LILAC — Learn from Internet: Log, Annotation, and Content

Tingshao Zhu
School of Information Science and Engineering
Graduate University of Chinese Academy of Sciences
Beijing, P.R. China 100080

Russ Greiner
Department of Computing Science
University of Alberta
Canada T6G 2E8

Bin Hu
School of Information Science and Technology
Lanzhou University
Lanzhou, P.R. China 730000

Abstract

This paper summarizes an user study designed to evaluate various models of how users browse the web while working on their day-to-day tasks, in their office or at home. We use these models to predict which pages contain information the user will find useful, and provide empirical data that these learned models are effective.

1 Introduction

While the World Wide Web contains a vast quantity of information, it is often difficult for web users to find the information they want. Nowadays, most users use search engines to find useful information on the web, but search engines can only work if the users have intuitions about what keywords will cause the search engine to produce the relevant information. Unfortunately, users are not always able to find the right query to locate the exact pages that satisfies their information need. We call a page that the user must examine to accomplish her current task an "information content" page, or "IC-page" for short.

There are now many recommender systems to help web users, based on technologies such as Association Rules (Agrawal and Srikant 1994), Sequential Patterns (Agrawal and Srikant 1995; Mannila, Toivonen, and Verkamo 1997) and Collaborative Filtering (Resnick et al. 1994; Herlocker et al. 2004). However, most existing web recommendation systems direct users to web pages, from within a single web site, that other similar users have visited. Our system *W*ebIC 3 differs from these other recommendation systems as it suggests IC-pages that may appear anywhere on the Web.

To do this, *W*ebIC identifies the context: which web pages relate to the user's current information need. We assume this corresponds to all of the pages since the start of the user's current browsing session, which we call the current "IC-session", which *W*ebIC estimates based on the sequence of pages the user has visited and the actions the user has applied to the pages (e.g., clicking the back button, following a hyperlink, etc.). Then, we extract the "browsing behavior" attributes of each word w encountered in the user's current IC-session - e.g., how often each word appears in the title of a page in this IC-session, or in the anchor of a link that was followed, etc.

The performance system will then use this information about a word to determine if it is important, and will then form a query based on the important words.

We considered 3 criteria, and used each to produce a labeled data set, consequently we trained three kinds of models : IC-word, IC-Relevant, and IC-Query.

IC-word For each word w in an IC-session, we compute each of these "browsing behavior" features, and also indicate whether w appears in the IC-page (i.e., IC-word) or not. We then train a decision tree to predict which words are IC-words given only their browsing properties.

IC-Relevant During the evaluation stage, the user is asked to pick out only those words that are relevant to their search task. The word list is produced based on all the words in the current session. Only those words that are chosen by the user will be labelled as "relevant", and a decision tree is trained to learn the browsing properties of these words.

IC-Query In the IC-word model, all the words in IC-page are IC-words. But obviously, some of them are just general words, and therefore less relevant to the page content. To learn the IC-Query model, we don't use all the IC-words, but only those words that can be shown to correctly locate the page by querying a search engine using these keywords in the query. We label these words as IC-Query words. To obtain these IC-Query words, we run a Query Identifier (QI) on each IC-page p, which will generate a set of words that can locate p by querying a search engine. The trained decision tree will predict the most likely words which could locate the IC-page by querying a search engine (e.g., Google.com).

According to the labelled log data, on average 3% of the pages visited were marked as IC-pages. To deal with this imbalanced dataset, we make use of a down-sampling technique (Ling and Li 1998) since it can produce a more accurate classifier, (i.e., decision tree C4.5 (see (Quinlan 1992))), than by over-sampling techniques (Japkowicz 2000).

We have implemented the browsing feature extraction and trained browsing behavior models (Zhu, Greiner, and Häubl 2003b; 2003a), and the empirical results are encouraging. The purpose of the LILAC(Learn from the Internet : Log, Annotation, Content) study is to collect "annotated web logs", and to investigate the effectiveness of various learning approaches for our recommendation task. That is, while standard web logs record which sites a user visits, we need to know more: in particular, we need to know which pages supplied some relevant information content - i.e., qualified as an IC-page. *W*ebIC records where each participant visits, and also his or her annotation, indicating which pages contained information content.

