Off-line Evaluation of Recommendation Functions

Tingshao Zhu[1], Russ Greiner[1], Gerald Häubl[2], Kevin Jewell[1] and Bob Price[1]

[1] Dept. of Computing Science, University of Alberta, Canada T6G 2E1
{tszhu, greiner, kjewell, price}@cs.ualberta.ca
[2] School of Business, University of Alberta, Canada T6G 2R6
Gerald.Haeubl@ualberta.ca

Abstract. This paper proposes a novel method for assessing the performance of any Web recommendation function (i.e., user model), M, used in a Web recommender sytem, based on an off-line computation using labeled session data. Each labeled session consists of a sequence of Web pages followed by a page $p^{(IC)}$ that contains information the user claims is relevant. We then apply M to produce a corresponding suggested page $p^{(S)}$. In general, we say that M is good if $p^{(S)}$ has content "similar" to the associated $p^{(IC)}$, based on the the same session. This paper defines a number of functions for estimating this $p^{(S)}$ to $p^{(IC)}$ similarity that can be used to evaluate any new models off-line, and provides empirical data to demonstrate that evaluations based on these similarity functions match our intuitions.

1 Introduction

While the World Wide Web contains a vast quantity of information, it is often time consuming and sometimes difficult for a Web user to locate the information she[3] finds relevant. This motivates the large body of research on ways to assist the user in finding relevant pages. There are, however, many Web user models that can generate recommendations, but how to evaluate their performance is a critical task. It is often costly, in terms of both time and finances, to evaluate such systems in user studies.

In this paper, we propose a novel method to evaluate the performance of these recommendation functions by an off-line computation. Our evaluation uses the data that we collected in a previous user study (Section 2). From this data, we developed several similarity functions that estimate the subject's evaluation of the suggested page. Our cross-validated empirical results verify that these similarity functions are good models of the user's judgment. Therefore, they can then be used to evaluate any new user models.

Section 1.1 discusses related work. Section 2 describes the "LILAC" user study that we conducted previously to acquire labeled session data. Section 3 outlines our ideas for how to identify these similarity functions, using this collected data.

1.1 Related Work

There is a great deal of research on generating recommendations for Web users. Some of these systems recommend pages either within a specified Web site [1], or based on

[3] We will use the female pronoun ("she", "her") when referring to users of either gender.

some specific hand-selected words [2]; while others seek useful pages from anywhere on the Web [9]. This paper introduces an off-line technique to evaluate such models. Below we summarize several alternative approaches, and discusses how they relate to our work.

Kobsa and Fink [3] simulate the workload to test the performance and scalability of the user model servers. In our research, we run the off-line evaluation by simulating the users' assessment, and focus on evaluating the relevance of recommended pages.

Weibelzahl and Weber [7] propose two methods to evaluate the accuracy of any predictive user model. However, their approach can only be applied to straightforward user models, which means it cannot be applied to the complex user models produced by machine learning algorithms. By contrast, our method can evaluate any user models.

Ortigosa and Carro [5, 6] describe how they infer evaluation by using some heuristics in an adaptive-course system. In our case, the evaluating functions have been verified by data from our user study, which indicate that they are consistent with the users' judgement.

2 User Study — LILAC

We developed a system, *W*ebIC (Figure 1) that observes the user as she browses the Web, and in particular, records 35 different "browsing properties" of the words that appeared on these visited pages (e.g., for each word w, did the user tend to follow links anchored with w, or did the user "back out of" pages whose title included w, etc. [8]). *W*ebIC then applies a trained classifer to these browsing properties, to identify which of these encountered words is "relevant" to the user's current information need; it then attempts to find pages that address these needs.

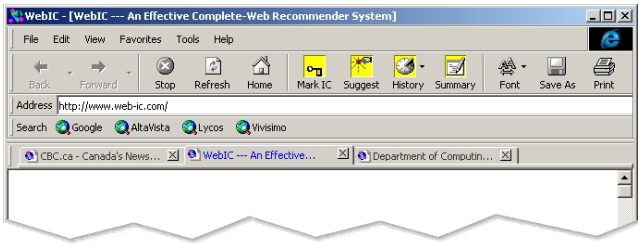

Fig. 1. *W*ebIC — An Effective Complete-Web Recommender System

The challenge in the *W*ebIC [8] research is finding a good classifer, for mapping the browsing properties of a word to a relevance score. To address this problem, we considered several models: the "Followed Hyperlink Word" (FHW) model as a baseline, and three "IC-models" trained from data, ICWord, ICQuery, and ICRelevant. Specifically, ICWord tries to identify the words will appear in the relevant page; ICQuery tries to identify the words that allow a search engine to locate the relevant page; and ICRelevant tries to predict the words explicitly selected by the user as being relevant.

