Search Personalization through Query and Page Topical Analysis

Sofia Stamou
Computer Engineering and Informatics Department
Patras University, GR-26500, Greece
stamou@ceid.upatras.gr

Alexandros Ntoulas
Microsoft Research
1065 La Avenida, Mountain View, CA 94043, USA
antoulas@microsoft.com

Abstract.

Thousands of users issue keyword queries to the Web search engines to find information on a number of topics. Since the users may have diverse backgrounds and may have different expectations for a given query, some search engines try to personalize their results to better match the overall interests of an individual user. This task involves two great challenges. First the search engines need to be able to effectively identify the user interests and build a profile for every individual user. Second, once such a profile is available, the search engines need to rank the results in a way that matches the interests of a given user.

In this article, we present our work towards a personalized Web search engine and we discuss how we addressed each of these challenges. Since users are typically not willing to provide information on their personal preferences, for the first challenge, we attempt to determine such preferences by examining the click history of each user. In particular, we leverage a topical ontology for estimating a user's topic preferences based on her past searches, i.e. previously issued queries and pages visited for those queries. We then explore the semantic similarity between the user's current query and the query-matching pages, in order to identify the user's current topic preference. For the second challenge, we have developed a ranking function that uses the learned past and current topic preferences in order to rank the search results to better match the preferences of a given user. Our experimental evaluation on the Google query-stream of human subjects over a period of one month shows that user preferences can be learned accurately through the use of our topical ontology and that our ranking function which takes into account the learned user preferences yields significant improvements in the quality of the search results.

Keywords: Personalized search, Web search, User preferences, Topical ontology, Topic-specific rankings

1. Introduction

The Web search engines serve millions of queries daily from thousands of eager users searching for a multitude of topics. Every user has a unique background and goal in mind and is searching for a specific


© 2008 *Kluwer Academic Publishers. Printed in the Netherlands.*

piece of information through the use of keyword queries. According to a study (Jansen et al., 2000), a significant number of such queries are typically under-specified and contain only a small number (between 1 and 3) of keywords. Such short queries are oftentimes marginally informative of the user's search intentions and may return search results that are not relevant to the user's information needs. In addition, certain queries are polysemous in nature since the same keyword can be used to express distinct topics or ideas. For instance, a video games fan may issue the query `quake` searching for a popular computer game, a scientist may issue the same query to search for data about earthquakes and a football fan might issue the exact same query (`quake`) to find information about a football team in California.

In order to address the variance in the information needs of people issuing queries, search engines are trying to personalize (or customize) their search results by returning results that are closely matching the interests of an individual user. Personalized search has a significant potential in providing the user with information that accurately satisfies her particular search intentions, but it entails two great challenges. The first challenge, involves identifying the interests of an individual user. This task is particularly challenging mainly due to the reluctance of the users to give explicit feedback about their search preferences. The second challenge is that, given that we have identified a given user's preferences, we need a way to retrieve search results that closely match those preferences. In this article, we present our work towards a personalized Web search engine and we discuss how we addressed each of the above challenges.

For the first challenge, instead of requiring explicit feedback from the user, we are implicitly learning her interests through the analysis of her past clickthrough data. Based on this analysis, we create a user search profile which can be later used in the personalized ranking of the search results. Typically, although implicit feedback does not require the user's direct involvement in the personalization process, one problem arising in a practical setting is that the user interests may change over time and therefore the learned profiles need to be updated in order to reflect the user's current interests. In this article, we study the dynamic identification of user search preferences based on both their past and current searches. We present a model that automatically captures the past preferences of a user, based on her click history. Our model can be updated to also account for the user's current interests based on the association between the user's query and her past topic preferences. Our main idea is to employ a topical ontology that we constructed from high-quality sources (such as WordNet (Fellbaum, 1998) and SUMO (Pease et al., 2002)) in order to compute a topic

to assign to each one of a user's previously visited pages as they are identified from the user's click history. Based on the topics of the pages visited for a given query, we estimate the similarity of the query to each of the topics from the ontology. Additionally, we examine the semantic association between a given query and the pages visited for that query in order to learn the topic that best describes a user's preference.

In order to address the second challenge, we have developed and implemented a ranking function that orders the search results based on the learned interests of a given user. Our main idea is based on determining how well every page within the search results matches the topics present within a given user's profile.

The remainder of our article is organized as follows. In the next section, we start by discussing some background for our work. In Section 3 we present our method for automatically identifying the topical preferences of a given user, and we discuss how we use the identified topics for personalizing the search results. In Section 4 we demonstrate the potential of our method by presenting the results of a study involving human subjects and the queries they issued to a real search engine during October of 2006. We review related work in Section 5 and we conclude in Section 6.

2. Background

As we discussed in the previous section, our approach in creating a personalized Web search engine is to use a topical ontology in order to determine the topics of the pages that each individual user has visited. Based on these identified topics, we can then build the user's preference model and proceed to rank the results according to this model.

In this section, we present the topical ontology that we have developed and that can be used towards this goal. In Section 3, we discuss an algorithm called DirectoryRank that determines how relevant a given Web page is to each of the topics within our ontology and we present our profile learning model.

2.1. The Topical Ontology

For our purpose of using an ontology to identify the general topics that might be of interest to the Web users, we choose to develop an ontology that would describe human perception of the most popular topics that are communicated in the Web data. Thus, we define our ontology as a hierarchy of topics that are currently used for categorizing Web pages. To ensure that our ontology would define concepts that are

representative of the Web's topical content, we borrowed the ontology's topical categories from the Dmoz Directory[1] and we further enriched them with conceptual hierarchies that we leveraged from existing ontological resources that have proved to be richly encoded and useful. The resources from which we leveraged the ontology's hierarchies are:

1. **WordNet.** WordNet (Fellbaum, 1998) is a large lexical network of almost 160,000 terms organized in nearly 118,000 cognitive synonym sets (synsets), each representing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. The resulting network of meaningfully related words and concepts can be effectively navigated and is freely available for download.[2]

2. **The Suggested Upper Merged Ontology (SUMO).** SUMO (Pease et al., 2002) is an upper-level formal ontology with 20,000 terms and 60,000 axioms written in the SUO-KIF language and it is publicly available[3] through the IEEE Web site. SUMO is a richly axiomatized ontology that combines several domain taxonomies and provides complete mappings to all WordNet synsets.

3. **The MultiWordNet Domains (MWND).** MWND (Bentivogli et al., 2004) is an augmented version of WordNet that assigns domain labels to the synsets of WordNet. Within MWND[4] each WordNet synset is assigned at least one domain label from the total set of 165 hierarchically-structured domains available.

Given that our goal was to augment Dmoz with WordNet as comprehensively as possible, we used SUMO and MWND as complementary lexical resources. These two resources were selected because they are considered to contain rich and qualitative content, they are freely available and they are directly mapped to WordNet. By employing these two additional resources, our goal was to ensure that we would be able to also link Dmoz categories that are absent from WordNet domains (e.g. Kids and Teens, Regional) with their corresponding WordNet concepts. For a thorough description of the methodology we adopted for building the ontology, we refer the interested reader to the work of (Stamou et al., 2007).

At a high level, the construction of our ontology involved a combination of both top-down and bottom-up approaches and aimed at

[1] http://www.dmoz.org
[2] http://www.cogsci.princeton.edu/~wn/
[3] http://ontology.teknowledge.com
[4] http://wndomains.itc.it

anchoring the WordNet lexical hierarchies to their corresponding Dmoz topical categories.

In order to enrich the Dmoz topics with their corresponding lexical hierarchies, we relied on the domain concepts in SUMO and MWND that are used to label the lexical elements in WordNet hierarchies and we selected those domain concepts that are hyponyms of the Dmoz topics as our ontology's middle level concepts (Mid-Level Ontology).

The first decision we had to make in the process of merging the Dmoz topics and the WordNet hierarchies into a common topical ontology was to settle on the relations to be used to map WordNet synsets to the Dmoz topic categories. To decide which of the three popular formal ontology relations (i.e. specialization, instantiation and membership) were of interest in each step, we relied on the SUMO and MWND structure from which we derived dependencies between concepts. More specifically, we started by identifying which of the SUMO or MWND domain concepts are hyponyms in WordNet of the Dmoz top level topics. Once we identified such concepts, we incorporated them in the ontology as sub-topics (i.e. middle level concepts) of the ontology's top level categories and we also imported the lexical elements from the WordNet hierarchies that were labeled with the identified middle level concepts (top-down merging).

For example, consider the Dmoz top level topic `Sports` and the SUMO domain concept `Swimming`. By looking up these two concepts in the WordNet ontology we can identify that they are semantically related. Therefore, we incorporate the SUMO domain concept `Swimming` as a middle level topic in our ontology (linked to the `Sports` top level category) and we append to our ontology (as bottom level concepts) the WordNet hierarchies that are labeled with the respective SUMO domain concept. We should note here that since `Swimming` is already a sub-topic of `Sports` in Dmoz (i.e. Sports→Water Sports→Swimming) and therefore we could have directly appended all WordNet lexical hierarchies labelled with `Swimming` under the respective Dmoz sub-topic. In this case however, since the topic `Water Sports` is not used to label any WordNet synset part of our ontology would not be represented in WordNet's lexical hierarchies. To avoid this, we use SUMO and MWND during the construction process to ensure that we have adequate coverage for each node in the ontology instead of merging simply by string matching between WordNet and Dmoz categories.

