Higher-Order Rank Analysis for Web Structure

Ikumi Horie
Tsuda College
2-1-1,Tsuda-machi,
Kodaira,Tokyo, Japan
horie@tsuda.ac.jp

Kazunori Yamaguchi
University of Tokyo
3-8-1, Komaba, Meguro,
Tokyo, Japan
yamaguch@graco.c.u-tokyo.ac.jp

Kenji Kashiwabara
University of Tokyo
3-8-1, Komaba, Meguro,
Tokyo, Japan
kashiwa@graco.c.u-tokyo.ac.jp

ABSTRACT

In this paper, we propose a method for the structural analysis of Web sites.

The Web has become one of the most widely used media for electronic information because of its great flexibility. However, this flexibility has led to complicated structures. A structure that differs from the typical structures in a Web site might confuse readers, thus reducing the effectiveness of the site. A method for detecting unusual structures would be useful for identifying such structures so that their impact can be studied and ways to improve Web site effectiveness developed.

We viewed the Web as a directed graph, and introduced a higher-order rank based on the non-well-founded set theory. We then developed higher-order rank analysis for detecting irregularities, defined as structures which differ from the typical structure of a target site. To test the effectiveness of our method, we applied it to several Web sites in actual use, and succeeded in identifying irregular structures in the sites.

Categories and Subject Descriptors

H.5.4 [**INFORMATION INTERFACES AND PRESENTATION**]: Hypertext/Hypermedia

General Terms

Algorithms, Experimentation, Theory

Keywords

Link analysis, non-well-founded set theory, AFA, Web graph, structural analysis

1. INTRODUCTION

With the accelerating rates of advance in both science and engineering, more information has become available to us electronically. Digitized information is particularly useful, compared to information in legacy formats, because it is easy to handle. However, its full value can best be realized when the information can be accessed through flexible media. The Web has become one of the most widely used media for electronic information because of its great flexibility. However, the ease of Web use is not always experienced by readers who lack any a priori knowledge regarding the structure of Web pages. By visiting pages over and over again, readers gradually come to understand how pages are typically related. If part of a Web site has an irregular structure, though, inexperienced readers can easily lose their way. Without thoughtful site design, the flexibility of Web pages can just confuse readers. Thus, a simple and uniform Web structure is crucial.

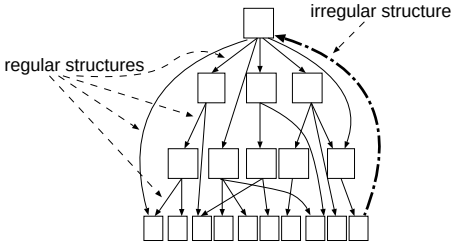

Figure 1: Regular and irregular structures

Fig. 1 shows examples of regular and irregular structures. The Web page structure in Fig. 1 is hierarchical for the most part, but the dotted link from the lower-right page to the top page is not found in a typical structure. Readers following such a link would find themselves on a page in an unexpected part of the Web site. In Fig. 1, regular structure is a hierarchy, and the structure which differs from the typical structure is irregular.

The best structure depends on the nature of the subject being considered, though, and it is difficult for site authors to determine the optimal structure in advance. As Web pages increase in number, their structure tends to become more and more complex. In such a situation it is hard for authors to keep Web structures uniform. Authoring tools can keep the Web structure uniform if the structure is decided upon before the site creation, but this is often not the case. As the Web grows, the structure has to diverge from the standard structure if the nature of new subjects calls for such divergence. Such a structure often becomes too complex. A method for detecting structures that differ from the typical structure of target Web pages is thus required.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
HT'05, September 6–9, 2005, Salzburg, Austria.
Copyright 2005 ACM 1-59593-168-6/05/0009 ...$5.00.

In this paper, we propose a method for detecting irregular structures. The method would be useful for identifying such structures so that their effect can be studied and ways to improve them developed. We focus on the analysis of the link structure in Web pages, which is independent of the page contents (e.g., text, images, and so on). To represent such a pure structure, we employ a directed graph consisting of nodes which represent Web pages, and arcs representing the links between pages. This is a suitable way to represent a structure without any spurious information. Then, we introduce a higher-order rank for identifying structures on the graph. The higher-order rank is the information as to how each node can reach nodes having no outgoing arcs.

We previously proposed a method for detecting irregular arcs [1, 2]. Even though that method succeeded in detecting several errors in a site, it has three weaknesses. 1) It is sensitive to the typical site structure, and we had to choose a proper scheme. 2) It cannot find any node that has multiple link errors. 3) It takes a long time because all of the arcs are subject to an irregularity check. Our new method overcomes these weaknesses. It needs no prior knowledge regarding the structure of a Web site to analyze the site. It can detect a node even if the node has more than one irregular arc. In addition, it needs less time because only nodes are subject to the irregularity check.

We used six sample Web sites for analysis: Google Japan, Yahoo! JAPAN, the Ministry of Foreign Affairs in Japan site, the Official Web site of the Prime Minister of Japan and His Cabinet, the World Lecture Hall, and the Sixteenth ACM Conference on Hypertext and Hypermedia site. The results of this analysis show the usefulness of our method.