In this paper, we summarize the user study that we have conducted to evaluate our browsing behavior models. Section 2 first discusses the experiment design, including the motivation and process management. Section 3 then describes *W*ebIC developed for this study, and how it was used in LILAC. Section 4 presents empirical results from LILAC. Section 5 finally summarize the lessons that we learn from LILAC and improvements for future user studies.

2 Experiment Design

This study, LILAC, aims at evaluating the browsing behavior models. The focus of the study is to gather data from people working on their day-to-day tasks, in their office or at home.

There are four models used in LILAC, Followed Hyperlink Word (FHW) which is used as a baseline model, IC-word, IC-Relevant, and IC-Query. The idea of FHW is to collect those words found in the anchor text of the followed hyperlinks in the session, and the similar concept — Inferring User Need by Information Scent (IUNIS), is presented in (Chi et al. 2001). As such, there is no training involved in this model. In all models, words are stemmed and "stop" words are removed prior to prediction. For all treatments, a user browses the web independently and may at some point either mark a page as an IC-page or request a "suggestion" from the model.

To guarantee the consistency of *W*ebIC and these model files, *W*ebIC forces a check for a new version of *W*ebIC, both on startup and during every hour of use. If a new version is available, *W*ebIC will upload the current user data, and then download the new model files, as well as the latest *W*ebIC, and finally exit. On startup, *W*ebIC also checks for user data files that are more than one week old and forces an upload if they are found, and then proceeds to download the models. Whenever new models have been trained and are available, *W*ebIC will be updated to force users to get the new model files and upload their data at regular intervals. Fig. 1 shows how the participators interacted with LILAC server.

To participate in the LILAC study, the subjects needed to install *W*ebIC (in Section 3) on their own computer, and browse their own choice of non-personal English language web pages. *W*ebIC will keep track of all the interactions — storing a record of the pages the user visited, as well as the evaluation data for the pages that they considered as relevant.

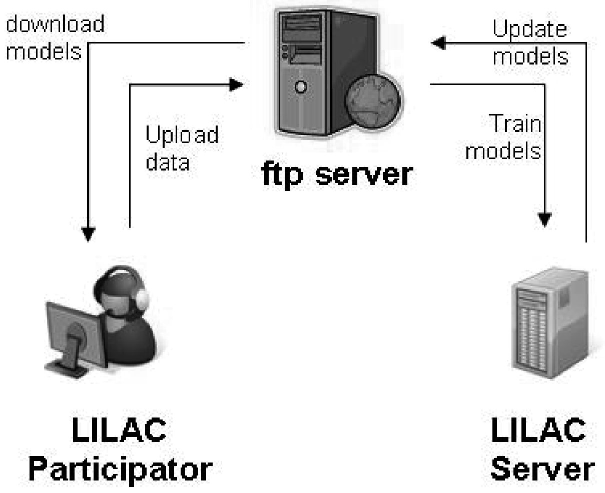


Figure 1: Interaction between Participators and LILAC Server

Whenever the subject discovers an IC-page, they are instructed to label this page as IC-page by clicking on the "MarkIC" button. Here, *W*ebIC will ask the subject to compare this page to an alternative page that it suggests using one of several models. In the case that the subject couldn't find a page that addresses their information need, they have the option of clicking on the "Suggest" button, which retrieves a recommended page for their review. As before, the subject will be asked to rate the usefulness of the suggested page with respect to their current information needs. In either case, the entire rating process should take no more than a few seconds to complete. Section 3.3 provides more details of the evaluation process.

There are five distinct steps involved in the LILAC study.

Pre-Test The main purpose of the pre-test is to make *W*ebIC more stable. As such, local colleagues were recruited to participate in the pre-test, using *W*ebIC for 2 hours each week, and reporting any bugs and any suggestions that would improve the usability of *W*ebIC.