In the cases were the SUMO or MWND domain concepts were not directly linked (i.e. neither was hyponym of the other) to any of the Dmoz top level topics, we relied on the hypernyms of the SUMO or MWND and we examined whether their generalized concepts relate to any of the Dmoz topics. The SUMO or MWND hypernyms that

are semantically linked to the Dmoz topics were incorporated in the ontology as middle level concepts and we also imported the WordNet hierarchies whose elements are labeled with the respective domain concepts (bottom-up merging).

For instance, consider the Dmoz top level category `Health` and the SUMO domain category `Disease` or `Syndrome`. Looking up the two concepts in WordNet hierarchies returns no direct link between them. However, going from the SUMO matching concept (`Disease`) one level up in the WordNet ontology shows that the direct hypernym of `Disease` (i.e. `Illness`) is an inherited hypernym of `Health` (the Dmoz topic). Therefore, the hypernym of the SUMO domain concept `Disease`, namely `Illness`, is incorporated in the ontology as a middle level concept linked to its respective top level category (i.e. `Health`). Similarly, the WordNet lexical hierarchies labeled with the SUMO domain concept `Disease` are incorporated in our ontology under the respective middle level category i.e. `Illness`.

Similarly, we tackled membership links by adopting a middle-out merging approach. For instance, the SUMO concept "Human" has no specialization (i.e. hyponymy) link to any of the Dmoz top level topics. In order to assess whether the given SUMO domain concept could be incorporated in our ontology, we work as follows. We examine the definition of the term "Human" in WordNet where it is described as "a living or extinct member of ...". We also examine the definitions of Dmoz domain concepts within WordNet and we determine that the Dmoz topic "Society" is defined in WordNet as "an extended social group...". From these definitions, we can infer that "human is a member of ..." and that "society is a social group (of people)". Therefore, we decided to link the SUMO concept "Human" to the Dmoz topic "Society" via a member-of relation and then append the WordNet synsets labeled with the domain "Human" to our ontology. In all merging steps, we used the SUMO and MWND domain concepts to connect the WordNet synsets and the Dmoz topics, when a direct relation between the latter could not be traced within our resources. These SUMO and MWND domain concepts formulated the middle level topics in our ontology, i.e. the sub-categories of the respective Dmoz topics.

At the end of this process, we came down to a total set of 489 middle level concepts, which were organized to the 16 Dmoz top level topics. The resulting upper level ontology (i.e. top and middle level concepts, 156 of which are Dmoz topics) is a directed acyclic graph with maximum depth 6 and maximum branching factor, 28 (i.e. number of children concepts from a node). Figure 1 shows a portion of our ontology for the Dmoz topic Arts.

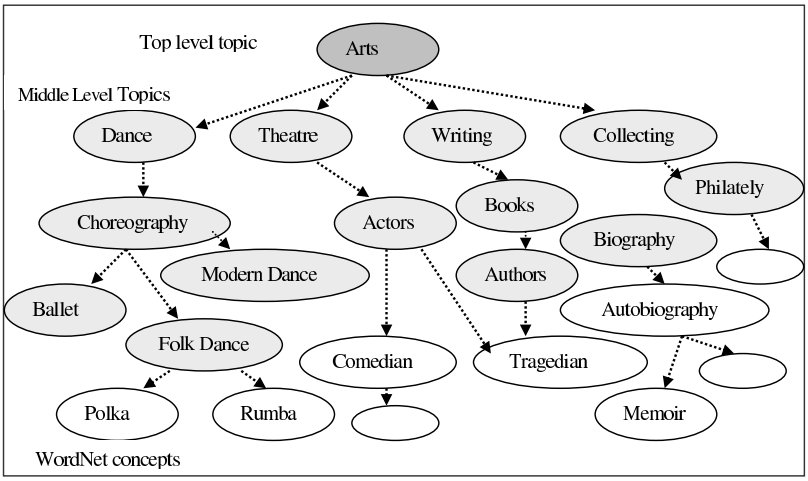

Figure 1. A portion of the ontology for the Dmoz topic Arts.

In the Figure, the dark-gray node on top illustrates the top-level ontology concept (Arts) that is borrowed from the Dmoz top level topics. The light-gray nodes represent the ontology's middle level concepts, which are borrowed either from a SUMO or a MWND categories and which are linked to their respective top level concepts through one of the following relation types: specialization, instantiation or membership. Finally, the lower level ontology concepts (depicted as white nodes) represent the lexical elements in WordNet ontology that are specialized concepts of the respective middle level concepts.

Having developed the ontology, one question is whether our ontology is sufficient for the task of identifying the Web pages' topics. Since the upper level topics covered in our ontology have been borrowed from the Dmoz Directory topics, we believe that the ontology covers the domain of Web topics reasonably well. Moreover, by relying on WordNet data to organize the concepts in our ontology, we deem that the ontology's hierarchies are meaningfully structured. Consequently, we believe that we have created a useful resource that can help us automate the task of identifying the Web pages' topical categories.

Although it is unlikely that a single ontology can meet the needs of all Web users and applications, we believe that providing the means to explore the ontology for the automatic categorization of Web pages into topic directories is less subjective than relying solely on individual categorization judgments.

Another issue we need to address at this point is the soundness of the approach that we took for merging SUMO and MWND into a common resource. Since both the above resources share a significant

amount of overlapping content, we did not encounter any significant difficulties while integrating SUMO and MWND hierarchies with the ontology's topics. In addition, WordNet data was manually checked throughout the entire merging process, in order to ensure the validity of the produced mergings.

Finally, although it might be argued that building the ontology is a burdensome and time-consuming task (it took us nearly 3 months to build and validate the ontology) that could be avoided through the use of a classification method, we believe that the ontology once it is built allows more flexibility as it abandons the need for (re-)training a classifier. This is also attested in a previous work we have carried out (Stamou et al., 2006) where we managed to automatically classify about 300,000 Web pages in the ontology's topics in nearly 6 hours (including data cleaning and pre-processing time) without any need for prior training or human involvement.

3. Ontology-based Personalized Search

Given the topical ontology that we constructed, we need a way of determining what are the potential topics of a Web page. These topics will be used later in conjunction with the topics of a user's profile to present her with personalized search results. We begin our discussion by presenting our approach on how we exploit a topical ontology along with a user's past queries and visited pages to identify the topics that this user is interested in.

Based on the topic importance values of the visited pages, we estimate the user's degree of interest in each of the identified topics. We then describe how we can succinctly learn the latent topic preferences of a user's current query for which there is no previous click history available. Roughly, this learning can be done based on the relationship between the user's past topic preferences and the semantic association between the current query and the query-matching pages. Finally, we present how we can exploit the user's learned preferences (past and current) during query-time to rank search results according to the preferences of this particular user.

3.1. Topical Analysis of Click History

The billions of pages available on the Web today may span a large number of topical categories. For example, the Dmoz Directory contains more than 590,000 topical categories which are used to organize pages from the Web. In a practical setting, a user is unlikely to be equally

interested in all of the existing topics on the Web. Rather, each user has a small number of topic preferences which she is interested in.

Our intuition here is that the topic preferences of each user are implicitly communicated through the queries she issues to search engines and the pages she views based on those queries.

One way to identify the topics of the pages that a given user is interested in is to use the pages' classification that is provided by a Web Directory (such as Dmoz). However, considering that Dmoz (one of the largest Directories) classifies only 0.03%[5] of the pages that are known to the search engines, we need a way to determine the topics that the user-visited pages belong to, regardless of whether they appear in Dmoz or not. Therefore, we need to employ a classification scheme that categorizes Web pages into a set of predefined topics. In this paper, we use a classification method which uses our topical ontology for estimating a suitable category to assign to every page. The main advantages that our approach exhibits over the existing classification schemes is that it manages to automatically detect the suitable topic(s) of a page without any prior classification knowledge. Moreover, in a recent study (Stamou et al., 2006) we found that our classification model managed to classify a large set of pages in their respective 156 topics with an overall accuracy of 70% compared to the accuracy of a Bayesian classifier that delivered for the same set of pages and topics an accuracy of about 66%.

To that end, in order to find the topics of the visited pages, we adopt the lexical chaining technique (Barzilay and Elhadad, 1997). A lexical chain is essentially a sequence of semantically related consecutive terms in a text and it is generated in a three-step approach: (i) select a set of candidate thematic terms from the page, (ii) for each candidate term, find an appropriate chain relying on a relatedness criterion among members of the chains, and (iii) if such a chain is found, insert the term in the chain. Thematic terms are terms which convey most of the topical information in texts (Gliozzo et al., 2004) and are typically the nouns and proper nouns within the content of the page.

The relatedness factor in the second step is determined by the type of the links that are used in WordNet for connecting the candidate term to the terms that are already stored in existing lexical chains. In case a candidate term is not represented in the WordNet concepts it is not used in the lexical chain construction process.

