, etc. Understanding the hyperlink semantic will provide better solution for Web content management and integration. For example, in actual Web data, a logical document (such as slides and BBS articles in one thread) discussing one topic is often organized into a set of pages connected via links provided by the page author as document structure links. In such a situation, a data unit for Web data integration and processing should not be a page but should be a connected sub-graph corresponding to one logical document. If we can identify the links pointing to the related contents, we will acquire better performance on focus crawling and computation of link analysis.

To solve the problems mentioned above, we explore intra-page semantic of the Web page and partition the whole Web page to a finer granularity based on the repetitive pattern discovery and clustering. Furthermore, we synthesize hyperlink semantic of Web pages, and define an information organizing oriented hyperlink semantic category. We propose the corresponding feature selection and quantification methods according to presentation and context of the hyperlink carrier. We employ C4.5 decision-tree to recognize the hyperlink semantic automatically. We evaluate our ideas by experiment, and the result proves its feasibility.

The reminder of this paper is organized as follows. In the next section, we briefly discuss the related work. Section 3 illustrates our analysis method of intra-page semantic structure and inter-page semantic structure in detail. Experiment evaluation and analysis are described in section 4. Last section presents our conclusion and future work.

2 Related Work

The content contained in one Web page no longer remains semantically cohesive with the increasing of complexity of the page's content and structure, and people begin to research how to segment the Web page according to its content. (Xiaoli Li 2002 and Ziv Bar-Yossef 2002) propose their respective approach to segment a Web page into fine-grained areas, but they all use very naive heuristic methods. (Shian-Hua Lin 2002)'s method of detecting informative content block in a Web page is lack of universality since it can only cope with tabular pages containing <Table> tag.

From the point view of Web information organizing, there is not much complete research on the hyperlink semantic analysis yet. (Keishi Tajima 1999, Yoshiaki Mizuuchiy 1999 and Wen-Syan Li 2001) propose the concept of logical information unit consisting of multiple physical pages as connected sub-graph corresponding to one logical Web document, and their hyperlink type analysis are only facing keyword-based Web IR and just considering routing function between the pages in the unit.

Compared with existing work, our intra-page semantic analysis is mainly based on automatic induction and applicable to all kinds of Web pages. Our inter-page semantic analysis is more complete and can adapt to more applications.

3 Intra-page and Inter-page Semantic Analysis of Web Pages

This section illustrates our analysis methods in detail.

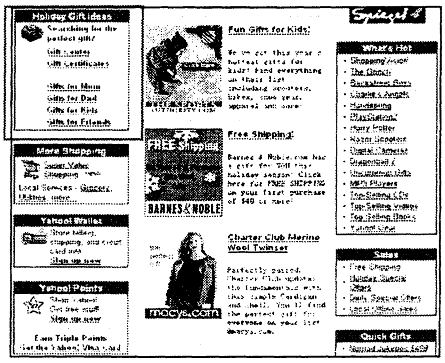
3.1 Intra-page Semantic analysis

Typically, the contents of a Web page encompass a number of related or unrelated topics. Each topic usually occupies a separate region in the page and each region is coherent topic area according to its content, and it is usually also a visual block from the display point of view. We define this kind of region as SEIB (semantic information block) and it is the foundation of Web intra-page semantic structure. However, the above definition of SEIB is only from the intuition of human without strict rules, in order to detect the SEIB in a Web page by machine processing, we need consider the approach in the syntax level, which materialize the intuitive definition of the SEIB into an actual algorithm.

3.1.1 illustrates the syntactic analysis that detects STIBs (structure information blocks) from the HTML tag token stream and DOM tree. 3.1.2 explains a method to derive SEIBs from STIBs using semantic clustering and to add semantic label MTB (Main Text Block) and RLB (Related Text Block) to the SEIBs.

3.1.1 STIB Detection Based on Repetitive Pattern Discovery

After observing and analyzing a lot of Web pages, we get some clues and guidance of Web page structure. We often find many different repetitive patterns within a Web page, and the instances (occurrence of the pattern) corresponding to each repetitive pattern are arranged orderly and compactly with the similar HTML syntax, same presentation style, similar meaning or function. Each occurrence of a pattern may represent an information item, and all of the information items form an information block, we call this kind of block as STIB (Structure Information Block). Therefore we can exploit the repetitive pattern to decompose a Web page into fine-grained areas for the intensive processing. The region in the top left corner of the page in the Figure 1 is a typical STIB contains the repetitive pattern. In the figure 2, we show the repetitive pattern and its corresponding instances of this STIB in the figure 1 by HTML tags.



