LinkGuide: Towards a Better Collection of Hyperlinks in a Website Homepage

A. Ammari and V. Zharkova
School of Informatics, University of Bradford
anammari@Bradford.ac.uk, v.v.zharkova@Bradford.ac.uk

***ABSTRACT*—A dramatic and continuous increase in the complexity and size of websites on the Internet makes rather difficult to build such websites with required information to be easily found. This study addresses an intelligent approach to design and implement the homepages of the various types of websites; such as commercial, portal, and search engine websites. In particular, the study aims to study the problem of how from a large set of hyperlinks available in those websites, to recommend a small optimal collection of hyperlinks, leading to the required information with the maximum speed and least efforts. The objective of this hyperlink recommendation is to maximize the efficiency, effectiveness, and utilization of the homepage of a Website. As a result, LinkGuide – a Web Mining based framework, is proposed to solve this hyperlink recommendation problem. The study discusses an overview of the design status of LinkGuide, and presents an evaluation criterion that will be used to measure the performance of the proposed framework.**

***Index Terms*—Hyperlink Recommendation, LinkGuide, Web Mining.**

I. INTRODUCTION

As the size and complexity of websites expands dramatically, it has become increasingly challenging to design websites where the Internet users can easily find information they seek. To address this challenge, we formally define a research problem in this area: Hyperlink Recommendation, and then propose a web mining-based framework: LinkGuide, as a solution.

There are the two major ways that the Internet users adopt to find the information they seek on the World Wide Web [5]: either by using search engines, or by clicking on hyperlinks. Research on the search engines focuses on improving recall and precision [5], [12], [13]. Our research, however, concentrates on improving the efficiency of the hyperlink – based searching approach. Normally, Web users click on a series of hyperlinks in order to find the information they seek. Hence, recommending appropriate hyperlinks for each one of them in the web pages they browse is crucial to improving their search efficiency and reaching the required webpage. In particular, our research focuses on placing the appropriate hyperlinks in the home page of a website, the entrance gate to a website.

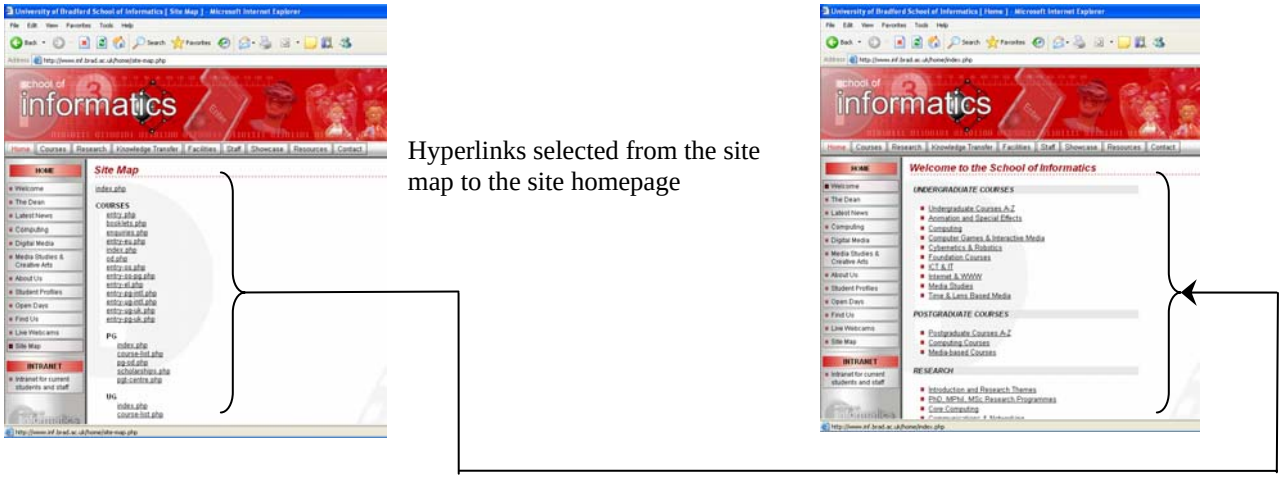

Figure 1: The hyperlink pool (left) and the homepage (right) of the University of Bradford's School of Informatics

A website homepage consists of hyperlinks selected from a hyperlink pool which is a set of hyperlinks pointing to top-level web pages [8]. Usually, the hyperlink pool of a website consists of hyperlinks listed in the site map page or the site index page. As shown in Fig. 1, hyperlinks in the homepage of the University of Bradford School of Informatics website

(http://www.inf.brad.ac.uk/home/index.php) are selected from its hyperlink pool. The hyperlink pool consists of hyperlinks in its site map page (http://www.inf.brad.ac.uk/home/site-map.php).

