A Recommendation Model Based on Site Semantics and Usage Mining

Sofia Stamou Lefteris Kozanidis Paraskevi Tzekou Nikos Zotos

Computer Engineering and Informatics Department, Patras University 26500 GREECE

{stamou, kozanid, tzekou, 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 three great challenges: (i) the accurate identification of the user interests in the sites' content (ii) the detection of the user goals in their site visits and (iii) the encapsulation of the user interests and goals into the sites' presentation and content.

In this paper, we study how we can effectively identify the user interests and goals in their site visits and we evaluate their correlation as this is exemplified in the users' navigational patterns and the site's semantic content and structure. Based on our findings we propose a novel recommendation mechanism that employs web mining techniques for suggesting customized site views that satisfy both the user preferences and goals. Our experimental evaluation shows that the user site interests and interaction goals can be accurately detected from the users' navigational behavior and that our recommendation model, which uses the identified user preferences and goals yields significant improvements in the sites' usability.

Categories and Subject Descriptors

H.5.4 [**Hypertext/Hypermedia**]: Navigation, User issues; H.3.1 [**Content Analysis and Indexing**] Linguistic processing; H.3.m [**Information Systems**]: Miscellaneous

General Terms

Experimentation, Measurement, Performance, Human Factors.

Keywords

Web usage mining, web site customization, user profiles, subject hierarchy, recommendations

1. 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 needs.

Adjusting the content or structure of web data to specific needs 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] [26] [27] 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 contents change and new pages are added that are not yet included in the web log [20]. To overcome such difficulties, researchers have proposed the exploitation of information about the content [11] [21] and/or the structure [9] 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 [12] [29].

The commonality in most of the existing site customization approaches is that they attempt to model the user interest as a set of topics weighted by their degree of preference. Although this method is successful for building general user profiles, nevertheless it is inadequate for deciphering the specific user goals in their site visits. To make the distinction between user interests and interaction goals clear, consider the following situation. User *A*, an engineer, is interested in *programming* and visits a number of sites to find information about her subject of interest. In each of her site visits though, the user intents to obtain different types of information such as download sample source code, read a document for an in-depth background in programming languages, etc. Evidently, if we were to provide that user with customized site views we would not only need to recommend her pages about programming, but also suggest pages that contain the exact type of information about programming that satisfies the user's goal.

In this paper, we extend previous works on site customization and we introduce a novel recommendation model that combines in new ways the sites' usage patterns and semantics so as to derive knowledge about both the users' site interests and interaction goals. 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. Moreover, it relies on the association between the sites' usage and structural data in order to detect the user goals that correspond to particular interests and 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 and goals in their site visits as these are exemplified in the user's navigational patterns. For computing the user interests in site visits our approach relies on a subject hierarchy and employs a number of heuristics for estimating both short-term and long-term interests. For identifying the user goals in site visits our approach relies on the sites' structural properties and mines the way in which content types influence the users' interaction with the site. The contribution of our user interests and goals detection model relies on the combination of site semantics and user behavioral features in an attempt to build site-specific user profiles as opposed to generic user profiles that most of existing profiling methods pursue.
- We show how we can explore the identified user interests and goals in order to detect within a site which pages are both interesting and useful to the user. The computations of interesting pages rely on (i) the pages' semantic content as this is processed and evaluated via the use of the subject hierarchy and (ii) the pages' structural properties and features as these are determined by the analysis of the pages' metadata annotations. Unlike existing site customization approaches that conflate interesting and useful pages our models makes a clear distinction between the two; based on the intuition that a page dealing with an interesting topic is not always useful to a particular user intention.
- We introduce a recommendation model that correlates the identified user interests and goals 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 an experimental 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 investigated the effectiveness of our model in identifying the user goals in their site visits based on the analysis of the previous user interactions with the sites' content. Finally, we evaluated the effectiveness of our recommendation system in improving the sites' usability and hence in ameliorating the users' navigational experience. Results indicate that although users with alike interests might have diverse goals in their site visits and vice versa, our approach manages to effectively identify them both. Moreover results indicate that recommendations based on rich user profiles (i.e. profiles that consider both preferences and goals) are valued higher compared to the recommendations that are determined from the user interests alone.

The rest of the article is organized as follows. We begin our discussion with a detailed presentation of our web site customization model. In Section 2.1, we present our method for the automatic identification of the user interests in their site visits, while in Section 2.2 we introduce a number of features for the prediction of the user goals in their site accesses. In Section 2.3 we describe how we process the web sites' structure and content in order to identify which of the site pages match the identified user interests and goals. Finally, in Sections 2.4 we introduce our recommendation mechanism. In Section 3 we evaluate the effectiveness of our approach and we discuss obtained results. In Section 4 we review related work and we conclude the article in Section 5.

