Applying Diversity and Novelty to Personalized Search[*]

Peter Lubell-Doughtie[†]
University of Amsterdam
lubell@science.uva.nl

ABSTRACT

Recent research into retrieving and reordering search results so that they cover a maximum number of topics and information needs has gained traction as a means to assist navigation through the explosion of available online information. This problem is also addressed by personalized search: the retrieving and reordering of search results biased to the preferences of a particular user. We investigate a search personalization system that builds a user model from the latent topics mined from their queries and clicked documents. We find that a user's clicked documents exist in a specific area of the latent topic space. By "diversifying" search results biased to our user model we can reorder search results to the user's preferences.

1. INTRODUCTION

The goal of advanced search methods and interfaces is to provide users with more effective ways in which to view and manipulate search results. The two types of advanced search we will focus on are facet-based search and personalized search. Facet-based search assigns a topic or set of topics to documents returned for a query and allows users to navigate these documents by navigating their topics. Topics may be assigned based on an external ontology, the relative frequency of document terms, a topic modeling algorithm, or some combination of these approaches.

Graph based approaches leverage a combination of assigned topics, a semantic graph linking topics and/or folksonomies, a hyperlink graph connecting documents, and social networking graphs. Personalization is produced by reordering topics based on user topic preferences as indicated through various log files and context sources, potentially as extensive as the word processor documents the user is currently browsing and their geographic location. A recurring theme throughout the research is that the advantages of personalization vary widely per user and query, with more benefit for users who conduct more searches over topics of which they are better informed.

Distinct from research into search personalization is research into alternative foundations for ad-hoc search result orderings which do away with the classic assumption that ordering should be strictly based on relevance to the query. This assumption – ensconced as the probabilistic ranking principle, or PRP – forms the theoretical basis for ranking systems such as TFIDF and BM25 as well as the popular evaluation measures MAP, MRR, and nDCG. Alternative approaches drop the assumption of independence between search results, made in the PRP, and attempt to increase the *novelty*: coverage of new topics, and *diversity*: coverage of multiple topics, in the result list's documents (below we will use the term diversity in reference to both diversity and novelty).

Research into diversity methods for information retrieval has led to algorithms which re-rank search results based on a topic's probability given a query, thereby emphasizing a weighted diversity amongst topics. Because diversity is a new area, researchers have not explicitly applied these algorithms to personalized search in any studies we've encountered. Excepting [18] in which diversity is used in relation to query reformulation applied to result re-ranking.

We propose to apply the topic biasing capabilities of diversity re-ranking to personalized search. We also suggest that result list scoring algorithms which are theoretically dependent upon the PRP are inappropriate for personalized search if (and because) they do not take into account the varying personal topic biases that ought factor into the scoring. By applying re-ranking algorithms, developed for optimizing against diversity scoring metrics, personalized search can be improved and this improvement can be accurately measured.

We proceed by presenting previous research into personalized search, diversity, and evaluation in Section 2. In Section 3 we describe multiple approaches to building user models from query logs and the re-ranking algorithms that form our method. Next we present a preliminary experiment in Section 4 and the results and analysis in Section 5. In Section 6 we further discuss our results, then we conclude and present future work in Section 7.

2. RELATED WORK

Personalized search and diversity use various techniques (including faceted search, topic modeling, and data-mining) to address multiple overlapping problems. The main problems addressed by personalized search are how to build a user model and how to rank or re-rank documents. The main problems addressed by diversity are how to categorize documents and how to rank or re-rank documents. Overlap occurs because user models are often built through categorized documents and because ranking may be performed with respect to categorization in both cases. We will review

[*] Written for Learning Systems Profile Project, Professor Maarten van Someren, 2nd Semester 2009-2010

[†] UvA Net ID: 6095445

different approaches to personalized search, briefly review search system evaluation, and then look at how diversity research can be applied to personalization.

2.1 Personalized Search

Personalized search aims to improve the presentation of search results by selecting which results to display and reordering results per user and query based upon a user model. Many approaches build user models by assigning a topic model to users through analysis of their previous queries and selected results as recorded offline in search logs and/or online as the user is searching. A topic model is then assigned to each document returned for a query and the documents are re-ordered based on some ranking function which uses a topic weighting derived from the user model. Research has shown that the effectiveness of search personalization depends on the user and the query, that click based personalization can consistently perform well while profile based personalization is unstable, and that considering contexts at different time scales is an important factor in improving performance [8].

There is significant overlap between the various methods used for personalized search. The literature reviewed below is divided based upon the primary source of information used in re-ranking. A thorough overview of personalized search is provided in Micarelli et al. [16].

2.1.1 Topics and Categorization

The Topic-Sensitive PageRank (TSPR) model is defined based upon a topic-driven random surfer model in which the user chooses a topic t from the user's topic preference vector $\mathbf{T}$ based upon a conditional distribution [17]. Next the user uniformly jumps to a page in topic t and continues by, with some probability, performing a random walk starting from this page or choosing a new topic according to $\mathbf{T}$ (this is equivalent to PageRank but with a teleportation vector biased by $\mathbf{T}$). Based on this model the authors learn $\mathbf{T}$ for each user from her visit probability vector $\mathbf{V}$, which can be calculated a priori from search engine logs. The relation used is

$$\mathbf{V} = \sum_{t \in T} T(t) \cdot \mathbf{TSPR}_t^{9/4} \tag{1}$$

where $T(t)$ is the to be calculated user preference for topic t and $\mathbf{TSPR}_t$ is the Topic-Sensitive PageRank vector for topic t, with the 9/4 exponent added based on research showing that this expression can compensate for the biasing in page visits search engine results cause through the influence of their rankings.

Based on Equation 1 the authors use maximum likelihood estimation to calculate $\mathbf{T}$. Given a query q, the authors rank a page p based on the topic vector according to

$$\sum_{t \in T} Pr(T(t)|q) TSPR_t(p) \tag{2}$$

where the probability of a topic given the query, $Pr(T(t)|q)$, is based upon a combination of user topic preferences and the topic of the query. In user evaluations ranking based on personalized TSPR or query biased personalized TSPR outperforms both simple TSPR and PageRank. The results also show that personalized TSPR outperforms query biased personalized TSPR on most tasks, but not significantly. The amount of improvement depends on users, with more improvement seen for more active users, what we would expect given that we will have more data on active users and should therefore have more representative user profiles.