Barzilay and Elhadad introduce a greedy disambiguation algorithm that constructs all possible interpretations of the source text using

[5] Dmoz contains approximately 4 million Web pages and the size of the indexable Web is estimated to be at least 11.5 billion pages (Gulli and Signorini, 2005).

lexical chains. However, (Song et al., 2004) noted some caveats in this disambiguation formula in avoiding errors because it does not discriminate between relation types that connect words to each other in the WordNet hierarchy.

To surpass this limitation, they propose a new disambiguation formula, which relies on a scoring function f, which provides a score indicating how likely it is that a given relation between two words is the correct one. Given two words, w_1 and w_2, their scoring function f via a relation r, depends on an association score between words, their depth in WordNet and their respective relation weight.

The association score ($Assoc$) of the word pair (w_1, w_2) is determined by the word co-occurrence frequency in a corpus and it is given by:

$$Assoc(w_1, w_2) = \frac{log(p(w_1, w_2) + 1)}{N_s(w_1) \cdot N_s(w_2)} \tag{1}$$

where $p(w_1, w_2)$ is the probability that the word pair (w_1, w_2) co-occurs in the corpus within a neighborhood size of 20 words (Turney, 2004) and $N_s(w)$ is a normalization factor, which indicates the number of WordNet senses that a word w has. In our work, we relied on the Web TREC corpus for computing the co-occurrence probability of the thematic words.

Given a word pair (w_1, w_2) their $DepthScore$ expresses the words' position in WordNet hierarchy and captures the fact that the lower a word is in WordNet hierarchy, the more specific its meaning. $DepthScore$ is defined as:

$$DepthScore(w_1, w_2) = Depth(w_1) \cdot Depth(w_2) \tag{2}$$

where $Depth(w)$ is the depth (i.e. number of levels from the root) of word w in WordNet.

Additionally, given a word pair (w_1, w_2) there is at most one semantic relation type $r(w_1, w_2)$ that connects w_1 and w_2. For a given relation type, semantic relation weights ($RelWeight$) have been experimentally fixed according to (Song et al., 2004) to 1 for reiteration, 0.2 for synonymy and hyper/hyponymy, 0.3 for antonymy, 0.4 for mero/holonymy and 0.005 for siblings. Finally, the scoring function f of w_1 and w_2 is defined as:

$$f(w_1, w_2) = Assoc(w_1, w_2) \cdot DepthScore(w_1, w_2) \cdot RelWeight(r(w_1, w_2)) \tag{3}$$

The score of lexical chain c_j is calculated as the sum of the score of each consecutive pair of words w_i and w_{i+1}, Formally:

$$Score(c_j) = \sum_{i=1}^{|c_j|-1} f(w_i, w_{i+1}) \quad (4)$$

To compute a single lexical chain for every downloaded Web page, we segment the latter into shingles, using the shingling technique, described in (Broder et al., 1997). To form a shingle, we group n adjacent words of a page, with $n = 50$, which roughly corresponds to the number of words in a typical paragraph. For every shingle, we generate and score lexical chains using the formula described above. In case a shingle produces multiple lexical chains, the chain of the highest score is regarded as the most representative shingle's chain, thus eliminating chain ambiguities. We then compare the overlap between the elements of the lexical chains of all adjacent shingles. Elements that are shared across chains are deleted so that lexical chains display no redundancy. The remaining elements are merged together into a single chain C_p, representing the contents of the *entire page*, and a new normalized $Score(C_p)$ for the resulting chain C_p is computed. This way we ensure that the overall score of every page's lexical chain is maximal. The elements of each chain are used as keywords for assigning the underlying pages in topical categories. We now introduce a model that automatically identifies the topics of Web pages.

3.1.1. *Identifying the Topics of Visited Pages*
In order to detect the topics of the pages that a user has previously visited, we follow the TODE categorization scheme, presented in (Stamou et al., 2007). The main idea is to map the thematic keywords of a page to the concepts of the ontology by traversing the ontology's matching nodes up to the root nodes. Recall that thematic words are disambiguated during the generation of the lexical chains, thus ensuring that every keyword is mapped to a single node in the ontology.

Traversal of the ontology hierarchies amounts to following the hypernym links of every matching concept until all their corresponding topics (first or second level) are retrieved. For short documents with very narrow subjects there might be a single matching topic. However, due to the richness of our ontology, it is often the case that some thematic words of a given page correspond to multiple topics.

To accommodate for the assignment of multiple topics, we compute a Relatedness Score ($RScore$) of every Web page to each of the ontology's matching topics. This relatedness score indicates the expressiveness of each of the ontology's topics in describing the Web pages' contents. Formally, the relatedness score of a page p (represented by the lexical chain C_p) to topic T of the ontology is defined as the product of the

page's chain $Score(C_p)$ and the fraction of words in the page's chain that are descendants (i.e. specializations) of T, which are noted as $d(T)$.

$$RScore(p,T) = Score(C_p) \cdot \frac{|C_p \cap d(T)|}{|C_p|} \quad (5)$$

The denominator is used to remove any effect the length of a lexical chain might have on $RScore$. $RScore$ values range between 0 and 1, with 0 corresponding to no relatedness at all and 1 indicating the category that is highly expressive of the page's topic. Finally, a Web page is assigned to the topical category T_k for which it has the highest relatedness score of all its $RScores$ above a threshold τ, with τ been experimentally fixed to $\tau = 0.5$. The page's (as expressed by chain C_p) indexing score is:

$$IScore(p) = max_{\forall T_k \in ontology} RScore(p, T_k) \quad (6)$$

Pages, with chain elements matching several topics in the ontology, and with relatedness scores to any of the matching topics below τ, are categorized in all their matching topics. By allowing pages to be categorized in multiple topics we aim at capturing the variety of topics for a given page (and subsequently for a given user) in scenarios where we don't have enough support for one single topic.

Our overall goal behind the use of threshold τ is to discriminate between pages that have a strong affinity to one single topic and pages that do not have such a strong affinity. Our empirical observations from examining the topics of the pages calculated using the $RScore$ metric above suggested that pages with a few (one or two) very predominant topics (i.e. $\tau > 0.5$) tend to focus on one single topic which is usually captured well by the topic with the highest $RScore$. In such cases the remaining topics (below τ) had small values and therefore we consider the topic with the highest value as conveying the most signal about the topic of a page. On the other hand, pages having all of their $RScores$ less than the threshold were usually either pages that discussed a variety of topics or did not contain enough content for our method to determine a predominant topic. In these cases, we maintain all topics in order to ensure that we capture the potential diverse interest of the pages (and later on the user). Overall, the use of threshold τ is driven by practical and efficiency reasons (i.e. by pruning the topics when $\tau > 0.5$ we don't have to consider as many topics when modeling the user interests.)

In the remainder of the paper we will use $\mathcal{T}(p)$ to represent the set of topics that a given page p has been assigned to and $\mathcal{P}(T)$ to represent the pages assigned under topic T.

3.1.2. *Topic Importance of the Visited Pages*
Within each topic of the ontology the pages are sorted on the basis of a *DirectoryRank* (DR) metric, which captures the importance of a page for a given directory topic. In the DR scheme that was presented in (Krikos et al., 2005) the importance of a page with respect to a given topic is perceived as the amount of information that the page communicates about the topic.

Intuitively, an important page in a category, is a page that has a high relatedness score ($RScore$) to the category's topic and that is semantically close (similar) to other pages, which in turn they are highly related to the given topic.

The semantic similarity between two pages p_1 and p_2 is determined by the degree of overlap between their thematic terms C_{p_1} and C_{p_2} respectively, i.e. the common thematic terms in p_1 and p_2. It is formally calculated as the Dice coefficient of the thematic terms:

$$Sim(p_1, p_2) = \frac{2 \cdot |C_{p_1} \cap C_{p_2}|}{|C_{p_1}| + |C_{p_2}|} \tag{7}$$

The DR metric defines the importance of a page in a topic to be the sum of its topic relatedness score and its overall relatedness to the fraction of pages with which it relates in the given topic. This way, if a page is highly related to topic T (i.e. has high $RScore$ for T) and also relates highly with many important pages in T (i.e. has high Sim values with other pages assigned to T), its DR score will be high.

Formally, consider that we want to calculate the DR score for page p for the hierarchy topic T. Assume that p has $RScore(p, T)$ and let $p_1, p_2, \ldots, p_n$ be the other pages in T with which p semantically relates with scores of $Sim(p, p_1), Sim(p, p_2), \ldots, Sim(p, p_n)$, respectively. Then the DR of p is given by:

$$DR(p, T) = RScore(p, T) + \frac{1}{n} \sum_{i=1}^{n} Sim(p, p_i) \tag{8}$$

where n corresponds to the total number of pages in topic T with which p semantically relates, i.e. has $Sim > 0$. In its simplest implementation, the $RScore$ and Sim scores participate equally in the DR calculation based on the intuition that a Directory user will not only prefer to see topically important pages but she is equally interested in viewing pages that are highly associated to other topically important pages. In (Krikos et al., 2005), the authors computed DR values for a large set of pages listed in 156 topics in the Open Directory and showed that DR orders pages substantially different from the standard PageRank measure, and that it also has a notable potential in improv-

ing the user browsing experience when looking for information about particular topics.

