International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 10 Issue 1 –OCTOBER 2014.

A Well Structured Website Design with Efficient Navigation

Abstract- Developing a well designed website is a challenging task. Although creating a site structure is normally the role of an information architect, the reality is that everybody from designers to website owners find themselves working on it. The web developers are able to design based on their own capabilities nevertheless of end user point of view. Although the navigation over the web pages is simplified by re-linking the web pages with hyperlinks, the modified structure is unpredictable. Here we provide the means to improve the website structure without actually modifying the overall page structure. Instead of re-linking the pages and confusing the users we introduce a new mathematical programming model which improves the user navigation effectively and allows subsequent changes without disturbing the current page structure. With experimental study and simulations over this work we state that this model will provide effective user navigation with limited modifications.

Index Terms— Web Pages, Hyperlinks, user navigation, web mining, mathematical programming

I. INTRODUCTION

Although most of us would agree when it comes to judging a website by its layout design, it's almost a certainty that any entrepreneur wouldn't disagree that having a good layout design is essential to its success. As much as you would love to visit a website with cool design, every online customer would also like to explore online sites that are visually attractive. In addition to its visual appeal, websites that were carefully designed have better features and that's the product of good planning by the entrepreneur- website owner. We tend to easily fall in love with such sites which is a good thing for those who are investing on good website layout designs. On the other hand, online sites with bad designs usually get the downside of their investment.

They normally find it hard to rise from the ground so to speak due to its unattractive layout. A primary cause of poor website design is that the web developers' understanding of how a website should be structured can be considerably different from those of the users. Such differences result in cases where users cannot easily locate the desired information in a website. This problem is difficult to avoid because when creating a website, web developers may not have a clear understanding of users' preferences and can only organize pages based on their own judgments. However, the measure of website effectiveness should be the satisfaction of the users rather than that of the developers. Thus, Webpages should be organized in a way that generally matches the user's model of how pages should be organized. Previous studies on website has focused on a variety of issues, such as understanding web structures finding relevant pages of a given page, mining informative structure of a news website, and extracting template from webpages. Our work, on the other hand, is closely related to the literature that examines how to improve website navigability through the use of user navigation data. Various works have made an effort to address this question and they can be generally classified into two categories to facilitate a particular user by dynamically reconstituting pages based on his profile and traversal paths, often referred as personalization, and to modify the site structure to ease the navigation for all users, often referred as transformation. In this paper, we are concerned primarily with transformation approaches. The literature considering transformations approaches mainly focuses on developing methods to completely reorganize the link structure of a website. Approaches, their drawbacks are obvious. First, since a complete reorganization could radically change the location of familiar items, the new website may disorient users. Second, the reorganized website structure is highly unpredictable, and the cost of disorienting users after the changes remains unanalyzed. This is because a website's structure is typically designed by experts and bears business or organizational logic, but this logic may no longer exist in the new structure when the website is completely reorganized. Besides, no prior studies have assessed the usability of a completely reorganized website, leading to doubts on the applicability of the reorganization approaches. Finally, since website reorganization approaches could dramatically change the current structure, they cannot be frequently performed to improve the navigability. Recognizing the drawbacks of website reorganization approaches, we address the question of how to improve the structure of a website rather than reorganize it substantially. Specifically, we develop a mathematical programming (MP) model that facilitates user navigation on a website with minimal changes to its current structure. Our model is particularly appropriate for informational websites whose contents are static and relatively stable over time. Examples of organizations that have informational websites are universities, tourist attractions, hospitals, federal agencies, and sports organizations. Our

