Effective Site Customization Based on Web Semantics and Usage Mining

Paraskevi Tzekou, Sofia Stamou, Lefteris Kozanidis, Nikos Zotos
Computer Engineering and Informatics Department
Patras University, 26500 GREECE

{tzekou, stamou, kozanid, zotosn}@ceid.upatras.gr

Abstract-**The explosive growth of online data and the diversity of goals that may be pursued over the web have significantly increased the monetary value of the web traffic. To tap into this accelerating market, web site operators try to increase their traffic by customizing their sites to the needs of specific users. Web site customization involves two great challenges: the effective identification of the user interests and the encapsulation of those interests into the sites' presentation and content. In this paper, we study how we can effectively detect the user interests that are hidden behind navigational patterns and we introduce a novel recommendation mechanism that employs web mining techniques for correlating the identified interests to the sites' semantic content, in order to customize them to specific users. Our experimental evaluation shows that the user interests can be accurately detected from their navigational behavior and that our recommendation mechanism, which uses the identified interests, yields significant improvements in the sites' usability.**

I. Introduction

Millions of people access the web daily for various reasons: find information, perform financial transactions, communicate with others, etc. Due to the explosive growth of the online data and the diversity of goals that may be pursued over the web, it is not surprising that web traffic has gained a high monetary value over the last years. To tap into this accelerating market, web site operators strive to improve the usability and user retention of their sites, by customizing the latter to the needs of their users. Web site customization is the process of modifying the information or services provided by a web site so as to meet the user interests.

Adjusting the content or structure of web data to specific interests has been an active field of research for several years. Some operators attempt to improve their sites based on the analysis of the web usage data. Most of these efforts [5] [17] [18] focus on extracting useful patterns and rules, using data mining techniques, in order to understand the users' navigational behavior so that decisions concerning site restructuring may then be made by humans. However, usage-based site customization can be problematic either when there is not enough data in order to extract patterns related to certain categories, or when the site content changes and new pages are added that are not yet included in the web log [15]. To overcome such difficulties, researchers have proposed the exploitation of information about the content [8] [16] and/or the structure [7] of web sites. In particular, they propose to combine site usage and content knowledge in order to dynamically modify the web sites. Mining web logs to discover knowledge about the user interests has also been addressed in the context of recommendation engines [9] [20].

In this paper, we extend previous works and we introduce a site customization model that combines in new ways the sites' usage patterns and semantics so as to derive knowledge about the users' site interests. Our model explores a built-in subject hierarchy for the semantic annotation of the sites' content as well as for the identification of the user interests in their site navigations. Based on the association between the sites' usage and content semantics, our model builds recommendations that aim at providing users with customized site views. The contribution of our work lies in the following:

- We introduce a novel approach for the automatic identification of the user interests as these are exemplified in the user's navigational patterns. Our approach relies on a subject hierarchy for computing the user preferences in site visits and employs a number of heuristics for estimating both short-term and long-term user interests.
- We show how we can explore the identified user interests in order to detect within a site which pages are interesting to the user. The computations of interesting pages rely on the pages' semantic content as this is analyzed, processed and evaluated via the use of the hierarchy.
- We introduce a recommendation model that correlates the identified user interests to their navigational patterns in order to predict which site pages a user would like to see in the recommendations of her future site accesses. Based on the predicted user preferences, our recommender customizes the sites' presentation accordingly..

To demonstrate the effectiveness of our approach in web site customization, we carried out a user study where we measured the accuracy of our model in capturing the user interests based on the semantic analysis of their navigational history. We also examined the effectiveness of our recommendation system in improving the sites' usability and hence in ameliorating the users' navigational experience. Results indicate the effectiveness of our approach in identifying the user interests automatically and prove the usefulness of the recommendations suggested to web users.

The rest of the paper is organized as follows. We begin our discussion with a detailed presentation of our web site customization model. In Section II.A, we present our method for the automatic identification of the user interests in their site visits. In Section II.B we describe how we process the web sites' content in order to identify which of the site pages match the identified user interests. Finally, in Section II.C we introduce our recommendation mechanism. In Section III we evaluate the effectiveness of our approach and we discuss obtained results. In Section IV we review related work and we conclude the paper in Section V.

II. WEB MINING FOR IMPROVED NAVIGATIONS

In our work, we introduce the use of a subject hierarchy for building models that represent both the user interests and the site semantics. Based on the combination of the user and the site models, we build recommendations in order to improve the users' navigations in the sites' contents. Figure 1 illustrates the functional architecture of our system.

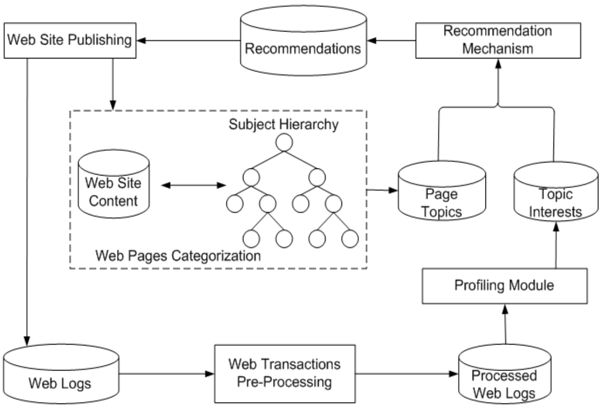

Fig. 1. System architecture.