In another topic-based approach to personalization, topic preferences are collected online as users browse. In [19] clicking a page indicates a preference for the page's topic and pages are re-ranked based upon the calculated preference for the page's topic with a bias against lower ranked – therefore assumed to be less relevant – pages. To balance against problems with greedily biasing towards previously seen topics, an exploration bonus is given to topics with a small probability of interest. A user study showed that *click utility*, estimated as the number of cumulative clicks, improved for the greedy re-ranking and even more for the exploitation and exploration based re-ranking, in which an exploration bias was used – this is line with work showing the non-optimality of greedy search [22].

The two previously mentioned methods rely on topics defined in the Open Directory Project (ODP)[1], an editor based hierarchical categorization of over 4.5 million websites. In [21] ODP topics are further enriched by using hyponyms from Princeton's WordNet [2] as subtopics. Based on this enriched topic hierarchy the authors construct a lexical chain of semantically related consecutive terms of the documents in a topic. These lexical chains define the DirectoryRank of a page, which is the page's relatedness score to a topic plus its relatedness score to all the pages within a topic.

Given a user topic preference profile determined from their browsing history, DirectoryRank is used to re-rank pages by weighting with respect to this profile. DirectoryRank was evaluated with and without personalization and the authors found that, although it varies per query and user, personalization outperforms. As expected, personalization was more effective for re-ranking documents when the query is within a topic the user is familiar with, and therefore has a more established and/or explicit topic preference for. The authors suggest an expanded, updatable, or otherwise dynamic ontology would likely improve performance and generalizability.

2.1.2 Query and Document Context

An alternative way to view and analyze search engine log files is as a set of discrete browsing paths. The size of, and division made to, these paths can be used to differentiate user interests across varying contextual dimensions. A recent study uses logs from a browser toolbar to extract *browse trails*, defined as the ordered URLs visited by a user, and *context trails*, defined as a terminal URL and the list of the 5 URLs preceding it [23]. The URLs' pages are then classified into the ODP topic hierarchy and user interests are determined through category label frequency calculated according to six contexts:

i. no context, assign labels to the trails' terminal pages and aggregate

ii. interaction context, assign labels to pages preceding the terminal URL and aggregated

iii. task context, collect search queries leading to this URL and then collect the other pages these queries return, finally aggregate labels from these pages

[1] Open Directory Project (ODP) http://www.dmoz.org/
[2] WordNet http://wordnet.princeton.edu/

iv. collection context, collect in-links for a URL and aggregate labels from these in-linked pages

v. historic context, aggregate labels for all pages visited within a certain time-period

vi. social context, aggregate the historic contexts from a subset of all users that visited the page

The authors' hypothesis is that by the *principle of polyrepresentation* the overlaps between multiple contexts will reduce uncertainty. Evaluations based on historic usage data found that the interaction context best predicts short-term user interests, the task context best predicts medium-term interests, and the historic and social contexts best predict long-term interests. Combined models were also used based upon linear averaging and results showed that for each time scale there is at least 1 combination that outperforms any isolated context, supporting the principle of polyrepresentation. This indicates that different combinations of contexts can best predict topic preferences at different time-scales.

Context can also be analyzed with respect to the sequence of queries issued during a search session. In [25] the context of the current query is determined from the previous query and the user's interaction with the documents returned for the previous query. Based on the previous query the context of the current query is defined by one of the following principles: *reformulation*, *specialization*, *generalization*, or *general association*.

In their research the authors manually analyzed search session logs to create a set of training examples for query contexts and then used these as features, in the *learning to rank* algorithm RankSVM [14], to re-rank documents. To evaluate which principles were effective the authors conducted a t-test on the difference between the probability a document is clicked given a principle is satisfied and given it is not satisfied. Results showed that only the third principle, *generalization*, is not effective, which is in part due to the small number of examples for this principle. Additionally, when evaluated based on number of clicks on returned documents for a query, the re-ranking algorithm improves on the baseline.

Additional research into query reformulation has shown link based analysis of consecutive queries and semantic modification patterns can discover relationships between queries not identified by term based analysis. In [12] the authors mapped queries from search engine log files onto linked data networks such as DBpedia[3], WordNet, and others. These resources were used to find the relationships between consecutive query terms, and iteratively refine them towards more relevant relationships. That the authors found informative semantic relationships between query terms shows the need for further research into query chains and query context based re-ranking.

2.1.3 Snippet Clustering and Facets

The features used in the context based approaches and the topic vectors used in Topic-Sensitive PageRank are both based on historical usage data and, as such, they can be calculated offline. A positive result of offline calculations is that ranking values can be relatively quickly computed for a query; a negative result is that personalization will be slow to adapt to changes in users' preferences. A more dynamic approach to personalization is based on clustering the web page snippets returned by search engines into a hierarchy and allowing users to navigate through this hierarchy.

In [9] web page snippets from the pages returned for a query by multiple search engines are clustered on the fly and used to generate meaningful labels without reference to an external ontology. The hierarchical clustering of these labels provide categories users can select and deselect to navigate through the search results. Beyond adaptability, three additional advantages of this approach are that it scales better than TSPR, allows for increased user privacy, and doesn't require training. The authors evaluated the general theme behind their approach and, in a user study, found that 85% of those surveyed described facet-based search as useful.

2.1.4 Link Analysis

Topic-Sensitive PageRank uses hyperlink graphs for the creation of topics or facets; this is one of many graphs available for search engine personalization. In the original presentation of the PageRank algorithm Brin and Page describe a *personalized PageRank* vector which used a set of pages chosen according to the user's interests as the pages in the teleportation vector [4]. Later work shows that the personalized PageRank can be more efficiently computed by restricting the web graph to a subset of pages the user is likely to be interested in and incrementally expanding this set based on the PageRank of frontier pages [11]. In this work the authors also apply their method to calculate a PageHostRank, which considers all pages within a host as a single node and forms a reduced graph: the host graph.

