Predicting User Preferences - From Semantic to Pragmatic Metrics of Web Navigation Behavior

Ion Juvina
Center for Content and Knowledge Engineering, Utrecht University, The Netherlands
Padualaan 14, De Uithof, 3584CH Utrecht
Tel: +31-30-253 6761
ion@cs.uu.nl

ABSTRACT
This paper aims at demonstrating a particular method to extract relevant information from navigation data. Participants had to perform web tasks while their navigation behavior was recorded. Then, they had to rate a list of presents based on the question: “how much would you like to receive this present for your birthday?” The semantic similarities (calculated with Latent Semantic Analysis) between users’ navigation paths and presents’ descriptions were used as implicit estimates of users’ preferences. Results show that users’ explicit preferences can be accurately estimated based on users’ navigation behavior. Consequently, we propose this method to be used in building user models for adaptive web applications.

Keywords
Web navigation, user modeling, implicit preferences.

INTRODUCTION
Navigation is a major part of user experience on the web [10]. This facility brought new problems for application designers and usability specialists; cognitive overload and disorientation are the main ones [3,11]. Web users often get more information than they need or are interested in, and they have to find their ways through complex and dynamic information spaces.

Research attempting at modeling cognitive mechanisms involved in web-navigation gains increasing influence in the HCI community. For instance, such a model, called CoLiDeS, is proposed in [8]. It explains how users parse and comprehend the content of a Web page and then select what action to proceed. This model uses Latent Semantic Analysis (LSA) [9] to estimate the semantic similarity between user goals and semantic entities (e.g. link descriptions) on web pages.

A related line of research aims at modeling the user’s navigation behavior in order to provide adaptive navigation support in web applications [4]. One proposed solution to information overload is information filtering based on a user model [12], that is, users should be provided only with relevant information according to their needs, interests, knowledge, etc.

Recommender systems are personalized applications that suggest possible user actions based on user interests and preferences. Recommender systems contain a model of user interests and preferences based on information provided by the user or inferred by the system based on user behavior. Using information provided explicitly by the user (ratings of documents or products) has some disadvantages [7,1]: it requires time and effort from the user; user interests change in time; giving ratings interferes with natural navigation; user might not be willing to disclose personal information.

User models based on information provided implicitly by the user release the user input effort as they gather information from user behavior during the natural use of the application [1]. Updating and improving such user models is possible, since each user action counts as a re-diagnose of the user [7]. Claypool [1] compared user interests inferred from navigation data with explicit user interest ratings and found that viewtime, scrolling and mouse movements and clicks are good indicators of user interests.

In our view, a navigator’s model could include:

- **Syntactic** information: How do users navigate? In which manner do they move across the information space? Which links are followed, in which order, what does the navigation graph look like (e.g. linear or nonlinear)?
- **Semantic** information: What is the meaning of the information that the user encountered during navigation, which of this information was processed/found relevant by the user?
- **Pragmatic** information: What are users’ goals and tasks? What are their interests, preferences, and (potential) actions?
As we will use this distinction in the context of modeling web navigation, ‘syntactic’ means structural, topologic information, ‘semantic’ refers to the content of visited pages, and ‘pragmatic’ information indicates what are the reasons and (expected) gains for the user of visiting certain pages.

OVERVIEW OF A PREVIOUS STUDY AND OBJECTIVES OF THE PRESENT STUDY
This paper aims at demonstrating a particular way to extract relevant information about users’ navigation behavior. This information can be used in building user models for adaptive web applications.

Based on easy-to-collect web logging data, syntactic, semantic and pragmatic information about users’ navigation behavior can be extracted. The relevance of this information for the use of web applications can be checked against data collected by other means, respectively, questionnaires, cognitive tests and performance measures.

In a previous study [5,6], we showed how syntactic and semantic information can be extracted from navigation data and how relevant such information is for the quality of user experience on the web. In particular, two specific navigation styles were related to users’ perceived disorientation and a semantic measure called path adequacy was used as an indicator of task performance.

Path adequacy was determined as a coefficient of semantic similarity between a navigation path and a task description. A navigation path is a concatenation of semantic entities that the user has encountered in her/his way toward a specific location. As semantic objects, one can consider, link anchors, page titles, page contents, URLs, clickable icons, banners and images etc. We used navigation paths composed of page titles. A high similarity means that the path a subject followed overlaps to a great extent with the task description.