model, however, may not be appropriate for websites that purely use dynamic pages or have volatile contents. This is because a steady state might never be reached in user access patterns in such websites, so it may not be possible to use the weblog data to improve the site structure. The number of outward links in a page, i.e., the out-degree, is an important factor in modeling web structure. Prior studies typically model it as hard constraints so that pages in the new structure cannot have more links than a specified out-degree threshold, because having too many links in a page can cause information overload to users and is considered undesirable. For instance, Lin uses 6, 8, and 10 as the out-degree threshold in experiments. This modeling approach, however, enforces severe restrictions on the new structure, as it prohibits pages from having more links than a specified threshold, even if adding these links may greatly facilitate user navigation. Our model formulates the out-degree as a cost term in the objective function to penalize pages that have more links than the threshold, so a page's out-degree may exceed the threshold if the cost of adding such links can be justified. We perform extensive experiments on a data set collected from a real website. The results indicate that our model can significantly improve the site structure with only few changes. Besides, the optimal solutions of the MP model are effectively obtained, suggesting that our model is practical to real-world websites. We also test our model with synthetic data sets that are considerably larger than the real data set and other data sets tested in previous studies addressing website reorganization problem. The solution times are remarkably low for all cases tested, ranging from a fraction of second to up to 34 seconds. Moreover, the solution times are shown to increase reasonably with the size of the website, indicating that the proposed MP model can be easily scaled to a large extent. To assess the user navigation on the improved website, we partition the entire real data set into training and testing sets. We use the training data to generate improved structures which are evaluated on the testing data using simulations to approximate the real usage. We define two metrics and use them to assess whether user navigation is indeed enhanced on the improved structure. Particularly, the first metric measures whether the average user navigation is facilitated in the improved website, and the second metric measures how many users can benefit from the improved structure. Evaluation results confirm that user navigation on the improved website is greatly enhanced.

II. RELATED WORK

The methods proposed by Mobasher et al. and Yan et al. create clusters of users profiles from weblogs and then dynamically generate links for users who are classified into different categories based on their access patterns. Nakagawa and Mobasher develop a hybrid personalization system that can dynamically switch between recommendation models based on degree of connectivity and the user's position in the site. For reviews on web personalization approaches, Web transformation, on the other hand, involves changing the structure of a website to facilitate the navigation for a large set of users instead of personalizing pages for individual users. Fu et al. describe an approach to reorganize webpages so as to provide users with their desired information in fewer clicks. However, this approach considers only local structures in a website rather than the site as a whole, so the new structure may not be necessarily optimal. Gupta et al. propose a heuristic method based on simulated annealing to relink webpages to improve navigability. This method makes use of the aggregate user preference data and can be used to improve the link structure in websites for both wired and wireless devices. However, this approach does not yield optimal solutions and takes relatively a long time to run even for a small website. Lin develops integer programming models to reorganize a website based on the cohesion between pages to reduce information overload and search depth for users. In addition, a two-stage heuristic involving two integer-programming models is developed to reduce the computation time. However, this heuristic still requires very long computation times to solve for the optimal solution, especially when the website contains many links.

Besides, the models were tested on randomly generated websites only, so its applicability on real websites remains questionable. To resolve the efficiency problem in, Lin and Tseng propose an ant colony system to reorganize website structures. Although their approach is shown to provide solutions in a relatively short computation time, the sizes of the synthetic websites and real website tested in are still relatively small, posing questions on its scalability to large-sized websites. There are several remarkable differences between web transformation and personalization approaches.

First, transformation approaches create or modify the structure of a website used for all users, while personalization approaches dynamically reconstitute pages for individual users. Hence, there is no predefined/built-in web structure for personalization approaches.

Second, in order to understand the preference of individual users, personalization approaches need to collect information associated with these users (known as user profiles). This computationally intensive and time-consuming process is not required for transformation approaches.

Third, transformation approaches make use of aggregate usage data from weblog files and do not require tracking the past usage for each user while dynamic pages are typically generated based on the users' traversal path. Thus, personalization approaches are more suitable for dynamic websites whose contents are more volatile and transformation approaches are more appropriate for websites that have a built-in structure and store relatively static and stable contents. This paper examines the questions of how to improve user navigation in a website with minimal changes to its structure. It complements the literature of transformation approaches

that focus on reconstructing the link structure of a website. As a result, our model is suitable for website maintenance and can be applied in a regular manner.