Pilot Study After the subjects have confirmed their participation in the study, we then randomly selected 10 subjects for a one-week pilot study. The purpose of the pilot study is to evaluate *W*ebIC and test the interface mechanisms before we launch the regular study. The comments and feedback from the pilot study were very helpful in identifying any potential problems or issues prior to the following study.

LILAC Study The five-week regular study is designed to qualitatively assess the different browsing behavior based models strength relative to a baseline model, and to gather data for further studies (e.g., recommendation, personalization, etc.). Ideally, we want to demonstrate the effectiveness of our models work on arbitrary pages taken from arbitrary sites on the Web. As such our goal is to test our models by actual users working on day-to-day tasks, no matter where they are and what they are working on.

Follow-up Survey The goals of the follow-up survey are 1) To improve our understanding of the hypothesis that a users browsing actions are informative about their needs; 2) To gain an assessment of how significant this source of information is relative to other sources of information; 3) To test the usefulness of our assumptions about how a users needs, page content and browsing actions can be represented and how relationships between these representations can be expressed; and 4) To evaluate how well *W*ebIC worked with real users during unrestricted browsing of the World Wide Web.

Subjects will be issued a unique ID number at the time of enrollment in the experiment. The ID will be entered into the *W*ebIC browser. All data uploaded from subject's computers will be stored under this identifier. LILAC administrator keeps a separate database for recording the correspondence between identifiers and subject names for the purposes of compensation. The researchers do not need and will not have access to this data and the administration will not have access to the web logs. All log data was treated as strictly confidential.

3 *W*ebIC — An Effective Complete-Web Recommender System

*W*ebIC is an IE-based browser, as shown in Fig. 2, and it has some extensions to facilitate the user study.

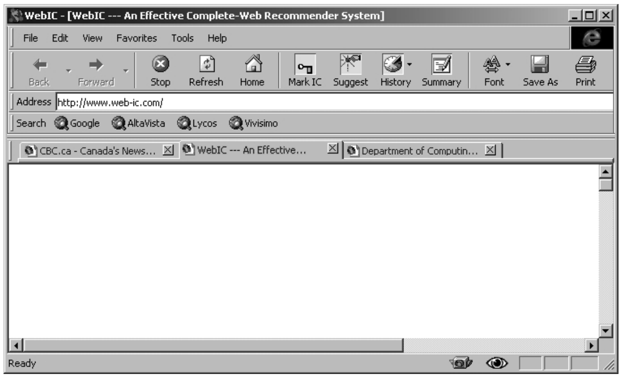


Figure 2: *W*ebIC — An Effective Complete-Web Recommender System

*W*ebIC computes browsing features for essentially all of the words that appear in any of the observed pages, and then use the model to predict the user's current information need. *W*ebIC then sends an appropriate query to a search engine (e.g., Google) to generate an appropriate IC-page. The query is selected from the predicted information need using the trained models.

3.1 Annotation

There are two purposes of the "Annotation" in *W*ebIC:

1. Just as AIE (Zhu, Greiner, and Häubl 2003b), by distributing *W*ebIC to people for their ordinary web browsing, we can collect annotated web logs which can be used to learn general user model or community specified model.

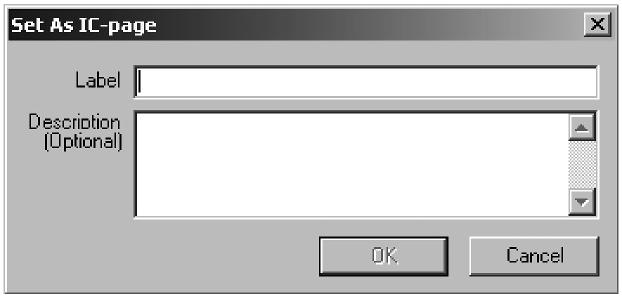


Figure 3: Annotation of *W*ebIC

2. Collect the log data for one specific user which can be used to build the personalized behavior model.

In *W*ebIC, we not only record the URL and time stamp of the browsing, but also the source of the web pages that the user has visited. For the web frame, we download all the involved frame pages. The reason we make the exact snapshot of the user's browsing is that if only the URL is recorded, maybe later the content will be changed or unavailable. We want to record the very accurate page sequence that the user has seen during the browsing. We also record the action sequence as well, that is, how the user reaches this page, such as, following the hyperlink in previous page, or type in the Address, etc..