In a high level, our method proceeds as follows. Given a site's web logs and the site's content, we pre-process log data in order to automatically identify the user interests. To do so, we address the log data processing from a classification perspective and we annotate every page visited within a site with an appropriate category from the subject hierarchy. We then combine the categories assigned to every visited page and the navigational patterns exemplified in the page visits, in order to estimate the user's degree of interest in each of the categories considered. The categories that appear to be the most interesting to the user are selected for representing that user's profile.

Having computed the user profiles, we proceed with the exploitation of the site's semantics. Our goal is to identify among the site pages the ones that are closer to the user interests. To enable that, we annotate every page in a site with an appropriate subject from the hierarchy and we compute the user's degree of interest in each of the pages. By employing the same hierarchy for representing both the user profiles and the sites' contents, we ensure consistency in the annotations given.

As a final step, our approach correlates the information obtained from the user's navigational behavior and the site semantics in order to recommend useful pages that the user may miss in her navigation, either because these are new pages and the user ignores their existence, or because these appear in deep site structures and they have a limited visibility.

A. Identifying the User Interests in Site Visits

Given the multitude of information that may be offered in a web site and the variety of interests between the different users who navigate into that site; we may assume that the success of a site customization system lies in the ability to distinguish between the different user interests. In this section, we present our approach towards user interests' identification through the semantic analysis of the users' navigation history.

To obtain information about the users' site navigations, we rely on the site's log files, out of which we extract data about a visitor's identity[1], time and date of access, complete navigation path, duration and frequency of visit, click-stream information and so on. We store this data to a transactions log database and we pre-process it in order to discover usage patterns as well as the underlying correlation between users and pages.

In particular, we download all the site pages that a user has visited, we parse them to remove markup, we apply tokenization, POS-tagging and we remove their stop words. Thereafter, we rely on the page's content terms[2] and anchor text in order to extract a set of keywords that will be used for characterizing the page's thematic content. The reason for considering also the page's anchor text for keyword extraction is the observation that in many cases the text around a link to a page is descriptive of its content [6]. In our approach, to identify keywords for a page *P*, we rely on the anchor text of other pages that point to *P*. However, to ensure that our approach is easy to implement and entails reasonable computational cost, we restrict the anchor text data within the site's level, i.e. we identify anchor text keywords for a site's page *P* based on the anchor text of other pages in the same site that point to *P*.

Having collected all the content terms in a page's text and anchor text, we weight them using the tf*idf formula in order to estimate how important is each of the keywords for the page's content. We sort the page's keywords and we retain the *n* (*n* =25%) most highly weighted terms. Based on these highly weighted keywords, we attempt to identify the page's thematic content.

For identifying the theme of a page's content, we rely on a subject hierarchy and a classification module that have both been developed in the course of an earlier study (cf. [19]) in order to automatically identify a suitable subject from the hierarchy to annotate the page's content. To enable the semantic annotation of every site page that has been visited by a user, we proceed as follows. We take the *n* most important keywords extracted from a page's content and anchor text and we map them to the hierarchy's nodes. The hierarchy that our model employs is discussed in [20] and it emerged after appending to each of the 16 top level categories of the Dmoz Directory [1] the WordNet [3] hierarchies whose root concepts are specializations of the respective Dmoz topics.

Having mapped the page keywords to their corresponding hierarchy nodes, we attempt keywords' disambiguation based on their semantic similarity values. In particular, we apply the Wu and Palmer similarity metric [21], which combines the depth of paired concepts in WordNet and the depth of their least common subsumer (LCS), in order to measure how much information the two concepts share in common. According to Wu and Palmer the similarity between two terms w_i and w_k is given by:

$$\text{Similarity}\,(w_i, w_k) = \frac{2 * \text{depth}\,\left(\text{LCS}\,(i,k)\right)}{\text{depth}\,(i) + \text{depth}\,(k)} \qquad \textbf{(1)}$$

Since the appropriate senses for w_i and w_k are not known, our measure selects the senses which maximize Similarity in

[1] There exist several techniques for uniquely identifying visitors, such as cookies, IP address, registration forms, the *identd* protocol specified in RFC 1413 [2], etc.

[2] Content terms are nouns, proper nouns, verbs, adjectives and adverbs.

order to annotate every keyword in the page with an appropriate sense.

Following keywords' disambiguation, our next step is to annotate the pages' content with an appropriate hierarchy topic, i.e. to classify every visited page to a suitable subject in the hierarchy. For page classification, we rely on the pages' keywords for which we explore both their importance weights to the pages' content and their topical categories. Considering that the keywords' importance weights are given by their tf*idf values, and that their topical categories can be easily derived from the topics that our hierarchy uses to label[3] the keyword matching senses, we can easily estimate the topic that is the most representative for the page's content as follows.

We group the disambiguated keywords of a page into topical clusters, with every cluster representing a different topic and containing all the keywords whose senses are annotated with that topic. We then rely on the keywords' importance weights in order to compute the average importance of the clusters' items. That is, we take the average importance weights of the keywords organized in a cluster in order to measure how representative is the topic of the cluster to the page's content. We then take the topic of the cluster whose elements exhibit the maximum importance average in order to annotate the content of the page. This way we annotate every visited page in a site with an appropriate topic from the hierarchy. The topics identified for the visited pages constitute the topical preferences of the user in the site. Figure 2 illustrates the user profiling process, i.e. how our approach identifies the site topics that interest the user.