In [15] graphs based on annotations in taxonomies and folksonomies are used as the link graph. The approach uses a Personalized PageRank which biases the teleportation step of PageRank towards a subset of pages, the pages within a specific facet, and determines the relationship between a page's facet-specific PageRank and the page's membership in that facet. This method is applied to queries by calculating the facet membership of the first n pages returned for a query. The user can interactively select facets and resubmit their search to produce a personalized re-ranking. The authors conducted experiments to test the accuracy of the categories extracted and found that their method produced high precision.

In [5] personalization is based upon a user's social network. Social networks are defined based on *familiarity*, that users know each other, and *similarity*, that users share topics of interest. To re-rank entities (which may be documents or other users) e, based on a query q, the system retrieves a ranked list of related users and related terms then calculates a score based upon an interpolation between a non-personalized score, $S_{np}(q,e)$ and the ranked list of related users $N(u)$ and related terms $N(t)$, for a user u.

Given a search profile defined as $P(u) = (N(u), T(u))$, search results are re-ranked as:

$$S_p(q,e|P(u)) = \alpha S_{np}(q,e) + (1-\alpha)[\beta \sum_{v \in N(u)} w(u,v) \cdot w(v,e) + (1-\beta) \sum_{t \in T(u)} w(u,t) \cdot w(t,e)] \tag{3}$$

where $w(u,v)$ is the relationship strength between the current user and the related user, $w(v,e)$ is the relationship

[3] DBpedia http://dbpedia.org/

strength between the related user and the entity, $w(u, t)$ is the relationship strength between the current user and the related term, and $w(t, e)$ is the relationship strength between the related term and the entity. The parameters α and β determine the interpolation weighting.

Equation 3 calculates the personalized part of the personalized score as an interpolation between the sum of the user's and entity's similarity to related users and the sum of the user's and entity's similarity to related terms. In evaluations performed through both an offline study and a user survey, the authors found that similar users are better predictors than familiar users, although the results varied per query and the results from the offline study disagreed with those from the user survey. The study highlights that ideally a system must decide on the correct search policy per user and query pair.

2.1.5 Ontological Profiles

In [20] user profiles are represented through an ontology based on the ODP topic hierarchy enriched by semantic networks based on the terms within the topic's documents. The concepts have annotated *interest scores* based on the user's browsing history, and weights between concepts are determined by the similarity between their subsumed documents. Dependent on these weights interest scores are updated for all concepts using a spreading activation algorithm. Once built, a re-ranking algorithm uses the ontology based profile to find the concept most similar to the query and then calculate rank scores as the similarity of a document to a query times the similarity of the concept to a query and the user's interest score in the concept. In experiments conducted using documents in the ODP the authors found that user interests converge over time and that the re-ranking algorithm improves precision and recall.

2.2 Evaluating Advanced Search

In reviewing the many approaches to personalized search it is clear that there is no standard or accepted method being used by the authors for evaluation of the results produced by their systems. Although many studies use the ODP as a starting point for topics (albeit perhaps enhanced through different methods), the datasets used are different in each research study; they are a combination of public and proprietary search log files as well as results from user studies sampled from individuals associated with the research sponsoring institution.

Beyond differences in the dataset used, the evaluation measures vary across studies, from precision and recall to number of clicks to user surveys and task specific probability measures. Analyzing this wide range of evaluation methods is compounded by the use of different data sets or survey procedures and reflects the varying and multiple goals that can be pursued through advanced search and personalized search. To allay problems with data source variance and comparison of result evaluation we refer to research into classifying the functionality of advanced search systems.

In [24] the authors present a method for evaluating the features implemented by faceted browsers and other exploratory search interfaces. Continuing previous research into characterizing interactive flow and automation, their analysis is based upon first identifying the features of the search system and how users can interact with them, and then for each feature calculating how well it supports a set of search tactics. A compiled plot of which and to what degree the system supports each search tactic provides insight into differences and similarities between search systems, as well as what their appropriate use cases may be.

The systems reviewed in Section 2.1 are less focused on interface design and more on algorithm design. Still, we can distinguish the snippet clustering [9] and PageRank facet [15] research from the rest as they explicitly support faceted browsing on the interface level. On an algorithmic level this is something of a false distinction because most of the systems reviewed could be fashioned to support facetted browsing, however, some could not. For instance, neither the social network based personalization [5] nor the personalize PageRanks [4, 11] produce topics that could be used as facets. The system we develop will support facetted search on an algorithmic level but it is not designed so that user explicitly interact with facets.

2.3 Diversity for Personalization

Research into diversity has simultaneously pursued improved evaluation measures for diversity and the design of algorithms which perform well on these evaluation measures. The most commonly used metrics are α-nDCG and P-IA. The α-nDCG metric was first introduced in [6], it is a modified version of normalized discounted cumulative gain which defines the probability a document is relevant as the probability there exists a topic in the intersection of the topics in the user's query and those in the document. P-IA refers to the "intent aware" precision metric, introduced in [1], which weights the precision by the number of documents within a category and the probability of the query being within that category (various other intent aware metrics are similarly defined). The experiments performed in [6] and [1] assume the topics are chosen and weighted to be specific to the query, however by choosing topics and weights specific to a user and query tuple we can produce a reordering personalized to topic preferences.

In the same paper defining intent aware metrics Agrawal et al. define the IA-SELECT algorithm which reorders a search result list to optimize coverage of the list's categories [6]. A similar but more flexible approach to reordering in order to optimize diversity is presented in [7]. In the Agrawal et al. approach there is no explicit method for adjusting the amount of diversity but one can be contrived by adjusting the importance of a document's rank to its probability of relevance, the importance of rank is inversely correlated with the importance of diversity and the amount of reordering. In [7] there is an explicit parameter which controls the influence of the original relevance in the reordering and can be used to adjust the amount of diversity. Search systems can use an adjustable diversity measure to control the amount of personalization, and especially to limit the amount of personalization in cases were we suspect it will degrade usefulness, perhaps because there is not enough information about the user.

3. METHOD

We assume we are operating in an information retrieval environment were queries are submitted as a sequence of terms and results are returned as an ordered set of pages that may be enhanced through latent topics. In order to design a search personalization system we must first circumscribe personalization as with respect to a session, user, or query and what evidence we will consider when forming the personalization. This evidence could include only user feedback, the session or user log files, the log files of similar users or sessions, or additional context information (e.g. geographic location, open browser windows or documents). Based on this data we will build a user model and a ranking or reranking algorithm which orders search results according to the user model.