Broder et al. studied the structural analysis of the Web as a graph [3], but their target was the entire Internet. They did not discuss the structures of Web pages at a single site. Botafogo and Shneiderman have used structural analysis of hypertexts to solve the "lost in hyperspace" problem [4, 5]. They analyzed and clustered hypertexts by using numerical measures and metrics which indicate the properties of a structure. McEneaney also used numerical measures and metrics to visualize and assess navigation [6, 7]. He analyzed the user movement in hypertext, and showed the relationship between resulting measures and performance in hypertext search tasks. Chen regarded three fundamental hypertext relationships — hypertext linkage, content similarity, and browsing patterns — as measures in generalized similarity analysis (GSA) [8]. He structured and visualized the Web by using these measures. The above studies, though, were focused on the overall structure of the Web. In contrast, we focus on Web pages at a single site and have analyzed the Web structure in detail.

Wang and Liu proposed a method to discover the typical structures of a collection of online documents, where the user specifies the minimum frequency of a typical structure [9]. They focused on a typical structure, though, while we focus on irregular links that differ from other structures.

Mukherjea and Foley constructed an overview diagram of the World Wide Web [10]. However, their diagram has no function to identify irregular links. Our method detects irregular nodes automatically.

Whitehead [11] compared the structure of hypertext through the containment model. He analyzed the data model, rather than the structure of the hypertext. In this paper, we analyze the structure of the hypertext.

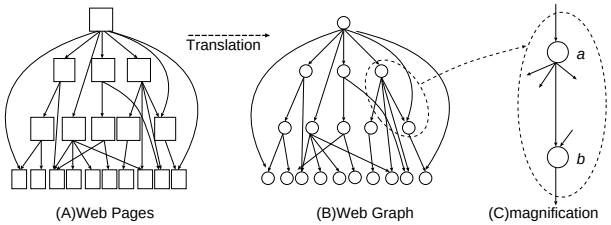

Figure 2: Web pages and Web graph

Amitay et al. proposed the 'connectivity sonar', which detects site functionality using structural patterns [12]. They automatically categorized sites into eight distinct functional classes. However, their study was not intended to detect irregular structures in a Web site.

This paper is organized as follows. In Section 2, we briefly explain our method and the basic notion. In Section 3, we report experimental results from the application of our method to actual Web sites, and evaluate the method based on these results. We then conclude in Section 4.

2. HIGHER-ORDER RANK ANALYSIS OF WEB STRUCTURES

In this section, we present the basic concepts used in our analysis. We translate Web pages into directed graphs (explained in Section 2.1.) For the analysis, we use a summary of the structure surrounding each node which is called a higher-order rank (defined in Section 2.2.) The higher-order rank is based on a non-standard set theory called AFA (introduced in Section 2.3.) A method to detect irregular nodes through the higher-order rank (explained in Section 2.4) is compared with our previous method (explained in Section 2.5,) and this comparison shows that our new method is superior (illustrated in Section 2.6.)

2.1 Graph Structure in the Web

In this paper, we are interested in only the structure of pages, so we use only the link information and ignore the other information on pages (text, symbols, images, and so on) [1]. In this respect, the Web can be regarded as a directed graph consisting of nodes, which represent Web pages, and arcs, which represent links between pages as shown in Fig. 2. In our analysis, Web pages are translated into directed graphs in this way. Therefore, we use some graph terminology in this paper, which we will explain below. If there is an arc from node a to node b, b is called a *child* of a, and a is called a *parent* of b. In Fig. 2, node a is the parent of node b and node b is the child of node a. An arc from node a to node b is denoted by $a \rightarrow b$. A node which has no child is called a *leaf*.

2.2 Higher-Order Rank

We view a leaf, which is a dead-end in navigation, as different from another node which has links to follow. In addition, we consider a node which has a leaf as its child as different from one which does not have a leaf as its child. To show the difference, we introduce a rank which indicates the minimum number of links from a node to a leaf.

[1]By treating the other information as atoms, we can represent it in set theory.

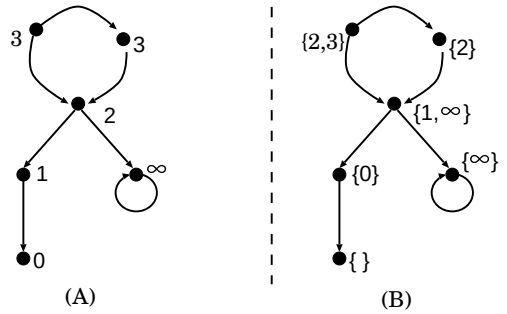

Figure 3: Higher-order rank: rank_0(A) and rank_1(B) on a Web graph

By definition, the rank of a leaf is 0, and a node which has a leaf as its child is of rank 1 as shown in Fig. 3(A). (The number at each node is the rank of the node in Fig. 3.) If no leaf is reachable from a node by following links, then the node's rank is ∞. The parent node with nodes whose rank is 1 and ∞ as children has a rank of 2, because 1 is less than ∞. In this way, the rank of any node can be determined.

The ranks can be combined to further discriminate nodes. A node of rank 1 should have a child of rank 0, but it may optionally have a child of rank 1 or rank 2. We view a node with a child of rank 1 as different from one with no child of rank 1. To show the difference clearly, we associate a node with rank_1, which is the set of ranks of its children. For example, if a node has children of rank 1 and rank 0, the rank_1 of the node is $\{1, 0\}$. If a node has children of rank 2 and rank 1, then rank_1 of the node is $\{2, 1\}$. For example, the rank_1 of the node whose rank is 2 is $\{1, \infty\}$ in Fig. 3(B). Similarly, we can associate a node with rank_2, which is the set of rank_1 of each of its children. For example, if a node has children of rank_1 $\{1, 0\}$ and rank_1 $\{2, 1\}$, then the rank_2 of the node is $\{\{1, 0\}, \{2, 1\}\}$. In general, rank_k is defined inductively as $\text{rank}_k(n) = \{\text{rank}_{k-1}(m) \mid (n \to m) \in R\}$ for $k > 0$ with the initial condition $\text{rank}_k(n) = \text{rank}(n)$ for $k = 0$. Now, rank_k is defined for $k \geq 0$. We call rank_k a *higher-order rank*.

