Predicting Where a Web User Wants to Go

Tingshao Zhu, Russ Greiner
Dept. of Computing Science
University of Alberta
Edmonton, Alberta
Canada T6G 2E1

Gerald Häubl
School of Business
University of Alberta
Edmonton, Alberta
Canada T6G 2R6

ABSTRACT
In this paper, we introduce our on-going research that uses the content of the user's observed clickstream to predict which web pages she wants to visit. Our method first identifies which words will be in these "information content" pages, and then uses these words to construct search queries to retrieve the relevant webpages. We present empirical evidence that this approach can work effectively.

1. Introduction

To explain the ideas, consider a student who wants to borrow reference books for her courses from the library. While browsing the library's web site, she finds pages that identify the books she wants to borrow. One sequence of webpages is shown in Fig. 1; the printer icon identifies which pages contain useful pointers.

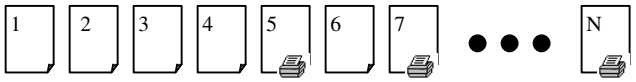

Figure 1: Book Search for New Term.

Notice that the student could accomplish her goal (identifying relevant books) by only visiting these "information content" (IC) web pages, as they are the pages that she must examine to complete her task. A recommender system that could recommend all-and-only these IC pages to the web user would clearly be very helpful.

In this paper, we propose a method for predicting these IC pages, based on the observable click stream. The framework takes advantage of machine learning algorithms to predict the IC pages from the web user's browsing history.

2. Learning Problem

To predict an IC-page, our predictor first tries to identify which words will be in such a page (so-called "IC words") and then uses these IC-words to construct search queries to send to a standard search engine (e.g., Google); see Fig. 2. (Ideally, the results of this query will be IC pages.)

This short abstract focuses on the first of these tasks, IC-word prediction --- that is, given a click stream, predicting which words will be in IC-pages.

We view IC-word prediction as a classification task, but instead of building a model based on specific words, our model is based on "browsing features" of words, such as: how many times a word is in the anchor text of hyperlinks, how many times a word is in the search keyword list, etc..

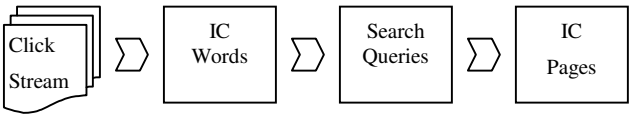

Figure 2: Framework for IC-Page Prediction.

After training, our system may find some patterns like, e.g., "any word that appears in the three consecutive pages will be in the IC page". Note that this is different from systems that produce association rules [1] – e.g., "*if page H has been visited, then page I will also be examined*," as those association rules can only predict pages that have been seen. By contrast, our method can make predictions about pages that have not been visited. Our approach also differs from systems that involve a set of *pre-defined* words – e.g., "*for all the words in page T, if "web" has weight 0.9, "software" has 0.8, … , then page T may be interesting to the user*," as we do not force the user to pre-define the words that might be relevant.

We are not looking for patterns based on specific words nor for a specific user, but rather for general patterns across different people, which we expect to be useful, even in an unfamiliar web environment.

3. Classifier Training/Testing

To learn, and subsequently evaluate, our system, we collected a set of *annotated web logs*; each a sequence of web pages that a user has visited, where each page is labeled with a bit that indicates whether or not that page is an IC page – i.e., essential to achieving the user's specific task goal.

We collected these annotated web logs in a laboratory study with 128 subjects. Each participant was asked to perform a specific task: identify 3 novel vacation destinations that s/he is interested in, and prepare a detailed plan for a vacation at each destination (including specific travel dates, flight numbers, accommodation, activities, etc.).

Participants were given access to our customized browsing tool (AIE – Annotation Internet Explorer), which recorded their specific web logs, and required them to provide the IC annotation.

After preparing the data, each word that occurs in a click stream is represented by a feature vector (indicating the

word's browsing features) plus an indicator of whether it appears in the IC page or not. We then train NaïveBayes (NB) classifiers. (NB is a simple belief net structure that assumes that the attributes are independent of one another, conditional on the class label [5].)

We ran our test on 7*20 subject groups, where each group involved *N = 2,3,…8,* subjects which were selected randomly from the subject list. For each *N,* we randomly selected 20 different groups. Note that we allowed overlap among these 20 groups. For each group, we built 10-fold training/testing datasets, and computed the median value of these 10 results as the final score for this group. (We used medians, because they are less sensitive to outliers than means.)