3.1 Formalizing a User Model

We classify the two approaches to user modeling that have been exploited in the research as *topical* and *ontological*. In the topical approach user interests are depicted as weights assigned to a list of topics and can be represented as a vector $\mathbf{t}$ where each integer $\mathbf{t}_i \geq 0$ is an unbounded frequency count or where each $0 \leq \mathbf{t}_i \leq 1$ is a real number specifying the users interest in topic i and $\sum_i \mathbf{t}_i = 1$, making the vector interpretable as a probability distribution. The topics may be based on an external and static list of topics or they may be latent topics, which may be unlabeled, created through a topic modeling algorithm. An elaboration of the topical model, moving towards an ontology, places topics within a hierarchy and computes the probability of a topic as the sum of the probabilities of all its subtopics.

An ontology based user model represents user preferences through a network of terms and topics defined as a weighted graph where nodes are words and edges are relationships between two words in which the relationship strength is denoted by the edge weight. Because phrases (ordered sequences of words) form and express very different semantic units and information needs, a word graph is not the ideal formalism for our purposes. To capture phrases we can use a structure similar to a hypergraph, in which edges are between ordered sets of nodes and the number of nodes in each ordered node set is of size $\leq k$, letting k be the maximum number of words in a phrase.

3.2 Extracting a User Model

Whatever formalism is used for the user model, the next step is to add meaning to this formalism representative of the user. Assume we have logs where each row contains a user u, session s, query q, set of returned documents for the query D, set of documents selected by the user $B \subseteq D$, set of documents ignored by the user $C \subseteq D$ such that $C \cap B = \emptyset$, and a time t. An ignored document $d \in C$ is usually defined as any not clicked document up to one document below the lowest ranked clicked document. This data is represented as a set of tuples $\{u, s, q, D, B, C, t\}$. For a document with n terms we say $D \ni d = \langle w_0, w_1, ..., w_n \rangle$ where $w_i \in L$ given L is a lexicon of possible terms. Terms used in hyperlinks are often especially significant to the content of a page, a set of integers $\{a_0, a_1, ..., a_m\}$ corresponding to indices in d such that $\forall i, j [0 \leq a_i \leq n \wedge a_i \neq a_j]$ can represent these linked terms. We can use a topic modeling algorithm, such as latent Dirichlet allocation [3], to assign a set of latent topics to documents within our document sets D, B, or C.

We now describe using this data to build a topical user model. Given that clicking on a document indicates that the topics and terms in the document are of interest to the user, and ignoring a document indicates the opposite, we can construct a topic interest vector for her by incrementing the weight of topics in the set B and decrement the weight of topics in the set C. We can consider the topic classification of each document as a point in n-dimensional space, where n is the number of topics. In building the user vector we will be attempting to find a point close to documents in B (selected) and far from those in C (ignored), one solution is to find the centroid of B and C then choose a point weighted close to B's centroid and far from C's.

Given $V(d|u)$ is the quality value of document d for user u, $w(d)$ is importance of document d with respect to user profiling, α is a parameter controlling the influence of ignored documents, and $\mathbf{t}_d$ is a topic vector classifying document d to predetermined topics from an external ontology, we can create a topic vector $\mathbf{t}_u$ for a user u as follows:

$$\mathbf{t}_u = \sum_{d \in B} V(d|u) w(d) \mathbf{t}_d - \alpha \sum_{d \in C} w(d) \mathbf{t}_d \quad (4)$$

Equation 4 is inspired by the personalization strategies in [8] and the diversity algorithm of [1]. Both the importance of a document $w(d)$, and the quality values $V(d|u)$ can be calculated as a function of the document's rank with respect to $|B|$.

An immediate problem is that if B and C are not separable in the topic space Equation 4 will ineffectively distinguish desired from ignored documents. We can analytically determine both whether a hyperplane separates – or how closely it approximates the separation between – the set of clicked and ignored documents, as well as how clustered the clicked and ignored documents are. We would expect that for more separable and more clustered documents our personalized vector will be more accurate.

The weighted centroid method also begs the question: why is a vector the best representation of user preferences? A topic preference vector representation assumes that a single point in topic space is the user's topical ideal and interest will increasingly diminish as a function of distance from this point. The user's topic interest may be better represented as multimodal or through disinterest at a point in topic space and with increasing interest as a function of distance from this point.

As a variation on the single weighted centroid method we can cluster the locations of the documents in B and C in topic space to produce a set of positive vector centroids $\{\mathbf{m}_{B,1}, ..., \mathbf{m}_{B,K}\}$ and negative centroids $\{\mathbf{m}_{C,1}, ..., \mathbf{m}_{C,L}\}$ representative of the user model. The fit of a novel document to the user model can be calculated as its closeness to the positive centroids and distance from the negative centroids. Consider calculating the personalized score of document d based on its topic vector $\mathbf{t}_d$:

$$s_{\mathbf{t}_d} = \sum_{i=1}^{K} (\mathbf{t}_d \cdot \mathbf{m}_{B,i}) w(\mathbf{m}_{B,i}) - \sum_{i=1}^{L} (\mathbf{t}_d \cdot \mathbf{m}_{C,i}) w(\mathbf{m}_{C,i}) \quad (5)$$

where $w(\mathbf{m})$ is the importance of the centroid $\mathbf{m}$, which could be uniform or a function of the number of documents used to calculate the centroid.

We can calculate the score at an even more detailed level by bypassing centroids and simply considering the documents in the user's click history themselves. If the set of centroids from the clicked and ignored documents is replaced with the topics of a set of clicked and ignored documents we can use a form equivalent to Equation 5 for scoring a new document. A computational argument against doing this is that the size of the clicked and ignored documents could be very large and is unbounded where as the number of centroids for these documents is much less than the number of documents and can be arbitrarily bounded. A statistical argument against doing this is that the centroids can be used to compensate for potential over-fitting and the presence of outliers that result from matching directly to each document's topics.