For more mathematically rigorous definitions, see Appendix D.

2.3 AFA: Anti-Foundation Axiom

We consider nodes having the same rank_k value as structurally equivalent with respect to the rank_k. We call a set of equivalent nodes an *equivalence class of* rank_k, or a *rank_k-equivalence class*, borrowing the mathematical terminology. The important fact regarding equivalence classes of rank_k is that an equivalence class of rank_{k+1} is a union of some equivalence classes of rank_k. In other words, by proceeding from rank_k to rank_{k+1}, each equivalence class remains the same equivalence class or is divided into equivalence classes. This division process terminates at some k; that is, equivalence classes of some rank_k are the finest equivalence classes defined by the higher-order rank. Another important fact is that these finest equivalence classes coincide with sets in the non-well-founded set theory based on the anti-foundation axiom (AFA) [13, 14]. Based on this fact, we adopted the AFA as the basis of the analysis by the higher-order rank.

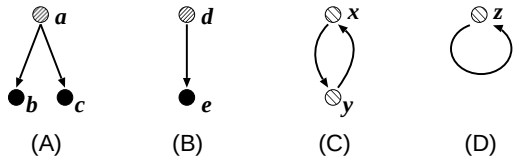

Figure 4: Sets in Standard graphical notation

The membership relation of set theory can be represented by a directed graph, where each set is represented by a node and a membership relation of $b \in a$ from a to b is represented by an arc a $\to$ b. We call this graph a *membership graph*. Conversely, we can view a directed graph $G = (N, R)$ where $R \subset N \times N$ as the AFA. Nodes of the graph are unified as long as no contradiction occurs under the AFA. We call this structure the *AFA structure* of the graph. Note that some nodes are equivalent in the AFA structure; that is, they correspond to the same set in the AFA. The definition of equivalence of nodes in the AFA structure can be found in Appendix C.

Sets a, b, and c such that $a = \{b, c\}$ are denoted as in Fig. 4(A). In set theory, the equality of sets is determined by their members. So, $b(= \{\})$ and $c(= \{\})$ are identical. A set which has no members is denoted by $\emptyset$. Leaves correspond to $\emptyset$ in the AFA. For d and e in Fig. 4(B), $e = \emptyset = b = c$ and $d = \{e\} = \{b\} = a$. In Fig. 4(C), we obtain $x = y$ by the AFA definition. Thus, sets x and y are identical to z in Fig. 4(D), which is the simplest circular set and is denoted by Ω.

In this paper, we first translate a Web graph into an AFA structure, and then apply an analysis based on the higher-order rank to the AFA structure. This is because the AFA structure is much smaller and easier to analyze than the Web graph.

2.4 Higher-Order Rank Analysis

To find irregularities in the Web structure, we use the division process of equivalence classes of the higher-order rank. We explain the idea using an example in Fig. 5. After the translation of Web pages, an AFA structure is generated as shown in Fig. 5(B).

The equivalence classes of rank_0 are $\{a\}, \{b, c, d, e, f\}, \{g\}$, and $\{h\}$. Here, $\{a\}$, $\{g\}$, and $\{h\}$ are singletons. For a rank_0-equivalence class which is a singleton, there is no other similar node with respect to a directed path from the node to a leaf. So, a, g, and h have irregular structures. Next, the equivalence classes of rank_1 are $\{a\}, \{c\}, \{b, d, e, f\}, \{g\}$, and $\{h\}$. That is, $\{b, c, d, e, f\}$ is divided into $\{b, d, e, f\}$ (of rank_1 $\{2\}$) and $\{c\}$ (of rank_1 $\{2, \infty\}$). So $\{b, d, e, f\}$ and $\{c\}$ have different structures, and because $\{c\}$ is smaller (having fewer elements), we suppose that $\{c\}$, instead of $\{b, d, e, f\}$, is irregular.

Based on this observation, we assume that a node in a singleton which is the result of the division of some higher-order rank is related to some irregularity. The method to find an irregularity based on this assumption is as follows:

1. Translate Web pages into an AFA structure.
2. Calculate the rank_0 of nodes and decide the equivalence classes.

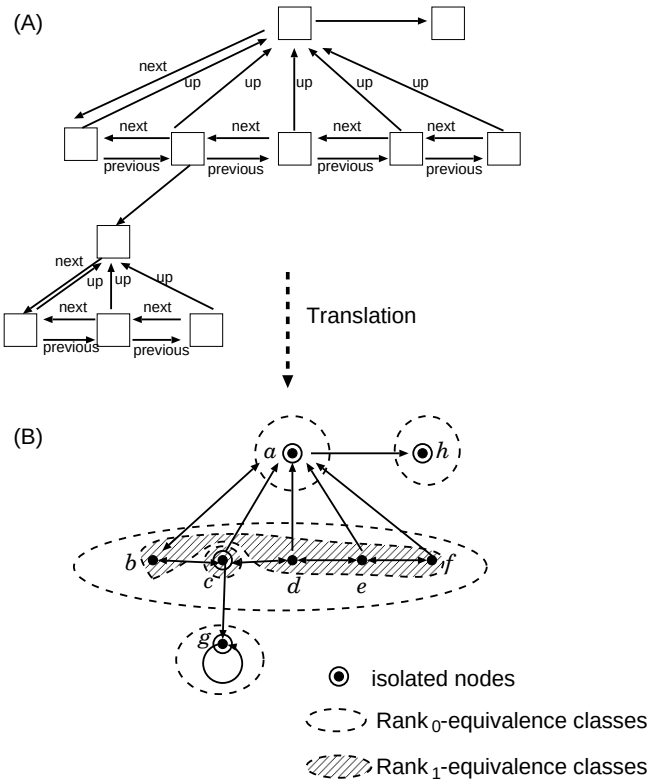

Figure 5: Higher-order rank analysis

3. In the rank$_0$-equivalence classes, select the nodes in the singletons.

4. Repeat the following for $k = 1, 2, 3, \ldots$, until all $rank_k$-equivalence classes are singletons.

   (a) Calculate the rank$_k$ of nodes and decide how the previous equivalence classes are divided.

   (b) Select the nodes in the newly generated singleton classes.

We call the nodes in the singletons found by this method *rank$_k$-isolated nodes* or only *isolated nodes*, and this method *higher-order rank analysis*.

2.5 Reduction Analysis

In comparison with the higher-order rank analysis, we briefly introduce our previous method: the reduction analysis [1, 2] here.

In this analysis, we dealt with sequences and hierarchies. Patrick [15] remarks that "three essential styles can be used to build a Web site: sequences, hierarchies, and webs." The third one poses few restrictions on the pattern of information use. In other words, webs do not have a regular structure. We focused on detecting an irregular structure by using the regular structure as a benchmark. Thus, our targets are sequences and hierarchies.

Regarding sequences, the AFA of a sequential structure is also a sequence, and the deviation of the structure is easily detected. So, we advance to a sequence with some additional structures whose analysis is not so straightforward. For such

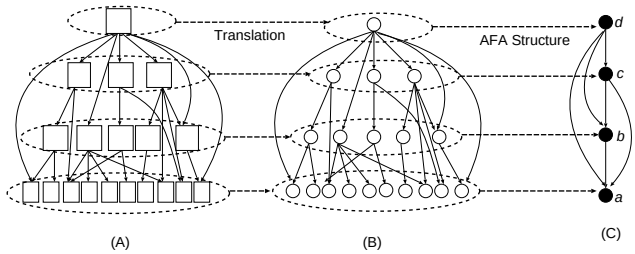

Figure 6: Hierarchical structure (representation in Web, graph, and AFA structure)

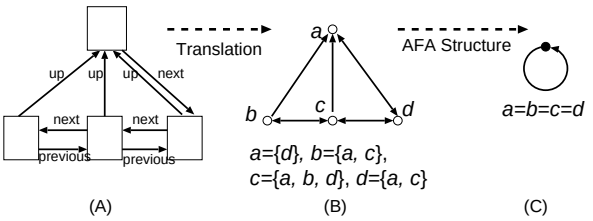

Figure 7: Linear structure (representation in Web, graph, and AFA structure)

a sequence, which we often encounter, we take a sequence with one entrance page as shown in Fig. 7.

Web pages without irregular structures are translated into simpler AFA structures like Figs. 6 and 7. In other words, an irregularity in a structure often makes the AFA structure more complex. By deleting some spurious links, we can unify the substructures in an AFA structure and obtain a reduced AFA structure. For example, the Web pages in Fig. 8(B) have a spurious link (the dotted arrow), unlike the more uniform Web pages in Fig. 8(A). The corresponding AFA structures in Figs. 8(C) and (D) show that by deleting the spurious link, we obtain a simpler AFA structure. We thus assume that links whose removal makes the AFA structure simpler are spurious. In this way, we can detect candidates which may be irregular links. We call this method *reduction analysis*.

We assess the significance of an irregular link candidate according to the magnitude of unification resulting from a change such as the deletion of that link. We use the following two schemes to measure this magnitude.

Hierarchical Structure Scheme: For a hierarchical structure, as shown in Fig. 9(A), unified sets (sets of nodes which are made equivalent by removing some arc) are in the forms $\{\{a, d\}\}$, $\{\{a, d\}, \{c, e\}\}$, ... We therefore employ the set inclusion relation $\subseteq$ between the unified sets as a measure and select the maximal unified sets. In this case, we regard $e \rightarrow d$ as an irregular link.

Linear Sequence Scheme: For a linear structure, as shown in Fig. 9(B), unified sets are in the forms $\{\{g, h\}\}$, $\{\{g, h, i\}\}$, $\{\{g, h, i, j\}\}$, $\{\{g, h, i, j, k\}\}$, ... In this case, we employ the set inclusion relation $\subseteq$ between members of the unified sets as a measure, and select the maximal unified sets. In the case of Fig. 9(B), $k \rightarrow l$ is regarded as an irregular link.

We then delete candidates with larger significance.