To annotate the IC-page, after examining each page, if the visited page qualifies as an IC-page, the user can click the "MarkIC" button in *W*ebIC, and if so, must give it a label and optionally produce a short summary specifying why it qualifies — e.g., "this page introduces all the tourist locales of Beijing". Fig. 3 shows the pop-up window to allow the user to input the label and description of the IC-page.

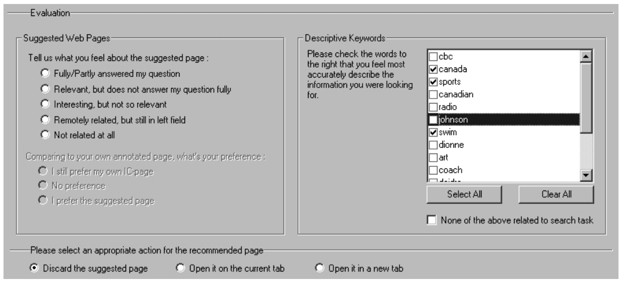
3.2 Recommending

For the recommendation generation, after watching the user's click stream (without annotation): First, extract the browsing features of the words in the current session, then apply the extracted patterns (either from individual or population models) to generate a synthesized query which is subsequently processed by a search engine (e.g., Google.com). Since the model is built on the "browsing behavior" of the words, and not the words themselves, it can be used in any web environment. As such, *W*ebIC can predict the useful pages no matter where the user is or what they are working on.

3.3 Evaluation

Here, whenever the user requests a recommendation by clicking the "Suggest" button, *W*ebIC will select one of its models randomly to generate a recommendation page, as shown in Figure 4. As one of the goals of the LILAC study is to evaluate our various models, this version of *W*ebIC will therefore ask the user to evaluate this proposed page.

Another goal of LILAC is to collect annotated web logs for future research; we therefore instructed these paid participants to click "MarkIC" whenever they found a page



Evaluation

Suggested Web Pages

Tell us what you feel about the suggested page :

○ Fully/Partly answered my question
○ Relevant, but does not answer my question fully
○ Interesting, but not so relevant
○ Remotely related, but still in left field
○ Not related at all

Comparing to your own annotated page, what's your preference :

○ I still prefer my own IC-page
○ No preference
○ I prefer the suggested page

Descriptive Keywords

Please check the words to the right that you feel most accurately describe the information you were looking for.

□ cbc
☑ canada
☑ sports
□ canadian
□ radio
□ johnson
☑ swim
□ dionne
□ art
□ coach

Select All | Clear All

□ None of the above related to search task

Please select an appropriate action for the recommended page

● Discard the suggested page ○ Open it on the current tab ○ Open it in a new tab

Figure 4: Evaluation Interface for LILAC

they consider to be an IC-page. After marking an IC-page, *W*ebIC will recommend an alternative web page as before (excluding the IC-page), as if the user had clicked "Suggest" here. Once again, *W*ebIC will then ask the user to evaluate this recommended page.

In order to evaluate the recommendation, *W*ebIC will ask the user to provide feedback in several key areas.

First, the user is instructed to "Tell us what you feel about the suggested page" in order to indicate whether the information provided on the page suggested by *W*ebIC was relevant to his/her search task. There are two categories of relevance evaluations: *related* and *not related at all*. We further divided the *related* category into four different levels, including "Fully answered my question", "Somewhat relevant, but does not answer my question fully", "Interesting, but not so relevant", and "Remotely related, but still in left field".

In the case of identifying an IC-page, *W*ebIC will suggest an alternative page, and ask the user to evaluate how the content of this page compares to the original IC-page. To accomplish this, the user needs to examine both the suggested page and their own IC-page, and they will be asked the question "Comparing to your own IC-page, what's your preference?". The user must select from one of the following three options: *I still prefer my own IC-page*, *No preference*, and *I prefer the suggested page*.