In our current work we are applying DR to the Web search paradigm and especially towards personalized search. Therefore, instead of measuring the importance of a page to some directory topic, we measure the importance of the page to the topic that is preferred by the user. That is, we weight DR values according to the likelihood that the topic considered is of interest to the user. We should note that studying the individual metrics employed to compute DirectoryRank, although interesting, are beyond the scope of our current study. The interested reader may refer to (Krikos et al., 2005) for more details.

3.2. Past Topic Preference Identification

In this section, we describe how our model identifies the topic preferences of a user based on her past click history. At a high level, our model relies on the topical categories of the pages visited for a query q and the topic importance (as expressed by the DR metric) of the visited pages, in order to estimate how related q is to each of the topics in our ontology.

More specifically, let us assume that a given user has issued query q and has visited the set of pages $V(q)$ for this query. Then, for each of the pages within $V(q)$, we identify one or more ontology topics, based on the approach presented in Section 3.1.1 and we calculate its topic importance according to the DR metric for the topics in our ontology. We repeat the process for all the k queries that the user has issued so far and thus we generate a table with tuples of the form:

$$\langle q_i, p_j, T_m, DR(p_j, T_m)\rangle \tag{9}$$

This table essentially records for every query q_i (total of k queries) that a user issued, the pages p_j that she clicked on (total of $|V(q_i)|$ pages for q_i) along with the topics and their DR value (total of $|\mathcal{T}(p_j)|$ for a given q_i, p_j pair) that a page was assigned to.

Based on the data stored in the table above, we can now determine for every query how interested the user was for each of the topics that has been encountered in her past click history. Formally, we consider that the metric expressing that q relates to a topic T is captured by the average DR values of the pages visited for q that are classified under T, and is given by:

$$\mathcal{I}_{QT}(q,T) = \frac{1}{|V'(q,T)|} \sum_{p_j \in V'(q,T)} DR(p_j, T) \tag{10}$$

where, $V'(q,T)$ is the set of pages visited for q that are assigned to topic T. Based on the above metric, we can now compute the degree of relatedness of each query to each of the candidate topics.

One issue that we should note here is that for simplicity we have so far assumed that there are no duplicate page visits. That is, if a user issues a query and clicks on a page, the user only clicks on the page once. To regulate the effect of multiple page visits, we may use the logarithm of the number of page visits N to a given page: $N' = log(1+N)$ and compute the importance (DR) of a page weighted by its visit frequency. Moreover, in case a query received no clicks from the user we omit it from the user profiling process. Finally, we utilize the weighted DR values of the pages visited for a query q in order to estimate the value $\mathcal{I}_{QT}$ above.

We now proceed to describe how we utilize the $\mathcal{I}_{QT}$ values of all the queries previously submitted by a user, in order to estimate the user's degree of interest in each of the topics that relate to her past search intentions. The user's interest on topic T is given by:

$$\mathcal{I}_P(T) = \frac{1}{k}\sum_{j=1}^{k}\mathcal{I}_{QT}(q_j,T) \tag{11}$$

Where q_j are the past queries for a given user that relate to topic T and k is the total number of past queries considered for that user. $\mathcal{I}_P(T)$ gives the user's degree of preference for some topic based on the relatedness between her past queries and the topic considered, so that the greater the relatedness between the queries and a topic, the stronger the user's interest in that topic.

The $\mathcal{I}_P$ values for each topic can be readily exploited for offering personalized rankings based on a user's past click history alone. However, our personalization model, being rather generic, attempts to also identify the topic preference of a user's current query, based on the intuition that a user's search interests might change over time. In the following section, we describe how we can utilize a user's past topic preferences for identifying the topic described in a user's current query, for which there is no click data available.

3.3. Current Topic Preference Identification

So far, we have described how we can automatically identify a set of candidate topics for describing the queries issued by a user based on the topical categories of the user's past click history. We have also presented how we can estimate the amount of relatedness ($\mathcal{I}_{QT}$) of a query to a candidate topic based on the DR values of the visited pages that correspond to the identified topic.

We now turn our attention to how we can accurately identify the topic preference of a user's current query, based on both the learned preferences of the user and the query itself. Based on the findings of (Teevan et al., 2007) that web users tend to submit the same queries multiple times across different web searches and that 40% of repeated searches are re-finding queries, the first step we take towards search personalization is to examine whether the user's current query has been previously submitted in the user's past searches. If so, we employ the topical category associated with that query as the latent interest of the user issuing the query without any further need for computing a new user profile.[6] On the other hand, in case the user's current query is a new one (i.e. it has not been recorded in that user's search history), our method proceeds as follows.

In deciding the most likely topic of a new query q among all candidate ontology topics, we begin by estimating the semantic similarity between the terms in q and the thematic terms in the pages listed under each of the topics in the ontology. Note that in estimating the topic preference underlying a new user query, we rely on a set of pages already classified and ordered in their respective ontology topics. These pages have been collected in the course of an earlier study (Krikos et al., 2005). To measure the semantic similarity between the terms in a query and the terms in the pages listed under each of the topics, we use the similarity measure presented in (Resnik, 1995), which is based on the hypothesis that the more information two concepts share in common, the more similar they are. The information shared by two concepts is indicated by the information content of their most specific common subsumer, i.e. their first common hypernym in the ontology. Formally, the semantic similarity between two words, w_1 and w_2 is given by:

$$ssim(w_1, w_2) = -logP(mscs(w_1, w_2)) \tag{12}$$

Where $P(mscs(w_1, w_2))$ is the probability of encountering one of the concepts within the common subsumers of w_1 and w_2 and the $ssim$ are normalized to take values between 0 and 1. The measure of the most specific common subsumer ($mscs$) depends on: (i) the length of the shortest path from the root to the most specific common subsumer of w_1 and w_2 and (ii) the density of concepts on this path. Based on the semantic similarity values between the query terms and the thematic

[6] Of course a repeated search might not always aim at the re-finding of information. In this case, it would be useful to the user to provide her with a diversified set of results, where, besides results based on her existing profile, portion of the results would belong to categories different from her existing profile in order to encourage "exploration" of other categories. Such a method is out of the scope of this work.

terms in a page, we compute the average query-page similarity (Sim_{QP}) for all pairs of query and page terms as:

$$Sim_{QP}(q,p) = \frac{1}{|t(q)||t(p)|} \sum_{\forall t_q \in t(q), \forall t_p \in t(p)} ssim(t_q, t_p) \qquad (13)$$

where $t(q)$ denotes the terms in query q and $t(p)$ denotes the terms in page p that have some degree of similarity to the query terms. Finally, we take the similarity values between the terms in a query $t(q)$ and the terms in all the pages listed under each of the ontology's topics and we compute the similarity of the query to any of these topics. We formally define a metric that indicates how interested the user is for topic T given the current query q for which no visitation data is available as:

$$\mathcal{I}_C(q,T) = \frac{1}{|\mathcal{P}(T)|} \sum_{p_j \in \mathcal{P}(T)} Sim_{QP}(q, p_j) \qquad (14)$$

Recall that $\mathcal{P}(T)$ is the set of pages under topic T.

In the previous section, we described how our model operates for learning a user's past topic preferences, based on the analysis of her past click history. In this section, we have so far presented how our model can infer the topic preference of a user based on the relationship between the user's query and the page contents.

We should note here that our method for identifying the user interests based on new queries is quite different from the method that our model uses for identifying the user interests based on the user's past queries. This is because for new queries (for which there is no click data available) we attempt to decipher the user interests before the user actually clicks on some search results. In other words, we attempt the automatic query-topic detection without relying on any implicit information communicated by the user via her clicking behavior. Therefore, we try to identify the user interests for new queries by relying primarily on the semantic similarity between the query terms and the terms in the pages listed under each of the ontology topics.

We now turn to discussing how we can put together the knowledge accumulated so far regarding the user's interests in both the past and current queries in order to determine her overall topic preference.

Given that we can learn the topics in which the user was interested in her previous searches and estimate a degree of interest for them ($\mathcal{I}_P$) and given also that we can estimate the topic that is hidden behind a new query issued by the user ($\mathcal{I}_C$), we can measure how interested the user is *overall* for the topic T given the current query q as follows:

$$\mathcal{I}(q,T) = \alpha \cdot \mathcal{I}_P(T) + (1-\alpha) \cdot \mathcal{I}_C(q,T) \qquad (15)$$

Note that Equation 15 gives weight α to the user's past estimated topical preference for the given topic T and $1-\alpha$ to the estimated topical preference of the current query. In this way, we can fine-tune our model to be more "conservative" (larger values of α) and to consider past topical preferences of higher importance, or we can tune it to be more "aggressive" (small values of α) and mainly focus on using the estimation of the topics for the current query.