3.2.1 *Contextualizing the User Model*

Because user interests occur along various timelines a topic preference vector constructed from the user's current session may be more relevant than one based on the user's entire browsing history. Let k in $\mathbf{t}_{u,k}$ be an integer specifying how many previous user sessions to consider such that $\mathbf{t}_{u,1}$ looks at only the previous session and, if the total number of sessions is n, $\mathbf{t}_{u,n} = \mathbf{t}_u$. (We could also use the second index to divide up time in an alternative manner, such as into the number of documents clicked, number of queries issued, percent of elapsed time, or amount of elapsed time.) Focusing on changes in information need and ignoring document selection we assume that the user does the following during her searching sessions:

1. formulate an information need
2. submit a query based on this information need
3. repeat step 1 or 2

This type of search behavior should be visible through a topical analysis of the user's query and click history. Assuming different information needs have demonstrably different topic distributions, a marked change in the topic distribution of clicked pages will occur when the user moves from step 3 to step 1 and will not occur when the user moves from step 3 to step 2. Referring back to the topic classification illustration, a shift in information need will be accompanied by a shift of the hyperplane separating clicked and ignored documents, and if our sample of user search history covers multiple information needs we'd expect a degradation in separability and a decrease in clustering of document topics. We can account for this shift in information need by modifying the topic preference vector.

3.3 Re-ranking

Our re-ranking algorithm takes a set of pages returned for a query and assigns scores to them using the user's topic preference vector $\mathbf{t}_u$. Taking the result diversification problem defined in [1] as a starting point, we define the problem of *result personalization* as follows: given a set of results D, returned for a query q, issued by a user u, a probability distribution over topics for the user and query $P(t|q, t_u)$, and the quality values for the documents $V(d|q, t, t_u)$, our goal is to determine a set $S \subseteq D$ such that $|S| = k$ which maximizes

$$P(S|q, \mathbf{t}_u) = \sum_i P(t_i|q, t_{u,i})(1 - \prod_{d \in S}(1 - V(d|q, t_i, t_{u,i}))) \tag{6}$$

the sum of the probability that some documents will satisfy topic i. Assuming $P(t_i|q, t_{u,i}) = t_{u,i}P(t_i|q)$ and $V(d|q, t_i, t_{u,i}) = t_{u,i}V(d|q, t_i)$ we reformulate Equation 6 as:

$$P(S|q, \mathbf{t}_u) = \sum_i t_{u,i}P(t_i|q)(1 - \prod_{d \in S}(1 - t_{u,i}V(d|q, t_i))) \tag{7}$$

If we have multiple user topic vectors representing preferences on different time scales we can compute the usefulness of S through an interpolation over multiple user topic vectors. Consider interpolating over the last session's topics and the global topic vector:

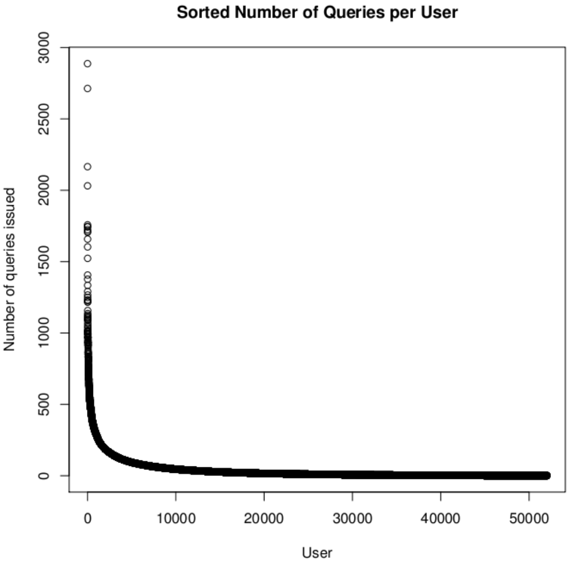

Figure 1: Number of queries per user sorted by number of queries.

$$P(S|q, u) = \alpha P(S|q, \mathbf{t}_{u,1}) + (1 - \alpha)P(S|q, \mathbf{t}_u) \tag{8}$$

This interpolation could be replaced with a product interpolation or an interpolation over a larger set of user topic vectors.

The assumption essential to personalization, and the one which we make in the above equations is that users will prefer documents with topics closer to their past topic preferences. We use a modified version of the IA-SELECT algorithm presented in [1] to create a personalized set of documents S which conforms to the objective defined in Equation 7 or 8. Our version of IA-SELECT calculates the quality value of documents as:

$$V(d|q, t_i) = \beta t_{d,i} + (1 - \beta)t_{d,i}^{1+\log(rank(d))} \tag{9}$$

where $0 \leq \beta \leq 1$ is used to control the amount of diversity with $\beta = 1$ supplying the maximum diversity and $\beta = 0$ the minimum, $t_{d,i}$ is the probability of topic i for document d, and $rank(d)$ is the original rank of document d.

4. EVALUATION

As a preliminary to the evaluation of a full search result re-ranking system we investigate to what extent a user's query logs can inform us of their topic preferences. We examine the AOL search log data[4] to build user topic profiles from a subset of the user's clicked data. We also test that our reordering algorithm reorders based on the user's topic preferences.

4.1 Data

We use collection 01 of the AOL search log data. The data is recorded as tuples of an anonymous ID, query, query

[4] AOL Data `http://www.gregsadetsky.com/aol-data/`

time, item rank, and click URL. We consider only queries for which the user has clicked a result. As can be seen in Figure 1 there is a strong power law relationship between users and the number of queries issued, with a relatively small number of users issuing many queries and most users issuing only a small number of queries.

4.2 Identify User Topics

To explore whether the users' clicked search results provide insight into determinable topic preferences we select a user (anonymous ID 13362448) who has a large number of records (1237) then retrieve the Yahoo! Search titles and summaries of the pages they clicked on. We then remove all strings 2 characters or less and apply a common words stoplist to the text of the query, page title, and page summary and use latent Dirichlet allocation (LDA) to create latent topics from this processed text. We create a user vector by summing per topic over the documents' topics and normalizing. We generate multiple user topic vectors by limiting clicks considered for the user topic vector to a subset of the user's total clicks.