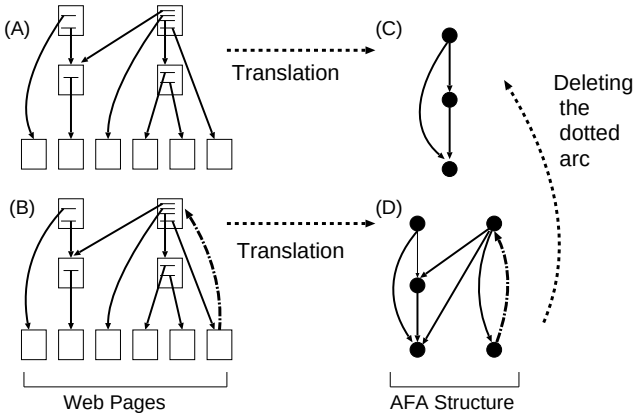

Figure 8: A spurious link and its effect on an AFA structure

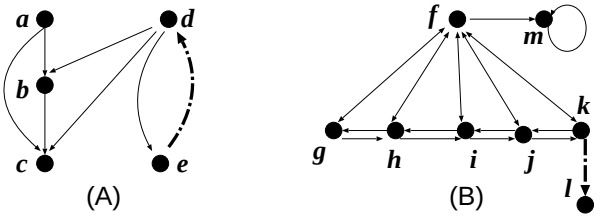

Figure 9: Hierarchical structure scheme (A) and linear sequence scheme (B)

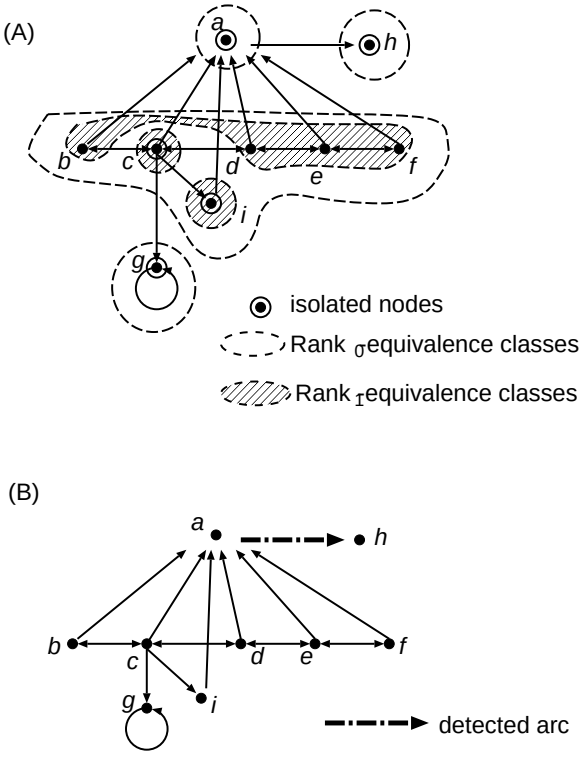

Figure 10: Comparison of higher-order rank analysis (A) and reduction analysis (B)

2.6 Comparison between Higher-Order Rank Analysis and Reduction Analysis

In this section, we compare higher-order rank analysis and reduction analysis with a sample shown in Fig. 10.

When we apply higher-order rank analysis to the sample, nodes a, c, g, h, and i are detected as shown in Fig. 10(A). In contrast, when we apply reduction analysis of a linear sequence scheme to the sample, only arc $a \rightarrow h$ is detected (Fig. 10(B)).

The start node a of arc $a \rightarrow h$ is detected by the higher-order rank analysis. Nodes c and i, which are detected by higher-order rank analysis, cannot be detected by reduction analysis. This is because node c (and its arcs) has multiple irregular arcs $c \rightarrow i$ and $c \rightarrow g$, simple (single arc) reduction analysis cannot find the arcs, and reduction analysis cannot find the arc $i \rightarrow a$ since node i is in neither a hierarchical structure nor a linear sequence.

Thus, higher-order rank analysis and reduction analysis produced different results, and higher-order rank analysis is superior to reduction analysis for the following reasons.

- It may detect some irregularity caused by multiple irregular arcs.
- No structure-dependent scheme has to be assumed.
- Isolated nodes can be found efficiently.

3. EXPERIMENTS

In this section, we discuss our experiments where we applied higher-order rank analysis to actual Web sites.

3.1 Data

We analyzed six Web sites in actual use: Google (Google Japan[2]), Yahoo (Yahoo! Japan[3]), Mofa (Ministry of Foreign Affairs Official Web site in Japan[4]), Kantei (Official Web site of the Prime Minister of Japan and His Cabinet[5]), WLH (World Lecture Hall[6]), and HT05 (Sixteenth ACM Conference on Hypertext and Hypermedia[7]).

Google is one of the most widely used search engines. Yahoo runs a directory service. Mofa and Kantei are Web sites for Japanese government organizations. WLH is an entry point to free online course materials from around the world. HT05 is the Web site for the Sixteenth ACM Conference on Hypertext and Hypermedia.

Every Web page was downloaded recursively using the wget command in at most 10 depth paths around Mar. 20, 2005. The basic values for each site are shown in Table 1.

Every Web site has the isolated node corresponding to the pages which has no links. Thus, it is omitted in the following examples.

3.2 Example: Google Japan

In this example, we show that our method can detect Web pages having structures which differ from those of most other pages.

[2]http://www.google.co.jp/
[3]http://www.yahoo.co.jp/
[4]http://www.mofa.go.jp/
[5]http://www.kantei.go.jp/
[6]http://web.austin.utexas.edu/wlh/
[7]http://www.ht05.org/

Site	Web Pages	Web Links	AFA Nodes	AFA Arcs
Google	1,645	5,684	136	1,161
Yahoo	12,798	14,076	50	2,563
Mofa	36,432	299,069	8,273	105,851
Kantei	43,449	100,388	3,276	18,583
WLH	2,647	34,994	2	2
HT05	124	948	2	2

Table 1: Web sites for experiments. Web pages and Web links show the respective number of Web pages and Web links. AFA Nodes and AFA Arcs respectively mean the numbers of nodes and arcs in an AFA structure.