III. METRIC FOR EVALUATING NAVIGATION EFFECTIVENESS

The Metric

Our objective is to improve the navigation effectiveness of a website with minimal changes. Therefore, the first question is, given a website, how to evaluate its navigation effectiveness. Marsico and Levialdi point out that information becomes useful only when it is presented in a way consistent with the target users' expectation. Palmer indicates that an easy-navigated website should allow users to access desired data without getting lost or having to backtrack. We follow these ideas and evaluate a website's navigation effectiveness based on how consistently the information is organized with respect to the user's expectations. Thus, a well-structured website should be organized in such a way that the discrepancy between its structure and users' expectation of the structure is minimized. Since users of informational websites typically have some information targets i.e., some specific information they are seeking, we measure this discrepancy by the number of times a user has attempted before locating the target. Our metric is related to the notion of information scent developed in the context of information foraging theory. Information foraging theory models the cost structure of human information gathering using theanalogy of animals foraging for food and is a widely accepted theory for addressing the information seeking process on the web. Information scent refers to proximal cues (e.g., the snippets of text and graphics of links) that allow users to estimate the location of the "distal" target information and determine an appropriate path . Users are faced with a decision point at each page; they use information scent to evaluate the likely effort and the probability of reaching their targets via each link and make navigation decisions accordingly. Consequently, a user is assumed to follow the path that appears most likely to lead him to the target. This suggests that a user may backtrack to an already visited page to traverse a new path if he could not locate the target page in the current path. Therefore, we use the number of paths a user has traversed to reach the target as a proximate measure to the number of times the user has attempted to locate one target. We use backtracks to identify the paths that a user has traversed, where a backtrack is defined as a user's revisit to a previously browsed page. The intuition is that users will backtrack if they do not find the page where they expect it. Thus, a path is defined as a sequence of pages visited by a user without backtracking, a concept that is similar to the maximal forward reference defined in Chen et al. Essentially, each backtracking point is the end of a path. Hence, the more paths a user has traversed to reach the target, the more discrepant the site structure is from the user's expectation.

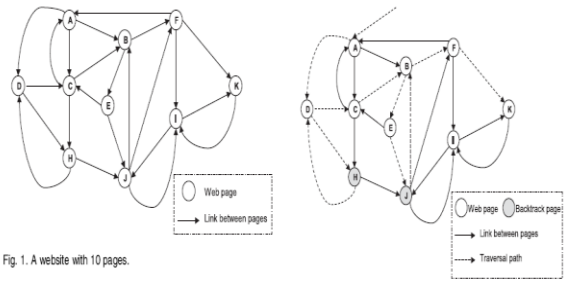

Fig. 1. A website with 10 pages.

IV. PROBLEM DESCRIPTION

Difficulty in navigation is reported as the problem that triggers most consumers to abandon a website and switch to a competitor. Generally, having traversed several paths to locate a target indicates that this user is likely to have experienced navigation difficulty. Therefore, Webmasters can ensure effective user navigation by improving the site structure to help users reach targets faster. This is especially vital to commercial websites, because easy navigated websites can create a positive attitude toward the firm, and stimulate online purchases, whereas websites with low usability are unlikely to attract and retain customers . Our model allows Webmasters to specify a goal for user navigation that the improved structure should meet. This goal is associated with individual target pages and is defined as the maximum number of paths allowed to reach the target page in a mini session. We term this goal the path threshold for short in this paper. In other words, in order to achieve the user navigation goal, the website structure must be altered in a way such that the number of paths needed to locate the targets in the improved structure is not larger than the path threshold. In the example shown in Fig. 2, the user has traversed three paths before reaching the target. An intuitive solution to help this user reach the target faster is to introduce more links There are many ways to add extra links. If a link is added from D to K, the user can directly reach K via D, and hence reach the target in the first path. Thus, adding this link "saves" the user two paths. Similarly, establishing a link from B to K enables the user to reach the target in the second path. Hence, this saves him one path. We could also insert a link from E to K, and this is considered the same as linking B to K. This is because both B and E are pages visited in the second path, so linking either one to K saves only one path. Yet, another possibility is to link C to F, a non target page. In this case, we assume that the user does not follow the new link, because it does not directly connect a page to the target. While many links can be added to improve navigability, our objective is to achieve the specified goal for user navigation with minimal changes to a website. We measure the changes by the number of new links added to the current site structure. There are several reasons that we should insert minimal links. First, minimizing changes to the