One additional issue that we need to address with Equation 15 is that it does not account for the potential semantic similarity that exists between the topic identified for a current query, i.e. a query for which there is not click history available, and the topics describing the user's past queries. In order to include this missing piece of information, we use the ontology in order to compute the semantic similarity between the current query topic and the past query topics. In general, the similarity between two topics T_1 and T_2 is determined by the maximum number of topics that subsume both T_1 and T_2 in the ontology, and is given by:

$$s(T_1, T_2) = 2 \cdot \frac{|common\ subsumers\ of\ T_1\ and\ T_2|}{|subsumers\ of\ T_1| + |subsumers\ of\ T_2|} \tag{16}$$

Having accounted for the potential inter-topic association, we can now generalize our model for learning the topic preferences of a user's current query q, as follows:

$$\mathcal{I}(q,T) = \alpha \cdot \mathcal{I}_P(T) + \beta \cdot \mathcal{I}_C(q,T) + (1-\alpha-\beta)\frac{1}{|\mathbb{T}_P|}\sum_{T_i \in \mathbb{T}_P} s(T,T_i) \tag{17}$$

where $\mathbb{T}_P$ are all the topics that have been identified as important for a given user's past queries. Note that if the past and new topic preferences are the same (i.e. the topic identified for describing a current query is among the topics that describe the user's past queries) their relatedness in the above formula will be one. Again, $\mathcal{I}$ is generic enough to allow for different weights to the past, current and related topic preferences. In general, the values of α and β need not be fixed but may vary depending on the application, the user, her search behavior and/or the nature of the query. In addition they can be adjusted dynamically for a given user over time using some sort of relevance feedback from the application. Determining a good model for α and β is out of the scope of this article and is an interesting future research avenue that we plan to investigate. In our current implementation and the experiments we present in Section 4 we have set $\alpha = \beta = 1/3$.

Based on the assumption that it might not be possible to learn the user's exact topic preferences relying on her visit history alone, our goal was to design our model in a way that its effectiveness in learning a user's topic preference is not totally dependent on the user's click history. We believe that we have achieved that by allowing new queries (for which there is no click data) to participate in the learning process. By doing so, we ensure that our model is generic enough to accommodate volatile topic preferences, i.e. the case that a user's interests change over time.

3.4. Ranking Search Results Using Topic Preferences

We now describe how we rank search results based on the user's topic preferences we have learned. Our approach is based on the framework proposed in (Krikos et al., 2005).

Given a query q, we examine the top-n (typically $n = 200$) results that would normally be returned by a search engine. Then, we assign a weight to every page in the result set according to how interested a user is to the m topics represented by her profile and how relevant each page in the result set is to the given topics. More specifically, for every page p within the result set we calculate the value:

$$R_P(p) = \frac{1}{m} \sum_{i=1}^{m} \mathcal{I}(q, T_i) \cdot DR(p, T_i) \tag{18}$$

That is, we average the DR values of the page weighted by the metric capturing the user's interest for each of the topics. In the above equation, $\mathcal{I}(q, T)$ represents how interested the user who issued q is in topic T based on her past topic preferences, the query itself and the relation between the past and the currently preferred topics. Additionally, $DR(p, T_i)$ indicates the global importance of page p in topic T_i and $R_P(p)$ is the personalized topic importance of p. We should note here that, when building a real system for personalized search, several parts of the whole R_P calculation (e.g. the DR computation, extraction of thematic words) can be performed offline thus speeding up the computation of R_P.

In the following section, we experimentally evaluate the effectiveness of $\mathcal{I}$, our personalization factor, in improving the quality of search results.

4. Experimental Evaluation

The performance of the personalization algorithm that we presented in the previous section depends on two things: first, our ability to determine the topics of the given queries successfully and, second, our ability to reorder the pages in a manner that is preferable to the users. In this section we discuss the experiments we have conducted to evaluate our proposed method for personalizing Web search and we discuss the obtained results.

We begin with the description of our experimental setup in Section 4.1. Then, in Section 4.2 we describe a user study we carried out to measure our ability to learn the topics of the issued queries accurately. In Section 4.3 we present the results from our user survey that measures the perceived quality of our personalized ranking method. Finally, we discuss our experiences on the performance of our system in Section 4.4

4.1. Experimental Setup

To evaluate the effectiveness of our technique we have recorded the query and click sessions of 11 users for a period of one month. More specifically, for the experiments that we present in the next sections we relied on the following data:

(i) The queries that 11 users issued during October 2006. To collect such data, we contacted 11 postgraduate students from our school and asked them to install a browser plug-in that we have implemented which recorded the queries that our subjects issued to Google. Overall in our data set we collected a total of 756 queries during the period of one month with 68.72 queries per user on average.

(ii) For every query issued and right before presenting the results to the user we asked our subjects to specify the general topic of interest communicated via that query. The users were presented with the text descriptors of the categories in the first 2 levels of the Open Directory that are common within our ontology (156 topics in total). The topics were presented to the subject in a hierarchical manner for easier navigation. On average, each of our subjects was interested in 1.8 (out of 16) different top-level topics per query and 48 (out of 156) second-level topics. Topic interests for each of the queries were explicitly determined by our study participants.

(iii) For every query issued by each user we recorded the pages that were clicked by the respective user. From this set of pages we generated

Table I. Experimental pages topic distribution.

Category	**no. of documents**	**no. of sub-topics**
Arts	28,342	18
Sports	20,662	26
Games	11,062	6
Home	6,262	7
Shopping	52,342	15
Business	60,982	7
Health	23,222	7
News	9,462	4
Society	28,662	14
Computers	35,382	13
Reference	13,712	10
Recreation	8,182	20
Science	20,022	9
Total	318,296	156

the set of *visited* pages by each user by keeping only pages that the user dwelled on for more than 10 seconds.

(iv) For every page that was visited by a given user as a result of a click, we asked the user to specify whether the visited page was relevant or non-relevant to the given query. For the relevance judgement we restricted to pages reachable directly from the search results and we did not consider any pages reachable after following more than one links from the search results. The relevance judgement was performed right after visiting each page and not after the termination of the user session.

(v) For every page used in the following experiments we calculated its topic as follows. If the page appeared in Dmoz under the 156 topics presented to the users we used the DMOZ topic as the topic of the page and we pre-computed its DirectoryRank (DR) value for the given topic. If the page did not appear in Dmoz we calculated the affinity of the page to the Dmoz topics based on its $RScore$ and $IScore$ values and according to the discussion in Section 3.1.1. After that we computed the page's DR for the assigned categories. Table I shows the distribution of the pages to the Dmoz top-level categories that we used in our experiment.

In the following sections, we describe how we used our experimental data to evaluate the accuracy of our model in determining the user topic preferences automatically and the effectiveness of our personalization method in improving the quality of search results.

4.2. Accuracy of Topic Preference Learning

One important component of our method is the identification of the topics that are of interest to the users. If we can identify the topics for a given query correctly we can then build a more accurate profile and we can then personalize the results more effectively. On the other hand if, through our click-based topic learning method, we learn topics that were completely different from what the users had in mind then our personalized ranking might not be very effective. In this section we study the accuracy of our method in learning through the click history of the users the topics that they are interested in.

To that end, we will use the topic preferences that were specified by each user after the submission of a query, as we described in Section 4.1. These topics will serve as the ground truth for evaluating the performance of our method. However, given the fact that a user is typically interested in a small number of topics per query, we will focus on comparing at most the top-3 topics from the ground-truth set with the set of topics identified from our method. In order to rank the topics from the ground-truth set, we consider that each viewed page contributes a vote to the topic where it has the highest DR and we order the user-specified topics based on the number of votes. In this way we get a ground-truth set $\mathbb{L}_a$ for the topic preferences.

For each of our experimental queries, we apply our model and, by exploiting the user click history, we compute the estimated ranking of topics $\mathbb{L}_e$, based on Equation 11. In order to evaluate the accuracy of our learning method, we used the OSim measure, reported in the work of (Haveliwala, 2002), which indicates the degree of overlap between the top k topics of the two sets of preferences for each of the queries. Formally, the overlap of two ranked sets $\mathbb{L}_a$ and $\mathbb{L}_e$ (each of size n_l) is given by:

$$OSim(\mathbb{L}_e, \mathbb{L}_a) = \frac{|\mathbb{L}_e \cap \mathbb{L}_a|}{n_l} \tag{19}$$

where $\mathbb{L}_a$ denotes the actual topic preference for a given query as this is determined by the ground-truth model, $\mathbb{L}_e$ denotes the estimated topic based on the user search history for that query as it has been estimated by our proposed model and n_l denotes the total number of topics considered for a given query.

Table II. Distribution of experimental queries across the different values for overlapping topics between the ground truth and the ones identified by our method.

OSim	**Fraction of queries**
1	10.0%
2/3	73.3%
1/3	13.3%
0	3.4%

Using the above formula we computed for each of the experimental queries the overlap between the topic preferences ranked in the top-3 positions for that query by the ground truth and our preference estimation model respectively.

The results are shown in Table II, where we report the fraction of queries (out of the 756) for which our method managed to correctly identify all their ground-truth topics (i.e. having $OSim = 1$), 2 out of 3, 1 out 3 or zero topics.

Overall, our method has a good potential in learning the user preferences that are latent behind their search queries. The overlap of the learned topics between our method and the ground truth is 0.632 on average, which means that our method managed to identify 2 out of 3 topics correctly most of the time. There were 76 queries in our dataset for which our method guessed correctly all 3 topics and 26 queries where it did not manage to identify any of the ground-truth topics.

In the evaluation that we have just described, the decision whether our method managed to identify the desired topics correctly is "binary": we either managed to identify the topic or we didn't. However, there may be topics which are semantically similar. In these cases if a user specified a given topic but our method did not manage to get this exact topic right we will not count it as a correct topic. However, we would like to capture the fact that there may be topics that we got wrong but are somehow semantically similar to the topic that we should have identified.