Holiday Gift Ideas

Searching for the perfect gift?

Gift Center

Gift Certificates

Gifts for Mom

Gifts for Dad

Gifts for Kids

Gifts for Friends

More Shopping

Yahoo! Shopping

Local Services - Grocers, Utilities, more

Yahoo! Wallet

Store billing, shipping, and credit card info

Sign up now

Yahoo! Points

Shop Yahoo!

Get free stuff

Sign up now

Earn Triple Points

Get the Yahoo! Visa card

Fun Gifts for Kids!

We've got this year's hottest gifts for kids! Find everything on their list including scooters, bikes, toys, apparel and more

Free Shipping!

Barnes & Noble.com has a gift for all this holiday season! Click here for FREE SHIPPING on your first purchase of $40 or more!

Charter Club Merino Wool Twinset

Perfectly paired. Charter Club updates the fundamentals with this simple Cardigan and shell. You'll find the perfect gift for everyone on your list @macys.com

What's Hot

- Shopping Vision
- The Grinch
- Backstreet Boys
- Charlie's Angels
- Handspring
- PlayStation2
- Harry Potter
- Razor Scooters
- Digital Cameras
- Dreamcast
- Unusual Gifts
- MP3 Players
- Top-Selling CDs
- Top-Selling Videos
- Top-Selling Books
- Yahoo! Gear

Sales

- Free Shipping
- Holiday Special Offers
- Daily Special Offers
- Local Offline Sales

Quick Gifts

- [illegible]

Figure 1. STIB and information Items in the Web page with repetitive patterns

Pattern:
<TR><TD><IMG><TD><TEXT><FONT><B><A><TEXT>
Instance Set:

- <TR><TD><IMG><TD><&nbsp;><FONT><B><A><Gift Center>
- <TR><TD><IMG><TD><&nbsp;><FONT><B><A><Gift Certificater>

. . .

- <TR><TD><IMG><TD><&nbsp;><FONT><B><A><Gifts for Kids>
- <TR><TD><IMG><TD><&nbsp;><FONT><B><A><Gifts for Friends>

Figure 2. Repetitive Pattern and Corresponding Instances

Since it is more convenient to discover repetitive patterns by token stream, we generate tag token stream to represent the Web page by pre-order traversing DOM tree. We also create a mapping table between the tag token in the stream and the node in the DOM tree. We construct a Suffix Tree with the tag token stream, then the automatic pattern discovery is adopted to induce all repetitive patterns from the Suffix Tree, filtering out improper ones (remove about 90% of the original ones so as to improve the efficiency of following processing), and generate the set of candidate patterns and corresponding instances. We design "non-overlap", "left diverse" and "compact" rules to refine and filter the repetitive patterns and instances. The details are omitted due to limitation of the paper length.

We can map each pattern and its corresponding instances back to the DOM Tree. Because instances of a pattern always have a same parent node in HTML DOM Tree, for each pattern we find a sub-tree containing all its instances in DOM tree, and the root of smallest sub-tree representing a region corresponding to the STIB in the Web page. The instances of each pattern are named information items. When all the STIBs are generated in a Web page, we organize them into a hierarchical structure according to their corresponding positions in the DOM Tree and then construct a STIB tree, so a Web page can be decomposed into STIBs. The STIB can be nested since an information item contained in a STIB also can be a sub-STIB itself.

(Chia-Hui Chang 2001) also use the repetitive pattern discovery for the information extraction. However, their methods can only handle pages with just one pattern, such as Web pages of tabular data. Our method can work on the more complex Web page, and find multi repetitive patterns and the hierarchy relations between them.

3.1.2 SEIB Generation Based on Clustering and Labelling

Although the STIB tree already gives us an initial partition of the page, but STIBs in different levels represent information blocks of different granularities. Therefore, before semantic clustering we have to select information blocks with appropriate granularity for clustering. We call these blocks with appropriate granularity as Basic Information Blocks.