2. Web Mining for Improved Navigations

In our work, we suggest the exploitation of the sites' structural properties and metadata for their annotation with content types. Depending on the type of information that the site pages contain and the way in which users interact with them, we can learn the user goals in their site visits. Moreover, 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.

In a high level, our method proceeds as follows. Given a site's web logs, we pre-process them in order to automatically identify the user goals and interests in their site visits. To derive the user goals, we parse the site's structural metadata in order to derive the content type of the visited site pages. We then mine the site's transaction logs in order to cluster the users' interaction with the site's content into one of the following types: (i) download a resource, (ii) obtain information and (iii) interact with a person, a program or an application. Finally, we match the content types associated to every visited page and the interaction types exemplified in the user visits in order to infer the latent user goals in their site accesses.

Furthermore, to identify the user interests in their site visits, we rely on the site's semantic content and we use a subject hierarchy in order to annotate every page visited within a site with an appropriate topical category. 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.

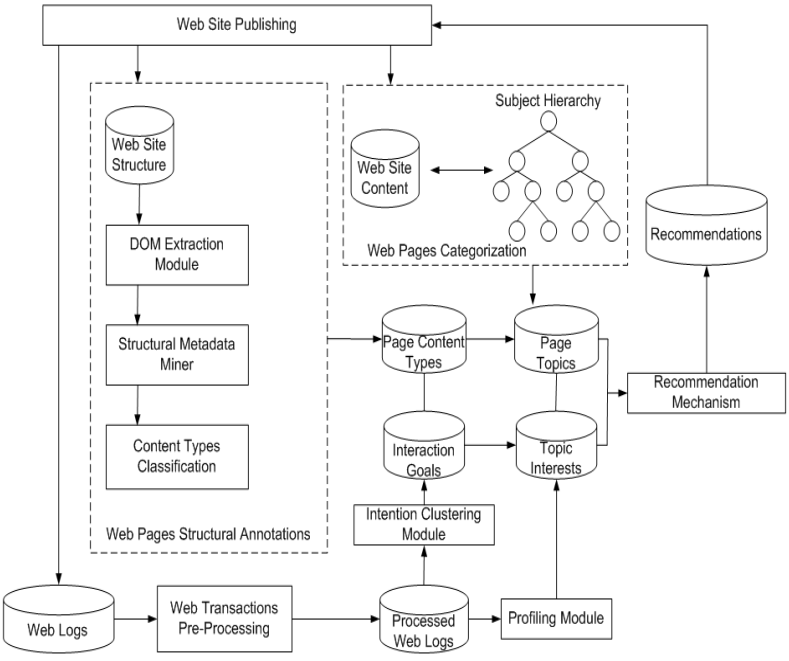

Figure 1. System architecture.

Having computed the user profiles we enrich them with information about the learnt user goals in their past site visits. This way, we derive rich user models that indicate not only the user preferences in the site's content but also the reasons why users interact with the site.

Following the identification of the user interests and goals in their site visits, we proceed with the exploitation of the site's structural and semantic content in order to identify which of the site pages might be useful and of interest to the user. In our model, an interesting page is a page that deals with a topic in which the user is interested in, whereas a useful page is an interesting page that contains the type of information that most likely satisfies the user's goal.

To identify interesting pages, 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.

Moreover, to identify useful pages, we annotate every page in a site with an appropriate content type depending on the pages' structural and metadata elements. The annotations given to represent the pages' content types are borrowed from the labels selected to describe the user goals, i.e. transactional, informational, resource.

As a final step, our approach correlates the information obtained from the user's navigational behavior and the site's structural and semantic content in order to recommend interesting yet 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.

2.1 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, clickstream 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[2], we apply tokenization, POS-tagging and we remove their stop words. Thereafter, we rely on the page's content terms[3] 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 [7]. 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. [28]) 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 [28] 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 [30], 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:

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

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

[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] Markup annotations are stored in a utility index for separate processing, as described in Section 2.2.

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

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[4] 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.

