KAGAMI: Web Rating Agent Based on Hyperlink Structure

Naonori Otsuka, Hironori Hiraishi, Fumio Mizoguchi
Department of Industrial Administration, Faculty of Sci. and Tech.,
Science University of Tokyo
Noda, Chiba, Japan
{ootsuka, hiraishi, mizo}@ims.sut.ac.jp

Abstract

In recent years, Internet users have been able to benefit from the increased amount of information in the World Wide Web (WWW) space. However, information retrieval efficiency has decreased because of the increasingly complicated WWW. A fundamental solution has not been achieved, though the information retrieval efficiency has been improved by browsing support tools such as the Web Information Access System (WIAS)[1][2][3]. Web sites must be improved to achieve a fundamental solution.

In this paper, we rate Web sites and clarify the problems using Web site rating agent KAGAMI[4] (Japanese for "mirror") and present an index for Web site improvement. This system autonomously collects data using a Web robot and rates sites using the site structure based on the WIAS framework. Additionally, this system generates a report by various visualization and supports for improving the Web site rating by this system.

1. INTRODUCTION

We have designed and implemented a Web site rating System, used it to rate Web sites, and aim at supporting improvement. In this chapter, we introduce related research and services as present conditions of Web site rating, describe the approach of this system, and explain how WIAS is applied to it. Search engine google's PageRank, has adopted the reference relation of a page as the valuation basis from the assumption that a good page is linked from good pages. This rating technique is supported by a large-scale database. Netscape's Web Site Garage[6] rates by compatibility whether it can be used in various browsers and use recommended HTML grammar, and stimulates Web site improvement. This rating technique employs expression by HTML. GIGA Information Group's Web Site ScoreCard[7] rates mainly by needs (usability, customer support, purchase procedures, etc.) in EC sites. This rating technique suits the practical purposes of EC. There are various similar Web site rating viewpoints. We established a rating basis according to user needs by applying research of browsing support. On this basis, we rate Web sites and support their improvement. We designed the web rating agent KAGAMI as a practical system.

1.2 WIAS

WIAS conducts research on efficient use of the Web through research on browsing support as in this paper. In research on browsing support, they discover information buried in huge and complicated Web sites and seek to improve information retrieval efficiency by presenting it in a form intelligible to users. One browsing support tool is WIAS, which attempts to improve the information retrieval efficiency of Web sites by mapping the Web site, searching and filtering information in the Web site and browser, and coordinating each element (see Fig. 1).

Figure 1. The image of WIAS
The visualization features adopted by WIAS are enumerated below.

1. The link structure of Web site is simplified and represented as a tree structure, which eliminates duplication.

2. The whole Web site can be understood by applying the technique of hyperbolic geometry called a "Hyperbolic Tree" [8] [9] [10] when the Web site is mapped.

3. It is possible to display two or more conditions of a page at once by adding parameters like color and height to the node expressing each Web page.

WIAS supports discovering the information on the target when a user performs searching, filtering, focal movement, and zooming interactively on the map of a Web site with these features.

The WIAS system thus adopts the user's viewpoint. Using data structure and visualization techniques to rate a website can be an effective index for website construction suited to a user's needs. Hence, we built our system by applying the framework of WIAS in this paper.

2. DESIGN

The Web site rating system KAGAMI designed in this paper attains its purpose by employing the features shown below.

- A rating system adopting a user's viewpoint is realized by using the framework of the browsing support tool WIAS.
- Exclusion of subjectivity and low cost are realized by processing mechanically a flow from data collection to analysis.
- A rating result is effectively fed back by WIAS by using various visualization expression modes such as maps and graphs.

In this chapter, we describe the practical design objectives for each item.

2.1 Use of the WIAS Framework

Although there are various viewpoints for rating Web sites, in this paper, we propose a rating method from the viewpoint of the user who performs browsing. In order to realize a rating method from the viewpoint of the user, we must consider a user's needs and action pattern. For reasons mentioned earlier, in this system, we achieve a rating system that reflects a user's needs and action pattern by reusing the information gathering technique and data structure adopted in WIAS.

2.2 Automation of Processing

Many conventional Web site rating systems perform rating by actually browsing. This takes manpower and time, and there is a problem with subjectivity. With this in mind, in this system, we automated a series of processes from automatic information gathering by the WIAS data collection agent to mechanical rating of the heuristics obtained from research of browsing support and visualization of results.

2.3 Visualizations of Rating Results

In rating a Web site, it is important that the user be able to understand the result intuitively even if he or she lacks special knowledge. This is because it is more important to understand a problem and to use it for improvement rather than rating.

The system designed in this paper consists of visualization of a radar chart that shows the balance of each parameter of the Web site, a map that shows an overview of the Web site, and a layered structure figure capturing the Web site as a tree structure (see Fig. 2).