To investigate this issue further we also evaluated the semantic similarity among the first three topics estimated for every query by the ground-truth system and our model respectively. To estimate the similarity between a query's actual and estimated topics, we employ the topic similarity (s) values given by the topic's similarity in the

ontology (see Equation 16) and we compute for every query the average similarity between each of the actual and estimated topics. We compute the semantic similarity $TSim$ between the sets of topics $\mathbb{L}_a$ (the ground truth) and the estimated set of topics $\mathbb{L}_e$ as:

$$TSim(\mathbb{L}_e, \mathbb{L}_a) = \frac{1}{n_l n_l} \sum_{\forall T_e \in \mathbb{L}_e, \forall T_a \in \mathbb{L}_a} s(T_e, T_a) \tag{20}$$

Based on the above equation, we compute the semantic similarity of topics from our method to the ground truth. Our results indicated that, our method has a a semantic similarity of 0.738, which means that, overall, the topics that were identified from our method were about 73% semantically similar to the ground truth topics. Considering that our method has 63% accuracy in learning the same topic for a query that a human has explicitly specified and given that our method has 73% accuracy in learning a close-matching to the ground truth topic for a query, we can derive that in 10% of the cases where our model did not succeed in identifying a manually-selected query topic, it still managed to deliver a topic that is semantically close to the one determined by our subjects for describing the intentions of the query.

Summarizing, our model has a promising potential in identifying suitable query topics for most queries. Recall that the user topic preferences are derived from the topics identified for the queries that the user has previously issued (Equation 11). Therefore, if our model manages to successfully identify a suitable topic for the user queries, it may also be capable of effectively ranking the search results based on the identified topics. In the next section, we investigate how the learned topic preferences participate in the personalization process and we study how they influence the quality of the search results.

4.3. Quality of Personalized Search

In this section, we experimentally measure the effectiveness of our personalization method in improving the overall quality of search results. To measure that, we used the data collected from our human survey (described in Section 4.1) and we compared the following ranking schemes.

1. **DirectoryRank (DR):** Given a query, we rank pages that match the query based on their importance to the topics under which they are listed within the Open Directory. In case the query-retrieved pages are not listed in the Open Directory, we pre-process them and classify them to the respective ontology topics as we previously described. This ranking method (Equation 8) does not take

into account the user preferences but favors pages which are very representative of a topic. Our goal here is to indirectly "identify" the topics of a query based on the returned results and present the user with very representative pages of such topics.

2. **Personalized DirectoryRank (R_P):** We rank the pages that relate to a given query based on Equation 18, which considers the estimated user preferences for the topics under which the query relevant pages are listed from the Open Directory. This ranking method uses both the user topic preferences from her click history along with the query terms to identify the likely topic of the query.

To measure the quality of each ranking scheme, we relied on a large pool of search queries that we had recorded during an earlier study, out of which we picked the 10 most frequently occurring queries that had not been issued by the users of our current study. We then presented those 10 queries to our 11 participants and we asked them to specify for each query the pages that they would consider informative/relevant to their own information need had they submitted such a query.

The pages presented to the users during this judging process were drawn from the set of pages that we preprocessed as we discussed in Section 4.1. For each query we determined the top-20 pages that contained most frequently the query keywords (since the set of pages is a union of search results and Dmoz pages they are globally popular on the Web anyway). Overall, we relied on a total set of 200 pages (20 pages per query).

The reason for focusing on the most frequent queries instead of random queries was to avoid very rare, and potentially specialized, queries where our subjects would not have a clear idea of what the intention might be. Additionally, we picked a different query set for our second evaluation (instead of relying on our participants' self-defined queries, previously collected) in order to measure our method's performance on queries that have not been explored during the user profiling process. Moreover, relying on a query set that was new to all our participants enabled us to ensure some level of consistency in the evaluation of results from the individual subjects.

We should note that our users were simply asked to indicate which of the displayed pages (if any) was in their opinion relevant to the intention of the query, without any restrictions on the order of selection. That is, the pages were presented in random order so as not to influence our subjects decisions, which were made on a simple yes/no (i.e. relevant/non-relevant) basis.

Additionally, considering that the queries were chosen by us on the ground that they were new queries to our subjects (i.e. they have not

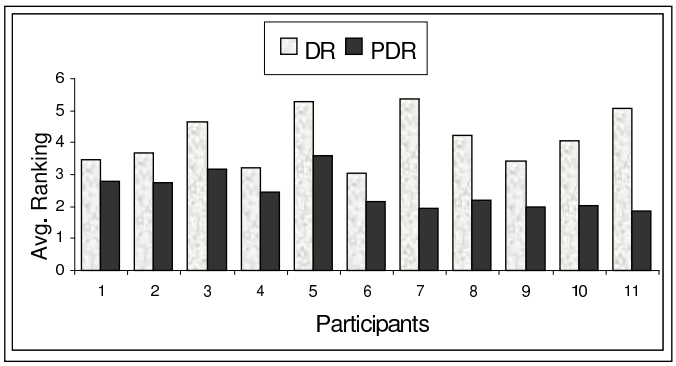

Figure 2. Average weighted rankings of the examined pages by participant. Lower values indicate improved search quality.

been explored in the course of modeling the topic preferences of our subjects), the level of "familiarity" to each query may vary from subject to subject. To account for this possibility, we also asked our users to indicate their "familiarity" to each of the queries that was presented to them.

More specifically, when a query was presented to a user (and before displaying the query results) we asked them to indicate how familiar they are to the respective query (thus capturing how likely they are in issuing such a query during a Web search). Familiarity was specified on a 10-point scale with 0.1 (i.e. 10%) meaning that the user is not very familiar with the query, while 1 (100%) meaning that the user is very familiar with the query.

For each of our participants and queries, we computed the weighted (by the level of the user familiarity to the query) average rank of the selected pages the level of the user familiarity to the query using the following formula (according to (Qiu and Cho, 2006)) under each ranking scheme:

$$AvgRank(u,q) = \sum_{p \in S} Rank(p) \cdot Pr(q|u) \tag{21}$$

Where S denotes the set of pages that a user (u) selected as relevant to the query q, $Rank(p)$ denotes the rank position (i.e. 1,2,3,...) of page p by the respective ranking scheme, i.e. either DR or R_P, and $Pr(q|u)$ gives the likelihood (i.e. the level of familiarity) the user issues query q in a web search. Smaller $AvgRank$ values indicate better results quality, i.e. higher position of the relevant pages in the list of search results.

In Figure 2, we have aggregated and averaged the results of Equation 21 by users to demonstrate the overall effectiveness of our personalized ranking scheme for each of our study participants. We can see that

Table III. Average level of user familiarity with the 10 experimental queries.

Participant	**Average Familiarity**
#1	0.6
#2	0.5
#3	0.5
#4	0.5
#5	0.6
#6	0.6
#7	0.8
#8	0.7
#9	0.7
#10	0.7
#11	0.8

our personalized *DirectoryRank* method outperforms *DirectoryRank* in all cases. However, obtained results demonstrate that, for some users, R_P improves the quality of search results more significantly than others. For example, for the 7th participant our personalized *DirectoryRank* scheme outperforms *DirectoryRank* by almost 64%, whereas for the 1st participant the improvement is around 19.5%. This implies that the performance of our personalization approach depends on the user's search patterns and the nature of their queries. In particular, if there are only a few topics suitable to describe a given query, then our model will be able to capture them accurately and it will increase the performance of our personalized ranking. The overall improvement of personalized *DirectoryRank* over the simple *DirectoryRank* is 40.78% on average for all participants.

Another issue that we examined is whether the variation in the users' ratings on the different rankings is influenced by the users' familiarity with the respective queries. Table III reports the average level of familiarity that each of our participants indicated for the set of queries that they examined.

A combined analysis of the results presented in Figure 2 and Table III implies that there is some underlying connection between the users' familiarity with the queries and their ratings to the query retrieved pages. In particular, we observe that the more familiar a user is to the underlying query, the higher she values the personalized results for the query. For example, for the 7th and the 11th participant, who were the most familiar with the examined queries (i.e. average level of

familiarity = 0.8), we observe that our personalized ranking scheme had an improvement of 34.3% and 32.3% respectively in delivering qualitative search results, compared to the performance of DR. On the other hand, for the 2nd, the 3rd and the 4th participants, who were not that familiar with the given queries (i.e. average level of familiarity = 0.5) personalized rankings slightly improved the search quality (i.e. by 9.4%, 14.6% and 7.3% respectively) compared to non-personalized rankings. Overall, we observe that the more familiar a user is to a particular query, the better she can determine her topic preference intended in the query. Therefore, she can better distinguish between personalized and non-personalized search results and value the former higher.

4.4. Discussion

So far our evaluation showed that our model (taken as a whole) has a good potential in personalizing search results according to the learned user preferences. In this paper our goal is to present a model for personalizing search results and to perform an overall evaluation of our model. However, the system that we have presented so far comprises of various components which may affect the overall performance. In this section we try to shed some light on the different parameters discussed so far.

First, our model relies on a topical ontology as the backbone infrastructure for identifying the topics of visited pages and thereafter for learning the topical preferences of the individual users. Since user profiling is greatly dependent on the classification accuracy in assigning a correct topical category to each of the visited pages, it naturally occurs that the accuracy of our personalized search mechanism depends on the classification performance. In other words, the greater the classification accuracy in detecting the correct topic of a page, the better the effectiveness of our system in deciphering the topical preference of the user.