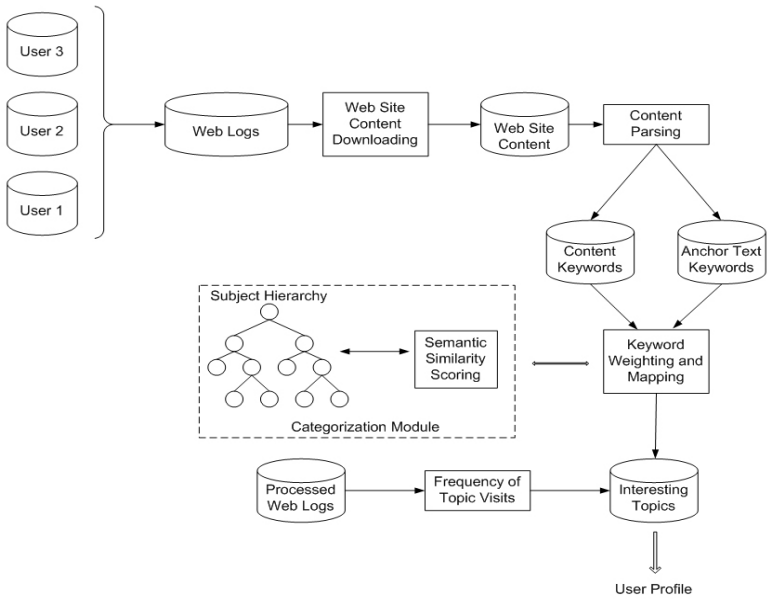

Figure 2. The user profiling process.

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[5] 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:

[4] Note that every node in our hierarchy is annotated with a suitable topical category borrowed from the top level topics in the Dmoz ontology.
[5] In our work, we define a user session as a delimited time-ordered set of user clicks to a single web server. A user session is also called a visit.

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

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:

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

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:

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

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.

2.2 Identifying the User Goals in Site Visits

In the previous section we presented our method for identifying what users are looking for in their web site visits, i.e. what are the topics that interest them in the site's content. In this section, we address the problem of understanding the user intentions in their site visits, i.e. finding out why users interact with web sites. Although the problem of identifying the user goals in their web accesses is not new, nevertheless existing works concentrate on the web search paradigm and they explore the user issued queries for detecting their search intentions [6] [15] [24]. In our work, we investigate the goal identification problem from the site access perspective and we try to decipher the correlation between user goals and web usage patterns.

The first step in this direction is to determine a set of possible goals that users might have in mind while visiting a web site. One way to go about this is to ask users explicitly describe their intentions. However, this is not only impossible in a practical setting but it also puts an extra burden that the user might be reluctant to take. Another approach is to implicitly lean the user goals based on the analysis of the site's usage and structural content.

In our work, we opted for the second approach and we have designed a generic framework that tries to identify the user goals in their site visits based on the sites' structural properties and usage mining. The core idea of our approach relies on the intuition that there are three general types of activities[6] that users perform while interacting with a site, namely:

[6] Activity types in user interactions are inspired from the search goal types described in [24].

(i) Navigate in the site's content in order to obtain information about a subject of interest. We denote the goal that characterizes this type of interaction as *informational*[7].

(ii) Access the site's content in order to obtain a particular resource (e.g. song, file) and not information. We denote the goal that characterizes this type of interaction as *resource*.

(iii) Visit a web site in order to interact with a person (e.g. chat) or a dynamic web service (e.g. online gaming, booking, etc.) We denote the goal that characterizes this type of interaction as *transactional*.

To identify which of the above goals or combination of goals the users pursue in their site visits, we process the site's structural data in order to annotate every visited page within a site with one of the following content types: informational, resource, transactional. For processing the site's structure we rely on the site's URLs and their organization (i.e. sitemap), the site pages' markup annotations as well as the displayed anchor tags that point to the site pages. Based on the above data, we firstly parse the site pages' URLs in order to identify which pages contain downloadable material. This can be easily derived from the URL suffix, i.e. pages with URLs ending in .exe, .mp3, .PDF or other popular file types are annotated as resource pages.

For the site pages whose URL is not directly informative of their content type, we explore their structural properties and anchor tags and look for features such as frames, audiovisual data, Java Scripts, flashes, calendars, forms, etc. Based on the findings of [16] that pages containing frames, calendars, badges, comments, archives, etc. are blog or forum pages and relying on the empirical knowledge that pages containing forms, search interfaces or multimedia data are interactive pages, we annotate pages of such structural content as transactional.

Finally, we parse the pages' markup annotations in order to detect the presence of links, images, documents, etc. within the pages' content. Pages containing simply links, body text, tables, images and figures are annotated as informational, whereas pages with embedded files are annotated as resource pages.

Having annotated every page in the site's transaction logs with an appropriate content type, we mine the users' site interactions in order to derive the user goals in their site visits. In particular, we rely on the site's log files out of which we extract information about the number of times a user has clicked on pages of the same content type. We then heuristically estimate the user's goals in her site visits by associating the visited pages' content types and the user's clicking behavior.

For our estimations, we group by content type the site pages that a user has visited in her recorded sessions and we compute the distribution of clicks on them. In particular, we firstly sort the visited pages in every group (i.e. informational, transactional or resource) in descending order of the number of clicks they have received from the user across her site visits. Afterwards, we create a histogram where the i^{th} bin corresponds to the number of clicks accumulated on the i^{th} page of the group. We further normalize the frequency values so that these values add up to 1. Given the click distribution on the pages of a particular content type, we can infer the goal for that site visits by investigating how that click distribution is skewed towards the pages' content types.