We design heuristic rules and traverse STIB Tree in pre-order to acquire appropriate blocks for clustering. For each STIB traversed and its information items, we take their size, ratio of anchor text length to all text length and information of neighboring nodes as measures to determine whether it is a block we need. All blocks acquired are divided into two types, i.e., text or link, according to the ratio of anchor text length to all text length. The detail of the heuristic rules is omitted due to limitation of the paper length.

Semantic clustering method is used to merge these appropriate Basic Information Blocks into SEIBs. In order to compute the semantic similarity between two Basic Information Blocks, each block is represented in the form of "bag of words", i.e. a set of <word, frequency>, and a stop-list is also used to remove meaningless words. The computation of semantic similarity is based on the conventional VSM (Vector Space Model) method. Moreover, because two adjacent blocks are more likely to be similar, the similarity between two adjacent blocks is doubled.

Clustering is performed on text blocks and link blocks respectively. A common method known as "partition clustering" is used to generate SEIBs, and the detail is as following.

- Sort blocks in descending order according to the size of blocks.
- Add the biggest block to current cluster.
- For each block in current cluster, compute the similarity to other blocks never clustered. Moreover, the similarity between two adjacent blocks is doubled.
- If similarity is above a threshold, the block never clustered can be added to current cluster. Repeat the above loop until each block is processed. Now, all blocks in current cluster are grouped into a SEIB.
- Select the biggest block from all blocks left as the seed of the new cluster. Repeat the above loop. If all of the Basic Information Blocks have been clustered into a certain SEIB, the procedure ends.

It should be noted that the clustering among blocks in above procedure is transmissible, i.e., if block A is similar to block B, and block B is similar to block C, then block A, B and C will be clustered whether A and C is similar or not.

For Web pages whose main contents are text paragraphs, such as a news content page, the most informative SEIB is the MTB (Main Text Block), and sometimes, there also exists a RLB (Related Link Block), which contains a group of hyperlinks semantically related to the content of MTB. MTB and RLB are critical part for many Web applications. The key steps of labeling method are illustrated as following.

- Check the ratio of anchor text length to all text length in the Web page. If it is below a certain threshold, then the Web page is most likely a page of with MTB. Otherwise, quit.
- Find the biggest text type SEIB in the Web page. If its size is above a threshold, it can be taken as the MTB. Otherwise, semantic clustering method is applied on text SEIBs to generate MTB.
- Based on the MTB, we select one link type SEIB which is most similar to it. If their similarity is above a certain threshold, this link type SEIB is taken as the RLB. Otherwise, no RLB exists.

It should be noted that all thresholds we used are empirical and can be adjusted according to the feedback of the result.

3.2 Inter-page Semantic analysis

Web page authors often use hyperlinks to tell various semantic relations between source and target pages. When author create a hyperlink, he often want the reader can not only understand meaning of anchor text or other hyperlink carrier, but also predict the content of the destination page from the context and presentation of hyperlink carrier. So the hyperlink carriers are usually presented with some conventionality, which is not dependent on the specific application field. Therefore it is possible to automatically extract hyperlink semantic information from the context and presentation of hyperlink carrier.

Automatic extraction of hyperlink semantic can be thought as hyperlink semantic classification. The hyperlink semantic, which has diverse presentation and is hard to be identified by some simple heuristic rules depict different functions and intent of page author to organize the information. Therefore, we use machine-learning method to automatically generate the judgment rule. There are many classification approaches, such as the decision-trees, the Bayesian network and the Neural Network. But for our research, the hyperlink semantic category and feature selection are most important aspects, and we chose well-known C4.5 decision tree machine learning method (Quinlan, J. R. 1992).

3.2.1 Hyperlink Semantic Category

Considering the requirements of Web information organization and the current technical feasibility, we propose a set of hyperlink semantic categories. These categories can be divided into two types, one is functional type such as advertisement link, copyright and privacy information link; another type represents the author's content organizing intent, such as document Structure Link, domain link, title link, related link and reference link. From point view of research, the functional link is relatively easy to identify, and the latter is more interesting. Our research will focus on content organizing link type. Following list is a brief description of each type.

