Navigation Analysis Tool based on the Correlation between Contents Distribution and Access Patterns

Hiroki Kato, Takehiro Nakayama, Yohei Yamane

Solution Development Center
Industry Solutions Company
Fuji Xerox Co., Ltd

430 Sakai, Nakai-machi, Ashigarakami-gun, Kanagawa, 259-0157,Japan

{hiroki.kato, takehiro.nakayama, yohei.yamane}@fujixerox.co.jp

Abstract

This paper proposes a tool to assist web site publishers in improving their web site design. The tool uses an analysis technique that discovers the gap between web site publisher expectations and user behavior. The former are extracted by measuring the inter-page conceptual relevance and/or link connectivity. On the other hand, the latter is extracted by measuring the inter-page access co-occurrence. By evaluating the correlation between them, the tool discovers pages which should be improved in terms of web site design. It further uses a visualization technique using polar coordinate system to assist publishers in inspecting user access patterns in these pages. This technique will assist in finding clues for improving discovered gaps. The effectiveness of the tool is validated by case studies.

1. Introduction

The World Wide Web has grown explosively since its creation. As web sites play a significant role in business, their utilization is indispensable for success in business. To make the best use of web sites, their constant maintenance, in consideration of user behavior, is needed.

Business web sites generally contain a wide range of topics to provide information for users who have different interests and goals. Web site publishers also have various goals such as merchandising and advertisement. Thus, it is a difficult task to maintain web sites so that they satisfy both requests.

Many techniques to improve the effectiveness of a web site have been proposed. Web traffic analysis is one approach, and there are several commercial tools available [1][15][17]. Most of these traffic analysis tools generate web traffic statistical reports, including access ranking, access path ranking (most of these path lengths are less than 4), and so on. Some tools use an interactive technique to analyze user access patterns. When an analyst selects a target page, these tools show how many users selected one page as a departure page, and how many users selected one page as a destination page.

Web site publishers use these tools to check the effectiveness of their web sites. To improve content structures, they have to find problems in their web sites by scrutinizing reports on their web sites. In addition, they have to consider which pages of their web site require improvement. However, there are few tools to assist web site publishers in these efforts. In this paper, we propose techniques to help web site publishers find ineffective pages and that provide clues for improving them.

1.1 Our Approach

On our assumption, web site publishers expect that conceptually related pages should co-occur in visits if their site is well designed. The user also expects to efficiently access web pages which are conceptually related. Thus, finding the gap between web site publisher expectations and user behavior suggests a set of pages that should be improved in terms of web site design. The visualization of user behavior within these pages will help determine how the content structure should be redesigned.

In this paper, we propose two techniques to find the gaps. One uses the correlation between path length and inter-page access co-occurrence for each pair of pages in the web site. We measure the path length by the number of hyperlinks required to go from one page to another. We measure inter-page access co-occurrence by modifying the vector space model [11]. Further we show these relations visually using x-y coordinate system. The other technique analyzes the web sites at a more abstract level. It uses the correlation between the concept represented in a hyper-structure and the user traversal ratio. The concept is a set of linked web pages which represent a topic. It generated on the basis of restriction of the inter-page conceptual relevance to the

page selected as the start page. We measure inter-page conceptual relevance by the vector space model based on word frequency. We measure the user traversal ratio of each page in the concept by the ratio of the number of users who visited the start page and the number of users who visited each page in the concept after visiting the start page. Next, we compute the correlation coefficient between user the inter-page conceptual relevance and user traversal ratio about each page in the concept. We measure the gap in the concept by this correlation value. A small correlation value shows that users do not follow the web site publisher expectations.

Further, we propose a visualization technique to analyze user access paths related to discovered gaps. The user access paths are visualized using a polar coordinate system. Using this system, we analyze user access paths, such as from what page they come, to what page they go. This visualization tool assists in finding clues for improving the structure of contents at a glance.

This paper is organized as follows: Section 2 describes preprocessing of access log. Section 3 describes techniques for discovering the gaps between the web site publisher expectations and user behavior. It also describes the visualization technique. Section 4 discusses the effectiveness of our approach by means of case studies. Section 5 discusses related work. Section 6 concludes the paper and suggests future study.

In this paper, we use Fuji Xerox's public web site and an online brokerage's web site as sources of experimental data.

2. Preprocessing

In data mining, preprocessing of data is an important task. User behavior is not recorded accurately in an access log of a web server because of proxy servers and web browser cache. There are many approaches toward collecting accurate access logs such as using applet [12], network monitoring [1], and event logging [5]. However, these techniques cannot be simply applied to access logs already stored in the web server.

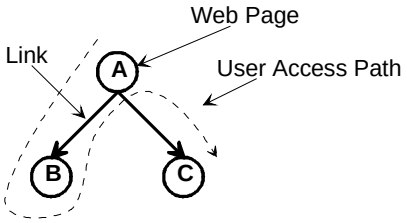

Figure 1: User access path

Extraction of user access patterns is essential task for finding gaps. In our techniques, user access patterns represent the time-sorted pages accessed by an identical user.