Given the website design principle that scrolling must be avoided in homepages [15], a well-designed homepage usually contains around fifty hyperlinks. However, the hyperlink pool of a typical Web site has at least several hundred hyperlinks. For example, the homepage of the University of Bradford School of Informatics website consists of 58 hyperlinks, while the hyperlink pool has 248 hyperlinks. It is computationally too expensive to exhaust all combinations of several dozen hyperlinks from a hyperlink pool with several hundred hyperlinks and find the one that is the most efficient in guiding Web surfers to the information they need. In the School of Informatics example, the number of combinations of selecting 58 hyperlinks from 248 hyperlinks is 2.28E +57.

In comparison, our proposed hyperlink recommendation framework, LinkGuide, incorporates both patterns extracted from the structure of a website and those discovered from the web server log file, which records valuable information about the behaviour of the website users. LinkGuide first employs web structure mining to discover the hyperlinks that lead to easily find the biggest number of top-level webpages in the site. Secondly, LinkGuide employs web mining techniques [10] to extract hidden patterns about hyperlink preferences in the website based on the navigational behaviour of the website users.

The rest of the paper is organized as follows: In the next section we review related work in Web Mining techniques from the literature. In section three we formally define the Hyperlink Recommendation problem based on our definitions of the effectiveness, efficiency, and utilization of a homepage; and we give an overview of a Web – Mining based framework, LinkGuide, as a solution; and in section four we conclude our paper and discuss the future work of LinkGuide implementation.

II. Related Work

Web mining is defined as a process of discovering and analyzing useful information from the Web. The web data is classified into content data, structure data, and usage data [20]. As a result, for each type of the data, a corresponding web mining method was developed. Web content mining is the process of automatically retrieving, filtering, and categorizing Web documents. A good web content mining survey can be found in [5]. As Web content mining typically makes use of only texts on web pages, valuable information implicitly contained in hyperlinks is overlooked. In contrast, web structure mining [4] infers useful patterns from the Web's link topology to help retrieve high quality web pages. HITS [9] and PageRank [3] are two widely used web structure mining algorithms. Web usage mining [20] is the process of applying data mining techniques to discover web access patterns from web log files. A web log file is a collection of the data that explicitly records the information related to users' searching behaviour in a web site. Useful attributes for the web usage mining in a web log include IP address, time, and URL, which explicitly describe who, at what time, accessed which Web page. Additional attributes include status of a HTTP request and the amount of bytes returned by a Web server.

Research in web usage mining can be classified into two groups: general purpose projects and specific-purpose projects. General purpose projects, such as [6] and [7], focused on web usage mining in general. An architecture and specific steps for web usage mining are proposed in [7] that presented a method to identify potentially interesting patterns in user preferences from the mining results (e.g., patterns in which unlinked web pages are visited together frequently). A new data mining capability to mine path traversal patterns from web logs is explored in [6]. Specific purpose projects focused on applications of web usage mining. Web usage mining can be used to improve organizations of websites. Adaptive website project [16] used web visiting patterns, learned from web logs, to automatically improve the organization and presentation of websites. Web usage mining to measure and improve the success of websites is exploited in [19]. Web usage mining can also be used to personalize users' web surfing experience. In [21], clusters of visitors who exhibited similar information needs (e.g., visitors who accessed similar Web pages) were discovered via web usage mining. These clusters could be used to classify new visitors and dynamically suggest hyperlinks for them. The techniques to learn user preferences from web usage data using data mining techniques, such as association rule mining are presented in [14]. Based on the learned preferences, dynamic hyperlinks could be recommended for active visiting sessions. The algorithm MINPATH is developed in [1] for automatic suggestions of useful shortcut links in real time in order to improve a wireless web navigation.