Domain link: The hyperlink points to a page related with a certain domain in broad sense concept. The domain can imply a specific topic, such as "Football WorldCup", also can express a broad topic consisting of several correlative topics, for example, sports containing football, basketball, tennis, etc. The anchor text is the highly condense of the content of the domain.

Title link: The main contents of link destination describe certain information concretely, and anchor text is the summary of such information.

Related link: Contents of link destination and contents of link source are semantically related, but the author doesn't regard this hyperlink as main part of source page.

Refer link: Link destination is a supplement explanation of a term in main part of link source.

Document structure link: Link source and link destination belong to one logical document, and always hold specific sequence. This kind of link just act as organizer of a document, link anchor text has no meaning to document content.

Analysis of such hyperlink semantic information may do help to Web information intelligent processing, such as: Web information filtering, integration and computing authority of pages.

3.2.2 Feature Selection and Quantification

We analyze the hyperlink semantic only based on link source page content since this approach will be more universal for different applications. According to the hyperlink semantic category and presentation conventionality of link carrier, we select following features to quantify for training and classification.

Some of selected features are direct and inherent.

Hyperlink carrier type: Hyperlink carrier can be TEXT or IMAGE, ANIMATION, most of hyperlink carriers are in TEXT format.

Length of anchor text: It is the length of the current anchor text string.

Font of anchor text: It is description of the font in the anchor text, such as <H1>, <H2>, <H3>, <H4>, <H5>, <H6>, <strong>, <b>, <em>, <i>, <u>, <font>, etc.

Size of the hyperlink carrier media: If hyperlink carrier is not the text, then carrier attribute is the size of display area of the media.

Hyperlink destination file type: HTML and not HTML type. We can identify this feature by suffix of the destination file name.

Keyword in anchor text: Whether there are some special keywords corresponding to specific semantic category in anchor text. Some special purpose link usually has fixed pattern, we classify these keywords into 4 categories: DS (document structure, such as "previous", "next", "more"); RL (related link, such as "Related", "So also"); CI (common information, such as "term of use", "company info", "contact"); NO (No above specific keywords).

Link destination display style: Whether the destination page is displayed in new window or current window, this feature can be obtained from 'A' tag's target attribute.

Other features are extracted and inferred by context dependant methods based on the DOM, STIB and SEIB.

Hyperlink source page type: Hyperlink source page can be classified into two types: content and list, content pages are those pages containing few hyperlinks and list pages are those pages containing lots of hyperlinks, this can be identified by the ratio of anchor text length to all text length in Web pages.

Hyperlink density in local area: the local area means the SEIB containing current hyperlink, and the density is ratio of anchor text length to all text length in this block.

Hyperlink list flag: This feature can tell us if current link is in a link list. We can detect whether current hyperlink is contained in an information item and the other information items of the same STIB also contain the hyperlinks in the same level.

Number of hyperlink in the list: if current hyperlink is in a list, the feature is the number of the links in such list.

Above features can be quantified to corresponding number or enumeration data type according to the feature definition.

3.2.3 Training and Optimization

In order to build training data set, we develop an aided GUI tool to mark the type of hyperlink semantic in the Web pages and save the marked result in XML files. We input this manually labeled result as training data to C4.5 classification system and generate the rules automatically.

An important problem in classification is the selection of the best features to use. It is usually happens that some features are more important than others, and some features may be irrelevant or redundant or even only carry the noise. If we can remove the unnecessary features, therefore, we can improve the classifier.

We study an approach named CSS (Combined Stepwise Selection) proposed by (Steven Salzberg 1992). This method uses the ideas of both SBS (Stepwise Backward Selection) and SFS (Stepwise Forward Selection) to search through the feature space. We exploit a similar method in our experiments.

Firstly, we decrease the original features one by one to the BFS (Base Feature Set) that can produce highest accuracy by a method similar to SBS. We introduce a threshold θ, which indicates how much we are willing to allow the accuracy to decrease each time a feature is eliminated. The method works as follows: remove one feature at a time from the set, and evaluate the classifier with reduced feature set by 3-fold cross validation. If the classifier works equally well without the removed feature, then we will assume that it may be ignored. If d is the number of features in the original set, we run d times experiments and delete the feature that causes the smallest decrease in accuracy, as long as that decease is less than θ.