We use heuristics to identify a session of an identical user[1](e.g., the interval of user requests). For the extraction of user access patterns, there is a technique using web site hyperlink structure[4]. This technique extracts the maximal forward references as user access patterns from user requests in each session. The maximal forward reference represents a set of page which a user visits without revisiting a previously visited page. In Figure1, if the user request of the session is A -> B -> C, the maximal forward references are {A, B} and {A, C}. {A, B, C} represents a set of web pages. One drawback of this method is that it equally treats two different user access patterns. For example, A->B->C and A->C->B are identical. Suppose page D to which page A refers. Most users whose request is A->B->C visit D, but users whose request is A->C->B do not. If the web site publisher expects user to visit D, this difference in user access patterns is an important clue for redesigning the content structure. Thus, we generate user access patterns according to the order of requests recorded in access logs. Furthermore, we reject duplicate accesses to the visited page, because we want to equally treat both users who use browser cache (browser back and forward buttons) and users who do not. For example, if a user request is A->B->A->C, we extract {A, B, C} as the user's access pattern and then consider the order of requests.

We use a link structure to extract web publisher expectations. However, if a web site uses frames, there is a difference between links the user can follow and links in the page that the user actually selected. Thus, we generate a link table (this is a list of the relationship between destination pages and departure pages) by considering the actual view shown to user. In generating user access patterns, we also remove accesses to pages that are referred by "FRAME SRC" tag.

Field Name	Data Type	Explanation
SESSION_NO	INT	Session identifier
USER_ID	VARCHAR	User identifier(client IP Address or user account name)
START_DATE	DATE	Start time of session
PATTERN	VARCHAR	Page sequence which represents user access pattern

Table 1: Field name and data type and explanation of LOG DB

[1] In the online brokerage site's log, users are identified by authorization.

Under these conditions, we convert access logs to LOG DB, which consists of “SESSION No”; “USER ID (sometimes IP address)”; “START DATE”; ”PATTERN” columns (Table 1). In this process, we remove accesses from search engine robots and proxy servers by means of heuristics (e.g., access to /robots.txt; exhaustive access in a short period). In the following analysis, our system uses this database.

3. Analysis of Gap between Web Site Publishers Expectation and User Behavior

In this section, we introduce techniques that discover gaps. We also introduce a visualization technique that finds clues for improving these gaps.

3.1 Hyperlink Topological Analysis

In this analysis, web site publisher expectations are assessed by the link distance between any two pages in the web site and user behavior is assessed by inter-page access co-occurrence. If a page pair which has co-occurred in many user access patterns has a long path length, creating a new hyperlink between this pair provides the user with easier access, and might enable the user who has given up navigation to visit from one page to another.

We compute the link distance (the number of hyperlinks required to reach from one page to the another) for each pair of pages. Obviously, there can be multiple paths between any two pages in the hypertext system. However, since considering all paths would be computationally expensive, we compute only the shortest path, taking into account the hyperlink direction.

We then compute the inter-page access co-occurrence for each pair of pages. Given LOG DB, we count USER IDs for each page and generate a list of USER IDs weighted by frequency. This frequency represents the number of sessions in which the same user visited the page. This list is viewed as a vector that represents page users. Finally, we measure the inter-page access co-occurrence (SimA) for each page pair using the following formula.

$$SimA(p_i, p_j) = \frac{\sum_{k=1}^{t} (\lambda_{ik} \cdot \lambda_{jk})}{\sqrt{\sum_{k=1}^{t} (\lambda_{ik})^2 \cdot \sum_{k=1}^{t} (\lambda_{jk})^2}}, (0 \le SimA(p_i, p_j) \le 1)$$

where λ_{ik} is the weight of the kth USER ID that visited p_i, and t is the number of distinct USER IDs found in LOG DB. SimA is 0 if one of the pages has never been visited. SimA is 1 if the same users visit both pages with the same frequency.

After computing two measures, we plot link distance versus the access co-occurrence for each page. Figure 2 shows a graph drawn by our system. The X-axis represents SimA and the Y-axis represents link distance. The page pairs plotted in the area at the right upper are likely to have problems because many clicks are required to visit both pages, which many users may think are related.

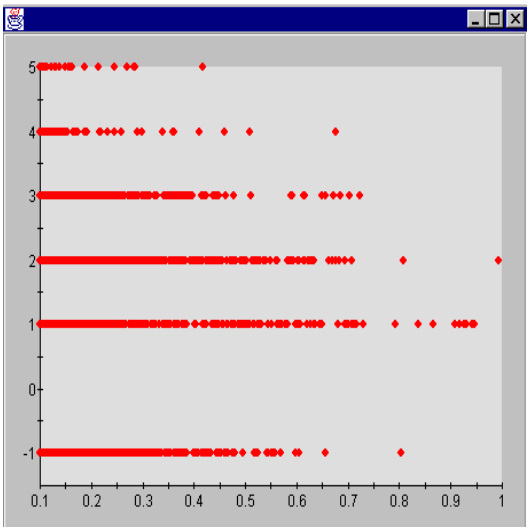

Figure 2: Link distance versus access co-occurrence of an online brokerage’s web site. For Y=-1, page pairs are plotted, which are not reachable from one page to another within the Link Distance = 5.