Intuitively, a highly skewed distribution suggests that pages of a particular content type are clicked by the user much more often than pages of other content types. Accordingly, the goal of the user visits should be the one that corresponds to that pages' content type, i.e. informational, resource or transactional. On the other hand, a flat distribution of clicks on pages of different content types suggest that the user's goal is mixed or absent (i.e. the user had no particular intention while visiting the site).

For example, Figure 3(a) shows the click distribution on site A (hhtp://planet-source-code.com) for pages of different content types. The rightmost bar in the figure shows that the primary goal for the user visiting site A is to obtain a resource (resource pages get 85% of user clicks) rather than obtain information (information pages receive 10% of user clicks) or interact with others (transactional pages get 5% of user clicks). A look at site A (Planet Source Code) reveals that most of its content concerns downloadable material (i.e. resource content), whereas there are some online tutorials (i.e. informational contents) as well as a postings forum where the users rate and comment on the offered material (i.e. transactional content). Based on the above observations and considering our findings in Figure 3(a), we realize that the predominant goal of a user visiting site A is to obtain a resource.

[7] Although in the web search paradigm there is a clear distinction between informational and navigational queries, in the site interaction model the two types of goals are conflated due to the lack of search queries.

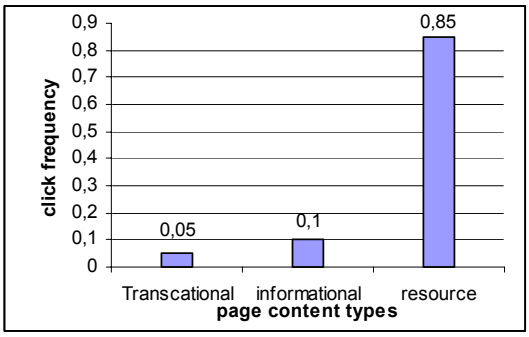

Figure 3(a). Click distribution for user goals in site A.

Alternatively, Figure 3(b) shows that the user visited site B (hhtp://techrepublic.com.com) either with a mixed goal or without a particular intention (no particular type of pages is clicked more often that others). Again a closer look at the contents of site TechRepublic reveals that it contains plenty of information about IT professionals (i.e. informational content), it has blogs and forums (i.e. transactional content) where users can exchange ideas, comments and experiences on IT topics and it also includes a downloads page (i.e. resource page) where user can obtain material for further usage beyond their site interaction. Therefore, a user visiting the above site might either pursue distinct goals at the same time or she might have no particular intention in mind while accessing the site.

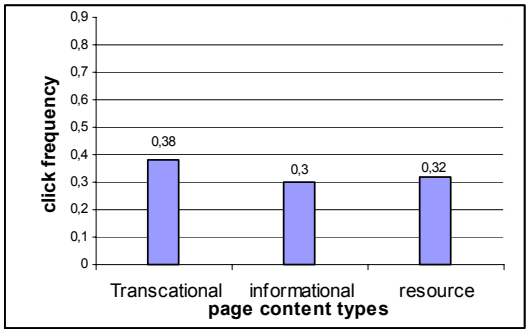

Figure 3(b). Click distribution for user goals in site B.

Based on the intuition that the user goals while visiting a site may be learned from the way in which she has previously interacted with the site's pages, our method attempts to identify what is the main intention of the user visits to a given site. To infer the goal that characterizes the user visits to this site's content, we use the following criterion: the user goal while accessing a particular site is primarily: (i) informational, if the distribution of clicks on informational pages is above threshold D, (ii) resource, if the distribution of clicks on resource pages is above D, (iii) transactional, if the distribution of clicks on pages of transactional content is above D, and (iv) mixed or unknown otherwise. The value of D, is experimentally set to D = 0.8 which practically implies that in order to be able to clearly identify the user's goal in her site accesses, the user should devote at least 80% of her site visits to pages whose content type meets that goal.

2.3 Identifying Interesting and Useful 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 (cf. Section 2.1). We have also proposed a method for identifying the user goals in their site interactions (cf. Section 2.2). 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. Moreover, the user may have varying intentions every time she interacts with highly interesting pages.

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.

Additionally, to guarantee that our model can effectively identify among the site interesting pages the ones that are likely to satisfy the user goals, we rely on the pages' structural content and we examine how different content types correlate to the user intentions.

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:

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

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:

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

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:

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

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.