Secondly, the BFP (Best Feature Pair) is selected from the BFS generated in the first step. We start with a small application of brute-force search, in order not to miss interactions between pairs of features. This idea behind this heuristic is that many interactions among features are probably pairwise. For instance, "Hyperlink carrier type" and "size of the hyperlink carrier" emerge at the same time. Therefore, we runs experiments with all subsets of 2 from BFS, and select the pair with best accuracy as the BFP.

Furthermore, we try adding each of the remaining features to BFP. The one that give the best improvement in accuracy is added to set, as long as the improvement is greater than θ. We then consider the remaining features, adding each one individually to the set of three features, and so on until the improvement is less than the threshold θ.

Finally we can use the above new feature set to generate the classification rules. In next section we describe the result of the classification experiment.

After the automatic classification, based on intra-semantic analysis we can reinforce our method with some heuristic rules. For instance, if an area in a Web page is identified as RLB in the intra-page semantic analysis, we can decide the hyperlink list in this block are related links. In addition, if there are hyperlinks in the same hierarchy of information items of a STIB, these hyperlinks should the same type, when the hyperlinks in the list of information items are classified into different classes, they should be identified as the class with highest proportion.

4 Experiment and Evaluation

We collect 100 Web pages as the experiment test set. These pages are collected from different kinds of typical Web sites include the company sites, government agency sites, e-commerce sites, the university sites and portal sites. About 30% of Web pages are homepage of the Web sites and others are internal pages in these Web sites.

4.1 Experiment of Intra-page Semantic Analysis

The intra-page semantic analysis method is applied to the above test set of 100 pages, and rate the result by manual work. We can score result as "good, normal and bad" three levels. The "good' grade means that the all extracted units are consistent with the annotator's decision; on the contrary, the "bad" grade means that most of the units are not recognized correctly or the main content units are not extracted well; and the "normal" grade shows only some small units are not identified correctly or main content units have only some trivial bugs. The test set is annotated by three persons, if these three annotators are unanimous we can easily get score; if three persons have two different opinions majority rule is followed; if three different opinions emerge we score "normal". Table 1 is the evaluation result.

Evaluation Unit		Good	Normal	Bad
Syntax & Structure	STIB	55%	35%	10%
	Info Item in STIB	58%	36%	6%
Semantic	SEIB	51%	36%	13%

Table 1 Evaluation result of intra-page semantic analysis

Type	Domain	Title	Related	Ref.	DS.
Number	3573	1386	1877	92	154
Precision	82.3%	75.8%	70.8%	84.2%	89.1%
Recall	86%	58.3%	63.4%	80.1%	71.4%

Table 2 Evaluation result of inter-page semantic analysis

Obviously, the effectiveness of STIB and SEIB generation is acceptable since the result rated above normal is over 80%. However, direct evaluation is subjective inevitably to some extent.

Furthermore, we want to consider whether intra-page semantic analysis can improve the performance of subsequent Web applications, such as Web IR. We collect about 3000 news pages from Yahoo News site. We create full text index for original pages and MTBs of these pages respectively, and use five query words to retrieve. As the average result, the precision increases substantially from 34% to 46.4% after SEIB extraction especially for MTB, while recall doesn't decrease much (from 100% to 95.8%), and many researchers believe that high precision is important even at expense of recall. As we have expected, intra-page semantic analysis, which acts as a pre-process, are able to efficiently improve the effectiveness of Web IR. (Xiaoli Li 2002, Shian-Hua Lin, Jan-Ming Ho 2002 and Ziv Bar-Yossef 2002)'s experiments also illustrate the extraction of information content blocks can improve the Web IR and DM application.

4.2 Experiment of Inter-page Semantic Analysis

We extract more than 7000 hyperlinks from the above test set with 100 pages and mark the type of each hyperlink by manual work. Because our research will focus on content organizing link type, we only deal with domain link, title link, reference link (Ref.), related link, and document structure link (DS.) in the experiment. We use the 3-fold cross validation and C4.5 decision tree to evaluate the hyperlink semantic classification. The average result is presented in Table 2.

In current test set, the number of sample pages of some hyperlink types, such as related link and document structure link, is too small, maybe we need extend and balance the training set to improve the quality of the marked sample in the future.