3.2 Concept Analysis

In this analysis, web site publisher expectations are assessed by inter-page conceptual relevance and user behavior is assessed by user traversal ratios. We don’t use SimA, because we consider the order of user access. One approach toward analyzing the gap is to plot inter-age conceptual relevance versus user behavior. In this section, we propose a technique for analyzing the web sites at a more abstract level.

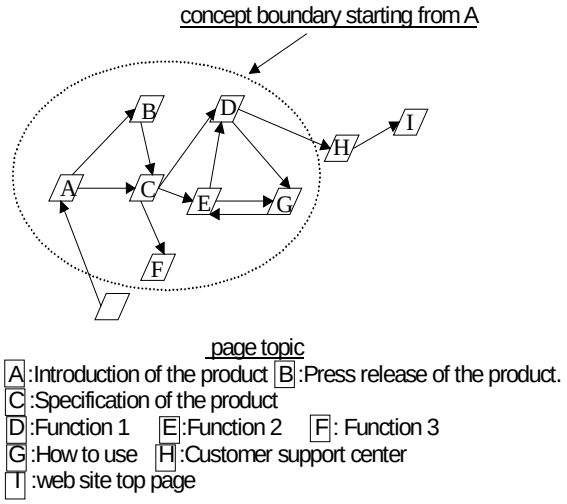

Figure 3: Example of concept representation

By analyzing at a more abstract level, web site publishers can more efficiently find and improve the gap. Page pair analysis (Hyperlink topological analysis) need to be scrutinized the relation of discovered page pairs which have gaps. On the contrary, in abstract level analysis, they concentrate to analyze the pages included the concept which has gap. However, this doesn't mean the abstract level analysis is superior to the page pair analysis. There are many types of gap which cannot to be found by the abstract level analysis. For example, the gap which cross two concepts is difficult to discover. Therefore, the abstract level analysis complements the page pair analysis.

Web site publishers represent the product or service information content as a set of hyperlinked pages, which we call a concept in this paper. In that case, web site publishers expect users access of all pages, since the user cannot understand the product or service correctly without accessing all pages. Thus, by evaluating the gap in each concept in the web site, the analyst will be able to inspect the effectiveness of a concept.

As shown in Figure 3, a concept is represented as a set of pages linked to each other. The concept is generated by means of the traversing links from a start page on the basis of restriction of the inter-page conceptual relevance to the start page. In this analysis, our system select start pages that have more access counts than a threshold value set by the analyst. We employ a vector space model to measure inter-page conceptual relevance. First, we obtain content words (nouns, verbs, and adjectives) by performing morphological analysis and stop word removal. Second, we compute the frequency of content words for each page. Third, we generate a list of content words weighted by this frequency. This list is viewed as a vector that represents page content:

$$p_i = (\omega_{i1}, \omega_{i2}, ..., \omega_{ik}, ..., \omega_{in})$$

where ω_{ik} is the weight of the kth content word, and n is the number of distinct content words found in the web site. Finally, we measure the inter-page conceptual relevance (SimC) for each page pair p_i and p_j using the cosine similarity formula as follows:

$$SimC(p_i, p_j) = \frac{\sum_{k=1}^{n} (\omega_{ik} \cdot \omega_{jk})}{\sqrt{\sum_{k=1}^{n} (\omega_{ik})^2 \cdot \sum_{k=1}^{n} (\omega_{jk})^2}}, (0 \le SimC(p_i, p_j) \le 1)$$

where SimC is 0 if one of the pages contains no content words. If the number of content words that appear in both pages is 0, the value of SimC is also 0. If two pages contain identical content words with the same frequency (i.e., the vectors of two pages are identical), the value of SimC is 1.

By traversing links from the selected start page with the restriction of $Sim(startpage, p_j) > Threshold_{sim}$ and $Length(startpage, p_j) > Threshold_{length}$, we extract a set of pages $\{startpage, p_0, p_1, ..., p_j\}$ as a concept. $Length(p_i, p_j)$ represents the path length between p_i and p_j. The similarity threshold $Threshold_{sim}$ controls content granularity. When it is higher, only closely related pages are extracted as a concept. When it is 0, all reachable pages are extracted as a single concept. Pages not reachable within the path length threshold $Threshold_{length}$ are ignored. This algorithm extracts some overlapped concepts, which means some pages are shared by more concepts. An analyst can control contents granularity by two parameters, a path length threshold and a similarity threshold. Though we could not get distinct explanation about the relation between the contents granularity and these parameters, we heuristically set appropriately parameters to get the concept at the granularity we expect.