Having estimated the degree to which each of the site pages meets the user interest, our next step is to trace which of the user interesting pages are useful, i.e. are most likely to satisfy the user's goal. To estimate that, we rely on the pages' content types derived by the analysis of their structural properties and the estimated user goals derived by the skewness of the user's click distribution on pages of specific content types in her past site accesses (cf. Section 2.2). In particular, we determine whether a page P is useful based on the following criterion:

$$\text{P} = \begin{cases} \text{useful} & \text{if content type of P} = \text{type of user goal} \\ \text{not useful} & \text{otherwise} \end{cases}$$

Based on the combination of the above features (i.e. pages' interestingness and usefulness), we generalize our model and we compute the probability that the user will prefer a particular page in a web site as follows:

$$\text{User Preference}(\text{P}) = \text{User Interest}(\text{P}) + \text{Usefulness}(\text{P}) \qquad (\mathbf{8})$$

Where Usefulness (P) denotes whether a page is useful or not and quantifies to:

$$\text{Usefulness}(P) = \begin{cases} 1 & \text{if P is useful} \\ 0 & \text{otherwise} \end{cases}$$

Under this model, we can identify which of the site pages are most likely to meet the user interests and satisfy the user intentions in their site visits. 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.

2.4 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 and goals Therefore, the recommendation system is responsible for deciding which site pages correlate to the user interests and intentions 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 *preferable-and-unvisited* and *preferable-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 [29] and we introduce a novel approach for predicting the users' preferred recommendations based on the analysis of their 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 accesses.

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:

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

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).

In a similar manner, our model predicts the type of user goals in their future site visits based on the skewness of the user's past click distributions on site pages of different content types. Upon the detection of a highly preferred type of pages, our model assumes the content of those pages to be useful to the user's intention and thus prioritizes them among the site interesting pages.

Under this approach, our model predicts the type of the user interests and goals 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 and their content type is the same to the type of the user's goal, our model recommends these pages regardless of the fact that these have already been seen by the user. Alternatively, if the value of F for all the site 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 and upon identification of user interesting pages; it predicts the type of the user interest and goal 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 and useful 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 goals and as such recommendations are dynamic in the sense that as our mechanism gets to learn the user interests and intentions, 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 and have different goals during different site visits. However, considering that the user's topic preference and visit goal might change during a single visit (i.e. session) or that the user might have more than one topic interests and visitation goals in the same visit, our model can be modified and use a sliding time window for generating recommendations.

3. 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 3.2 we describe a simulation-based experiment to estimate the accuracy of our model in offering customized site views according to the user interests and goals. Finally, in Section 3.3 we present the results from a human study that measures the perceived usefulness of our recommendation mechanism.

3.1 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 the following information for each of their sessions: (i) their preferred topic, (ii) the web site pages that interested them the most and (iii) their pursed intention in each of their site visits.

More specifically, 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. We also asked our subjects to denote their site access goals using one of the following types: informational, resource or transactional. Moreover, to denote qualitative pages we asked our subjects to rate every page in a site, using scores ranging from 1, meaning "not valuable at all" up to 5 "highly valuable[8]". 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, and we stored the cleaned data in a RDBMS server. Table 1 shows statistics on our experimental data.

We downloaded all the pages from each of the sites in our dataset, we processed them following the steps described in Sections 2.1 and 2.2 and we computed for every page a suitable topic from the hierarchy in order to model the page's semantic content and a suitable type in order to model the page's structural content. Thereafter, we processed our users' site transaction logs in order to mine their navigational patterns and derive the user interests and goals in each of their site visits.

[8] In our work, we deem a page to be valuable to some user if it satisfies both the user's interests and pursued goals.

Table 1. 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

In particular, we computed for each of our subjects and for each of their visited sites, their most preferred topic, their visitation goal, the sites' pages that were most correlated to their interests and goals 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, to mine the user intentions and to build recommendations for each of our subjects.

3.2 Accuracy of User Interests and Goals Identification

In this section, we measure the accuracy of our method in identifying the user site interests and goals as well as 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.

3.2.1 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 on 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 [33] 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 in Section 2.1). 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 rely on:

$$E(T_i) = \frac{|T_i - T|}{|T|} \qquad (9)$$

where T denotes the user's actual topic interests (i.e. baseline estimation) and Ti denotes the topic interests identified by our model. Figure 4 shows the results.

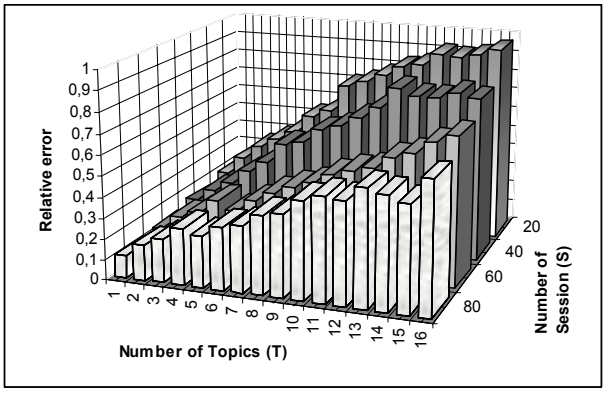

Figure 4. Relative errors in estimated topic interests.