Third, the user will be asked to select informative "Descriptive Keywords" from a short list of words that *W*ebIC predicted as relevant words. The information collected here will be used to train predicting models as described in Section 4.

Finally, the user will be asked to "Please select an appropriate action for the recommended page" from one of three choices: *Discard the suggested page*, *Open it on the current tab*, and *Open it in a new tab*. Analysis of these data will allow us to evaluate the user's impression of the suggested page.

4 LILAC Study

As stated above, the goal of LILAC is to determine whether the browsing behavior models are able to recommend pages that add value to user's browsing experience. To convert the raw log data to be suitable for training models, several steps must be done for preprocessing (Zhu, Greiner, and Häubl 2003a). Briefly the process requires that we first identify IC-sessions, then extract the "browsing features" of each word in the session, and compute its "label", and finally train different models to generate IC-pages.

4.1 LILAC Subjects

A total of 104 subjects participated in the five-week LILAC study, of which 97 resided in Canada, and 7 resided in the USA. Table 1 gives the detailed information for all LILAC subjects.

Canada		
	Edmonton	82
	Montreal	7
	Toronto	3
	Longueuil	2
	Calgary, St. Alberta, Beaumont	1 (each)
USA		
	PA, AL	2 (each)
	CA, NC, MA	1 (each)

Table 1: Location of LILAC Subjects

After the study, the subjects were required to provide more personal information to process their payment. 98 of the subjects provided the required information to get paid. Among them, 47% are female and 53% are male. The age distribution of the subjects is shown in Table 2.

Range	Number
$18 - 20$	23
$21 - 25$	41
$26 - 30$	20
$31 - 35$	11
$over 35$	3

Table 2: Age of LILAC Subjects

The subjects are representative in that they are from different places across North America, and have different experience of using the Web. This will help us test our models in a more meaningful way, and make the results more promising.

4.2 Browsing Behavior

To extract browsing features, we consider all words that appear in IC-session, removing stop words and stemming (Porter 1980), then compute 35 attributes for each word((Zhu, Greiner, and Häubl 2003b)). Note that when we train the models, we do not use the words themselves, but instead just these browsing features values derived from the words.

4.3 Empirical Results

In LILAC, a total of 93,443 Web pages were visited. Over this period of time, the users marked 2977 IC-pages and asked for recommendations by clicking the "Suggest" button 2531 times.

At the conclusion of the study, we collect the evaluation results of these IC-models and the baseline model, and the overall result is shown in Figure 5. The bars show the relative percentage of the five evaluation responses for each model.

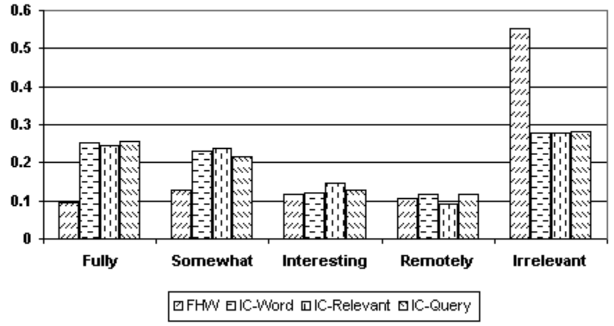


Figure 5: Overall Results of LILAC-1

From Figure 5, we conclude that each of the three different IC-models perform better than the baseline model, i.e., FHW. This result validates our basic assumption that we are able to provide useful recommendations by integrating the user's browsing behaviors into the prediction.

4.4 Follow Up Survey

The goal of the follow-up survey is to improve our understanding of *W*ebIC's hypothesis that user browsing actions are informative about user needs, to gain an assessment of how significant this source of information is relative to other sources of information, to test the usefulness of our assumptions about how user needs, page content and browsing actions can be represented and how relationships between these representations can be represented, and finally, how well the specific implementation of these theories in the form of the *W*ebIC prototype worked with real users during unrestricted browsing of the world wide web.