Throughout our work and during the evaluation of the personalization process we relied on the categories coming from the Dmoz hierarchy. By using a topical hierarchy that is manually constructed our hope is that we can capture in a better way the topical interests of the users. In addition, by enriching such a hierarchy with WordNet concepts and by using them when determining the topics of the pages we can achieve good classification accuracy (Stamou et al., 2006). It is likely that under a different classification scheme or topical ontology, our model would perform differently but we believe that it could still be directly applied without the need for any modifications.

One contribution of our model is its ability to consider the current user query during the user-profiling process rather than rely exclusively on the user's past search history. When we determine the degree of interest to a given topic for a given user we employ the past user interests, the topical category of the user's current query and the semantic relevance between past and current topic preferences. The way in which these metrics are combined certainly plays a role in the personalization of the search results in the sense that if past user preferences are valued higher than present ones, then the pages that match better the previously preferred topics will inevitably be given priority over the results that are closer to the user's current topic interests. Fine-tuning the contribution or even the formalization of these parameters is a task that depends on a variety of factors such as the user search behavior, the nature of her queries, the underlying search engine or whether we would like to allow the parameters to change over time. This is a task that we plan to investigate in the future, but we should point out the need for data available for a period longer than one month that we used for our current study.

5. Related Work

There has been previous work in personalizing web search. One approach to personalization is to have users explicitly describe their general search interests, which are stored as personal profiles (Pazzani et al., 1996; My Yahoo!, 2007). Many commercial systems rely on personal profiles to personalize search results by mapping Web pages to the same categories. Personal profiles, specified explicitly by the users have also been used to personalize PageRank (Aktas et al., 2004; Jeh and Widom, 2003). For example (Jeh and Widom, 2003) present a framework for restricting the bias vector during computations of PageRank. This framework relies on explicitly defined personalized views, which are employed as partial vectors at query time for personalizing search results. In (Aktas et al., 2004) personalized PageRanks are computed based on the user profiles explicitly specified by the users. Our work is different from the above approaches in that our method does not require users to be directly involved in the profile building process.

There also exist many works on the automatic learning of a user's preference based on the analysis of her past clickthrough history (Chen and Sycara, 1998; Pretschner and Gauch, 1999; Sugiyama et al., 2004) and past queries (Shen and Zhai, 2003; Speretta and Gauch, 2004). In (Pretschner and Gauch, 1999) 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. Moreover, in (Pretschner and Gauch, 1999) the authors limit their approach to the Web browsing paradigm and unlike our method they do not account for the semantic correlation between the pages and the issued queries. On the other hand, in (Chen and Sycara, 1998), multiple TF-IDF vectors are generated, each representing the user's interests in one area. In (Sugiyama et al., 2004) 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. In (Liu et al., 2002) the authors propose the mapping of queries to topical categories that are likely to be related to the user interests for personalizing Web search results. (Sun et al., 2005) explore the correlation between users, their queries and search results clicked, to model user preferences. (Ma et al., 2007) suggest the use of a topic hierarchy for modeling the user search interests. Although their model differs from ours in several parts, their experimental evaluation demonstrated the potential of ontologies in the search personalization process. There are several factors that determine what makes a page interesting to the user (Agichtein et al., 2006), such as the time spent on a page in conjunction to the page's length (Gauch et al., 2003), the points of focus on a page (Joachims et al., 2005), email, and/or bookmark of the page for future reference (Teevan et al., 2005) and so forth. Likewise (Teevan et al., 2005) employ rich models of user interests, built from both search-related information and information about the documents a user has read, created and/or emailed. Moreover, (Fox et al., 2005) explored how implicit measures of user interest (such as time spent on a page, clickthrough, user activities, etc.) can be used to develop predictive user models. Their experimental setup relied on a non-laboratory setting similar to the one employed in our study. Obtained results showed that the combination of implicit measures for building user profiles contributes towards the accurate prediction of the user satisfaction.

In a recent study (Dou et al., 2007) experimentally evaluated the performance of different personalization strategies and observed that personalized search has different effectiveness on different queries. Therefore, they suggest that although click-based personalized strategies work well, they could become more reliable if they are combined with other profile-based personalization techniques. Recently, (Chirita et al., 2007) explored desktop data for creating a Personal Information Repository that is used to represent a rich source of profiling information. Although, our work shares a common motivation with the work in (Chirita et al., 2007), nevertheless our implementation is different, since we rely on the user's recorded search behavior rather than on their personal

collection of data. Moreover, our approach is different in that user topic preferences are determined based on the content of real web pages and as such they represent the particular interests of the user while interacting with the web, which may be different to the general user interests communicated via the data she stores in her workstation.

Although most of these works mainly focus on learning the user interests from the analysis of her past click history, in our work we focus on learning the user's preferences not only from her previous searches but also from the topics that are hidden behind her current queries. That is, our approach accounts for non-stationary queries and addresses the user changing interests. In this scope, we perceive our work to be complementary to precious studies on automatic learning the user's search interests. The contribution of our approach is that we combine in novel ways components that have been previously used or proposed by others as well as the fact that we suggest the exploitation of an enriched topical ontology for personalizing web search.

Researchers have also proposed ways to personalize Web search by modifying PageRank to account for user personal preferences. In (Richardson and Domingos, 2002) the PageRank vectors are tailored based on query terms but not by individual users. (Haveliwala, 2002) introduces the Topic-Sensitive PageRank scheme, which personalizes PageRank values by giving different weights to pages; one for each topic listed in the Open Directory. Although Topic-Sensitive PageRank has a significant potential in improving the search result quality, nevertheless it does not fully address the problem of automatic learning the user interests. Recently, (Qiu and Cho, 2006) proposed a formal framework for learning the user's interest and used that knowledge for further improving the search quality of the Topic-Sensitive PageRank. Our work differs from these studies in that pages are characterized by their DirectoryRank values, which are determined by the pages' content importance to the topics considered rather than their links connectivity on the Web graph. It will be interesting though for a future study to see how our user preferences learning model can be applied to personalize search based on the pages' Topic-Sensitive PageRank.

6. Conclusion

In this paper we investigated the search personalization problem and we presented an ontology-based framework which automatically learns the user's search interests based on the combined analysis of the user's past clickthrough data and current queries.

In particular, we first proposed the use of a topical ontology for identifying the topic importance of Web pages and associate them with user clicks on search results. Then, based on this association we presented a method for actually learning the user interests based on both the user-issued queries and their relationship to the user's past topic preferences. Finally, we proposed a method to rank search results based on the learned user interests. To evaluate the effectiveness of our approach, we conducted an experimental study, where we used our method to learn the interests of 11 real users.

We then compared the performance of our approach in ranking results for our subjects to the performance of DirectoryRank, a ranking scheme that orders pages in terms of their importance for particular topics. Our experimental results, indicate improvement in the search quality – about 41% improvement over the topic-specific rankings – demonstrating the potential of our approach in personalizing search. In practice, our study showed that the search results returned by our method are both relevant to the user interests and of good quality.

We now discuss some advantages that our approach exhibits compared to other personalization techniques. First, considering that our method relies not only on the user's past click data but it also considers the user's current query in the profiling process, we believe that it can work well in the dynamic environment of a web search engine. A significant advantage of our approach is that it uses a built-in topical ontology to compute one or more suitable topics for describing the page contents as well as for representing the search profiles of the users. For a detailed evaluation on the ontology's efficiency in automatically categorizing pages, we refer the interested user to the work of (Stamou et al., 2007). Therefore, our model could extend other personalization schemes, which operate upon already classified pages, and be successfully applied to pages with yet-unknown topics (such as dynamic pages). Moreover, our method is not tightly integrated with our ontology, but it can be easily deployed with a different more generic or more specific ontology. Although our model can operate with any ontology chosen, the only pre-requisite is that the ontology is enriched with WordNet hierarchies. Given that WordNet is publicly available and there exist numerous applications that use it, we do not consider its exploitation as a problem. Moreover, there exist WordNets for several natural languages, most of which provide mappings to the English WordNet. The SUMO ontology also supports several languages other than English. Therefore, we believe that our ontology building model can be explored for creating ontologies for other languages and that our personalization method that uses the topical ontology can be useful for personalizing non-English search results.

In the future, we plan to extensively evaluate the performance of our method in personalizing Web search by comparing it to the performance of other rankings schemes, such as the Topic-Sensitive PageRank formula (Haveliwala, 2002), or the Personalized Topic-Sensitive PageRank scheme (Qiu and Cho, 2006). In addition, we plan to expand our framework to take rich models of user interests into account, such as the user activity while visiting a page (e.g. length of stay on a page, points of focus etc.), email, bookmark information, etc.

References

Agichtein, E., E. Brill, S. Dumais, and R. Ragno: 2006, 'Learning User Interaction Models for Predicting Web Search Result Preferences'. In: *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Seattle, WA, pp. 3–10.

Aktas, M., M. Nacar, and F. Menczer: 2004, 'Personalizing PageRank Based on Domain Profiles'. In: *Proceedings of WebKDD 2004: KDD Workshop on Web Mining and Web Usage Analysis*. Seattle, WA.