Since our work is mainly aimed to develop a web recommender approach, research on the recommendation systems is much related to this work. Content-based filtering and collaborative filtering are approaches to realizing recommendation systems. The content-based filtering approach has its root in information retrieval [17]. Recommendation systems the use this approach, such as NewsWeeder [11], recommended objects to a user based on the comparison between the contents of the objects and the

user's profile. Recommendation systems the use the collaborative filtering approach, such as Ringo [18], recommended objects to a user because other users with similar tastes liked these objects. The Fab system described in [2] is based on the combination of the content-based filtering approach and the collaborative filtering approach.

Research that employ only usage information, such as the work described in [6], did not consider the information contained in the structure of a web site, which could be valuable in extracting the hyperlink set that lead to the largest number of important pages in a website. On the other side, [8] gave high preferences to each of every two structurally linked hyperlinks if those hyperlinks are heavily accessed together in the same session by the same user. Such a high rank for the initial hyperlink that leads to the terminal hyperlink can be exaggerated simply because the user may not need the initial hyperlink at all if the terminal hyperlink was placed in the same page of that of the initial hyperlink.

Research in Web structure mining that combines the Web content and Web structure information is also related to our study. For example, HITS values web pages that point to many other pages that offer good contents. Our framework, LinkGuide, also values the pages that point to many other relevant (good) pages. However, good pages in LinkGuide are measured using the web usage information, such a frequency that the pages are visited together. LinkGuide considers both structurally related and unrelated pages.

III. Overview of the LinkGuide Framework

A. Problem Definition

It was established [8] that a web homepage is to be designed to meet the following conditions: (a) can effectively facilitate web surfers in locating the information stored in the website, (b) has a limited number of hyperlinks placed in the homepage, and (c) is frequently visited by web surfers when they search for information in the Web site. Therefore, three evaluation metrics to measure the quality of a web homepage are adopted in our study: **effectiveness**, **efficiency**, and **utilization**, which address the three previous points (a), (b), and (c). All three metrics should be calculated from Web log files. A web log file can be broken down into sessions with each session representing a sequence of consecutive web accesses by the same visitor. For the convenience of readers, important notations used in this study are defined in Table 1.

Effectiveness of a homepage is measured as the degree of ease in finding user-sought, top-level pages from the homepage. We adopt the definitions of [8] for the effectiveness of a single web session (1) and the effectiveness of a whole web log file (2) as follows:

$$effectiven\,ess(S_j) = \frac{\left|UHL(S_j) \cap HL\right|}{\left|UHL(S_j)\right|} \quad (1)$$

$$effectiven\,ess(\log) = \frac{\sum_{j=1}^{s} effectiven\,ess(S_j)}{s} \quad (2)$$

The efficiency of a homepage is considered [8] to be the number of hyperlinks the user clicked in the homepage and in the level two pages (pages that are pointed by homepage links) to the total number of hyperlinks that exist in the homepage. We see that this definition leaves a possibility that homepage efficiency may still be very high when the user clicks only links that exists in level two pages without being interested in hyperlinks in the homepage itself. To provide more accurate reflection to the efficiency of the homepage itself, we modified the efficiency definition to be the number of hyperlinks the user clicked only in the homepage to the total number of hyperlinks that exist in the homepage. The modified session – level efficiency (3) and log file – level efficiency (4) are defined as follows:

$$efficiency\,(S_j) = \frac{\left|UHL(S_j) \cap PHL\right|}{\left|PHL\right|} \quad (3)$$

$$efficiency\,(\log) = \frac{\sum_{j=1}^{s} efficiency\,(S_j)}{s} \quad (4)$$

Utilization of a homepage measures how often a homepage is visited. A well designed Web site homepage should attract users to utilize it for finding what they seek in other pages. Reference [8] measured the utilization by counting the number of user-sought, top-level web pages that can be easily found from the homepage. This definition includes the hyperlinks exist not only in the homepage, but also in the web pages that are directly pointed to by the homepage. To give a more accurate reflection to the utilization of the homepage itself of a website, utilization in this study will be measured by counting the number of user-sought, top-level Web pages that directly exist in the homepage, as follows:

$$utilizatio\,n(\log) = \sum_{j=1}^{S} \left|UHL(S_j) \cap PHL\right| \quad (5)$$

We formally define the Hyperlink Recommendation problem as follows:

Given: (1) The hyperlink pool (H) of a web site; (2) The number of hyperlinks to be placed in the homepage of the website (N), where: $N < |H|$.
Process: Recommend N hyperlinks from the hyperlink pool H to include in the homepage.
Objective: Maximize the effectiveness, efficiency, and utilization of the homepage.