4.3 Potential Reordering

To evaluate the benefits of reordering using the personalization vector we resubmit a user query and reorder the search results based upon the user's topic vector. The anonymous user we have chosen is apparently interested in dog racing, specifically greyhound racing. Many queries concern race results and information about betting on dog races. We resubmit the user's query "hollywood greyhound racing results," which on its original submission received 4 clicks on documents ranked 1, 11, 18, and 29. The low rank of some clicked documents show the potential benefits of re-ranking. Because the returned results will have changed since 2006 we'll be more interested in comparing the current rank versus the reordered rank. We do not diversify the search results and let $\beta = 0$ in Equation 9.

5. RESULTS

5.1 Queries Indicate Topics

To produce a data set that we can easily visualize we performed LDA using 3 latent topics. Figure 2 shows a 3D scatter plot of each query and clicked document pair for a specific user, with a document's location in each dimension corresponding to the weight of the topic represented by that dimension. If the user was uniformly and randomly selecting the documents they visited for their search queries we would not expect a signficant pattern or clustering. This is not the case. The data shows a clustering, such that the dotted linear regression plane can adequately characterize the data. The user generally prefers documents not in latent topic 1 and in either latent topic 3 or latent topic 2 but not both.

After having determined that the user's document pairs form a pattern in latent topic space we now test how well assigning documents to topics is distinguishing amongst them. We compute the specific topic that a document belongs to as the topic for which it has the highest score. Referring to Figure 2 most documents will be assigned to topics 3 or 2. Figure 3 displays the results of a principal components analysis of our user's documents. Red documents belong to topic 1, green documents to topic 2, and blue documents to topic 3. The first panel shows the first principal component

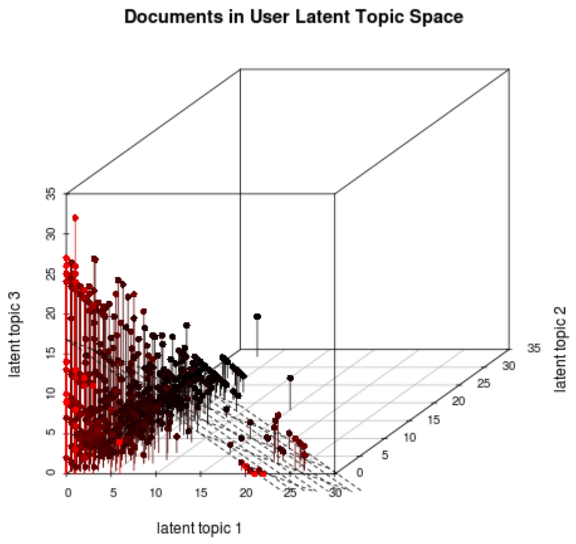

Figure 2: Scatter plot of the location of each query and document the user clicked on for that query in topic space. The coordinates correspond to the number of times words in each document were assigned to each topic. Coloring from brighter red to black indicates a corresponding low to high value in latent topic 2. The dotted plane is a fitted linear regression model.

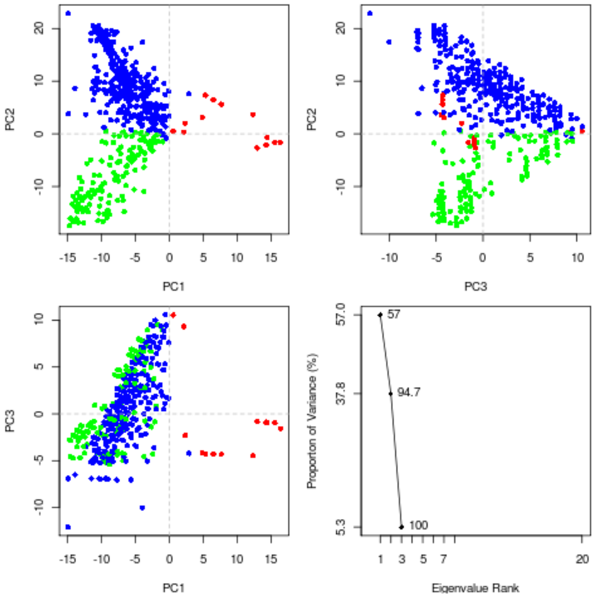

Figure 3: Principal components analysis of the topics assigned to the user's query and document pairs. Each point represents a query and document, the color of the point indicates the topic it has been assigned to.

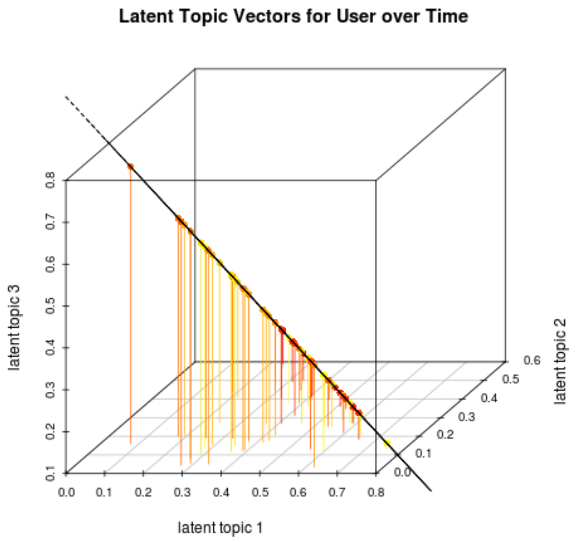

Figure 4: Points indicate the location of the user vector in topic space over time. The color of the point indicates the document click time, points that are redder are earliest in time and as points get yellower and lighter they are later in time.

versus the second principal component and is able to separate the three topics. All of the documents in topic 1 are in the first and fourth quadrant, while nearly all of topic 2's documents are in the third quadrant and all of topic 3's in the second quadrant.

5.2 User Topics over Time

In Figure 4 each point represents the user topic vector for a different set of 20 queries and clicked documents ordered by time. What appears as a black line is a plane based on a linear model fit to the data. As can be seen, all points representing user topic vectors can be fit to a 2D plane. The color of the points indicates their location in time with redder points being earlier and yellower points being later. Earlier points appear to prefer latent topic 1 and dislike latent topic 3, while intermediate points are more neutral and the later points show the opposite.

In Figure 5 we present a graph of the fitted linear model against the residuals of the linear model, this plot is the 2D projection of the points in Figure 4 onto their shared plane. This plot shows some cluster by time period but no clear temporal trend of change in topics. Next we'll look at changes in document subsets over time.