current structure can avoid disorienting familiar users. Second, adding unnecessary links can lead to pages having too many links, which increases users' cognitive loads and 0makes it difficult for them to read and comprehend. Third, since our model improves site structures on a regular basis, the number of new links should be kept at minimum such that the links in a website in the whole course of maintenance do not expand in a chaotic manner. There are cases where users could have reached the targets through existing links, but failed to do so in practice. One reason could be that these links are placed in inconspicuous locations and hence are not easily noticeable. Another reason might be that the labels of these links are misleading or confusing, causing difficulty to users in predicting the content at the target page. As a result, Webmasters should focus on enhancing the design of these existing links before adding new links. Our model considers this issue by placing a preference on the selection of such existing links.

V. COMPUTATIONAL EXPERIMENTS AND PERFORMANCE EVALUATIONS

Extensive experiments were conducted, both on a data set collected from a real website and on synthetic data sets. We first tested the model with varying parameters values on all data sets. Then, we partitioned the real data into training and testing data. We used the training data to generate improved site structures which were evaluated on the testing data using two metrics that are discussed in detail later. Moreover, we compared the results of our model with that of a heuristic.

5.1 Real Data Set

5.1.1 Description of the Real Data Set

The real data set was collected from the Music Machines website (http://machines.hyperreal.org) and contained about four million requests that were recorded in a span of four months. This data set is publicly available and has been widely used in the literature . Before analysis, we followed the log preprocessing steps described in to filter irrelevant information from raw log files. These steps include: 1) filter out requests to pages generated by Common Gateway Interface (CGI) or other server-side scripts as we only consider static pages that are designed as part of a website structure, 2) ignore unsuccessful requests (returned HTTP status code not 200), and 3) remove requests to image files (.gif, .jpg, etc.), as images are in general automatically downloaded due to the HTML tags rather than explicitly requested by users.

We utilized the page-stay time to identify target pages and to demarcate mini sessions from the processed log files. Three time thresholds (i.e., 1, 2, and 5 minutes) were used in the tests to examine how results changes with respect to different parameter values. Furthermore, we adapted the algorithm proposed in to identify the backtracking pages in mini sessions, which are then used to demarcate the paths traversed to reach the targets.