Figure 11: Typical Web page with a side menu at Google Japan

Figure 12: Web page without a side menu at Google Japan

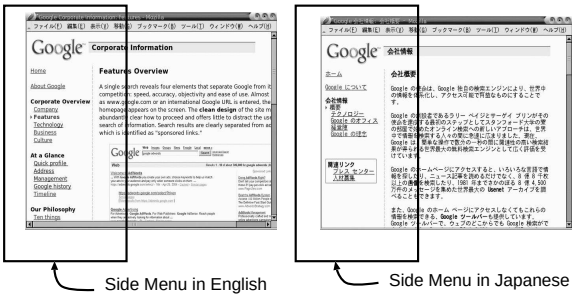

Figure 13: The side menu of Japanese pages differed from that of English pages.

For this site, the number of Web pages was 1,645, and the number of nodes in the AFA structure was 136. The number of links was 5,684 and the number of arcs in the AFA structure was 1,161.

There were three rank_0-equivalence classes, five rank_1-equivalence classes, and 12 rank_2-equivalence classes. Two rank_2-isolated nodes were found.

- One node of rank_2-isolated nodes corresponded to six Web pages, which unusually had no navigational links. Although almost all pages at this site uniformly had the side menu (Fig. 11), these six pages had only links to the top page of Google Japan and no side menu (Fig. 12).
- The other node of rank_2-isolated nodes corresponded to four pages. These pages were English pages and the corresponding Japanese pages with inconsistency in the interlink between them (Fig. 13). Higher-order rank analysis detected the inconsistency in the link structures.

32 rank_3-equivalence classes and 15 isolated nodes were found in the rank_3-equivalence classes. These isolated nodes in rank_3-equivalence classes correspond to special pages. For example, one of the rank_3-isolated nodes corresponded to the site map (Fig. 14). This page had links to all index pages in all categories. In addition, this page had more links than other standard pages.

During the course of the analysis, we found that the directory structure of http://www.google.co.jp/ is identical to that of http://www.google.co.jp/intl/ja/. For example, http://www.google.co.jp/press/zeitgeist.html is equivalent to http://www.google.co.jp/intl/ja/press/zeitgeist.html. The equivalence of these two pages was detected because they had the same directory structure, and the corresponding directories or files were identified in the AFA structure.

3.3 Example: Yahoo Japan

This example shows that higher-order rank analysis detects the index pages representing categories of the directory service.

We analyzed 12,798 Web pages and 14,076 links in Yahoo Japan, and translated them into an AFA structure of 50 nodes and 2,563 arcs. There were two rank_0-equivalence classes and three rank_1-equivalence classes. One of the rank_1-equivalence classes was a singleton; in other words, one rank_1-isolated node was detected.

The original page of the rank_1-isolated node was the top page of Yahoo Japan. This page was detected because the top page of Yahoo pointed to more pages than other pages did.

There were five rank_2-equivalence classes and one of them was a rank_2-isolated node. As shown in Fig. 15(A), the pages corresponding to the rank_2-isolated nodes were unusual in that they had no side menu. Except for these pages, all pages of that category had a side menu on the left side like Fig. 15(B).

Three nodes were isolated in 10 rank_3-equivalence classes. All pages corresponding to the rank_3-isolated nodes were index pages of the categories. This shows that each category had its own Web structure.

3.4 Example: Ministry of Foreign Affairs Official Web site in Japan

In this example, higher-order rank analysis found older papers which had different structures.

Figure 14: Site map at Google Japan. This page had more links than other pages.

Figure 15: An exceptional one without a side menu (A) and a standard page with a side menu (B) on Yahoo JAPAN

We used 36,432 pages and 299,069 links on the Ministry of Foreign Affairs Official Web site in Japan. The number of nodes and arcs in the AFA structure in this case were 8,273 and 105,851, respectively. Most pages have a top menu as shown in Fig. 16. There were seven rank_0-equivalence classes, 31 rank_1-equivalence classes, and two rank_0-isolated nodes and four rank_1-isolated nodes.

- Two nodes of rank_0-isolated nodes corresponded to 190 pages, which were old papers from before the year 2000. They had links to only themselves in Fig. 17 and such structures were different from most others.
- Four nodes of rank_1-isolated nodes corresponded to four pages, respectively. These pages had no standard top menu and typical structures. One page of them was the top page of subcategory in 1996(Fig. 18.)

Probably they were created before the main style was established.

3.5 Example: Official Web site of the Prime Minister of Japan and His Cabinet

This example shows that old papers and index pages of categories can be detected through higher-order rank analysis.

The Official Web site of the Prime Minister of Japan and His Cabinet had 43,449 pages and 100,388 links. Its AFA structure had 3,276 nodes and 18,583 arcs. The number of rank_0-equivalence classes was 55. We found 25 rank_0-isolated nodes.

Figure 16: Standard page with a top menu at Mofa

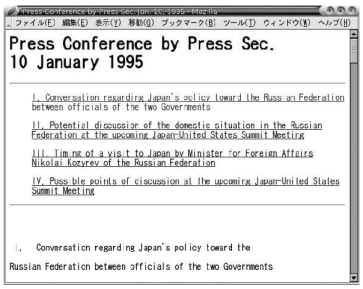

Figure 17: Part of an old paper at Mofa

- One node corresponded to 200 pages which have a link to a top page without top menu. The structure of these had no top menu and differed from that of other standard pages, which had a top menu (Fig. 19).
- 24 nodes corresponded to 24 pages respectively, which were a part of simple sequential papers with no navigational links (Fig. 20).

These pages were old papers from the cabinet of a former prime minister.