Figure 6 shows four plots each representing the topics of a subset of 300 documents from the user's query logs. The documents' distributions in topic space are all similar but not the same, as best shown by the different linear regression models which fit the data. The adjusted R^2 values of the linear regression models from the top row left to right to the bottom row left to right are $0.4091, 0.655, 0.6996, 0.6249$. The first set of documents is somewhat of an outlier in that it has a low R^2 value. For reference, the R^2 value for the collection as a whole is 0.6054.

5.3 Potential Reordering

In our reordering experiment we found that of the 4 clicked documents for the user's query ”hollywood greyhound racing results” only 2 documents are now on the Yahoo! Search result list. These documents are originally returned at position 12 and 39, after re-ranking they are moved to position 10 and 36, respectively. This is a positive result showing that these clicked documents are ranked higher towards user preferences, as opposed to being ranked the same or lower.

6. DISCUSSION

We have seen that a user's clicked documents are contained within a discernible region in topic space. Given the similarity in data per user we'd expect this result to generalize over the documents clicked by other users, producing for each user a preferred topic distribution. There are many different ways to make use of the user's documents' locations is topic space to form a user model. The approach we use in our experiments is only the simplest method: generate a topic vector for the user based on their documents' topics.

As shown through different means in Figures 4 and 6, the documents which the user is interested in, and the personalized topic vectors that will result from analyzing these documents, change over time. We can image the linear model plane fitted to document topics, represented in various different orientations in Figure 6, as it would be in an animation were each new click of a document slightly moves the plane in a way that, although discrete, will appear as a continuous representation of the evolution of the user's preferences.

In Figure 6 we see a noticeably different representation but still to be investigated is whether the animation just described will result in orientational discontinuities as the user switches from one information need to the next. The sliding

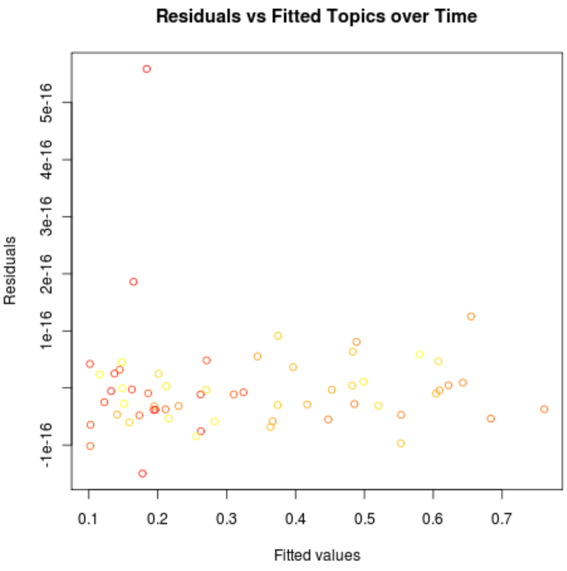

Figure 5: Residuals of fitted linear model for user topics over time. A projection of user topic vectors into the linear model plane. Redder colors occur earlier in time while yellower colors occur later in time

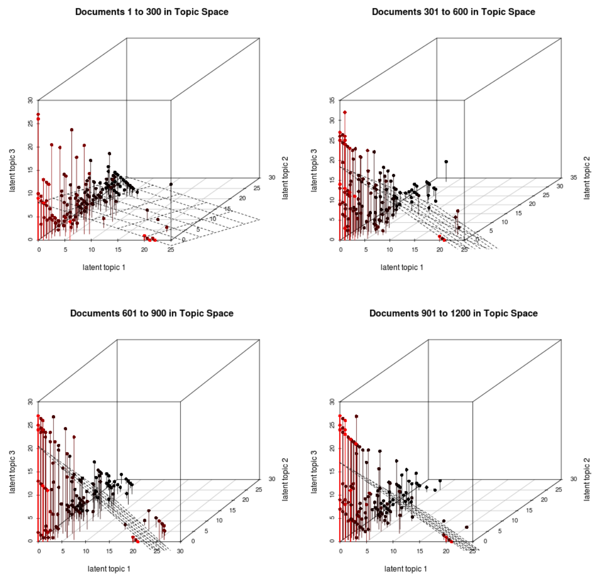

Figure 6: Document distribution in topic space for different subsets of the user's clicked documents. The plane is the best fit linear model for the data.

window of user topic vectors, displayed in Figures 4 and 5, do show a topical "migration" that we expect is connected to changing information needs of the user. The changes shown add to previously seen experimental justifications for personalizing search with respect to context according to session(s) or time scale.

6.1 Reordering

There is too little data to determine what value reordering provides the user with. We stopped short of a more detailed study into reordering based on the AOL data because the analysis of search result orderings from 2006 based on new data from 2010 is unlikely to produce reliable results. The minimal finding, that reordering of results produced a new ordering in line with user preferences, provides motivation for the development of a more complete system to apply to a more complete data set or with which we can perform a user study.

Diversity, in terms of reordering results without respect to rank, was set to zero in the evaluation we ran. This phrasing is somewhat misleading because (1) setting diversity to zero does not result in ignoring diversity, search results are still re-ranked in order to "diversify" with respect to the users topic preferences and (2) diversity in this case is a qualified diversity with respect to the user topic preference, not the traditional meaning of equalizing result distribution amongst potential topics. To address (2) we could use "personalized" but this undermines the point that personalization is performed through a biased diversification. We can ameliorate the confounding of control over diversity and personalization by adding the personalization vector back into the conditional and defining $V(d|q, t_i, t_{u,i})$ in Equation 6 so that the contributions of personalization according to the user topic preference and diversity over the corpus' topics are separately parameterized.

6.2 Negative examples and Multiple Centroids

Because our data did not contain negative examples of user preference – ignored documents – we did not make use of these in building the user model. Model building incorporating negative examples will be left to future work. We note that search query log data is often in the form $\{id, query, time, clicked\ document, clicked\ document\ URL\}$ and therefore retrieving negative examples ranges from challenging to unfeasible, unless data is collected by a system built especially for evaluation purposes or proprietary access is granted.