After concept extraction, we compute user traverse ratio $Tr(p_0, p_j)$ as follows.

$$Tr(p_0, p_i) = \frac{u_{0i}}{u_0}, (0 \le Tr(p_0, p_i) \le 1)$$

where u_{ij} is the number of users who visited p_j after p_i, u_i is the number of users who visited p_i and p_0 represents a start page of a concept

Finally, we compute the correlation coefficient $R_{sim,tr}$ between inter-page conceptual relevance and user traverse ratio for each concept as follows:

$$R_{sim,tr} = \frac{s^2_{sim,tr}}{\sqrt{s^2_{sim} \cdot s^2_{tr}}} \qquad (-1 \le R_{sim,tr} \le 1),$$

where

$$s^2_{sim,tr} = \sum_{i=1}^{m} \left(Tr_i - \overline{Tr}\right)\left(SimC_i - \overline{SimC}\right),$$

$$s^2_{sim} = \sum_{i=1}^{m} \left(SimC_i - \overline{SimC}\right)^2,$$

$$s^2_{tr} = \sum_{i=1}^{m} \left(Tr_i - \overline{Tr}\right)^2,$$

where $SimC_i$ is the value of conceptual relevance for the ith page in a concept, Tr_i is the value of user traverse ratio for the ith page in the concept, and m is the number of page in the concept. A small correlation value indicates that users do not follow web site publisher expectations of the concept.

3.3 Access Path Visualization

In this section, we describe the visualization technique for interactive analysis of user access patterns. User access paths are visualized by means of a polar coordinate view. In the polar coordinate system, the analyst-selected target page is plotted at the origin (0,0), the user traverse ratio is converted to an angle by multiplying it by 2π, and the user path length from the target page is represented by radius. The user path length is the number of pages between the target page and the marked page in their access patterns generated using the technique described in Section 2. Pages accessed after the target page are shown in the upper half of the circle; pages accessed before the target page are shown in the lower half of the circle.

ID	Sequence
01	A, B, C, D
02	A, B
03	A, B
04	B, C
05	C, B

Figure 4: Example of user access patterns

First, we extract the "USER ID" and "user access pattern" from LOG DB. Second, we generate a Suffix Tree [16] from user access pattern. For example, the suffix tree shown in Figure 5 is generated from data shown in Figure 4. Third, we extract a sub tree starting from the target page selected by an analyst, then calculate user backward paths from the suffix tree.

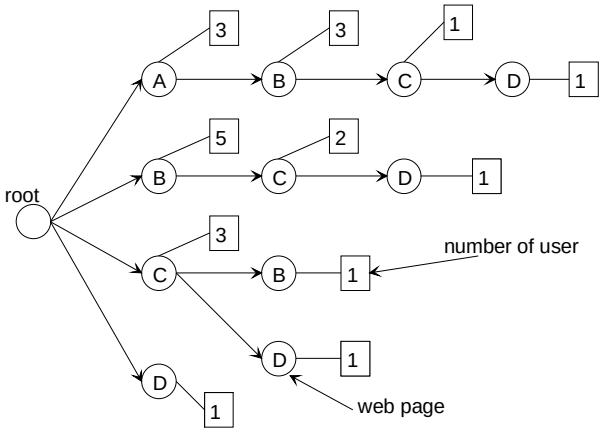

Figure 5: Suffix Tree

Figure 6 represents the results in the case of B selected as the target page. This figure shows that 60% of the users visited B via A , and that 20% user visited D via a page from B. The user path length is 2 from the target page to D.

Considering user behavior to and from the target page, the number of visited pages increases in proportion to the distance from the target page, and the ratio of the number of users who visited the pages decreases. From the polar coordinate system, the analyst can easily inspect pages distanced from the target page, because the area representing visited pages increase in proportion to distance from the target page. This tool enables the analyst to find the characteristic path at a glance.

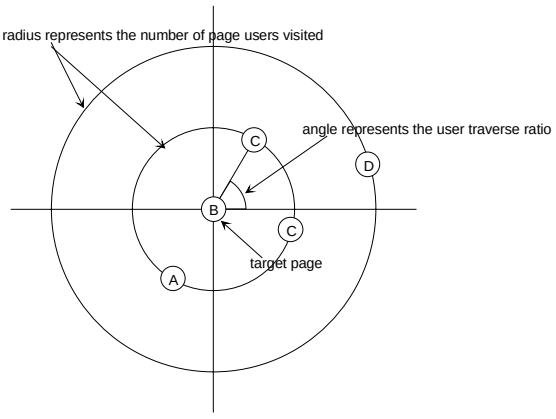

Figure 6: Polar coordinate system for user path visualization

4. Case study

In this section, we describe and discuss some cases that indicate the advantages (and disadvantages) of our technique.

