Estimating Users’ Interest in Web Pages
by Unobtrusively Monitoring Users’ Normal Behavior

Jude Shavlik
Jeremy Goecks

Computer Sciences Department
University of Wisconsin-Madison
Madison, WI 53706
shavlik@cs.wisc.edu

Abstract
For intelligent Web browsers attempting to learn a user’s interests, the cost of obtaining labeled training instances can be prohibitive because the user must directly label each training instance, and few users are willing to do so. We have been developing an approach that circumvents the need for human-labeled pages. Instead, we learn “surrogate” tasks where the desired output is already being measured by modern operating systems and Web browsers, such as the number of hyperlinks clicked on a page, the amount of scrolling performed, and the number of CPU cycles used. Our assumption is that some weighted combination of these easily obtained measurements will highly correlate with the user’s interests. In other words, by unobtrusively “observing” the user’s behavior we are able to automatically construct labeled training examples for learning useful functions with which we can estimate the user’s interest in a Web page. We report the results of a pilot study.

Introduction
Much research has been devoted to the task of developing intelligent agents that can learn a user’s interests (a “profile”) and find information in the World Wide Web (WWW) based on such profiles (e.g., Lang, 1995; Pazzani, Muramatsu, & Billsus, 1996; Joachims, Freitag & Mitchell, 1997; Shavlik, Calcari, Eliassi-Rad & Solock, 1999). Given a particular web page or hyperlink and a particular user, the agent’s task is to predict the user’s interest in that page or hyperlink. If the agent predicts that the user will be very interested in the page, such an agent can retrieve this page for later viewing by the user.

A central question in the topic of learning user profiles concerns how the learning algorithm obtains training examples. Expecting the user to manually rate many pages is not desirable. Previous work (Liberman, 1995; Joachims et al., 1997) has investigated employing passively recorded user-navigation behavior (i.e., analyzing hyperlinks clicked, hyperlinks passed over, and patterns of user behavior) as surrogate measures for predicting user interest in a page. We are extending this work by unobtrusively recording user-navigation behavior and additional user actions (such as scrolling and mouse activity). We plan to utilize a combination of these measurements as a surrogate for user interest in a page. We hypothesize (but have not yet tested) that these measurements correlate well with the user’s true interests. Since we can collect these automatically generated training examples during the user’s normal use of their web browser, our approach allows for the collection of large training sets without burdening the user.

Figure 1 clarifies our approach. The overall goal is to learn a mapping from web pages to a specific user’s interest. Rather than directly learning this mapping, which would require the user to rate many web pages, we instead break this task down into two parts. First, learn to map from a web page to some properties that browsers and operating systems can measure; this allows us to collect large numbers of training examples without burdening the user. This first step is the topic of our current research.

Copyright © 2000, American Association for Artificial Intelligence (www.aaai.org). All rights reserved.
Second, learn a mapping from these surrogate measurements to the user’s interests. We hypothesize that this mapping can be learned from some (human) test subjects, and the results directly applied to additional users. In other words, new users would never need to rate web pages. Instead, a “user-independent” model of the second mapping would be learned “once and for all.” This second step is a topic for our future research.

We have demonstrated (Goecks & Shavlik, 2000) that it is feasible to learn some promising surrogate measurements. We review those results in this article.

**System Overview**

Our current system involves two phases: data collection and machine learning. Our goal is predict our surrogates for user interest solely from the HTML contents of a web page. Such ability could be used to, say, have a computer program wander the web at night, collecting pages likely to be of interest to the user when he or she wakes up.

**Step 1. When the user visits a web page, record (a) the HTML contents of the page; and (b) normal actions performed by the user on the page.**

Various user actions that appear likely to correlate with the user’s interest in a page include the number of hyperlinks clicked on by the user, the amount of scrolling the user performed, and the amount the user moved the mouse over the currently displayed web page. Thus, when the user navigates to a new page, the agents records (a) the text of the HTML text; (b) the number of hyperlinks the user clicked on; (c) the amount of user scrolling activity; and (d) the amount of user mouse activity. Of these, (a) creates the input and (b)-(d) constitute the outputs of our training examples.

When the user returns to a page he or she has previously visited recently, it is logical that the agent should add the user actions recorded during this return visit to the user actions recorded during the previous visit. This makes sense because often a user navigates back-and-forth to many pages from a single “start” page, and the user is likely very interested in this start page. Creating a new training example for the page each time and recording only the user actions for that particular visit does not capture the user’s true interest in the page in this situation. Hence, a novel aspect of our approach is that our agent sums the actions of the user on each page visited over a finite period of time. The active time for an instance is this finite period of time during which the user’s actions on a page are summed together (active time is currently set to 20 minutes). By employing the concept of an instance active time, the agent attempts to capture the user’s true interest in a page, whether the user performs many actions on a page in one visit or the user revisits a page many times and performs just a few actions on the page during each visit.

**Step 2. Apply a machine learning algorithm to the labeled training examples collected in Step 1.**

Recall that our agent’s primary goal is, given the HTML text of a web page, to predict the amount of normal user actions on the page (i.e., our surrogates for user interest). In other words, using the learning algorithm, the agent is to learn to successfully predict the number of hyperlinks the user will click on, the amount of scrolling and mouse activity the user will perform.