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. $1 \leq T \leq 5$) 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. $6 \leq T \leq 10$), 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.

3.2.2 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 [14] between our estimated ordering of page recommendations and the ideal recommendations' ranking. Note that in our experiment, we deduced the ideal recommendation rankings from the data that our participants supplied (cf. Section 3.1), that is their preferred topics and most interesting pages in their site visits. Formally, the Kendall's distance metric (τ) between two ordered lists of recommendations is given by:

$$\tau(E_K, A_K) = \frac{\left|(i, j): i, j \in R, E_K(i) < E_K(j), A_K(i) > AK(j)\right|}{|R| \cdot |R - 1|} \quad (10)$$

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 [33] 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 5 shows the differences in the recommendations' ordering for k=10, i.e. when considering the top 10 recommended pages[9].

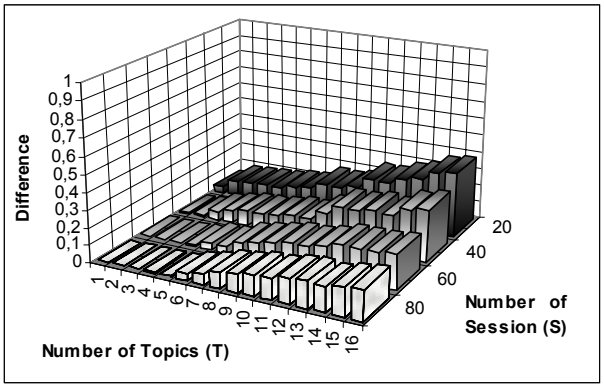

Figure 5. Ordering differences of top 10 recommendations.

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 $1 \leq T \leq 7$[10] 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.

3.2.3 Accuracy of Estimated User Goals

Having evaluated our model's accuracy in identifying the user interests in their site visits, our next step is to assess the amount of log data that our model requires for the accurate estimation of the user goals in their site accesses. In our evaluation, we relied on the synthetic user sessions, previously generated by simulation and we generate a sequence of L clicks done by the user on the pages visited in every session. Note that our experimental pages span all three content types. Once we generate synthetic user clicks, we compute the user's primary goal across sessions, following the method described in Section 2.2. To evaluate the accuracy of our model in learning the user's goal, we measure the relative error between the users' actual (i.e. baseline) and learnt goals. Note that the baseline goal detection model assumes equal distribution of clicks for every goal (i.e. equal distribution of clicks on pages of different content types). For our comparison, we rely on:

$$E(G_i) = \frac{|G_i - G|}{|G|} \qquad (11)$$

where G denotes the user's actual goals (i.e. baseline estimation) and Gi denotes the user goals estimated by our model. In our evaluation we experimented with different types of goals the user pursues in her site visits (i.e. the user has a single goal in all her site accesses, the user pursues two different goals in her site accesses and the user pursues all three goals in her site visits.

Obtained results indicate that as the number of the user pursued goals in a site increase, so does the number of clicks that we need to collect about that user's past clickthrough behavior in order to predict the user's future intentions in the site's contents. Figure 6 shows the results when all 3 types of goals are considered, i.e. when the user intents to visit pages of all three content types considered.

---

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

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

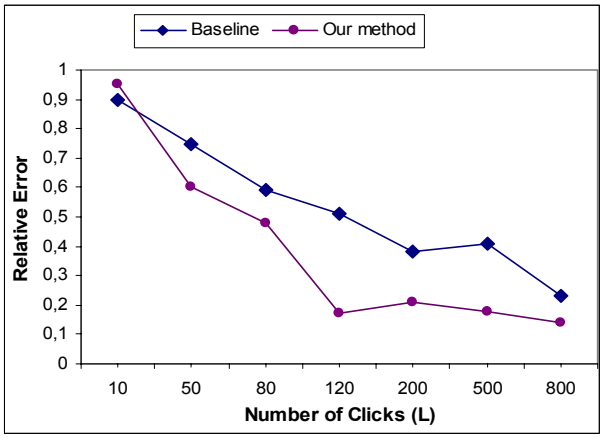

Figure 6. Comparison of relative errors in estimated user goals for 3 types of goals across the site visits.

From the graph we can see that when the user has multiple intentions (i.e. all 3 types of goals) in her site visits, our model can estimate the user goals with 83% accuracy after considering only 120 clicks from a user's past site accesses. We can also see that the relative error of our model is significantly smaller from the baseline estimation, indicating that our method is effective in learning the primary user goal in her site accesses.

Although not schematically illustrated, our method yields an even better learning accuracy (i.e. 94%) when two goals are pursued by the user using a sample size of 100 clicks. In overall, results demonstrate that we can effectively learn the user's intention in the site visits using a small sample size of clicks from the user's click history on the site pages.