In order to measure the semantic similarity between navigation paths and task descriptions we used Latent Semantic Analysis [9]. Path adequacy significantly correlated with return rate ($r=-0.48$), spatial ability ($r=0.36$), and task performance ($r=0.47$).

From this previous study we concluded that, when given a specific task, user’s navigation paths composed of page titles as semantic entities can be used to characterize the effectiveness of user’s navigation behavior. In addition, this study raised some questions to be object of further investigation, such as:

- Could we extract pragmatic information from user’s navigation data? In other words, could we infer what the user is likely to do?
- Are there any other semantic entities that can be logged and used to more accurately characterize navigation behavior? For instance, it is presumable that the description of a link leading to a specific page is more informative than the title of that page.

To answer these questions the study presented here has two objectives:

- To investigate the possibility of extracting pragmatic information from navigation data. In particular, we will focus here on characterizing user preferences. We consider this to be a first step in the more difficult task of inferring what users intend to do. The approach, though, is likely to be similar: navigation paths (containing semantic information) can reveal what information was available to the user and what decisions s/he made during the navigation session. Based on this information and on some additional knowledge about the hypermedia infrastructure being navigated (what tasks are supported), it is possible to infer what the user is intending to do (pragmatic information). Such inferences must be validated against users’ explicit statements about their actions or intentions. Validated inferences can be subsequently implemented as user modeling techniques in adaptive applications.

- To fine-tune the use of semantic metrics in characterizing users navigation behavior. In particular, we will investigate the use of link descriptions instead of page titles as semantic features of users’ navigation paths.

METHOD AND PROCEDURE
The first step in our investigation was prompting realistic web navigation behavior. The participants had to perform two generic web navigation tasks intended only at allowing users to (implicitly) express their preferences. Our assumption was that users would spend more time inspecting items that they find interesting, regardless what task they are doing. For this reason, the two navigation tasks were called masking tasks to differentiate them from the task that we expected users to (implicitly) do, that is to spend more time on items that they find to be of interest. The order of tasks was counterbalanced.

In a second session, 2 weeks after the navigation session, participants’ preferences and interests were explicitly measured based on online forms.

Task 1: Evaluation of media portals
Participants were asked to evaluate 6 large media websites based on a given set of criteria. The evaluation criteria were presented to participants before the navigation session and they were instructed to fill in the evaluation forms at the end of the session. The actual ratings were not taken into consideration in this study. The time allocated for this task was 65 minutes.

Task 2: Finding presents
Participants had to search for birthday presents to be given to persons with specified roles (mother, father, partner, best
friend, colleague) in a given set of online shopping websites. The time allocated for this task was 45 minutes.

**Measuring explicit preferences and interests**

Participants had to rate a list of presents based on the question: “how much would you like to receive this present for your birthday?” on a 5 point Likert scale. The list contained 24 presents selected randomly from the offer of several online shops. The list was intended to cover a large range of possible preferences. Each present was introduced by the aid of an image and a short description intended to be both informative and attractive.

In addition, participants chose from a list of interest categories the ones that are the most applicable to them, respectively. The list of interest categories was conceived by combining several catalogues found on popular portals such as yahoo. The list was composed of the following categories: Entertainment, Health, Music, Business, Computer, History, Finance, Shopping, Sports, Travel, Science, and Politics. This data about participants’ interests was used only as supplement in interpretation and reduction of data about user preferences.

**Subjects**

A number of 25 subjects participated in this study. They were not selected based on any specific criterion because we intended to gather a heterogeneous sample (not only students) representative for the population of Internet users.

**ANALYSES AND RESULTS**

As stated above, one of the objectives of our investigation was to find a finer way to semantically characterize users’ navigation paths. Specifically, we used link descriptions instead of page titles as semantic entities that navigation paths are composed of. For each participant a document was created with descriptions of links that were selected (clicked) by that participant during the navigation session. The available link descriptions were sorted in the descending order of the inspection time of the pages they pointed at (destination pages) and only a limited set was selected (approximately 100 words). The document created in such way for a particular user (referred here as user’s navigation path) was hypothesized to indicate user’s preferences and interests.

LSA semantic similarity coefficients were calculated between participants’ navigation paths, on one hand, and descriptions of presents and interests categories, on the other hand. These LSA coefficients were considered as implicit estimates of participants’ explicit ratings of presents and selections of interest categories.