We did not explore the idea of using clustered document centroids as representative of user preferences. This was inspired partly by the possibility of evaluating negative examples. Additionally, the documents don't appear to form clusters in Figure 2. Analysis in higher dimensions, with different users, or using different topic modeling methods may lead to clusters; using multiple centroid models will be left to future work.

7. CONCLUSIONS

Based on previous work in search personalization we have successfully used search logs to build a user model by extracting latent topic distributions from clicked documents and summing over these latent topic distributions. Because we only use three topics the experiments let us visualize the user's topic preference in these low dimensions and provide evidence that extracting topics in higher dimensions will produce greater insights into user preferences for queries or document lists. The experiments were preliminary in nature and provide justification for future work with more extensive data or a more complete system in which a higher dimensional topic space is used and perhaps topic labels are generated (e.g. by using the *learning to cluster* method in [26]).

7.1 Future Algorithms

Our modification to the IA-SELECT algorithm changed only the scoring measure and did not address the underlying assumptions made by that algorithm. Specifically, it maintains its greedy nature, which is justifiable for a one-time non-interactive reordering but less justifiable in an interactive setting were we expect query reformulation or facet navigation to occur. Future work will explore the interactive setting and optimizing the information gain produced by the reordered set of search results. One approach would be to focus on increasing topic diversity regardless of user preferences until enough click data concerning user topic preferences is collected. This could be implementing using the full conditional of the document quality measure as mentioned in Section 6.1.

The link analysis aspect of personalization, as in Topic-Sensitive PageRank and other web-graph-based algorithms, is not addressed by our research. It is commonly accepted that the random surfer model, upon which PageRank, Personalized PageRank, and Topic-Sensitive PageRank are based, is false. The user does not choose the links she takes from a uniform distribution but instead based on their context and her topic preferences. In [2] the authors replace the random jump by one based on textual analysis of the web page. In future work we will extend this further and replace a jump based on textual analysis with one based on this analysis relative to estimated user preferences.

7.2 The Semantic Graph

In our method we chose a topical user model over an ontological user model not based on merit but based on convenience, especially as related to integration with previous work in diversity. In future work that personalizes based upon the web graph a user model formed from a semantic network may be the preferred structure if computations between the two graphs are easier or more efficient to perform than computations between a graph and a vector. The semantic graph would also provide a straightforward method for increasing the weight of hyperlinked terms, another factor we have not dealt with. Lastly, it would be interesting to pursue a method using the semantic graph for solely topic diversification, a so far unexplored approach.

References

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *WSDM '09*, 2009.

[2] Yinghui Xu andKyoji Umemura. Literal-matching-biased link analysis. In *AIRS*, pages 153–164, 2004.

[3] D. Blei, A Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[4] S. Brin and L. Page. The anatomy of a large-scale hypertextual (web) search engine. *Computer Networks*, 30:107–117, 1998.

[5] D. Carmel, N. Zwerdling, I. Guy, S. Ofek-Koifman, N. Har'el, I. Ronen, E. Uziel, S. Yogev, and S. Chernov. Personalized social search based on the userÕs social network. In *CIKM '09*, 2009.

[6] C. Clark, M. Kolla, G. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *SIGIR '08*, pages 659–666, 2008.

[7] Z. Dou, K. Chen, R. Song, Y. Ma, S. Shi, and J. Wen. Microsoft research asia at the web track of trec 2009. In *Proceeding of TREC 2009*, 2009.

[8] Z. Dou, R. Song, and A Wen, J. Large-scale evaluation and analysis of personalized search strategies. In *WWW 2007*, 2007.

[9] P. Ferragina and A. Gulli. A personalized search engine based on web-snippet hierarchical clustering. In *WWW2005*, 2005.

[10] N. Fuhr. A probability ranking principle for interactive information retrieval. *Information Retrieval*, 2008.

[11] D. Gleich and M. Polito. Approximating personalized pagerank with minimal use of web graph data. *Internet Mathematics*, 3(3):257–294, 2007.

[12] V. Hollink, T. Tsikrika, and A. de Vries. Semantic vs term-based query modification analysis. In *Proceedings of the tenth Dutch-Belgian Information Retrieval Workshop*, 2010.

[13] V. Hollink and M. van Someren. Optimal link categorization for minimal retrieval effort. In *Proceedings of 6th Dutch-Belgian Information Retrieval Workshop*, 2006.

[14] Thorsten Joachims. Optimizing search engines using clickthrough data. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142, New York, NY, USA, 2002. ACM.

[15] C. Kohlschütter, P. Chirita, and W. Nejdl. Using link analysis to identify aspects in faceted web search. In *SIGIR '06*, 2006.

[16] A. Micarelli, F. Gasparetti, F. Sciarrone, and S. Gauch. Personalized search on the world wide web. *The adaptive web: methods and strategies of web personalization*, pages 195–230, 2007.

[17] F. Qiu and J. Cho. Automatic identification of user interest for personalized search. In *WWW2006*, 2006.

[18] F. Radlinksi and S. Dumais. Improving personalized web search using result diversification. In *SIGIR '06*, 2006.

[19] K. Sia, S. Zhu, Y. Chi, K. Hino, and B. Tseng. Capturing user interests by both exploitation and exploration. Technical report, 2006.

[20] A. Sieg, B. Mobasher, and R. Burke. Web search personalization with ontological user profiles. In *CIKM '07*, 2007.

[21] Sofia Stamou and Alexandros Ntoulas. Search personalization through query and page topical analysis. *User Modeling and User-Adapted Interaction*, 19(1-2):5–33, 2009.

[22] M. van Someren, S. Hagen, and V. Hollink. Greedy recommending is not always optimal. *Lecture Notes in Artificial Intelligence 3209*, pages 148–163, 2004.

[23] R. White, P. Bailey, and L. Chen. Predicting user interests from contextual information. In *SIGIR '09*, 2009.

[24] M. Wilson, M. Schraefel, and R. White. Evaluating advanced search interfaces using established information-seeking models. In *SIGIR '07*, 2007.

[25] B. Xiang, D. Jiang, J. Pei, X. Sun, E. Chen, and H. Li. Context-aware ranking in web search. In *SIGIR '10*, 2010.

[26] H. Zeng, Q. He, Z. Chen, W. Ma, and J. Ma. Learning to cluster web search results. In *SIGIR '04*, 2004.