5.2 Synthetic Data Sets

In addition to the real data set, synthetic/artificial data sets were generated and considered for computational experiments to evaluate the scalability of our model with respect to the size of the website and the number of mini sessions. For this reason, the artificial website structures and mini sessions were generated to have similar statistical characteristics as the real data set. For instance, the average outdegree for pages in the real website is 15, so the link structure for the artificial website was generated in a way such that each page contained 15 links on average. Three websites consisting of 1,000, 2,000, and 5,000 webpages were constructed. Our approach for generating the link structure is similar to that described in. Particularly, to generate the link structure that contains 1,000 pages, we first constructed a complete graph of 1,000 nodes (pages) and each directed edge was assigned a random value between 0 and 1. Then, we selected the edges with the smallest 15,000 values to form the link structure, resulting in an average out-degree of 15 for this website. In a similar manner, we generated the link structures for other artificial websites. The mini sessions were generated in a slightly different manner. Specifically, we directly generated the set of relevant candidate links for each mini session instead of creating the user's traversal path. As a result, this allows us to directly apply the model on synthetic data sets. The sets of relevant candidate links in synthetic data sets has similar characteristics with those from the real one, comprising one to five relevant candidate links per each relevant mini session, with each link being randomly selected. Each of the three artificial websites was tested with 10,000, 50,000, 100,000, and 300,000 mini sessions, resulting in 12 different categories. In each category, three data sets were generated and the results were averaged over the three sets. We note that the synthetic data sets considered in this paper are significantly larger than those used in related papers, which report results based on synthetic data sets with at most 200 pages and 3,200 links The math programs for the synthetic data were coded in AMPL and solved using CPLEX/AMPL 11.1.1 on a PC running Windows 7 on a 3.4 GHz processor. We experimented the model with two out-degree thresholds, i.e., C ¼ 20 and C ¼ 40, and two multipliers for the penalty term, i.e., m ¼ 0 and m ¼ 5, on

Evaluation Results on Improved Website Using Number of Paths Per Mini Session for $T = 5$ Min

Multiplier for penalty term (m)	Avg. no. of paths in improved Web site and no. of new links needed (in parenthesis)					
	Out-degree threshold C=20			Out-degree threshold C=40		
	b=1	b=2	b=3	b=1	b=2	b=3
0	1.335	1.589	1.785	1.335	1.589	1.785
	(5,794)	(1,145)	(467)	(5,794)	(1,145)	(467)
1	1.346	1.632	1.815	1.349	1.650	1.827
	(5,794)	(1,166)	(482)	(5,813)	(1,214)	(502)
5	1.346	1.639	1.855	1.351	1.680	1.840
	(5,794)	(1,182)	(514)	(5,839)	(1,399)	(555)

each synthetic data set. Noticeably, the times for generating optimal solutions are low for all cases and parameter values tested, ranging from 0.05 to 24.727 seconds. This indicates that the MP model is very robust to a wide range of problem sizes and parameter values. Particularly, the average solution times for website with 1,000, 2,000, and 5,000 pages are 0.231, 1.352, and 3.148 seconds. While the solution times do go up with the number of webpages, they seem to increase within a reasonable range.

5.3 Evaluation of the Improved Website

In addition to the extensive computational experiments on both real and synthetic data sets, we also perform evaluations on the improved structure to assess whether its navigation effectiveness is indeed enhanced by approximating its real usage. Specifically, we partition the real data set into a training set (first three months) and a testing set (last month). We generate the improved structure using the training data, and then evaluate it on the testing data using two metrics: the average number of paths per mini session and the percentage of mini sessions enhanced to a specified threshold. The first metric measures whether the improved structure can facilitate users to reach their targets faster than the current one on average, and the second metric measures how likely users suffering navigation difficulty can benefit from the improvements made to the site structure. The evaluation procedure using the first metric consists of three steps and is described as follows:

1. Apply the MP model on the training data to obtain the set of new links and links to be improved.
2. Acquire from the testing data the mini sessions that can be improved, i.e., having two or more paths, their length, i.e., number of paths, and the set of candidate links that can be used to improve them.
3. For each mini session acquired in step 2, check whether any candidate link matches one of the links obtained in step 1, that is, the results from the training data. If yes, with the assumption that users will traverse the new link or the enhanced link in the improved structure, remove all pages (excluding the target page) visited after the source node of the matching candidate link to obtain the new mini session for the improved website, and get its updated length information.