Among the five hyperlink types evaluated, the result of Domain, Reference and Document structure are much better, because the presentation and context features of these types are much more distinctive. For example, the document structure hyperlinks usually appear with some special anchor strings, such as "previous", "next", etc., so they can be well distinguished. The features of title and related hyperlink are very similar except that the related hyperlink is often close with some specific keywords, such as "See also" and "Related". Although we consider the intra-page semantic information block boundary, but don't detect keywords in the current information block, the results of these two hyperlink types are not so good.

In general, from above experiments we think it is feasible to infer the inter-page semantic structure of Web pages based on the context and presentation of hyperlinks. Of course there is still much space for the improvement, we can adjust the definition of the hyperlink semantic category, and then optimize the feature selection. Moreover, we can enhance the integration with the intra-page semantic analysis.

The inter-page semantic analysis is a fairly new exploring research, and we don't apply it on subsequent application yet.

5 Conclusion

In this paper, we explore intra-page semantic structure of the Web pages and propose a method with the integration of structural analysis and semantic clustering which segments a Web page into semantic information block as a fine-grained granularity. According to requirements of Web information organizing, we define a preliminary specification of Web hyperlink semantic category and propose corresponding feature selection and quantification methods based on the context and presentation of the hyperlink source page, and then exploit a C4.5 decision-tree method to classify hyperlink semantic type and analyze the inter-page semantic structure. From the experiment shown in Section 4, we find that the result proves our method is feasible for machine processing. In ongoing work, we want to make full use of such semantic information by applying them on Web applications. We also plan to explore hyperlink semantic structure with topology-based method and apply it on the practical application, such as Web site logical structure analysis.

References

Tim Berners-Lee, James Hendler and Ora Lassila 2001. The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific American, May 17, 2001.

Xiaoli Li, Bing Liu, Tong-Heng Phang, Minqing Hu 2002. Using Micro Information Units for Internet Search. CIKM'02, November 4-9, 2002, McLean, Virginia, USA.

Ziv Bar-Yossef and Sridhar Rajagopalan 2002. Template Detection via Data Mining and its Applications. In Proceedings of the WWW2002, May 7-11, 2002, Honolulu, Hawaii, USA.

Soumen Chakrabarti, Mukul Joshi, Vivek Tawde 2001. Enhanced Topic Distillation using Text, Markup Tags, and Hyperlinks. SIGIR'01, September 9-12, 2001, New Orleans, Louisiana, USA.

Shian-Hua Lin, Jan-Ming Ho 2002. Discovering Informative Content Blocks from Web Documents. SIGKDD '02, July 23-26, 2002, Edmonton, Alberta, Canada.

Sergey Brin and Lawrence Page 1998. The anatomy of a large-scale hypertextual Web search engine. In Proceedings of the WWW7, 1998.

Jon M. Kleinberg 1999. Authoritative Sources in a Hyperlinked Environment. Journal of the ACM, 46(5):604-632, 1999.

Keishi Tajima, Kenji Hatano, Takeshi Matsukura, Ryouichi Sano, Katsumi Tanaka 1999. Discovery and Retrieval of Logical Information Units in Web. Proc. of Workshop on Organizaing Wep Space (WOWS'99) in conjunction with ACM DL'99 Berkeley CA USA, Aug. 1999.

Yoshiaki Mizuuchiy, Keishi Tajima 1999. Finding Context Paths for Web Pages. Hypertext 99 Darmstadt Germany.

Wen-Syan Li, K. Selc,uk Candan, Quoc Vu, Divyakant Agrawal 2001. Retrieving and Organizing Web Pages by "Information Unit". In Proceedings of the WWW10, May 1-5, 2001, Hong Kong.

Chia-Hui Chang, Shao-Chen Lui 2001. IEPAD: Information Extraction Based on Pattern Discovery. In Proceedings of the WWW10, May 1-5, 2001, Hong Kong.

Quinlan, J. R. 1992. C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, CA, 1992.

Steven Salzberg 1992. Improving Classification Methods via Feature Selection. Technical Report TR JHU-92/12, Department of Computer Science, Johns Hopkins University, Baltimore, MD21218.