![Figure 2. Various visualization which constitutes KAGAMI](image)

3. IMPLEMENTATION

This chapter describes implementation of a system based on the design objectives stated in the preceding chapter. As the preceding chapter has stated, this system implements the base module by using the framework...
of WIAS. We built a platform-independent system by using JAVA, a multiplatform implementation language. The configuration of this system is shown in Fig. 3.

![System Configuration Diagram]

Figure 3. System configuration

This system consists of five modules as shown in Fig. 3. The control module, tree composition module, and tree-drawing module were implemented by extending the WIAS framework. The rating module defines various parameters from the know-how acquired from researching WIAS browsing support. The output module outputs the result in HTML and an image file that can be perused on the Web. The process flow of each module is explained in detail below.

3.1 Control Module

The control module is the central module. It receives the analysis demand (route URL of a target site) from a user and requests each module to perform the necessary processing. The processing flow of a control module is as follows.

1. The analysis demand (URL for rating) from a user is received, and a system is started.
2. The System is initialized by a properties file.
3. URL information received by process 1 is handed over to a tree composition module, and the tree data constituted based on it is received.
4. The tree data received by process 3 is handed over to a tree-drawing module, and the image data generated based on it is received.
5. The tree data received by process 3 is handed over to a rating module, and every parameter is made to rate based on it.
6. The image data received by process 4 is handed over to a result output module that outputs an image file and an HTML file.

We concentrated all processing on the control module. This made reconfiguration to the automation system of WIAS, which is an interactive tool, easy.

3.2 Tree Composition Module

A tree composition module patrols the inside of the Web site from the tree composition demand (input of URL) from a control module, collects required information, and stores it by the tree structure. In this system, the method of collecting the Web site links by width priority search is adopted as the composition of the tree. Therefore, an overview can be held when all information cannot be acquired, since the site space is huge.

The generation process is shown below.

1. The URL is taken out and downloaded from the queue in an entry.
2. Tags, such as "A" and "FRAME", are discovered and link information is extracted.
3. A relative path is modified to an absolute path, and an external link and an entered link are deleted.
4. If the link is unique, it will added to an entry.

The structure of the Web site is defined as a tree structure that eliminates duplication by repeating these processes recursively.

3.3 Tree Drawing Module

A tree drawing module draws the map of a Web site overview from the tree data constituted in the tree composition module. In this system, it is necessary to perform the visualization that can express an overview of the target with various scales of a Web site. We do this by applying the algorithm of the hyperbolic geometry also adopted by WIAS. The conversion process is shown below (see Fig. 4).

1. A node is arranged on a concentric circle focusing on a route.
2. The nodes arranged in process 1 project from the coordinates of process 1 to a hyperbolic plane.
3. The nodes arranged in process 2 project on a unit circle from a hyperbolic plane.

Finally, when drawing to the canvas on system, it displays information in arbitrary sizes by expanding the coordinates on a unit circle.
3.4 Rating Module

A rating module rates Web sites using various parameters from the tree data constituted in the tree composition module. In this system, we assumed the data acquired by WIAS could be utilized to rate a Web site and adopted the data as a rating parameter (Table 1).

<table>
<thead>
<tr>
<th>Structure</th>
<th>Rating to the layered structure of a link</th>
</tr>
</thead>
<tbody>
<tr>
<td>Link</td>
<td>Rating to the navigation in a site</td>
</tr>
<tr>
<td>Media</td>
<td>Rating to use of a picture file</td>
</tr>
<tr>
<td>Portal</td>
<td>Rating to an external link</td>
</tr>
<tr>
<td>Scale</td>
<td>Rating to the scale of a site</td>
</tr>
<tr>
<td>Update</td>
<td>Rating to informational freshness</td>
</tr>
</tbody>
</table>

Table 1. Web site rating parameter

Specifically, the following procedure was employed. First, we count the number of the elements corresponding to each parameter. Second, we apply to rate at the model of the Web site assumed to be ideal. Third, we rate a site by assigning a set of weights as heuristics from the knowledge of the Web site obtained from research of the browsing support like WIAS. In the paper, we introduce the details of "Structure" rating, which is the most unique rating basis.

3.4.1 Structure rating

Structure rating captures the structure of a Web site as a tree structure to eliminate duplication and rates the degree of the spread. This is the greatest feature of this rating basis because it is unrealizable in the Web page rating by HTML parsing performed in the past.