Table 1: Summary of Notations used in the study

Notation	Description
S	the number of sessions in a Web log
S_j	a session, for $j = 1, 2, \ldots, s$
H	the hyperlink pool of a Web site
UHL (S_j)	user-sought, top-level Web pages in S_j
l	the number of hyperlinks in a Web site
L_j	a hyperlink, for $j = 1, 2, \ldots, l$
PLj	the set of hyperlinks in the Web page pointed to by ***Lj***
PHL	the set of hyperlinks in a Web homepage
EHL	the set of hyperlinks in the Web pages pointed to by ***PHL***
HL	the set of hyperlinks in both ***PHL*** and ***EHL*** $= PHL \cup EHL$
N	the number of hyperlinks to be placed in the homepage
effectiveness (S_j)	$effectiven\ ess(S_j) = \frac{\lvert UHL(S_j) \cap HL \rvert}{\lvert UHL(S_j) \rvert}$
effectiveness **(log)**	$effectiven\ ess(\log) = \frac{\sum_{j=1}^{s} effectiven\ ess(S_j)}{s}$
efficiency (S_j)	$efficiency(S_j) = \frac{\lvert UHL(S_j) \cap PHL \rvert}{\lvert PHL \rvert}$
efficiency **(log)**	$efficiency(\log) = \frac{\sum_{j=1}^{s} efficiency(S_j)}{s}$
utilization **(log)**	$utilizatio\ n(\log) = \sum_{j=1}^{S} \lvert UHL(S_j) \cap PHL \rvert$

B. The LinkGuide Framework

There is a need to develop a Web Mining – based solution to solve the hyperlink recommendation problem with the least possible computational cost. We propose a Web Mining – Based framework, LinkGuide, to solve the hyperlink recommendation problem. LinkGuide will mainly incorporate data patterns that are extracted from two main types of resources: Data patterns extracted from the structure of a website (hyperlinks), and data patterns extracted from the log file of the website. The first type of data patterns represents structure data patterns (The hyperlink hierarchy of a website), which requires a web structure mining approach to be extracted. The second type of data patterns represents usage data patterns (patterns stored in web log files that describe web surfers' navigational behaviour), which requires a web usage mining approach to be extracted. Therefore, our proposed mining framework, LinkGuide, will be designed and implemented to combine the techniques of two main web mining methods: web structure and web usage mining.

Fig. 2 outlines a flow chart of the LinkGuide framework. LinkGuide will first extract structural patterns from the hyperlink pool of the website by parsing the pages that are pointed by all the hyperlinks in the site map. This will generate a list of hyperlink pairs that are structurally linked. Each pair will contain two hyperlinks, ***Li*** and ***Lj***, ***Li*** is the initial hyperlink and ***Lj*** is the terminal hyperlink in this structure relationship. LinkGuide will then apply an association rules mining technique to web log files to determine hyperlink pairs that are heavily accessed together. A heavily accessed hyperlink pair will have a support value that is larger than a predefined threshold [8].

Based on the two previously defined relationships between hyperlinks in a website: Structure Relationship, and Access Relationship, LinkGuide will categorize hyperlink pairs into the following main four groups:

1. Hyperlink pairs in Group A: This group of hyperlinks indicates that both a structure relationship and an access relationship hold between two hyperlinks. Hyperlink relationships in this group will be considered in our approach.
2. Hyperlinks in Group B: This group of hyperlinks indicates that an access relationship, but not a structure relationship, exists between two hyperlinks. Hyperlink relationships in this group will be considered in our approach.