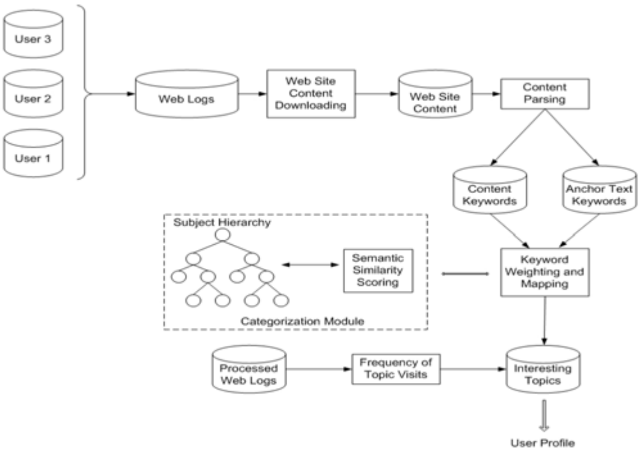

Fig. 2. The user profiling process.

So far we have presented how we can automatically identify a set of topics for describing the user's interests within a site, based on the topical categories of the site's pages that the user has visited. We now turn our attention on how we can utilize the user's navigational behavior, in order to estimate the user's degree of interest in each of the topics that describe the content of the visited site pages. To enable that, we rely on the site's transaction logs from which we collect the data recorded in the user sessions[4] and we preprocess it in order to extract statistical information about the user's site visits.

The information that we collect from the user sessions summarize to: (i) the number of times the user has clicked on each of the site pages in every session, (ii) the frequency with which the user visits pages of the same topic across sessions and (iii) the duration of every session. Based on the above data, we can estimate the degree of the user interest to each of the topics discussed in a site's content.

For our estimations, we group by topic the site pages that a user has visited in each session and we compute the degree to which each topic was of interest to the user in every session. Formally, the degree of the user's interest in topic T_A in a single session is determined by the fraction of pages in the user's visit that are categorized under T_A, given by:

$$\text{Visit Interest}\,(T_A) = \frac{|\,\#\text{ of visits to pages assigned to } T_A\,|}{|\,\#\text{ of total visits}\,|} \qquad \mathbf{(2)}$$

Based on the above formula, we can compute the probability that the user was interested in a particular topic in each of her site visits. Intuitively, the degree of a visit's interest to some topic indicates a short term preference of the user to the site's contents. More specifically, this topic preference probability help us deduce the user interests in a site's content at a given time interval. As such it does not suffice for accumulating knowledge about the user's general preferences in the site's contents. By general preferences, we mean the topics of the pages that the user regularly visits in her site accesses.

To account for the user's general interests, we again rely on the information collected from the user's sessions and we compute the degree to which a topic is preferred by the user across her site visits. Formally, the degree of the user's general topic preference in the site's content is given by the frequency with which the user visits pages of the same topic across her site interactions, as:

$$\text{Site Interest}\,(T_A) = \frac{1}{|S|}\sum_{T_A \in ST}^{S} \text{Visit Interest}\,(T_A) \qquad \mathbf{(3)}$$

Where S is the total number of sessions recorded a user's site transaction logs and ST is the set of topics discussed in the pages visited across the user's sessions. The user's site interest values give us perceptible evidence about the degree to which each site topic is generally preferred by the user and gives us some intuition about the long-term interests of the user, specified in the site's context.

Another useful indicator in deriving the user interests in a site's topics concerns the amount of time the user has spent on the site pages categorized under each of the topics. Based on the intuition that the more time the user devotes for reading pages dealing with a particular topic, the greater the user's interest in that topic, we estimate the user's interest in a topic heuristically as a weighted sum of the user's site interest in the topic and the normalized number of seconds the user spent reading pages about the topic:

$$\text{User Interest}\,(T_A) = \text{Site Interest}\,(T_A) + D(T_A) \qquad \mathbf{(4)}$$

[3] Note that every node in our hierarchy is annotated with a suitable topical category borrowed from the top level topics in the Dmoz ontology.

[4] In our work, we define a user session as a delimited set of user clicks to a single web server. A user session is also called a visit.

Where Site Interest (T_A) denotes the general user preference in topic T_A and D(T_A) denotes the normalized number of seconds that the user spent reading pages about $T_{A.}$

Based on the above formula, we can compute the degree to which each of the site's topics is of interest to the user. The topics that the user is interested in together with their degree of interestingness constitute the user's profile that our site customization model employs for recommending site views that match the given profile.

B. *Identifying Interesting Site Pages*

So far we have presented our approach towards the automatic identification of the user's topic interests within a site's contents and we introduced a scoring function for quantifying the degree of the user interests. However, although a user may be strongly interested in a particular topic, she may be less interested in some pages even if these pertain to her preferred topic.

To ensure that our customization model will be capable of identifying among a set of topic relevant pages, the ones that are of true interest to the user, we rely on the subject hierarchy and we compute the degree to which every page within a site correlates to the user interests.