We conducted a five-week user study, "LILAC" (Learn from the Internet: Log, Annotation, Content), both to learn the parameters for these IC-models, and also to evaluate their performance. During the study, each of the 100+ participants was required to install *W*ebIC on their own computer and then browse their own choice of web pages.[4]

During her browsing, the user may push the "Suggest" button to ask *W*ebIC to recommend a page, $p^{(S)}$. At other times, she may find a page $p^{(IC)}$ that satisfies her current information need. As part of this study, whenever she encounters such a page, she is asked to click the "MarkIC" button in the *W*ebIC browser to indicate that this an "Information Content page" (i.e., "ICpage"). *W*ebIC would then suggest an alternative page $p^{(S)}$ as chosen by one of the IC-models. In either case, whenever *W*ebIC recommends a page, the user is then asked to characterize how well this suggested page satisfied her current information need: "*Fully* answered my question", "*Somewhat* relevant, but does not answer my question fully", "*Interesting*, but not so relevant", "*Remotely* related, but still in left field", or "*Irrelevant*, not related at all". Figure 2 shows the overall results of these evaluation responses; each bar show the relative percentage of one evaluation response for one model.

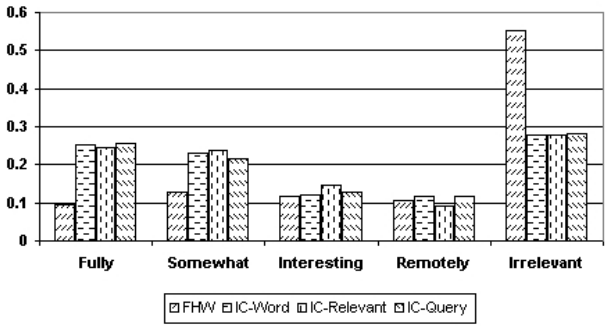

Fig. 2. Overall Results of the LILAC Study

We ran the Wilcoxon test on each possible pair of models, which confirmed that each of our trained IC-models performed better than the baseline FHW model. These results confirm our basic assumption that we are able to provide useful recommendations by integrating the user's browsing behaviors into the prediction.

3 Off-line Evaluation

An effective "similarity function" $s(\, p_1, \, p_2 \,)$, over a pair of Web pages, should return a large value iff p_1 and p_2 are similar in content. By definition the $p^{(IC)}$ found by the user satisfied her information need, as did any $p^{(S)}$ that was evaluated as "Fully". Therefore,

[4] We requested they use only English language pages. We also provides ways to turn off the data capture part of *W*ebIC when dealing with private information — e.g., email or banking.

we would expect $s(\, p^{(IC)},\, p^{(S)}\,)$ to return a large value iff $p^{(S)}$ was evaluated as a "Fully" page, and otherwise to return a small value.

The challenge is learning such a similarity function from the MarkIC data. Here, we only consider the two extreme kinds of suggested pages: "Fully" (S_+) and "Irrelevant" (S_-). We basically want a function $s(\,\cdot,\,\cdot\,)$ that has a significant difference between the values of $s(\, p_i^{(IC)},\, p_i^{(S_+)}\,)$ and $s(\, p_i^{(IC)},\, p_i^{(S_-)}\,)$. Below we propose three different similarity functions, which use W_{IC} and W_S to denote respectively the bags of words in $p^{(IC)}$ and $p^{(S)}$, after removing stop words and stemming.