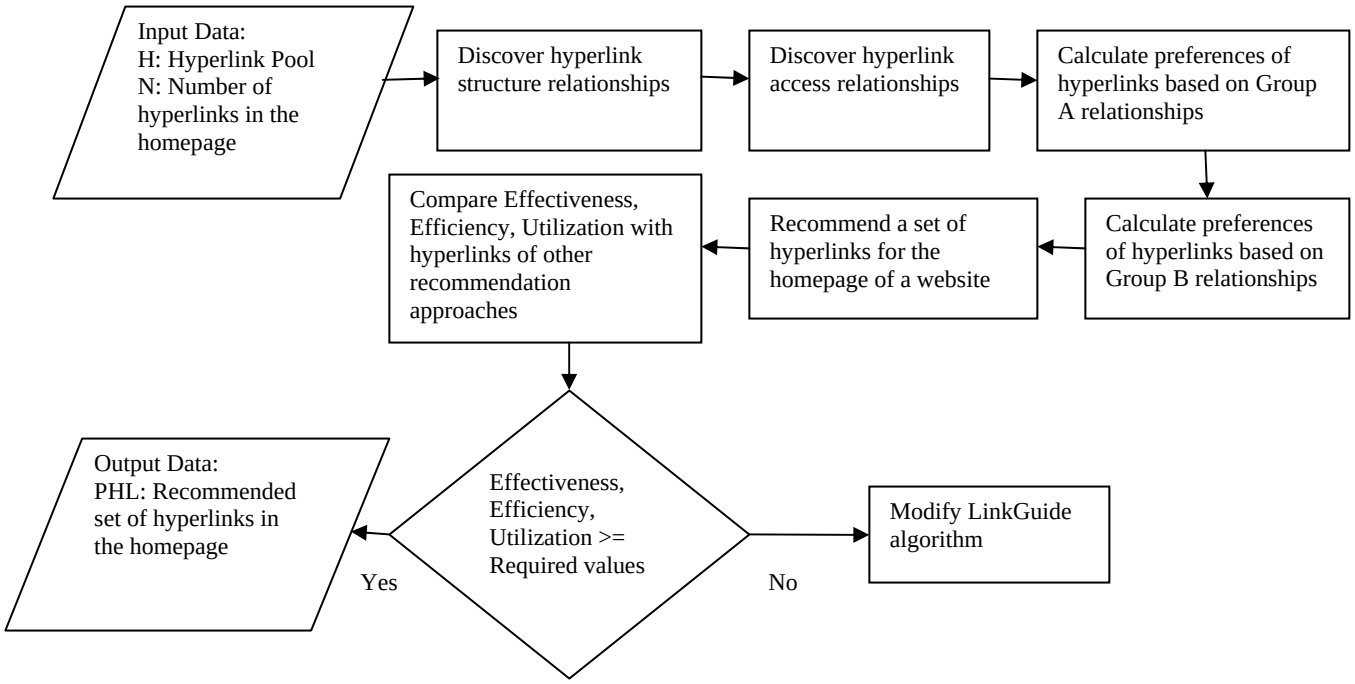

Figure 2: An abstract Flow Chart of LinkGuide

3. Hyperlinks in Group C: This group of hyperlinks indicates that a structure relationship, but not an access relationship, exists between two hyperlinks. This group reveals that the web page pointed to by the initial hyperlink in a structure relationship contains a rarely-visited hyperlink, which is the terminal hyperlink, in the structure relationship. Hence, a Group C relationship reveals a design problem with internal pages. As hyperlink recommendation focuses on choosing hyperlinks for a website homepage; our approach will not consider hyperlink relationships in this group.
4. Hyperlinks in Group D: This group of hyperlinks indicates that there is neither a structure relationship, nor an access relationship that exist between any two hyperlinks. Our approach will not consider hyperlink relationships in this group.

IV. CONCLUSION AND FUTURE WORK

In this paper, we have formally defined a hyperlink recommendation problem and propose the web mining – based framework, LinkGuide, as a solution. LinkGuide applies the two web mining techniques. Web structure mining is used to extract patterns related to how hyperlinks are connected in the site map of the website, whereas web usage mining is used to extract patterns related to the navigational behaviour of the users of the website. Implementation of the LinkGuide framework is needed as a future vindication of the study. For the experimental data, we will perform the experiments on the University of Bradford School of Informatics website since it is large and sufficiently generates web log files to allow us to apply the web usage mining process and do comparisons of different hyperlink recommendation approaches.