3.3 Quality of Site Customizations

To measure the quality of our site customization approach, we used the data collected from our human survey (Section 3.1) and we evaluated our model's effectiveness in offering valuable 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 valuable recommendation.

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

$$\text{Quality}(P) = \sum_{P \in S} \text{User Preference}(P) \bullet F \qquad (12)$$

Here S denotes the pages in a site, F denotes the probability that the user will revisit P and User Preference (P) denotes the probability that user will prefer page P. The degree of the User Preference in P that our model estimates is given by equation 8, while the actual user preference 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 valuable these are to their interests and goals. Note that we also asked our subjects to indicate whether they had a persistent or ephemeral interest in each of the pages considered.

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 [33] 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 $\geq$F and $\leq$F values for persistent and ephemeral page interests, respectively.

Based on the above formula, we evaluate our model's accuracy in identifying valuable pages to the site users by comparing the pages' estimated quality to their actual quality, denoted by our subjects. Figure 7, reports the average quality values of the pages in the sites that our subjects have visited during the reporting period.

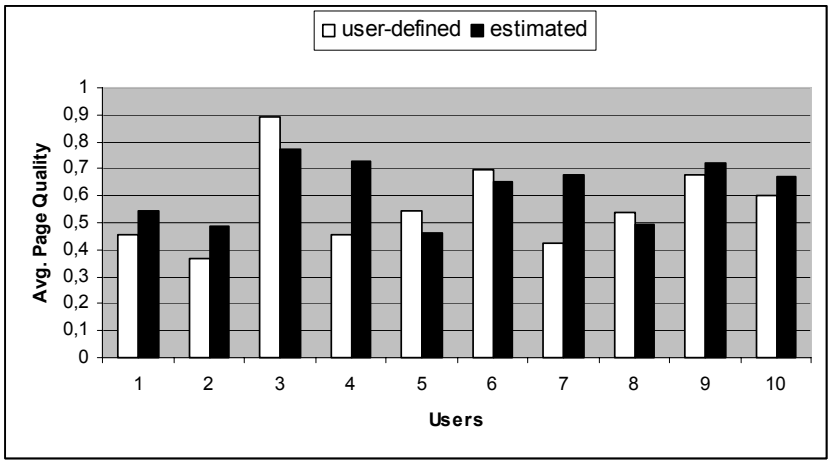

Figure 7. Comparison between the user-defined and our model-estimated average quality of the site pages that users have visited.

The bars on the horizontal axis represent the 10 subjects in our study. For every study participant there are two bars associated: the white bar shows the average quality that the user defined for the pages she has visited across her recorded site interactions and the dark bar shows the average quality that our model estimated for that user's visited pages. Scores are normalized to sum up to one and they are aggregated by users in the sense that for all the pages in the web sites that a user has visited we measured both their average actual and estimated quality. Results demonstrate that our model is quite effective in estimating valuable pages in the site's content, achieving an overall accuracy of 94% in judging the pages' contents as useful as humans would. In particular, we observe that for six of our subjects, our model valued slightly higher the visited pages quality compared to the user defined scores while for the remaining four participants the quality values that our model computed for the visited pages are relatively lower than the user indicated values, but still the estimated page's quality is close their user-perceived quality. The overall difference between the actual and the estimated page quality across all our subjects is 0.561, which translates to a low (i.e. 5.6%) disagreement between the pages a user would like to see in her recommendations and the pages that our model would recommend to that user.

Results indicate that our site customization model, which attempts to provide site visitors with valuable recommendations, has a significant potential in accurately estimating the value that every page in a site has to the user interests and goals. Based on the estimated quality 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.

4. Related Work

Many researchers have proposed ways of customizing Web sites to the needs of specific users [5] [9] [11] [12] [20]. One approach to personalization is to have users explicitly describe their general search interests, which are stored as personal profiles [21] [34]. Many commercial systems rely on personal profiles to customize the site's presentation by mapping web pages to the same categories with the ones representing the user interests. There also exist many works on the automatic learning of the user preference based on the analysis of their past clickthrough history [9] [24] [29]. In [24] for instance, a user's preference is identified based on the five most frequent topics in the user's log data. Our work is different from this approach in that we consider all possible topics that describe a user's click history. On the other hand, in [9] multiple TF-IDF vectors are generated, each representing the user's interests in one area. In [29] the authors employ collaborative filtering techniques for learning the user's preference from both the pages the user visited and those visited by users with similar interests. For an overview we refer the reader to the work of [31].

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 [11] [19]. In this respect, researchers have explored the use of ontologies in the user profiling process. We refer the reader to the work of [13] for an overview on the role of ontologies in the site customizations. There also exist studies [17] [21]

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 [18] the authors used an ontology and utilized the Wu & Palmer similarity measure [30] 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.