Given that the user interests are represented as a set of hierarchy topics weighted by their degree of interestingness, our aim is to represent the site pages in an analogous manner, i.e. as a set of topics weighted by their degree of topic relevance. Based on the above data, we can approximate the degree of the user interests in particular pages as a function of the of the user's interests in the site's topics and the pages' topic relevance values.

To represent the pages within a site as a set of topics weighted by their topic relevance scores, we firstly need to identify the topics of every page in a site and thereafter compute the degree to which each of the pages relates to the identified topics. To do that, we pre-process pages as discussed in Section 2.1 and we rely on the subject hierarchy in order to disambiguate their weighted keywords. We then group every page's keywords into topical clusters and we rely on the average keyword's importance weights in order to estimate the importance that the items in every cluster have to the cluster's topic. Formally, the average importance of keywords k (denoted as Wk) in a lexical cluster of *m* items (denoted as Cm) to some topic T of the cluster is:

$$\text{Avg. Importance}(\text{Cm}, \text{T}) = \frac{1}{|\text{m}|} \sum_{\text{k}=1}^{\text{k}=\text{m}} \text{Wk}(\text{T}) \qquad \textbf{(5)}$$

where m is the total number of lexical items in the cluster. Having computed the average importance weights of the page's keywords to their respective clusters, we employ the average importance of each of the identified clusters as a measure for indicating the degree of the page's relevance to the respective cluster's topic. Formally, the page's relevance to a topic T is the sum of importance over all its keywords' whose topic label is T:

$$\text{Relevance}(\text{P}, \text{T}) = \sum_{\text{Cm} \in \text{P}} \text{Av. Importance}(\text{Cm}, \text{T}) \qquad \textbf{(6)}$$

Based on the above, our model represents the pages within a site as a set of topics weighted by their degree of relevance to the page's keywords. We now describe how we combine the topical categories computed for the site's pages and the topical interests identified in a user's site visits in order to estimate the degree to which each of the site pages' might be of interest to the user.

To measure how interesting is a page P that relates to some topic T to the user with some interest in T, we rely on the correlation between the page's relevance to T and the user's interest in T, as:

$$\text{User Interest}(\text{P}) = \text{User Interest}(\text{T}) + \text{Relevance}(\text{P}, \text{T}) \qquad \textbf{(7)}$$

Based on the above formula, we can compute the probability that page P, which relates to topic T will be of interest to the user who has some interest in T. Next, we describe how our model selects which pages to recommend to a user during her site visits, in an attempt to improve the user's interaction with the site's content.

C. *Building Recommendations*

As mentioned before, the aim of a recommendation system is to suggest web site users with a set of pages that are deemed relevant to their interests. Therefore, the recommendation system is responsible for deciding which site pages correlate to the user interests and based on this decision to present the site's contents accordingly.

However, the greatest challenge that a recommendation system has to address is the so-called *portfolio effect* problem, i.e. how to ensure that the pages it recommends are not already seen by the user. An ideal recommender must be able to distinguish between *useful-but-unvisited* and *useful-but-visited* pages, and it must also infer whether the user wants to see new pages (i.e. unvisited), old pages (i.e. visited) or both.

In the course of our study, we have built a recommendation mechanism that tries to minimize the impact of the portfolio effect problem without asking for the user involvement. To tackle the first difficulty, i.e. to distinguish between visited and non-visited pages we obviously rely on the site's transaction logs where we record the time and frequency of page accesses by a user. Based on this data, we can easily identify the pages that the user has already visited (either recently or at some point of time) and exclude them from the recommendations offered. Alternatively, we could exclude only recently visited pages, based on the intuition that pages not accessed for a long time might be of interest to the user either because she may not recall having visited them or because the page's contents have been updated with data that the user has not seen in her previous visits.

But, the greatest challenge is to guarantee that the recommendations offered to the users will meet their expectations. In other words, we need to ensure that visited pages will be excluded from the recommendations only when the users do not wish to visit pages already seen. Likewise, we need to ensure that visited pages will not be excluded from the recommendations only if the users whish to revisit site pages.

To tackle the above difficulties, we rely on the distinction between persistent and ephemeral user interests, discussed in [20] and we introduce a novel approach for predicting the user's preferred recommendations based on the analysis of her navigational patterns. In our work, we perceive a user's interest in some page P to be persistent if the user regularly visits P across her site transactions, while we perceive the user's interest in P to be ephemeral if the user visited P arbitrarily in some past site visits. Given the separation between persistent and ephemeral user interests in the site pages, we can predict the interests characterizing the users' visits based on the following criterion:

$$\text{Interest(P)} = \{\text{ephemeral} \quad \text{if avg. number of clicks on P} < \text{F}$$
$$\{\text{persistent} \quad \text{if otherwise}$$

where F value is selected based on the expected distribution of clicks for persistent and ephemeral interests in the site pages.

Based on the intuition that the way in which a user interacts with a site demonstrates some stereotypical patterns (e.g. site pages visited in the same sequence) we speculate that their modeling can help us predict the patterns of the user's future interactions. In other words, if a user tends to revisit some of the site pages in the same sequence or when browsing site pages that deal with a particular topic, then the user will keep revisiting them in her future site accesses.