REFERENCES

[1] Anderson, C., Domingos, P., And Weld, D. 2001. "Adaptive Web navigation for wireless devices," in *Proceedings of the 17th International Joint Conference on Artificial Intelligence*, Seattle, WA, 879–884.

[2] Balabanovic, M. And Shoham , Y. 1997. "Fab: Content-based collaborative recommendation," in *Commun. ACM* 40, 3, 66–72.

[3] Brin, S. And Page, L. 1998. "The anatomy of a large-scale hypertextual Web search engines," in *Proceedings of the 7th International World Wide Web Conference*, Brisbane, Australia, April.

[4] Chakrabarti, S., Dom, B. E., Kumar, S. R., Raghavan, P., Rajagopalan, S., Tomkins, A., Gibson D., And Kleinbirg, J. 1999. "Mining the Web's link structure," in *IEEE Comput*. 32, 8, 60–67.

[5] Chakrabarti, S. 2000. "Data mining for hypertext: a tutorial survey," in *ACM SIGKDD Explorations* 1, 2, 1–11.

[6] Chen, M., Park, J., And Yu, P. 1996. "Data mining for path traversal patterns in a Web environment," in *Proceedings of the 16th International Conference on Distributed Computing Systems*, Hong Kong, China, May.

[7] Cooley, R., Tan, P., And Srivastava, J. 1999, "WebSIFT: the Web site information filter system," in *Proceedings of*

the Web Usage Analysis and User Profiling Workshop, San Diego, CA, August.

[8] Fang, X. And Sheng, O. 2004. "LinkSelector: A Web Mining Approach to Hyperlink Selection for Web Portals," in *ACM Transactions on Internet Technology*, Vol. 4, No. 2, May 2004, Pages 209–237.

[9] Kleinberg J. 1998. "Authoritative sources in a hyperlinked environment," in *Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms*, San Francisco, CA, Jan.

[10] Kosala, R. And Blockeel, H. 2000. "Web mining research: a survey," in *ACM SIGKDD Explorations* 2, 1, 1–15.

[11] Lang, K. 1995. NewsWeeder "Learning to filter net news," in *Proceedings of the 12th International Conference on Machine Learning*, Lake Tahoe, CA.

[12] Lawrence, S. And Giles, C. L. 1998. "Searching the World Wide Web," in *Science* 280, 98–100.

[13] Lawrence, S. And Giles, C. L. 1999. "Accessibility of information on the Web," in *Nature* 400, 107–109.

[14] Mobasher, B., Cooley, R., And Srivastava J. 2001. "Automatic personalization based on web usage mining," in *Comm. ACM* 43, 8, 142–151.

[15] Nielson, J. 1999. "User interface directions for the Web.Comm," ACM 42, 1, 65–72.

[16] Perkowitz, M. And Etzioni, O. 2000. "Towards adaptive web sites: conceptual framework and case study," Artif. Intell. 118, 1–2, 245–275.

[17] Salton, G. 1968. "Automatic Information Organization and Retrieval," McGraw-Hill Inc.

[18] Shardanand, U. And Maes, P. 1995. "Social information filtering: algorithms for automating word of mouth," in *Proceedings of ACM CHI'95*, Denver, CO, May, 210–217.

[19] Spiliopoulou, M. And Pohle, C. 2001. "Data mining to measure and improve the success of web sites," in *Data Mining Knowl. Disc.* 5, 1–2, 85–114.

[20] Srivastava, J., Cooley, R., Deshpande, M., And Tan, P. 2000. "Web usage mining: discovery and applications of usage patterns from web data," in *SIGKDD Explorations* 1, 2, 1–12.

[21] Yan, T., Jacobsen, M., Garcia-molina, H., And Dayal, U. 1996. "From user access patterns to dynamic hypertext linking," in *Proceedings of the 5th International World Wide Web Conference*, Paris, France, May.