There were 99 equivalence classes and 16 isolated nodes at rank_1. The 16 nodes correspond to 5,043 pages with irregular structures. Fig. 21 shows the three pages of them. They were the page of the constitution of Japan(Fig. 21(A)), the information page for kids(Fig. 21(B)), and the index page of the subcommittee meeting for the Act for protection of computer Processed Personal Data held by Administrative Organs (Fig. 21(C)). These pages seem to have followed an independent policy regarding Web structure.

3.6 Example: World Lecture Hall and HT05

In this example, we show that if all pages follow the same style, higher-order rank analysis correctly detects no pages.

As to the World Lecture Hall, we examined 2,647 pages and 34,994 links at this site, and the number of nodes and that of arcs in the AFA structure (Fig. 22) were two in both cases. Higher-order rank analysis did not find any isolated nodes. In this site, all the original pages (Fig. 23) had links to 'Top page', 'Find Course', 'Browse by Area', 'Advanced Search', 'About WLH', and so on. The node whose rank was 0 corresponded to the pages referred to by the pages in the World Lecture Hall. The node whose rank was 1 corresponded to the pages in the World Lecture Hall.

Figure 18: Top page of a subcategory at Mofa

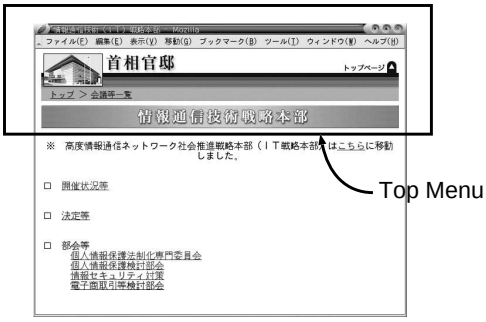

Figure 19: A standard Kantei page with a top menu

Higher-order rank analysis did not find any isolated nodes on the HT05 site, because all pages had the same structure as shown in Fig. 24.

4. CONCLUSION

In this paper, we have proposed a new method using higher-order rank for structural analysis of Web structure. We applied the higher-order rank analysis to actual Web sites, and successfully found notably irregular structures which differ from the typical structure of the Web sites.

Now, we are extending the functionality of our method to suggest possible corrections. Moreover, we plan to apply our method to the characterization of hypertext structures.

5. REFERENCES

[1] Ikumi Horie and Kazunori Yamaguchi, "Structural Analysis for Web Documentation Using the Non-Well-Founded Set," Fifteenth ACM Conference on Hypertext and Hypermedia, pp.42-43, 2004.

[2] Ikumi Horie and Kazunori Yamaguchi, "Structural Analysis for Web Documentation by the Non-Well-Founded Set," International Conference on Cybernetics and Information Technologies, Systems and Applications: CITSA, pp.210-215, 2004.

[3] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, Janet Wiener, "Graph structure in the web", In Proceedings of the 9th International World Wide Web Conference, pp. 247-256, 2000.

[4] Rodrigo A. Botafogo and Ben Shneiderman, "Identifying aggregates in hypertext structures," Hypertext'91 Proceedings, pp.63-74, 1991.

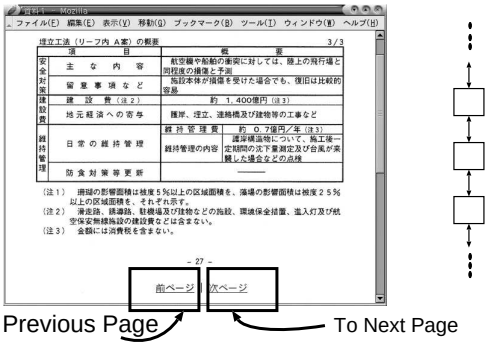

Figure 20: A part of sequential papers at Kantei

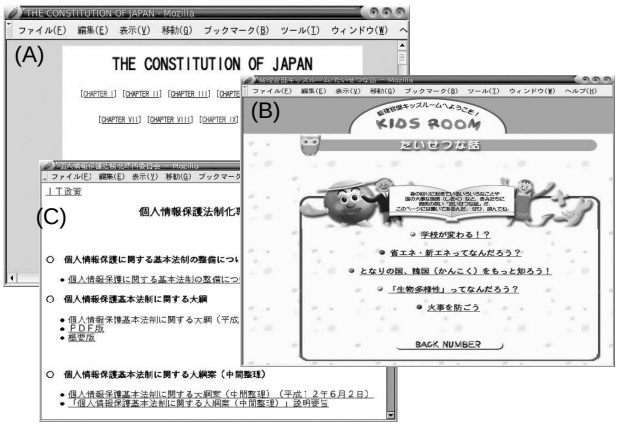

Figure 21: The page for the constitution of Japan (A), the top page for kids (B), and the index page of a subcommittee meeting(C)

[5] Rodrigo A. Botafogo, Ehud Rivlin, Ben Shneiderman, "Structural Analysis of Hypertexts: Identifying Hierarchies and Useful Metrics," ACM Trans. Inf. Syst. 10(2), pp.142-180, 1992.

[6] John E. McEneaney, "Graphic and numerical methods to assess navigation in hypertext," Int. J. Hum.-Comput. Stud. 55(5), pp.761-786, 2001.

[7] John E. McEneaney, "Visualizing and Assessing Navigation in Hypertext," Hypertext 1999, pp.61-70, 1999.

[8] Chaomei Chen, "Structuring and Visualising the WWW by Generalised Similarity Analysis," Hypertext 1997, pp.177-186, 1997.