To estimate the expected distribution of clicks in the user's future transactions with a site, we rely on the analysis of the user's past clicks distribution on the site's pages and proceed as follows. We sort the visited pages in a site in the descending order of the number of clicks that they have received from the user. We then compute the average click distribution on the site's pages and we heuristically set the threshold value of F (i.e. the expected distribution of future clicks on each site page).

Under this approach, our model predicts the type of the user interests in the site's pages and depending on that prediction, it makes decisions about whether to include a particular site page in the recommendations or not. In particular, if the value of F for some visited pages in a site is above the threshold, our model recommends them regardless of the fact that these have already been seen by the user. Alternatively, if the value of F for all the site's visited pages is below the threshold, our model recommends only new (i.e. unvisited) pages to the user. Note that in both cases, our model primarily relies on the pages' interestingness to the user preferences (see previous section) and upon identification of user interesting pages; it predicts the type of the user interest in their contents.

In a similar manner, and by employing different heuristics and threshold values, one could accommodate the case that the user wants to re-visit some but not all of the frequently revisited pages. However, we defer this investigation for a future study.

Following the process presented above, our model selects the pages to recommend to the user and orders recommendations in a way so that pages with the highest probability of being interesting show up first on the list of recommendations. The user can then interact with the recommendations by clicking on any of the suggested pages. The recommendations that our model suggests rely on the learnt user interests and as such recommendations are dynamic in the sense that as our mechanism gets to learn the user interests, it explores the accumulated knowledge in the subsequent user recommendations. In the current implementation of our model, the recommendations suggested change per user session, since the same user might be interested in different topics during different site visits. However, considering that the user's topic preference might change during a single visit (i.e. session) or that the user might have more than one topic interests in the same visit, our model can be modified and use a sliding time window for generating recommendations.

III. EXPERIMENTS

We now discuss the experiments we conducted in order to evaluate the effectiveness of our proposed site customization model and we present obtained results. We first describe our experimental setup. Then in Section III.B we describe a simulation-based experiment to estimate the accuracy of our model in offering customized site views according to the user interests. Finally, in Section III.C we present the results from a human study that measures the perceived usefulness of our recommendation mechanism.

A. Experimental Setup

To evaluate the effectiveness of our site customization approach, we implemented a browser plug-in that records the users' navigational behavior in their Web site visits. We then recruited 10 postgraduate students from our school, who were informed about our study and volunteered to install the plug-in and supply us with information about their Web transactions.

During their participation in the survey, our subjects were asked to keep a diary of preferences in their web site visits in which to record for each of their sessions, which was their preferred topic and which were the site pages that interested them the most in each of their site visits. To denote topic preferences we asked our subjects to use the text descriptors of the 16 top level Dmoz topics that are used to label the concepts in our hierarchy. Moreover, to denote interesting pages we asked our subjects to rate every page in a site, using scores ranging from 1, meaning "not interesting at all" up to 5 "very interesting". Finally, we asked our subjects to indicate whether they had a persistent or ephemeral interest in each of the site pages.

We used the log files collected from our subjects' web accesses for a period of two months, we cleaned them from hits that were redirected or caused errors, we removed records accounting to non-textual Web accesses and we stored the cleaned data in a RDBMS server. Table 1 shows statistics on our experimental data.

TABLE I
STATISTICS ON THE EXPERIMENTAL DATASET

Collection period	March-April 2007
# of users	10
# of sites visited	168
# of log files	2,981
Avg. # of pages visited per site	34.3
Avg. # of hits per day	295

We downloaded all the pages from each of the sites in our dataset, we processed them following the steps described in Section 2.1 and we computed for every page a suitable topic from the hierarchy in order to model the page's semantic content. Thereafter, we processed our users' site transaction logs in order to mine their navigational patterns and derive the user interests in each of their site visits.

In particular, we computed for each of our subjects and for each of their visited sites, their most preferred topic, the sites' pages that were most correlated to their interests as well as a number of recommendations for customizing the presentation of their visited sites. The computations of these values were performed on a workstation with a 2.4GHz 2 CPU and 2GB of RAM and took roughly 50 hours to pre-process our experimental data, to estimate the user's topic and page preferences and to build recommendations for each of our subjects.

B. Accuracy of User Interests Identification

In this section, we measure the accuracy of our methods in identifying the user site interests and in recommending useful pages to the site users. In this respect, we are primarily concerned with both the accuracy of our method and the amount of log data it requires for estimating accurate user profiles.

1) ***Accuracy of Identified User Interests***: To measure the effectiveness of our model in identifying the user topic interests in their site navigations, we relied on a synthetic dataset that we generated by simulation based our experimental site transaction logs.

In our implementation, the number of sessions in the user's site accesses is fixed to S as an experimental parameter and we assign to every session page visits as follows. We set a random set of topics the user is interested in every session to T and we distribute an equal number of page visits V to each of the topics in every session. In our implementation we set V=35 based on the findings of [24] that the majority of sessions is 34 pages or longer. Once we generate a user's sessions we compute the user's interest in each of the topics across sessions (equation 3). Note that under our implementation the user's interest is equally distributed across the topics in each of the sessions and interest values are normalized to sum up to one. Based on the above, we derive a baseline topic interest estimation which we compare it against the estimations derived by our model.