4.1 Case 1: Overall analysis of web sites

We applied our hyperlink Topological Analysis to actual web sites. In Figures 7, the graph at left shows data for the Fuji Xerox web site; that at right shows the data for an online brokerage. The horizontal line indicates average value of link distance. Another line is a fit to the plot using least squares regression, where we can observe a negative correlation, as expected. In the graph at left, there are isolated points in right center (y = 4). This area indicates many people accessing page pairs in the same session, but they had to click many times. So, these page pairs cost users too much effort. An intuitive solution is to create a new hyperlink between them, or to shorten the path.

Comparing the two graphs, the graph at right seems to be well organized, because there are less isolated points in the right center area than in graph at left. This means that the online brokerage visitors could access page pairs having high co-occurrence, using fewer clicks that the Fuji Xerox visitors needed. One reason is that all web pages have detailed menu at both the top and the left using frames in the online brokerage web site. Therefore, the user can access to the pages they want to visit with the fewer clicks. Further, users of the online brokerage's web site have the distinct purpose of trading.

To the best of our knowledge, there are no tools to indicate the overall performance of the web sites in terms of the effectiveness of user traverse. If the effectiveness of overall web site is clear, the analyst can know which pages they should investigate to improve site effectiveness.

Case 2: Hyperlink topological problem

We examined page pairs in the upper right area in Figure 2. As described above, they have long path length but co-occur in visits. The dashed boxes in Figure 8 show an example of these pairs in the online brokerage site.

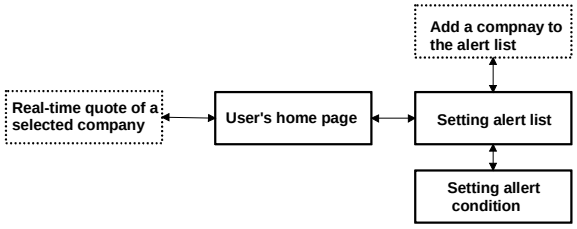

Figure 7 :Web Site's structures

In Figure 8, each box represents a web page and two pages connected by an arrow are hyperlinked to each other. Two dotted boxes represent the discovered page pair. The value of link distance was 3, and the value of access co-occurrence was 0.65.

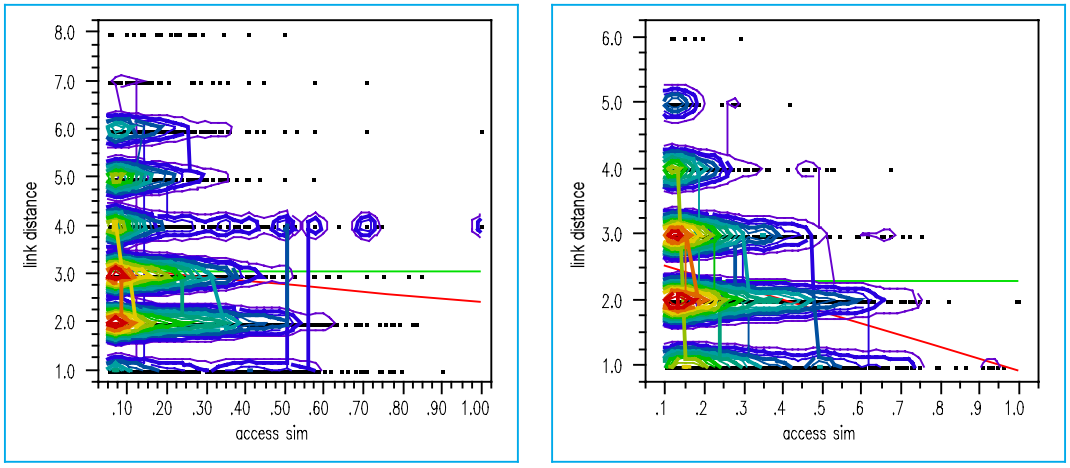

Figure 8: Example of analysis: At left, the Fuji Xerox web site; at right, the online brokerage web site. Since patterns were no readily identifiable in the original plot due to marker density, we used quantile density contours at 5% interval (These graphs were generated by JMP (SAS Institute, Inc:))

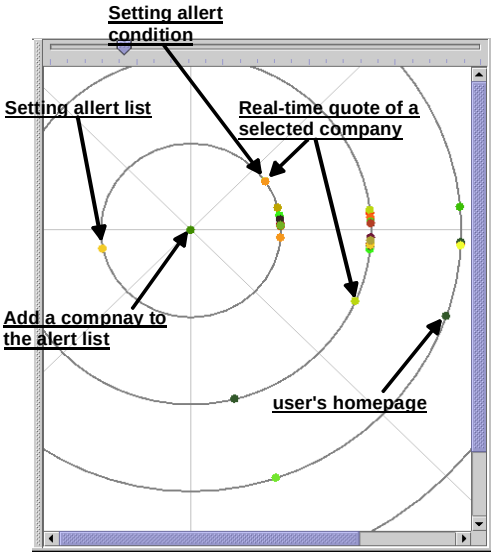

Figure 9: User access paths in the polar view

This web site provides a customized homepage for each user, and users can add their favorite company to the alert list shown in their homepage. The alert list has two functions as follows.