VI. DISCUSSION

6.1 Mini Session and Target Identification

We employed the page-stay timeout heuristic to identify users' targets and to demarcate mini sessions. The intuition is that users spend more time on the target pages. Page-stay time is a common implicit measurement found to be a good indicator of page/document relevance to the user in a number of studies. In the context of web usage mining, the page-stay timeout heuristic as well as other time-oriented heuristics are widely used for session identification, and are shown to be quite robust with respect to variations of the threshold values. The identification of target pages and mini sessions can be affected by the choice of page-stay timeout threshold. Because it is generally very difficult to unerringly identify mini sessions from anonymous user access data, we ran our experiments for different threshold values. Generally, increasing the threshold will result in fewer mini sessions with proportionally more mini sessions having a large number of paths while decreasing the threshold will have the opposite effect (see Table 7). In other words, increasing time threshold would decrease the number of relevant mini session for small path thresholds but could increase the number of relevant mini sessions for large path thresholds. As a result, we observed that as the time threshold increase, the number of new links decreases for b ¼ 1 and 2, but increases for b ¼ 3. While the results did change slightly, we showed that our model succeeded in finding the minimal number of links that can be used to improve user navigation substantially.

6.2 Choice of Parameter Values for the Model

6.2.1 Path Threshold

The path threshold represents the goal for user navigation that the improved structure should meet and can be obtained in several ways. First, it is possible to identify when visitors exit a website before reaching the targets from analysis of weblog files. Hence, examination of these sessions helps make a good estimation for the path thresholds. Second, surveying website visitors can help better understand users' expectations and make reasonable selections on the path threshold values. For example, if the majority of the surveyed visitors respond that they usually give up after traversing four paths, then the path threshold should be set to four or less. Third, firms like comScore and Nielsen have collected large amounts of client-side web usage data over a wide range of websites. Analyzing such data sets can also provide good insights into the selection of path threshold values for different types of websites. Although using small path thresholds could result in more improvements in web user navigation in general, our experiments showed that the changes (costs) needed increase significantly as the path threshold decreases.

6.4.2 Out-Degree Threshold

Webpages can be generally classified into two categories index pages and content pages. An index page is designed to help users better navigate and could include many links, while a content page contains information users are interested in and should not have many links. Thus, the out-degree threshold for a page is highly dependent on the purpose of the page and the website. Typically, the outdegree threshold for index pages should be larger than that for content pages. For instance, out-degree thresholds are set to 30 and 10 for index and content pages, respectively, in the experiments in. Since out-degree thresholds are context dependent and organization dependent, behavioral and experimental studies that examine the optimal outdegree threshold for different settings are needed. In general, the out-degree threshold could be set at a

small value when most webpages have relatively few links, and as new links are added, the threshold can be gradually increased. Note that since our model does not impose hard constraints on the out-degrees for pages in the improved structure, it is less affected by the choices of out-degree thresholds as compared to those in the literature.

VII. CONCLUSIONS

In this paper, we have proposed a mathematical programming model to improve the navigation effectiveness of a website while minimizing changes to its current structure, a critical issue that has not been examined in the literature. Our model is particularly appropriate for informational websites whose contents are relatively stable over time. It improves a website rather than reorganizes it and hence is suitable for website maintenance on a progressive basis. The tests on a real website showed that our model could provide significant improvements to user navigation by adding only few new links. Optimal solutions were quickly obtained, suggesting that the model is very effective to realworld websites. In addition, we have tested the MP model with a number of synthetic data sets that are much larger than the largest data set considered in related studies as well as the real data set. The MP model was observed to scale up very well, optimally solving large-sized problems in a few seconds in most cases on a desktop PC. To validate the performance of our model, we have defined two metrics and used them to evaluate the improved website using simulations. Our results confirmed that the improved structures indeed greatly facilitated user navigation. In addition, we found an appealing result that heavily disoriented users, i.e., those with a higher probability to abandon the website, are more likely to benefit from the improved structure than the less disoriented users. Experiment results also revealed that while using small path thresholds could result in better outcomes, it would also add significantly more new links. Thus, Webmasters need to carefully balance the tradeoff between desired improvements to the user navigation and the number of new links needed to accomplish the task when selecting appropriate path thresholds. Since no prior study has examined the same objective as ours, we compared our model with a heuristic instead. The comparison showed that our model could achieve comparable or better improvements than the heuristic with considerably fewer new links. The paper can be extended in several directions in addition to those mentioned in Section 6. For example, techniques that can accurately identify users' targets are critical to our model and future studies may focus on developing such techniques. As another example, our model has a constraint for out-degree threshold, which is motivated by cognitive reasons. The model could be further improved by incorporating additional constraints that can be identified using data mining methods. For instance, if data mining methods find that most users access the finance and sports pages together, then this information can be used to construct an additional constraint.