Our agent currently employs a fully connected, three-layer neural network as its learning algorithm. Our choice to initially use a neural network was made largely out of convenience. We are currently evaluating other learning algorithms for this task.

We base our representation of the web page on the bag-of-words representation (Salton, 1991); this representation is compatible with neural networks and empirically has proven quite successful in information retrieval (Salton, 1991; Lang, 1995; Pazzani et al., 1996). Actually, our agent uses an enhanced version of the basic bag-of-words representation. We equip our agent with the ability to handle “key phrases,” which are simply phrases the agent is to look for as opposed to single words.

Finally, we address the issue of acquiring keywords for a particular user. Numerous means come to mind, such as analyzing the user’s homepage or examining the text of web pages visited by the user and using a heuristic function (e.g., information gain; Mladenic, 1996) to choose the keywords from those pages’ text. The focus of this research is independent of the means used to choose keywords, and thus for simplicity in our initial experiments the agent is simply provided with a list of keywords. A “production-quality” agent would require a more sophisticated method of choosing keywords.

Each of the network output units corresponds to a particular user action; thus, there are three output units for the network. The output unit for hyperlinks clicked represents the fraction of the page’s hyperlinks that are clicked. E.g., a value of 0.1 means that the user clicked 10% of the current page’s hyperlinks. The outputs units for scrolling activity and mouse activity represent counts of the events reported by Microsoft’s IE 4.0 browser, scaled by 100. We train our network using the standard backpropagation algorithm.

**Experimental Methodology**

To determine how accurately our agent can predict these normal user actions, we performed a 10-fold cross-validation experiment. For each fold we use “early stopping” when training; 10% of each fold’s training data is used as a tuning set. After training is complete, the agent restores the network that performed best on the tuning set. We then record the agent’s accuracy on the fold’s test set. We use the root-mean-square method to measure the accuracy of the network’s predictions.

The keywords for this pilot study were related to machine learning; we used 107 key words and phrases. To collect the data, one of us (JG) browsed the WWW while our agent observed and collected data. The browsing attempted to simulate a normal individual interested in reading about various topics and research projects in
machine learning; however, browsing was not performed with any particular questions or issues in mind. A total of 200 pages were visited over a couple of days; 150 of these pages were related to machine learning, and the remaining 50 were not. On each page about one fifth of the hyperlinks were clicked, there were about 40 scrolling-related events, and there were about 70 mouse-related events.

An effort was made to visit a broad range of pages and perform a number of different navigation actions, including the use of the forward/back buttons, clicking on “dead” links and links to download files, and browsing FAQ’s and ftp sites. Although only one experimental subject was used, we believe that the data collected for this experiment is representative of a “real-world” data sample that might be obtained from an ordinary user.

Table 1 summarizes these results (recall that the average values for these three measurements were about 0.20, 0.40, and 0.70, respectively). The data shows that the three surrogates can be reasonably accurately predicted.

Table 1. Results of our cross-validation. The agent collected a total of two hundred (200) instances and used ten (10) train/test folds. For each fold, the agent set aside 10% of the data as a tuning set for “early stopping” of the network’s training; the agent trains on the train’ (train instances minus tune instances) set for 100 epochs, chooses the network state that performs best on the tune set, then measures accuracy on the test instances.

<table>
<thead>
<tr>
<th>Action Predicted</th>
<th>RMS Error on Test Sets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hyperlinks Clicked</td>
<td>0.07 ± 0.02</td>
</tr>
<tr>
<td>Scrolling Activity</td>
<td>0.05 ± 0.02</td>
</tr>
<tr>
<td>Mouse Activity</td>
<td>0.13 ± 0.03</td>
</tr>
</tbody>
</table>

Future Work

A number of questions have arisen from our initial work on this agent. Our agent would greatly benefit from improved methods for detecting user actions. The three user actions currently detected by our agent provide a good “first-order” approximation to the user’s activity on a web page, but one can imagine additional properties to automatically measure, such as CPU time expended on a page by the browser or amount of typing the user does within a web page.

We also plan to perform experiments using human subjects in order to investigate the machine learning of the right-hand box in Figure 1 (i.e., the mapping from our surrogate measurements to users’ true interests).

Conclusions

We briefly presented and evaluated the design of an agent that employs a standard neural network to learn a user’s interests in the WWW. The key aspect of our system is that it unobtrusively measures normal actions performed by the user on a web page. It uses these measurements as a surrogate for the desirable, but too burdensome to collect, measurements of the user’s interest level. Hence, rather than explicitly labeling the interest level of WWW pages, users implicitly label web pages by the actions they perform on them. Our cross-validation experiment suggests that the agent can learn to predict, to a reasonable accuracy, the surrogate measurements of user interest that we investigated.

Acknowledgements

This research is partially supported by NSF Grant IRI-9502990 and a Vilas Associates award (to JS) from the University of Wisconsin. This work is part of JG’s undergraduate thesis.

References