- The user can check whether quotes on companies have reached the expected price.
- The user can trade stock of listed companies with a single click

To add a company to the alert list, the users must follow the link referring to the *Setting alert list* page and then select the link referring to the *Add a company to the alert list* page. If the user desires the web site to inform them that their favorite stock has reached the price they expect, they can set the alert condition.

Figure 9 is the result of user access pattern visualization when we selected the *Add a company to the alert list* page as the target. From this result, 13% of the users accessed the *Real-time quote* page before the target page, 22% accessed that page after the target page, and 22% also accessed the *Setting alert condition* page after the target page. This shows that many users need the information given in the *Real-time quote* page in order to use the alert list. These users may feel this inconvenient because they must return to their homepage to check the quote. Furthermore, impatient users might not use this function. That may be one reason why the ratio of users who use the alert list is less than the web site publisher expected.

To improve the effectiveness of this web site, the web site publisher should add hyperlinks referring to the *Real-time quote* page (as shown in Figure 10) or add the information in the Real-time quote page to the *Add a company to the alert list* and *Setting alert condition* pages (as shown in Figure 11). These solutions should reduce user disorientation.

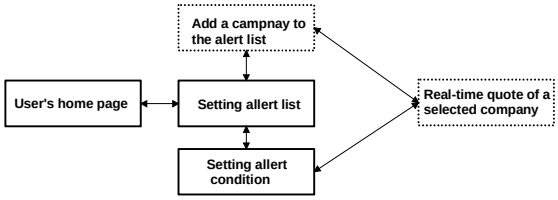

Figure 10:Redesign by means of adding hyperlinks.

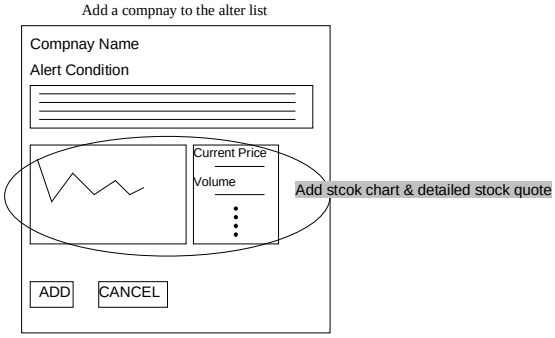

Figure 11:Redesign by means of adding the information in the *Real –time quote* page

This case study shows the advantage of our tool for a detailed analysis of the gap as well as gap discovery. By using our system, analysts can visually analyze user access paths to identify problem pages and find clues to improving their web site.

In this web site, each web pages are generated dynamically on the basis of user profiles and template files. However, the roles of web page (user's home page, real-time quote, etc) are not changed. Thus, we analyzed user accesses to template files.

4.2 Case 3: Concept Analysis

Concept analysis revealed a concept scoring high correlation coefficient between the inter-page conceptual relevance and the user's traversing ratio, and another concept scoring a low value. Figure 12 shows the site structure of these concepts. When we applied the concept analysis, the value of the similarity threshold was 0.3 and the value of path length threshold was 5. In this figure, each box represents a web page and an arrow represents a hyperlink. In this figure, there are two concepts and these concepts describe the identical product. According to the pages which user visited before, user is led to the different pages (Page1, Page2).

The concept starting from page1 received a low correlation value. In this concept, page 1 has high conceptual similarity with "Press release", "Usabil-

ity", "Copy function", but has a relatively low similarity with "Spec and price". Using access path visualization, we found that 28% of users visited "Space & price" and the ratio of users who visited other pages was lower. This is the reason why this concept received a low correlation value. In this topic, the web site publisher wanted to inform user the function and usability, but users wanted to know only the spec and price. We found that this concept did not work as the publisher expected. Our technique also gives a low score to those concepts which are accessed in the same way as this concept. This case study shows the advantage of our tool to evaluate web pages at a more abstract level, and to discover a concept which should be redesigned.

On the other hand, the concept starting from page2 scored a high correlation value. In this concept, most users accessed page2 only. Since page2 includes the information given in the "Spec & price" page, visitors seem to have achieved their goal by visiting page2. However, this correlation value does not indicate the gap between the web site publisher expectations and user behavior because the correlation value scored a high value by accident for the low traverse ratios. Further, our concept analysis technique cannot extract the web site publishers expectations when they designed the page so that they could satisfy users' request with the single page. In another concept, our technique also scored a high value since most users who visited the start page of the concept did not visit its other pages.