ACKNOWLEDGMENTS

The authors would like to thank the editors and the anonymous reviewers for their insightful comments and helpful suggestions, which have resulted in substantial improvements to this work.

REFERENCES

[1] Pingdom, "Internet 2009 in Numbers," http://royal.pingdom. com/2010/01/22/internet-2009-in-numbers/, 2010.
[2] J. Grau, "US Retail e-Commerce: Slower But Still Steady Growth," http://www.emarketer.com/Report.aspx?code=emarketer_ 2000492, 2008.
[3] Internetretailer, "Web Tech Spending Static-But High-for the Busiest E-Commerce Sites," http://www.internetretailer.com/ dailyNews.asp?id = 23440, 2007.
[4] D. Dhyani, W.K. Ng, and S.S. Bhowmick, "A Survey of Web Metrics," ACM Computing Surveys, vol. 34, no. 4, pp. 469-503, 2002.
[5] X. Fang and C. Holsapple, "An Empirical Study of Web Site Navigation Structures' Impacts on Web Site Usability," Decision Support Systems, vol. 43, no. 2, pp. 476-491, 2007.
[6] J. Lazar, Web Usability: A User-Centered Design Approach. Addison Wesley, 2006.
[7] D.F. Galletta, R. Henry, S. McCoy, and P. Polak, "When the Wait Isn't So Bad: The Interacting Effects of Website Delay, Familiarity, and Breadth," Information Systems Research, vol. 17, no. 1, pp. 20- 37, 2006.
[8] J. Palmer, "Web Site Usability, Design, and Performance Metrics," Information Systems Research, vol. 13, no. 2, pp. 151-167, 2002.
[9] V. McKinney, K. Yoon, and F. Zahedi, "The Measurement of Web-Customer Satisfaction: An Expectation and Disconfirmation Approach," Information Systems Research, vol. 13, no. 3, pp. 296- 315, 2002.
[10] T. Nakayama, H. Kato, and Y. Yamane, "Discovering the Gap between Web Site Designers' Expectations and Users' Behavior," Computer Networks, vol. 33, pp. 811-822, 2000.
[11] M. Perkowitz and O. Etzioni, "Towards Adaptive Web Sites: Conceptual Framework and Case Study," Artificial Intelligence, vol. 118, pp. 245-275, 2000.
[12] J. Lazar, User-Centered Web Development. Jones and Bartlett Publishers, 2001.
[13] Y. Yang, Y. Cao, Z. Nie, J. Zhou, and J. Wen, "Closing the Loop in Webpage Understanding," IEEE Trans. Knowledge and Data Eng., vol. 22, no. 5, pp. 639-650, May 2010.
[14] J. Hou and Y. Zhang, "Effectively Finding Relevant Web Pages from Linkage Information," IEEE Trans. Knowledge and Data Eng., vol. 15, no. 4, pp. 940-951, July/Aug. 2003.
[15] H. Kao, J. Ho, and M. Chen, "WISDOM: Web Intrapage Informative Structure Mining Based on Document Object Model," IEEE Trans. Knowledge and Data Eng., vol. 17, no. 5, pp. 614-627, May 2005.
[16] H. Kao, S. Lin, J. Ho, and M. Chen, "Mining Web Informative Structures and Contents Based on Entropy Analysis," IEEE Trans.Knowledge and Data Eng., vol. 16, no. 1, pp. 41-55, Jan. 2004.