This rating basis defines a rating formula having a structure that widens toward the end structure based on the assumption that a deeper layer contains more information in a Web site built intentionally (Fig. 5). This system rates the simple increase by comparing the number of nodes between contiguous layers, and the exponential increase by comparing the number of nodes of the target layer with the number of nodes of all higher layers. This system determines the final result by suitably weighting those rating (formulas 1, 2, and 5).

\[
Score = \alpha \cdot \Sigma_{i=1}^{N} f_i + (1 - \alpha) \cdot \Sigma_{i=1}^{N} g_i \tag{1}
\]

\[
f_i = \begin{cases} 
1 & (l_i \leq l_{i+1}) \\
0 & (l_i > l_{i+1}) 
\end{cases} \tag{2}
\]

\[
g_i = \begin{cases} 
1 & (l_i \geq \Sigma_{j=1}^{i-1} l_j) \\
0 & (l_i < \Sigma_{j=1}^{i-1} l_j) 
\end{cases} \tag{3}
\]

\(\alpha\): Weight of rating between adjacent layer

\(f_i\): The check of the simple increase by comparison of the number of nodes between contiguity layers

\(g_i\): The check of the exponential increase by comparison of the number of nodes with the whole higher layers

\(l_i\): Number of Node of layer \(i\)
3.5 Result Output Module

A result output module generates image file (GIF) such as a radar chart of rating and map of Web site, and HTML file for storing and supplementing the image files. Thereby, a result can be browsed by the general browser.

4. PRACTICAL APPLICATION

This chapter presents a real example that shows the usefulness of the system designed in this paper. The target of this system is a Web site where various forms exists. Therefore, broad sampling that eliminates subjectivity is required in order to show the usefulness. In order to satisfy such a demand, we provided the Web Rating Service and accepted Web site widely. The “Web Rating Service” and a successful example are outlined below.

4.1 Web Rating Service

Web Rating Service accepts a Web site to be rated using the Web site rating system designed and implemented in this paper, and publishes the result on a Web site ¹ (Fig. 6).

This service mechanically processes a series of activities from registration of a request to presentation of a rating result on a Web site. Therefore, we built the software JAVA environment base shown in Table 2 and implemented it.

<table>
<thead>
<tr>
<th>OS</th>
<th>Solaris 7.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Java</td>
<td>Java version 1.3.0</td>
</tr>
<tr>
<td>Servlet</td>
<td>Java Servlet 1.2</td>
</tr>
<tr>
<td>Web Server</td>
<td>Apache version 1.3.14, Jserv version 1.1.2</td>
</tr>
</tbody>
</table>

Table 2. Execution environmental software of Web Rating Service

<table>
<thead>
<tr>
<th>Service term</th>
<th>‘00/11/11-’00/11/21</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of accesses</td>
<td>about 15,000 HIT</td>
</tr>
<tr>
<td>Volume of data</td>
<td>about 7.7GB</td>
</tr>
<tr>
<td>Number of sites</td>
<td>682 site</td>
</tr>
<tr>
<td>Number of pages</td>
<td>about 800,000 pages</td>
</tr>
<tr>
<td>Average pages per site</td>
<td>about 1166 pages</td>
</tr>
<tr>
<td>Average volume per site</td>
<td>about 11.5MB</td>
</tr>
<tr>
<td>Maximum number of pages</td>
<td>32,214 pages</td>
</tr>
<tr>
<td>Maximum volume site</td>
<td>about 650MB</td>
</tr>
</tbody>
</table>

Table 3. Outline of Web Rating Service

Figure 6. Web Rating Service

This service was used by many general users after having been introduced in newspapers ² and the press. It was able to collect rating data of various Web sites. The outline of this service is shown in Table 3.

This service was able to rate various Web sites over ten days as shown in the table. This suggests that the reliability and the scalability of this system are sufficient.

---

¹ Information Media Center, Science University of Tokyo, http://www.imc.sut.ac.jp/

4.2 Successful Example

We rated several hundred Web sites with the Web Rating Service. Here we introduce two successful examples (Figs. 7 and 8).

The Web site in Fig. 7 was initially rated as having informational bias. However, a distinct improvement was seen as a result of reconfiguration after rating by this system.

The Web site of Fig. 8 was seen to lack information when first rated by this system. However, much detailed information was added after rating by this system, and it became a good site with balance.

As these results show, this system not only supports improvement but also functions as a Web site rating system. Efficient use of the Web, which is the purpose of this paper, thus becomes realizable.

Figure 7. Success example 1: The improvement from a map
5. CONCLUSIONS

In this paper, we designed the Web site rating system KAGAMI and automated a series of activities from information gathered from Web sites to analyze and visualize results. We defined the rating basis from the viewpoint of a user by applying the framework of browsing support tool WIAS. This system not only rates Web site but also supports their improvement. We implemented the Web Rating Service; we were able to demonstrate its performance as a system by having a large number of general users use this system. The usefulness was illustrated with a successful example.

6. FUTURE WORK

In this paper, we rated Web sites on a heuristic rating basis based on WIAS research results. However, since Web sites are very fluid targets, it is important to be able to adopt a new rating basis at any time.

As future work, we need to incorporate the concepts of Web usability[11][12] and Web accessibility[13][14] by general users as a Web site rating basis. Furthermore, we must provide support for improving Web sites by cooperating with machine learning systems. In this system, more general-purpose use will be attained, and realistic rating and support for improving Web sites will become realizable.

References