5. Related Work

Techniques to improve the effectiveness of web sites can be broadly classified as either user-side assistance or publisher-side assistance. The former helps users browse the hypertext structure effectively. For example, Footprints[18] visualizes past user behavior to help the user navigate effectively. Lieberman[8] tracks user browsing behavior to anticipate items of interest by exploring the hyperlinks from the user's current position. Publisher-side assistance presented in this paper, on the other hand, helps the web site publisher construct an effective site structure.

Discovering user behavior trend is essential to assist the web site publisher in understanding user preferences for further redesign activities. In the research fields, many discovery techniques have been developed. For example, Chen et al. [4] convert access log data a set of consecutive references to find frequent navigation patterns. The Borges and Levene [2] model accesses log data in the form of a directed graph in which nodes correspond to pages and arcs to hyperlinks. This graph is weighted and the weight of an arc represents probabilities that reflect user interaction with the web site. Further, association rules developed in the data mining field [6] are modified to extract navigation patterns from the graph. WUM [13] extracts interesting navigation patterns in terms of statistical and /or structural properties represented by the language MINT.

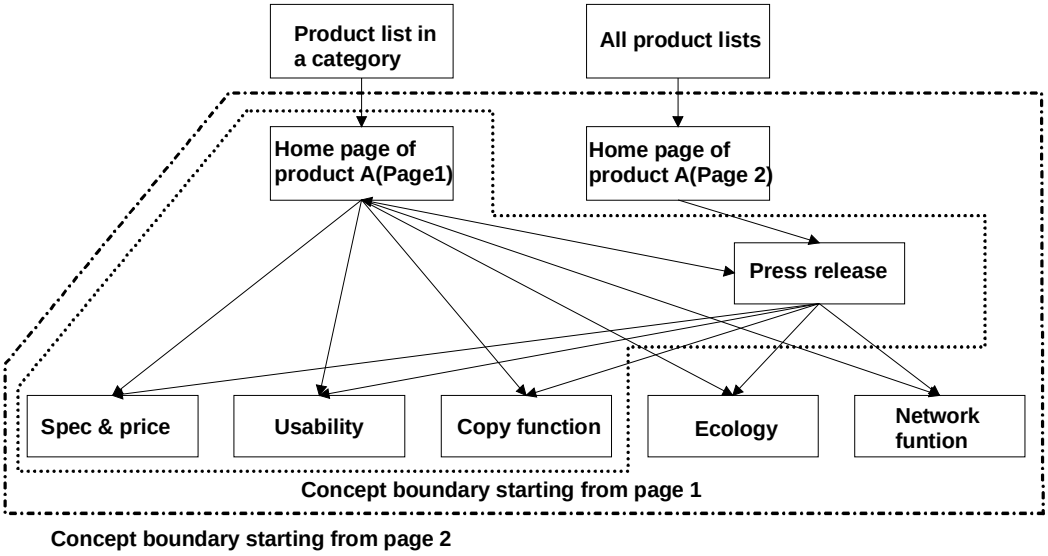

Figure 12: Hyperlink structure of two concepts

These techniques basically attempt to find access paths that a number of users have followed. However such navigation patterns only indicate how a web site of current design is being used. Thus, the web site publisher still needs to interpret these patterns to improve site quality. To assist in interpretation of frequent navigation patterns, some sort of feature that can characterize these patterns must be employed. The use of user classification is one approach, in which frequent navigation patterns are interpreted differently on the basis of importance of the user. Spiliopoulou et al. [14] compare navigation patterns of customers (web site users who have purchased something) with those of non-customers. This comparison leads to rules on how site topology should be improved to turn non-customers into customers. Another approach uses general shopping steps in an online store. Gomory et al. [7] suggest *micro-conversion rates* to evaluate the merchandising effectiveness of an online store and a visualization technique based on starfield model. Visualization helps publishers apprehend content efficiency. However, these techniques have only been applied to online stores.

Another approach uses hyperlink connectivity. Perkowitz and Etzioni [10] find clusters of pages that tend to co-occur in visits but that are not connected. For each cluster, an index page consisting of hyperlinks to the pages in the cluster is generated. In this way, more effective traversal between these pages is achieved. Their approach is similar to ours, except that their approach lacks techniques for evaluating infrequent navigation patterns. In other words, they extract no navigation patterns that should be frequent. Our techniques introduce the concept of gap analysis to discover pages that are less attractive.

6. Conclusion and Future Work

We have presented two techniques for discovering gaps between web site publisher expectations and user behavior. In the hyperlink topological analysis tool, publisher expectations are assessed by path length, and user behavior by measuring inter-page access co-occurrence. Further, plotting the both on the same graph reveals the gap, and we have shown the graph helps web site publisher find web site problems. In the concept analysis tool, publisher expectations are assessed by concept and user behavior by the ratio of user traverse. This method is effective for web site that contain product advertisements, but is not effective when the web site does not have a set of pages which describe some topics, such the site of an online-trading company. After finding the problem area by these methods, the web site publisher analyzes user access paths in detail by means of a visualization tool and finds the clues to improving their web sites.