To generate the training and testing data, we randomly selected positive and negative vectors of equal size. This required duplicating the remaining positive vectors by random repeating until an equal number of positive and negative training samples were obtained.

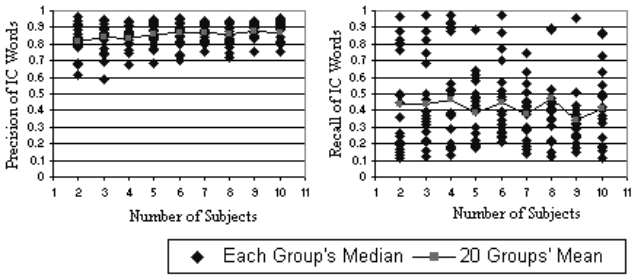

Figure 3: IC-Word Prediction.

The average prediction accuracy was around 65-70%. To better understand this, we computed prediction and recall values. The results are shown in Figure 3.

4. Evaluation

Clearly, it would also be useful to predict these pages early --- e.g., from Fig. 1, it would be better to recommend the IC page (page 5) after the user has traversed only pages 1 and 2, rather than wait until the user has reached page 4. We therefore define an evaluation method based on these two objectives.

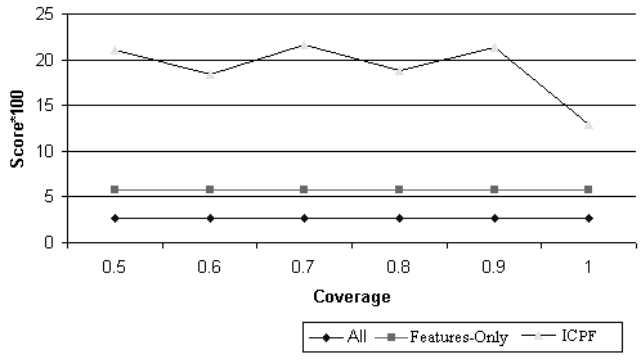

Figure 4: Evaluation Results.

We compare our method with other two very simple techniques: let the IC words be (1) all words in the input clickstream or (2) all feature words, which are those words enclosed by some specific HTML tags, such as ``a", ``title", ``b", ``h1", etc.. Fig. 4 shows that our approach did significantly better than both of these methods.

We propose that there exists a general model of goal-directed information search behavior on the web. Our model is not based on one specific web site or a specific set of words, and we expect that it can be applied in many different circumstances. Due to its high level of abstraction, our model is able to capture general behavioral patterns that apply to broad classes of users and environments.

5. Related Work

As we mentioned above, IC page prediction implicitly requires inferring what information the user wants. Chi et al. [3] identify "information need" from the context of the hyperlinks that the user followed, and *define* this to be the information that the user wants. Unfortunately, this is based on the very strong assumption that the hyperlink context must correspond to the user's intention.

The Choo et al. [4] experiment collected feedback from web users in order to infer their information seeking mode, which is too general to help individual users in individual sessions, and it is still difficult to infer the pattern of how to locate IC pages and what these pages look like.

Billsus and Pazzani [2] tried to learn what kinds of news will be interesting to a person by applying a Naive Bayes classifier to a Boolean feature vector representation of the candidate story. Unfortunately, in their research, there was no explicit feedback from the subject; instead they only inferred the interestingness of the news story from the viewer's actions, such as channel changes.

6. Future Work

We are currently working on ways to construct effective search queries from IC words, and are investigating more efficient ways to predict IC words. Our system for predicting IC pages will reliably help users reach the pages that satisfy their information needs.

REFERENCES

1. R. Agrawal and R. Srikant. Fast algorithm for mining association rules. In Proc. Of the 20th Int'l Conference on Very Large Databases, Santiago, Chile, Sept.. 1994.
2. D. Billsus and M. Pazzani. A hybrid user model for news story classification. In Proceedings of the Seventh International Conference on User Modeling (UM '99), Banff, Canada, 1999.
3. E. H. Chi, P. Pirolli, K. Chen, and J. Pitkow. Using information scent to model user information needs and actions on the web. In ACM CHI 2001 Conference on Human Factors in Computing Systems, pages 490--497, Seattle WA, 2001.
4. Chun Wei Choo, Brian Detlor, and Don Turnbull. A behavioral model of information seeking on the web – preliminary results of a study of how managers and it specialists use the web. Proceedings of the 61st Annual Meeting of the American Society for Information Science, pages 290--302, Pittsburgh, PA, Oct 1998.
5. Richard O. Duda and Peter E. Hart. Pattern Classification and Scene Analysis. Wiley, New York, 1973.