Barzilay, R. and M. Elhadad: 1997, 'Using Lexical Chains for Text Summarization'. In: *Intelligent Scalable Text Summarization Workshop (ISTS'97), ACL, Madrid, Spain.*

Bentivogli, L., P. Forner, B. Magnini, and E. Pianta: 2004, 'Revising WordNet Domains Hierarchy: Semantics, Coverage, and Balancing'. In: *Proceedings of COLING 2004 Workshop on Multilingual Linguistic Resources*. Geneva, Switzerland, pp. 101–108.

Broder, A. Z., S. C. Glassman, M. S. Manasse, and G. Zweig: 1997, 'Syntactic Clustering of the Web'. *Computer Networks* **29**(8–13), 1157–1166.

Chen, L. and K. Sycara: 1998, 'WebMate: a Personal Agent for Browsing and Searching'. In: *Proceedings of the Second International Conference on Autonomous Agents*. Minneapolis, MN, pp. 132–139.

Chirita, P. A., C. S. Firan, and W. Nejdl: 2007, 'Personalized Query Expansion for the Web'. In: *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Amsterdam, The Netherlands, pp. 7–14.

Dou, Z., R. Song, and J.-R. Wen: 2007, 'A Large-scale Evaluation and Analysis of Personalized Search Strategies'. In: *WWW '07: Proceedings of the 16th International Conference on World Wide Web*. Banff, Alberta, Canada, pp. 581–590, ACM.

Fellbaum, C.: 1998, *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

Fox, S., K. Karnawat, M. Mydland, S. Dumais, and T. White: 2005, 'Evaluating Implicit Measures to Improve Web Search'. *ACM Transactions on Information Systems* **23**(2), 147–168.

Gauch, S., J. Chaffee, and A. Pretschner: 2003, 'Ontology-based Personalized Search and Browsing'. *Web Intelligence and Agent Systems* **1**(3-4), 219–234.

Gliozzo, A., C. Strapparava, and I. Dagan: 2004, 'Unsupervised and Supervised Exploitation of Semantic Domains in Lexical Disambiguation'. *Computer Speech and Language* **3**(18), 275–299.

Gulli, A. and A. Signorini: 2005, 'The Indexable Web is more than 11.5 Billion Pages'. In: *Special Interest Tracks and Posters of the 14th International Conference on World Wide Web*. Chiba, Japan, pp. 902–903.

Haveliwala, T. H.: 2002, 'Topic-sensitive Pagerank'. In: *Proceedings of the Eleventh World Wide Web Conference*. Honolulu, HI, pp. 517–526.

Jansen, B. J., A. Spink, and T. Saracevic: 2000, 'Real Life, Real Users, and Real Needs: a Study and Analysis of User Queries on the Web'. *Information Processing & Management* **36**(2), 207–227.

Jeh, G. and J. Widom: 2003, 'Scaling Personalized Web Search'. In: *Proceedings of the 12th International Conference on World Wide Web*. Budapest, Hungary, pp. 271–279.

Joachims, T., L. Granka, B. Pan, H. Hembrooke, and G. Gay: 2005, 'Accurately Interpreting Clickthrough Data as Implicit Feedback'. In: *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Salvador, Brazil, pp. 154–161.

Krikos, V., S. Stamou, P. Kokosis, A. Ntoulas, and D. Christodoulakis: 2005, 'DirectoryRank: Ordering Pages in Web Directories'. In: *WIDM '05: Proceedings of the 7th Annual ACM International Workshop on Web Information and Data Management*. Bremen, Germany, pp. 17–22, ACM.

Liu, F., C. Yu, and W. Meng: 2002, 'Personalized Web Search by Mapping User Queries to Categories'. In: *CIKM '02: Proceedings of the Eleventh International Conference on Information and Knowledge Management*. McLean, VA, pp. 558–565, ACM.

Ma, Z., G. Pant, and O. R. L. Sheng: 2007, 'Interest-based Personalized Search'. *ACM Transactions on Information Systems* **25**(1), 5.

My Yahoo!: 2007, 'My Yahoo! `http://my.yahoo.com`'.

Pazzani, M. J., J. Muramatsu, and D. Billsus: 1996, 'Syskill & Webert: Identifying Interesting Web Sites'. In: *Proceedings of the 13th National Conference on Artificial Intelligence and 8th Conference on Innovative Applications of Artificial Intelligence*, Vol. 1. Portland, OR, pp. 54–61.

Pease, A., I. Niles, and J. Li: 2002, 'The Suggested Upper Merged Ontology: A Large Ontology for the Semantic Web and its Applications'. In: *Working Notes of the AAAI-2002 Workshop on Ontologies and the Semantic Web*. Edmonton, Canada.

Pretschner, A. and S. Gauch: 1999, 'Ontology Based Personalized Search'. In: *Proceedings of the 11th IEEE International Conference on Tools with Artificial Intelligence*. Chicago, IL, pp. 391–298, IEEE Computer Society.

Qiu, F. and J. Cho: 2006, 'Automatic Identification of User Interest for Personalized Search'. In: *Proceedings of the 15th International Conference on World Wide Web*. Edinburgh, Scotland, pp. 727–736.

Resnik, P.: 1995, 'Using Information Content to Evaluate Semantic Similarity in a Taxonomy'. In: *Proceedings of the International Joint Conference on Artificial Intelligence*. Montreal, Quebec, Canada, pp. 448–453.

Richardson, M. and P. Domingos: 2002, 'The Intelligent Surfer: Probabilistic Combination of Link and Content Information in PageRank'. In: *Advances in Neural Information Processing Systems 14*. Cambridge, MA, pp. 1441–1448, MIT Press.

Shen, X. and C. Zhai: 2003, 'Exploiting Query History for Document Ranking in Interactive Information Retrieval'. In: *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Toronto, Canada, pp. 377–378.

Song, Y. I., K. S. Han, and H. C. Rim: 2004, 'A Term Weighting Method Based on Lexical Chain for Automatic Summarization'. In: *Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics*. Seoul, Korea, pp. 636–639.

Speretta, M. and S. Gauch: 2004, 'Personalizing Search Based on User Search Histories'. In: *Proceedings of the 2004 CIKM Conference on Information and Knowledge Management*. Washington, D.C.

Stamou, S., A. Ntoulas, and D. Christodoulakis: 2007, 'TODE: An Ontology Based Model for the Dynamic Population of Web Directories'. *Data Management with Ontologies: Implementations, Findings and Frameworks, published by Idea Group Inc.* pp. 1–17.

Stamou, S., A. Ntoulas, V. Krikos, P. Kokosis, and D. Christodoulakis: 2006, 'Classifying Web Data in Directory Structures.'. In: *Proceedings of the 8th Asia Pacific Web Conference*. Harbin, China, pp. 238–249.

Sugiyama, K., K. Hatano, and M. Yoshikawa: 2004, 'Adaptive Web Search Based on User Profile Constructed without any Effort from Users'. In: *Proceedings of the 13th International Conference on World Wide Web*. New York, NY, pp. 675–684.

Sun, J. T., H. J. Zeng, H. Liu, Y. Lu, and Z. Chen: 2005, 'CubeSVD: a Novel Approach to Personalized Web Search'. In: *Proceedings of the 14th International Conference on World Wide Web*. Chiba, Japan, pp. 382–390.

Teevan, J., E. Adar, R. Jones, and M. Potts: 2007, 'Information Re-retrieval: Repeat Queries in Yahoo's Logs'. In: *SIGIR '07: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Amsterdam, Netherlands, pp. 151–158, ACM.

Teevan, J., S. T. Dumais, and E. Horvitz: 2005, 'Personalizing Search via Automated Analysis of Interests and Activities'. In: *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Salvador, Brazil, pp. 449–456.

Turney, P.: 2004, 'Word Sense Disambiguation by Web Mining for Word Co-Occurrence Probabilities'. In: *Proceedings of the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL-3)*. Barcelona, Spain, pp. 239–242.

Authors' Vitae

Sofia Stamou

Dr. Sofia Stamou is a senior researcher in the Databases Laboratory and adjunct lecturer in the Computer Engineering and Informatics Department at the University of Patras, where she teaches Language Technology. Dr. Stamou received her B.A. in Philosophy from the University of Ioannina in 1999 and her M.S. and Ph.D. degrees in Computer Science from the University of Patras in 2002 and 2006 respectively. Dr. Stamou has published several articles in international journals and conferences and has served as a program committee member for several conferences in the area of Language Technology. Her research interests

are in the evaluation, study and practical application of semantical analysis and text processing techniques in the World Wide Web.

Alexandros Ntoulas
Dr. Alexandros Ntoulas is a Researcher at Microsoft Research in Mountain View, California. He received a Ph.D. degree in Computer Science from the University of California Los Angeles (UCLA) in 2006 and a B.Sc. from the Computer Engineering and Informatics Department (CEID) of Patras University, Greece in 2000. His area of expertise is Databases and Web information systems and his research interests are in the study of systems and algorithms that facilitate the monitoring, collection, management, mining and searching of information on the World Wide Web. He is the co-founder of Infocious (now Lingo Semantics) and the recipient of the Best Paper Award for the ICDE 2005 conference. The work presented in this article was performed while he was at UCLA.