To evaluate the accuracy of our topic identification approach, we measure the relative error for our estimated site interests compared to the baseline estimation, which assumes equal weights for every topic. For our comparison, we use:

$$E(T_i) = \frac{|T_i - T|}{|T|} \qquad \textbf{(8)}$$

where T denotes the user's actual topic interests (i.e. baseline estimation) and Ti denotes the topic interests identified by our model. Figure 3 shows the results.

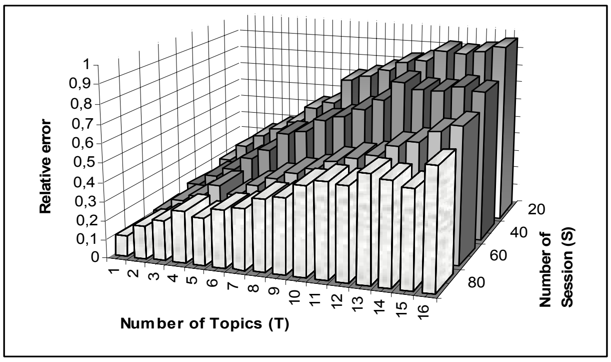

Fig. 3. Relative errors in estimated topic interests.

From the figure, we see that at the same T value, as the number of session S increases, the relative error of our method decreases. This practically implies that the more sessions considered about a topic, the less the relative error in estimating the degree of the user's interest in that topic.

For example, we can see that when the number of topics in which the user is interested in T=5, the relative error of our method when considering 20 sessions is 0.45 and it goes down to 0.25 when S=60. Moreover, we observe that when users are interested in a relatively small number of topics (i.e. 1≤T≤5) our method estimates their interests with an overall accuracy of 66.6% when considering only 20 sessions, which goes up to 73.2% when 40 sessions are considered. On the other hand, we observe that when the user is interested in many topics (i.e. 6≤T≤10), our method achieves an overall estimation accuracy of 62.6% when considering more sessions, i.e. S=80. Analogously, when the user is interested in more that 10 topics, our method has only a 46.2 overall accuracy in estimating the user interests through the examination of 80 user sessions. This practically implies that for a large number of different user interests, we need to collect a large amount of user logs until we can effectively estimate the degree of the user interests.

Compared to the baseline estimation, our method can effectively identify the user interests in their site navigations when interests vary between 1 and 7 topics with an overall estimation accuracy of 66.6% and considering only 40 user sessions. Therefore, we may conclude that our user interests' identification method needs only a small amount of log transactions to effectively estimate the degree of the user interests in the sites' contents, when interests span a relatively small (i.e. up to 7) number of topics.

2) ***Accuracy of Recommendations***: In this section, we experimentally investigate the accuracy of our method in recommending useful pages to the web site users, based on their identified topic preferences. To measure this accuracy, we again generate synthetic data for user navigations and estimate the degree of the user's interest in specific pages. Based on our estimations, we build recommendations and we sort them in terms of the pages' interestingness to the user preferences.

To evaluate the accuracy of our model in picking the most interesting pages to recommend to web site users, we rely on the Kendall's distance metric [11] between our estimated ordering of page recommendations and the ideal recommendations' ranking. Formally, the Kendall's distance metric (τ) between two ordered lists of recommendations is given by:

$$\tau\,(E_K, A_K) = \frac{\left|(\,i, j):i, j \in R, E_K(i) < E_K(j),\ A_K(i) > AK(j)\right|}{|R| \bullet |R - 1|} \qquad \textbf{(9)}$$

Where E_K denotes the ordered list of the top-k recommended pages estimated by our method, A_K denotes the ordered list of the top-k recommended pages computed from the user's actual topic preferences, R is the union of E_K and A_K and (i, j) is a random pair of distinct pages. τ values range between 0 and 1, taking 0 when the two orderings are identical. Given that most users visit on average 35 pages in their site navigations [24] we set the value of k between 5 and 40. We believe that the choice of k is reasonable based on the intuition that web site users would not like to see too few or too many pages in the recommendations offered. Figure 4 shows the differences in the recommendations' ordering for k=10, i.e. when considering the top 10 recommended pages[5].

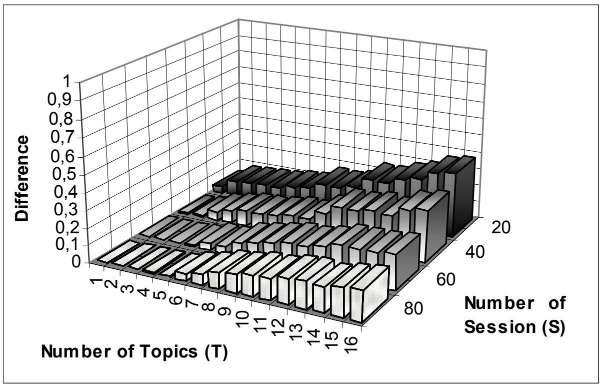

Fig. 4. Ordering differences of top 10 recommendations.