In the survey, we would like to find out a bit about the participators use the Internet and how they felt about using *W*ebIC. According to the follow-up survey, we found that about 75% participators agree that *W*ebIC is able to find useful information even sometimes they cannot find by themselves, and they would like to continue to use *W*ebIC if it is available after the study.

5 Summary

Our browsing behavior models identify relevant words based on the browsing features, and the models are independent of any particular words or domain. To evaluate the performance of browsing behavior models, we have conducted user study — LILAC, to integrate these models into an IE-based browser, and collect users' annotations and evaluations from their day-to-day task, either in office or at home. To facilitate the study, we have implemented *W*ebIC, which is based on IE, and has the same interface as IE.

The five-week study demonstrates that browsing behavior models work better than base-line model (i.e., FHW), and 72% of the suggested pages are relevant to the current search task.

The big concern of LILAC is the privacy. Since *W*ebIC can record almost everything while the subject was browsing the web. It also records some privacy information, for example, credit card number, password, etc. To avoid this scenario, we first instruct the subjects not to use *W*ebIC for any confidential web browsing. In *W*ebIC, we did not record any web requests under `https`.

Currently, the LILAC subjects have to annotate IC-pages explicitly, which is not what users usually do while they browse the web. Therefore, they might have to change their behavior to adapt to the study. For future user study, we need to design the experiment to make the annotation a normal operation under some special scenarios, or maybe other applications.

6 Acknowledgement

The authors gratefully acknowledges the generous support from Canada's Natural Science and Engineering Research Council, the Alberta Ingenuity Centre for Machine Learning (http://www.aicml.ca), and the Social Sciences and Humanities Research Council of Canada Initiative on the New Economy Research Alliances Program (SSHRC 538-2002-1013). We also thanks for the help from: David Woloschuk, Tong Zheng, Michael DeMarco, Xiaoyuan Su, and Louise Whyte.

References

Agrawal, R., and Srikant, R. 1994. Fast algorithms for mining association rules. In *Proc. of the 20th Int'l Conference on Very Large Databases (VLDB'94)*.

Agrawal, R., and Srikant, R. 1995. Mining sequential patterns. In *Proc. of the Int'l Conference on Data Engineering (ICDE)*.

Chi, E.; Pirolli, P.; Chen, K.; and Pitkow, J. 2001. Using information scent to model user information needs and actions on the web. In *ACM CHI 2001 Conference on Human Factors in Computing Systems*, 490–497.

Herlocker, J.; Konstan, J.; Terveen, L.; and Riedl, J. 2004. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems* 22(1):5–53.

Japkowicz, N. 2000. The class imbalance problem: Significance and strategies. In *Proceedings of the 2000 International Conference on Artificial Intelligence (ICAI '2000)*.

Ling, C., and Li, C. 1998. Data mining for direct marketing problems and solutions. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98)*. New York, NY: AAAI Press.

Mannila, H.; Toivonen, H.; and Verkamo, A. 1997. Discovery of frequent episodes in event sequences. *Data Mining and Knowledge Discovery* 1(2):259–289.

Porter, M. 1980. An algorithm for suffix stripping. *Program* 14(3):130–137.

Quinlan, R. 1992. *C4.5: Programs for Machine Learning*. San Mateo: Morgan Kaufmann Publishers.

Resnick, P.; Iacovou, N.; Suchak, M.; Bergstorm, P.; and Riedl, J. 1994. Grouplens: An open architecture for collaborative filtering of netnews. In *Proceedings of ACM 1994*

Conference on Computer Supported Cooperative Work, 175–186. Chapel Hill, North Carolina: ACM.

Zhu, T.; Greiner, R.; and Häubl, G. 2003a. An effective complete-web recommender system. In *The Twelfth International World Wide Web Conference(WWW2003)*.

Zhu, T.; Greiner, R.; and Häubl, G. 2003b. Learning a model of a web user's interests. In *The 9th International Conference on User Modeling(UM2003)*.