Our study is also related to research on identifying the user goals in their web interactions. However, the vast majority of existing works [25] [6] [15] concentrate on the web search paradigm and attempt to predict the user goals through the analysis of both the user issued queries and the user's click behavior for these queries. In our work we are inspired by these studies but we take a step further and claim that users have different goals in mind not only while searching the web but also when navigating in web sites' content. Under this approach, we extend the notion of pursued user goals to the web site visit paradigm and we suggest a novel approach for the prediction of the user intentions based on their site usage and the pages' structural properties.

Finally, our study touches upon issues regarding the identification of the web sites' functionality. Several researchers have proposed the exploitation of the sites' structural content [16] [21] in order to assist users acquire the desired information from the web. In this respect, researchers have proposed two main approaches for extracting the structure of the web, namely the microscopic analysis [7] that decomposes web content into clusters of related pages and the composition of clusters of neighboring pages based on their local connectivity as the latter is determined via the use of web APIs or web services. Although our work does not rely on an in-depth analysis of the sites' structural properties nevertheless it addresses the user goals detection task from the perspective of the users' visitation patterns on pages of specific content types based on the analysis of the pages structural content.

5. Concluding Remarks

In this article, we have proposed a novel recommendation model that aims at providing users with customized site views based on the association between the sites' usage patterns and the site pages' structural and semantic content. In particular, we first proposed a user model to formalize user's interests in web site pages and correlate them with users' visitations in the site's content. Our model leverages a subject hierarchy for identifying the pages' thematic content as well as for identifying the user preference in the site pages. Based on the correlation between the user interests and the pages' topics, we present a heuristic approach to actually learn the users' profiles.

Moreover, we introduced an intuitive model for predicting the user goals in their site visits. Our model relies on the site pages' structural properties and explores the users' interaction with the site in order to enrich the estimated user profiles with information about the user's intentions in their site visits. Finally, we introduced a novel recommendation mechanism that correlates the identified user interests and goals 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.

We have conducted experiments to evaluate the effectiveness of our model. In one the 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 experiment, using the same synthetic data, we found that for a relatively small amount of user clickthrough history (containing 120 past clicks on the site's pages), our model can accurately predict the user goals in their site visits even if user intentions span multiple types of interaction.

In a real-life experiment, we applied our method to estimate preferable pages for 10 subjects and we assessed the effectiveness of our model in estimating valuable recommendations to web site users. Obtained results demonstrated that, on average, our approach successfully estimates which pages might be preferable to the site users. In overall, the innovation of our approach lies on the following: our model makes a distinction between topical interests and pursued goals in a user's site visits and attempts to build the user's profile based on their combination. To the best of our knowledge our work is the first reported attempt that aims at encapsulating visit intentions in the web site customization process. In addition, unlike existing recommendation systems that prioritize interesting and unvisited pages in

their suggestions, our model is different in the sense that it relies on the viability of the user interests and considers the case that a user might want to re-visit an already seen page in her future site interactions. We are currently improving our model towards accounting for across-site user-specific information in order 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 and pursues alike goals.

6. 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] Broder A. A Taxonomy of Web Search. In SIGIR Forum, 36(2), 2002.

[7] Broder A., Kunar R., Maghoul F., Raghavan P., Rajagopalan S., Strata R., Tomkins A., Wiener T. Graph Structure in the Web: Experiments and Models. In Proceedings of the 9th WWW Conference, 2000.

[8] 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.

[9] Chen L., Sycara K. Webmate: a personal agent for browsing and searching. In Proceedings of the 2nd Intl. Conference on Autonomous Agents & Multiagent Systems,pp. 132-139, 2004.

[10] 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.

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

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

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

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

[15] Lee U., Liu Z., Cho J. Automatic Identification of User Goals in Web Search. In the WWW Conference, 2005.

[16] Lindermann Ch., Littig L. Coarse-grained Classification of Web Sites by their Structural Properties. In Proceedings of the WIDM Workshop, 2006.

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

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

[19] 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.

[20] 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.

[21] Murata T. Extraction of Structural Information from the Web. In L. Wang and Y. Jin (Eds.): FSKD, LNAI 3641, 2005.

[22] Pazzani M., Muramatsu J., Billsus D. Syskill & Webert: Identifying Interesting Web Sites. In Proceedings of the 13th National Conference on Artificial Intelligence, Portland, pp. 54-61, 1996

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

[24] Pretschner A., Gauch S. Ontology-based personalized search. In Proceedings of the 11th IEEE Intl. Conference on Tools with Artificial Intelligence, pp. 391-398, 1999

[25] Rose D., Levinson D. Understanding User Goals in Web Search. In the WWW Conference, pp. 13-19, 2004.

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

[27] 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.

[28] 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.

[29] 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.

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

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

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

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

[34] Yahoo! Inc. *MyYahoo* http://my.yahoo.com