We can see that our method has a significant potential in building useful recommendations to the web site users, even in cases that the user's topic interests cannot be precisely identified. In particular, we observe that when the user is interested in 5 topics (i.e. T=5) and there are 20 sessions considered about the user, the ordering of the recommendations given by our system has a distance of 0.1 compared to the ordering of the actual recommendations. This implies that only 10% of the pairs (i.e. 1 out of the 10 pairs considered) in our recommendations are reversed compared to the true recommendations. Moreover, for 1≤T≤7[6] the average distance between the estimated and the actual recommendations, when 40 sessions are considered, is 0.06, which implies that our model has 94% accuracy in estimating useful recommendation lists compared to the actual recommendation lists when a small number of user sessions is considered.

C. *Quality of Site Customizations*

To measure the quality of our site customization approach, we used the data collected from our human survey (Section III.A) and we evaluated our model's effectiveness in offering useful recommendations to web site users. In our evaluation, we relied on the collected transaction logs and we computed for each of our subjects and visited sites the degree to which each of the site pages would make a useful recommendation.

To evaluate the effectiveness of our model in identifying useful pages in the sites' contents, we compared the pages' estimated usefulness to the pages' actual usefulness, as the latter is explicitly indicated by our subjects. In our comparisons, we relied on the following formula for measuring the pages' usefulness.

$$\text{Usefulness}(P) = \sum_{P \in S} \text{User Interest}(P) \bullet F \qquad \textbf{(10)}$$

Here S denotes the pages in a site, F denotes the probability that the user will revisit P and Usefulness (P) denotes the usefulness of P by the given method. The degree of the User Interest in P (User Interest (P)) that our model estimates is given by equation 7, while the actual user interest in P is explicitly indicated by our participants. Recall that while recruiting our subjects we asked them to rate each of the site pages according to how interesting these are and we also asked then to indicate the type of their interest.

In order for our model to predict the type of the user interest in each of the site pages, we relied on the pages' expected click distribution and we set the threshold value F=0.44, based on the findings in [24] that the average page revisit rate is roughly 44%. That is, pages with a predicted click distribution above F are pages in which our model predicts a persistent user interest. Moreover, to quantify the type of the user interest indicated by our subjects, we set the value of F above 0.44 for the pages in which our subjects attributed a persistent interest, and below 0.44 otherwise. Finally, we equally distributed the ≥F and ≤F values for persistent and ephemeral page interests, respectively.

Based on the above formula, we evaluate our model's accuracy in identifying useful pages to the site users by comparing the pages' estimated usefulness to their actual usefulness, denoted by our subjects. Figure 5, reports the average differences between the estimated and the actual usefulness values of the pages in the sites that our subjects have visited during the reporting period. The bars on the horizontal axis represent the 10 subjects in our study. Scores are normalized to sum up to one and they are aggregated by users in the sense that for all the pages in the sites that a user has visited we measured both their average actual and estimated usefulness and we report the difference between the two. Average differences between the actual and the estimated pages' usefulness help us evaluate the overall effectiveness of our model in identifying useful pages in the sites' contents. That is, the lower the average difference values between the pages' actual and estimated usefulness, the better the performance of our model in identifying valuable recommendations to the site visitors.

Results show that our model is very effective in estimating useful pages in the sites' contents, achieving an overall accuracy of 0.763. As we can see, our method can accurately estimate the usefulness that most of the site pages have to particular user interests. Results indicate that our site customization model, which attempts to provide site visitors with valuable recommendations, has a significant potential in accurately estimating how useful is every page in a site to the user preferences. Based on the estimated usefulness values for the site pages, our model decides which pages to recommend to the site visitors, as well as the recommendations' ordering so as to enable customized site views for individual users.

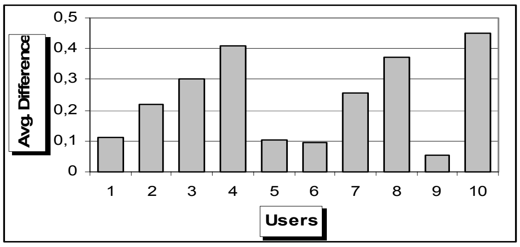

Fig. 5. Average difference values for the actual and the estimated usefulness of the site pages that users have visited.

[5] We obtained similar differences in orderings for larger k values.

[6] For that number of sessions and topics our model has a topic interest estimation accuracy of 66.6%.

IV.RELATED WORK

Many researchers have proposed ways of customizing Web sites to the needs of specific users [5] [7] [8] [9] [15]. For an overview we refer the reader to the work of [22]. Most of these efforts use data mining techniques in order to extract useful patterns and rules from the users' navigational behavior and based on these patterns they modify the site's content and structure so as to meet specific user interests. In the last years, there has been a surge of interest into enabling semantically-driven site modifications [8] [14]. In this respect, researchers have explored the use of ontologies in the user profiling process. We refer the reader to the work of [10] for an overview on the role of ontologies in the site customizations. There also exist studies [12] [16] that examine the sites' content in order to extract specific features. Such features are integrated in the customization process, so as to retrieve similarly characterized content. In this direction, researchers have investigated the problem of providing Web site visitors with recommendations that relate to their interests [4]. Recommendation systems match the user activity against specific profiles and provide every user with a list of recommended hypertext links. In [13] the authors used an ontology and utilized the Wu & Palmer similarity measure [21] for estimating the impact of different concepts on the users' navigational behavior. However, there are no results reported on the impact of the recommendation process. In our work, we extend the approaches suggested by other researchers and we combine them in novel ways in an attempt to build a recommendation mechanism that explores a subject hierarchy not only towards the identification of the user interests, but also towards the evaluation of the identified preferences in terms of time persistence. Moreover, we introduce some novel measures for the estimation of the user interests, which we deem complementary to existing ones.