In the hyperlink topological analysis, backward moves are not taken into account. Therefore an analyst cannot know whether user use direct path or undirected path (use back button, etc). In the case study of 4.1, the detailed menu seemed to bring the online brokerage site better result. In Fuji Xerox web site, analysts can judge whether some pages should be added such an menu to link each other, if they know users use many back button. The access path visualization technique helps analysts scrutinize this question.

In the concept analysis, topic extraction does not work as well as expected. One reason is that our technique does not have any mechanism to specify the starting page of a topic. Therefore, our technique extracts concepts which started from “press release”, etc. Another reason is that the disadvantage of our technique employs the vector space model for the measurement of publisher expectations. To accurately reflect publisher expectations, we plan to use metadata. If web site publishers utilize metadata, by which means page content is explicitly described, a topic can be extracted accurately. The evaluation of gap by the correlation coefficient does not work well in some cases. We consider that here, too, metadata will prove indispensable when we evaluate gaps.

The readability of our visualization tool is directly affected by out-degree in each page. In the case of Fuji Xerox Web site, the average out-degree is about 5. In recent study of structural properties of web graph (Broder et al [3]), most (over 80%) pages have 5 or less out-links. In our case study, we could read the result of our analysis with the restriction of radius = 5(ie, max path length = 11). Furthermore, we could effectively analyze user behavior under this restriction.

Nakayama et al [9] presented a technique how to apply quantitative data obtained through a multiple regression analysis that predicts hyperlink traversal frequency from page layout features. By combining this technique with the presented tools, the web site publisher is provided suggestions on layout improvement to modify user behavior, which they found as the origin of gap by the access path visualization tool.

7. Acknowledgements

This study was conducted as part of the Fuji Xerox Document Diagnosis project led by Toru Tanaka. We also gratefully acknowledge helpful suggestions by Yoshihiro Ueda.

8. References

[1] Accrue. http://www.accrue.com .

[2] J. Borges and M. Levene. Mining Association Rules in Hypertext Databases, in: Proc. the 4th International Conference on Knowledge Discovery and Data Mining, 1998.

[3] A.Broder, R.Kumar, F.Maghoul, P.Raghavan, S.Rajagopalan, R.Stata, A.Tomkins, and J Wiener. Graph Structure in the web, in: Proc WWW9, 2000.

[4] M. S. Chen, J. S. Park, and P. S. Yu. Data Mining for Path Traversal Patterns in a Web Environment, in: Proc. the 16th IEEE International Conference on Distributed Computing Systems, 1996.

[5] M. P. Etgen, and J.Cantor. What Does Getting WET (Web Event-logging Tool) Mean for Web Usability? 5th Conference on Human Factors & the Web, 1999.

[6] U. M. Fayyad, G. Piatetsky-Shapir, P. Smyth, and R. Uthurusamy ed. ADVANCES IN KNOWLEDGE DISCOVERY AND DATA MINING.AAAI press, 1996.

[7] S. Gomory, R. Hoch, J. Lee, M. Pdolaseck, and E. Sconberg. Analysis and Visualization of Metrics for Online Merchandising, in Proc WebKDD'99, 1999.

[8] H. Lieberman. Letizia: An Agent That Assists Web Browsing, in: Proc. the International Joint Conference on Artificial Intelligence, 1995.

[9] T. Nakayama, H. Kato, Y. Yamane. Discovering the Gap Between Web Site Designers' Expectations and Users' Behavior, in: Proc WWW9, 2000.

[10] M. Perkowitz and O. E tzioni. Adaptive Web Sites: Automatically Synthesizing Web Pages, in: Proc. the 15th National Conference on Artificial Intelligence, 1998.

[11] G. Salton. Developments in Automatic Text Retrieval, Science, Vol.253, 1991.

[12] C. Shahabi, A.M.Zarkesh, J.Adibi and V.Shah. Knowledge Discovery from users web-page navigation, in: Proceedings of the 7th International Workshop on Research Issues in Data Engineering, 1997.

[13] M. Spiliopoulou and L. C. Faulstich. WUM: A Tool for Web Utilization Analysis. In EDBT Workshop WebDB'98, Springer Verlag. extended version in LNCS 1590,1998.

[14] M. Spiliopoulou, C. Pohle, and L. C. Faulstich. Improving the Effectiveness of a Web Site with Web Usage Mining, in: Proc. WebKDD'99, 1999.

[15] SurfAid. http://surfaid.dfw.ibm.com/home

[16] E. Ukkonen. On-Line Construction of Suffix Trees,Algorithmica 14, pp246-260,1995.

[17] WebTrends. http://www.webtrends.com

[18] A. Wexelblat and P. Maes. Footprints: History-Rich Web Browsings, in Proc RIAO, pp75-84, 1997.