In order to achieve the main objective of this study (predicting user preferences based on navigation data) we checked if the hypothesized implicit preferences (calculated as LSA similarities between navigation paths and presents’ descriptions) are able to predict explicit preferences (ratings of presents). Factor analysis was used to reduce the data about both implicit and explicit preferences. Implicit preferences were reduced to 7 continuous variables (factors) and explicit preferences were reduced to one categorical variable with 3 values corresponding to 3 classes of presents that were highly rated by 3 distinct groups of participants, respectively. In addition, a 4th class was composed of 5 presents that were not distinctively preferred by any one of the three groups of participants.

Each group of participants was described by their most preferred presents. Thus,

- participants in the first group mainly prefer to get as birthday presents: The Matrix Reloaded DVD; The Shrek DVD; a Personal Digital Assistant (PDA); and a ticket to the Marlboro Masters racing circuit.
- participants in the second group mostly prefer to get as birthday presents: a subscription to The Economist; the book “Globalization and Its Discontents”; the book “Putin’s Russia”; the book “A Short History of nearly Everything”; a cultural travel guide; and the book “The Templar Revelation: Secret Guardians of the True Identity of Christ”.
- participants in the third group would like to get as birthday presents: the book “Harry Potter 5”; a Lighted Tavillo Table; a diamond vase; a Samsonite suitcase; and the DVD “Ancient Egypt”.

One could attempt to interpret the preferences of each of the three groups in more general terms. For example, group 1 can be seen as mainly preferring modern technology and the culture associated with it (fast cars, sophisticated devices, science fiction movies), group 2’s members are mainly interested in social and political aspects of our daily life, while group 3’s members are interested in stylish and popular things and values. Whereas such extrapolation exercise might be possible and useful, it is not our focus here. Since we only use this as a criterion to validate the inferred users preferences, any user segmentation can be used. We use a categorical criterion (user groups instead of individual users) only to reduce variability that is inherently beyond our control. The goal is just to illustrate an approach. Once the approach is validated and fine-tuned, the same principle can be applied to model smaller and smaller user segments or even individual users.

The information about participants’ selections from the interests list was used in data reduction of both implicit and explicit preferences (implicit: similarities between navigation paths and interests descriptions; explicit: subjects’ choices). Besides the main contribution in ensuring the robustness of factor analysis as a data reduction technique, the information about users’ interests was useful in interpreting the results, even though in a negative way. Thus, the three explicit groups differentiate among themselves by NOT being interested in different topics: members of group 1 are not interested in politics, group 2’s members are not interested in...
entertainment, shopping and health, and group 3's members are not interested in computers.

Discriminant analysis was used to predict the probability of a present to be preferred by one of the three groups of users. The 7 factors that indicate users' implicit preferences were used as classifiers. When all the presents are considered, results show that 83.3% of the presents’ probability to be preferred by a certain group of users (or by no one) is correctly estimated based on users navigation data. When neutral presents (not clearly preferred by any group) are eliminated, the accuracy of prediction increases to 100%.

CONCLUSION
We have presented a particular way to extract user-relevant information from web navigation data. LSA semantic similarity coefficients were calculated between participants’ navigation paths, on one hand, and descriptions of presents and interests categories, on the other hand. These LSA coefficients were considered as implicit estimates of participants’ explicit ratings of presents and selections of interest categories.

Specific to our approach is taking into consideration not only data at a web page level as in [8], but also data about the concatenation of semantic entities (the set of link descriptions) that users encounter along their navigation paths and the time users spend on pages connected with those entities. Extracting this information was facilitated by the semantic richness of available path data, which was judged against users’ explicit preferences.

We have seen that a particular type of pragmatic information extracted from navigation data (what would users prefer) can be used in predicting users explicit preferences. This is a first step in extracting other types of pragmatic information from navigation data, such as information about users’ propensity toward specific activities (e.g. buying a book versus watching a sports event).

The tasks used in this study are generic enough to ensure a broad generalization scope and ecological validity. Task 1 was performance-oriented (evaluation of media portals) and task 2 was leisure-oriented (online shopping). They brought about a large variability of users’ implicit preferences, which is to be expected in realistic contexts of use [2] and it is possible to be handled by the aid of statistical analysis, as we have shown.

The results of this study have important practical implications. Based on the method of inferring user preferences introduced here, a user model can be built and maintained in real time by an adaptive web application. The application can be programmed, based on such a model, to recommend navigation paths that are likely to be found useful or attractive by a particular user.

ACKNOWLEDGMENTS
We would like to thank Ellert van den Broek, Vincent van der Linden, and Koen Buurman for their invaluable support in data collection.

REFERENCES