V. CONCLUDING REMARKS

In this paper, we have proposed a site customization approach that explores a subject hierarchy for building user profiles as well as for identifying the thematic content of the sites' pages. Based on the association between user profiles and page semantics, our model mines the users' navigational patterns in the sites' contents in order to identify useful pages to recommend to the site visitors. We have conducted experiments to evaluate the effectiveness of our model. In one experiment using synthetic data, we found that for a relatively small number of user sessions (i.e. 40), our method yields a significant accuracy in estimating the user interests in the sites' contents and that our recommendation model, which relies on the user interests has a great potential in estimating valuable recommendations to the sites' visitors. In another real-life experiment, we applied our method to estimate useful pages for 10 subjects and we assessed the effectiveness of our model in estimating valuable recommendations to web site users. Obtained results demonstrate that on average our approach successfully estimates which pages might be useful to the site users. In the future, we plan to expand our framework to take across-site user-specific information into account so as to provide users with valuable recommendations from different sites with similar content or from sites in the contents of which the user has similar interests.

REFERENCES

[1] Open Directory Project (ODP): http://dmoz/org

[2] RFC. Identification Protocol. http://www.rfc-editor.org/rfc/rfc1413.txt

[3] WordNet: http://wordnet.princeton.edu

[4] Baraglia R, Silvestri F. An Online Recommender System for Large Web Sites. In Web Intelligence Conference, 2004.

[5] Berendt B., Spiliopoulou M. Analysis of Navigation Behavior in Web Sites Integrating Multiple Information Systems. In VLDB Journal, 9: 56-75, 2000.

[6] Chakrabarti S., Dom B., Gibson D., Kleinberg J., Kumar R., Raghavan P., Rajagopalan S., Tomkins A. Mining the Link Structure of the World Wide Web. In IEEE Computer, 32(6), 1999.

[7] Coenen F., Swinnen G., Vanhoof K., Wets G. A Framework for Self Adaptive Websites: Tactical versus Strategic Changes. In the WEBKDD Workshop, Boston, MA, 2000.

[8] Dai H., Mobasher B. Using Ontologies to Discover Domain-Level Web Usage Profiles. In Workshop on Semantic Web Mining, 2002.

[9] Eirinaki M., Vazirgiannis M., Varlamis I. SeWeP: Using Site Semantics and a Taxonomy to Enhance the Web Personalization Process. In the SIGKDD Conference, 2003.

[10] Eirinaki M., Mavroeidis D., Tsatsaronis G., Vazirgiannis M. Introducing Semantics in Web Personalization: The Role of Ontologies. In LNAI 4289, pp. 147-162, 2006.

[11] Kendall M., Gibbons J. Rank Correlation Methods. Edward Arnold, London, 1990..

[12] Jin X., Zhou Y., Mobasher B. A Maximum Entropy Web Recommendation System: Combining Collaborative and Content Features. In the ACM KDD Conference, 2005.

[13] Kearney P., Anand S. Employing a Domain Ontology to Gain Insights into the User Behavior. In Workshop on Intelligent Techniques for Web Personalization, 2005.

[14] Middleton S.E., Shadbolt N.R., De Roure D.C. Ontological User Profiling in Recommender Systems. In ACM Transactions on Information Systems, 22(1): 54-88, 2004.

[15] Mobasher B., Dai H., Luo T., Sung Y., Zhu J. Discovery of Aggregate Usage Profiles for Web Personalization. In the Web Mining for E-Commerce Workshop, 2000.

[16] Perkowitz M., Etzioni O. Adaptive Web Sites. In Com. of ACM, 43(8):152-158, 2000.

[17] Spiliopoulou M. Web Usage Mining for Web Site Evaluation. In Communications of the ACM, 43(8): 127-134, 2000.

[18] Srivastava J., Cooley R., Deshpande M., Tan P.N. Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. In SIGKDD Explorations, 1(2): 12-23, 2000.

[19] Stamou S., Krikos V., Ntoulas A., Kokosis P., Christodoulakis D. Classifying Web Data in Directory Structures. In the 8th APWeb Conference, pp. 238-249, 2006.

[20] Sugiyama K., Hatano K., Yoshikawa M. Adaptive Web Search Based on User Profile without any Effort from Users. In the WWW Conference, pp. 675-684, 2004.

[21] Wu Z., Palmer M. Web Semantics and Lexical Selection. In the 32nd ACL Meeting, 1994.

[22] Eirinaki M., Vazirgiannis M. Web Mining for Web Personalization. ACM Transactions on Internet Technology, 3(1): 1-27, 2003.

[23] Anderson C., Horvitz E. Web Montage: A Dynamic Personalized Start Page. In the WWW Conference, 2002.

[24] Obendorf H., Weinreich H., Herder E., Mayer M. Web Page Revisitation Revisited: Implications of a Long0term Click-stream Study of Browser Usage. In CHI Proceedings, 2007.