ITM: Information Theoretic Measure[4] $\quad s_{ITM}(p^{(IC)},\, p^{(S)}) = \frac{|W_{IC} \cap W_S|}{|W_{IC} \cup W_S|}$

Recall: ICWord Recall $\quad s_{Rec}(p^{(IC)}, p^{(S)}) = \frac{|W_{IC} \cap W_S|}{|W_{IC}|}$

avRankTFIDF: Mean of Ranks of the Common Words' TFIDF (ranks all the words in an ICpage based on TFIDF weights, from the highest to the lowest, and returns the mean of ranks of the words in $W_{IC} \cap W_S$)

$$s_{Rank}(p^{(IC)}, p^{(S)}) = \frac{\sum_{w \in W_{IC} \cap W_S} \text{TFIDFRank}_{(w \in W_{IC})}}{|W_{IC} \cap W_S|}$$

For each s we compute similarity scores using the collected sample sessions, and then perform the Mann-Whitney test to determine whether there is a significant difference between the "Fully" and "Irrelevant" cases. The results (Table 1) shows that the each of the similarity functions can detect a significant difference. For *avRankTFIDF*, the smaller the rank, the higher the similarity between pages.

Table 1. Mann-Whitney Test on Different Similarity Functions

	Hypothesis	Confidence Intervals	p
ITM	Fully>Irrelevant	>0.027	<0.0001
Recall	Fully>Irrelevant	>0.045	<0.0001
avRankTFIDF	Fully<Irrelevant	<-4.554	0.0055

3.1 Validating Similarity Functions on LILAC Data

Next, we analyzed these functions on the LILAC data without using any evaluation labels, to determine whether the results are consistent with our previous conclusions (Section 2), based on evaluation labels directly.

For around one-quarter of the MarkIC sessions, *W*ebIC selected the baseline FHW model. The similarity between the user's ICpage $p^{(IC)}$ and this proposed $p^{(S_{FHW})}$ page can be computed using each of these similarity functions $s \in \{s_{ITM},\, s_{Rec},\, s_{Rank}\}$. We then identify three new recommended pages off-line, one using each of the IC-models (i.e., ICWord, ICRelevant, and ICQuery). We can compute the overall similarity $s(\, p^{(IC)},\, p^{(S_\chi)}\,)$ where $\chi \in \{ICW,\, ICR,\, ICQ\}$. Similarly, for each MarkIC session using any of the IC-models, we can also run the FHW model on the same session to produce a new recommended page $p^{(S_{FHW})}$, and then compute $s(\, p^{(IC)},\, p^{(S_{FHW})}\,)$.

Table 2. Wilcoxon Test on LILAC MarkIC Session Data using different similarity functions

Hypothesis →	FHW<ICWord	FHW<ICRelevant	FHW<ICQuery
ITM	<0.0001	0.0002	<0.0001
Recall	0.087	0.0213	0.003
avRankTFIDF	0.0057	<0.0001	<0.0001

To verify that the off-line evaluation can acheive the same conclusions as obtained previously (i.e., each IC-Model is better than FHW), we use the Wilcoxon test on the correlated samples, and view a p-value less than 0.05 as supporting each claim. Table 2 shows the p values of each hypothesis given a similarity function. This data indicates that both the ITM and Rank (avRankTFIDF) functions can detect a significant difference between FHW and any of IC-Models, which is consistent with the overall results that were based on evaluations directly from LILAC.

4 Conclusion

We propose a novel method to assess the performance of Web user models off-line, which can infer the evaluation by an off-line computation. In particular, we can take advantage of the previously annotated Web logs in the LILAC study to evaluate any new user models. We have developed several similarity functions to approximate the subject's evaluation of the suggested page. By applying the similarity functions to the LILAC data, we find that the results based on these similarity functions are consistent with the evaluations made by the subjects directly in LILAC.

References

1. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. VLDB'94
2. Billsus, D., Pazzani, M.: A hybrid user model for news story classification. UM'99
3. Kobsa, A., Fink, J.: Performance evaluation of user modeling servers under real-world workload conditions. UM'03
4. Lin, D.: An information-theoretic definition of similarity. ICML'98
5. Ortigosa, A., Carro, R.: Agent-based support for continuous evaluation of e-courses. SCI 2002, Volume 2., Orlando, Florida (2002) 477–480
6. Ortigosa, A., Carro, R.: The continuous empirical evaluation approach: Evaluating adaptive web-based courses. UM'03 163–167
7. Weibelzahl, S., Weber, G.: Evaluating the inference mechanism of adaptive learning systems. UM'03
8. Zhu, T.: http://www.web-ic.com/.
9. Zhu, T., Greiner, R., Häubl, G.: An effective complete-web recommender system. WWW'03