[9] Wang and Huiqing Liu, "Discovering Typical Structures of Documents: A Road Map Approach," SIGIR'98, pp.146-154, 1998.

[10] Sougata Mukherjea, James D. Foley, "Visualizing the World-Wide Web with the Navigational View Builder," Computer Networks and ISDN Systems, 27(6), pp.1075-1087, 1995.

[11] E.James Whitehead, Jr, "Uniform Comparison of Data Models Using Containment Modeling," Hypertext 2002, pp.182-191, 2002.

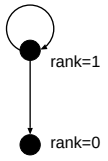

Figure 22: AFA structure of the World Lecture Hall and HT05

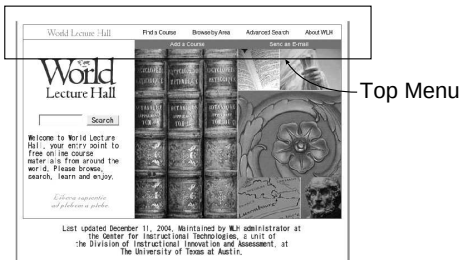

Figure 23: Top page of the World Lecture Hall

[12] E. Amitay, D. Carmel, A. Darlow, R. Lempel, and A. So, "The Connectivity Sonar: Detecting Site Functionality by Structural Patterns," Forteenth ACM Conference on Hypertext and Hypermedia, pp.38-47, 2003.

[13] Keith Devlin, "The Joy of Sets: Fundamentals of Contemporary Set Theory," Springer Verlag, 1993.

[14] Peter Aczel, "Non-well-founded Sets: CSLI Lecture Notes Number 14," Stanford, 1988.

[15] Patrick J. Lynch , Sarah Horton, "Web Style Guide: Basic Design Principles for Creating Web Sites," Yale Univ Pr, 2002.

[16] Robert Paige and Robert E. Tarjan, "Three partition refinement algorithms," SIAM J. Computer, Vol.16, No. 6, pp.973-989, 1987.

APPENDIX

In this appendix, we briefly introduce the mathematical background of our tools.

A. NON-WELL-FOUNDED SET THEORY

By analyzing the link structures of Web pages through set theory, we can identify pages which have the same transitive referential structures. However, because the referential structure often includes cycles, we cannot use standard set theory in which cyclic membership is inhibited explicitly by the axiom of foundation. We therefore use the non-well-founded set theory in which a set may contain itself transitively. There are variants of the non-well-founded set theory, and the one we adopted is based on the anti-foundation-axiom (AFA) [13, 14]. We refer to the non-well-founded set theory itself as AFA.

B. MEMBERSHIP GRAPH

The membership relation of set theory can be represented by a directed graph by representing each set by a node, and a membership relation $b \in a$ by an arc $a \to b$. Conversely, we

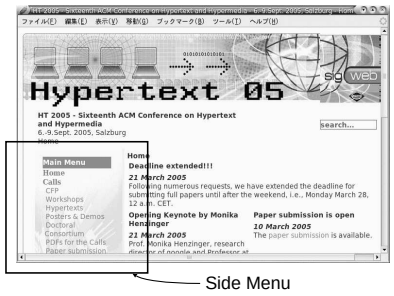

Figure 24: Top page of the HT05

can consider a directed graph $G = (N, R)$, where $R \subset N \times N$, as the AFA. Note that some nodes are equivalent in the AFA; that is, they correspond to the same set in the AFA. Identifying equivalent nodes in the AFA is an operation we used frequently, and this operation is carried out using either the coarsest stable partition in Appendix C or the higher-order rank in Appendix D depending on the purpose.

C. COARSEST STABLE PARTITION

A family $\mathcal{F}$ of subsets of N is called a *partition* on N if the following two conditions are satisfied:

1. Any $A, B \in \mathcal{F}$ satisfy $A \cap B = \emptyset$ or $A = B$.
2. The union of the elements in $\mathcal{F}$ is N.

A partition $\mathcal{F}$ is said to be a *refinement* of a partition $\mathcal{G}$ if any element in $\mathcal{F}$ is a subset of some element of $\mathcal{G}$. Note that all the partitions on N form a lattice with respect to the refinement relation.

For a membership graph $G = (N, R)$, a partition $\mathcal{F}$ is called *stable* if and only if, for any $A, B \in \mathcal{F}$, $B \subset \{b \mid (a \to b) \in R, a \in A\}$ or $B \cap \{b \mid (a \to b) \in R, a \in A\} = \emptyset$. For any directed graph, there always exists a unique coarsest stable partition with respect to the refinement relation. The equivalence relation induced by the coarsest stable partition coincides with the equivalence relation of nodes in the AFA. That is, nodes in the same partition of the coarsest stable partition are equivalent in the AFA, and vice versa.

The coarsest stable partition can be calculated in $O(m \log n)$ where m is the number of arcs and n is the number of nodes[16].

D. HIGHER-ORDER RANK

For a leaf node a, which has no child, let $\mathrm{rank}(a)$ be 0. For a node a which has a leaf in its descendants, let $\mathrm{rank}(a)$ be the length of the shortest directed path to some leaf node. For a node a which has no leaf among its descendants, let $\mathrm{rank}(a) = \infty$.

Next we define a higher-order rank rank_k for a nonnegative integer k. Let rank_0 be rank. For a positive k, let $\mathrm{rank}_k(a)$ be $\{\mathrm{rank}_{k-1}(b) \mid (a \to b) \in R\}$.

Proposition: For a sufficiently large k, the partition of N induced by rank_k coincides with the coarsest stable partition by R.

Because of the space limitation, we omit this proof.