Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2020) 28: 1631 – 1643

doi:10.3906/elk-1909-9

Selective personalization and group profiles for improved web search personalization

Samira KARIMI MANSOUB[1,*], **Gönenç ERCAN**[2], **İlyas ÇİÇEKLİ**[1]
[1]Department of Computer Engineering, Faculty of Engineering, Hacettepe University, Ankara, Turkey
[2]Institute of Informatics, Hacettepe University, Ankara, Turkey

Received: 02.09.2019 • **Accepted/Published Online:** 13.01.2020 • **Final Version:** 08.05.2020

Abstract: Personalization is a common technique used in Web search engines to improve the effectiveness of retrieval. While personalizing some queries yields significant improvements in user experience by providing a ranking in line with the user preferences, it fails to improve or even degrades the effectiveness for less ambiguous queries. A potential personalization metric could improve search engines by selectively applying personalization. One such measure, click entropy uses the query history and the clicked documents for the query, which might be sparse for some queries. In this article, the topic entropy measure is improved by integrating the user distribution into the metric, robust to the sparsity problem. Furthermore, a topic model-based ranking for the personalization method is proposed using grouped user profiles. Experiments reveal that the proposed potential prediction method correlates with human query ambiguity judgments and the group profile-based ranking method improves the mean reciprocal rank by 8%.

Key words: Personalized web search, topical user model, latent Dirichlet allocation

1. Introduction

Personalizing web searches by reranking the retrieved documents concerning a user's interests is adopted by many search engines today. Personalization of broad and ambiguous queries yields a better user experience. For instance, for the query "*test*", if the user issuing the query is a medical professional, results relevant to medical tests should be preferred over tests for evaluating students in educational institutes. On the other hand, for other queries with a more clear and specific meaning, ranking methods without any personalization are more effective [1]. A measure able to estimate the potential for personalization can enable selective application of personalization and improve the overall effectiveness of the search system.

Different measures are used to determine the potential for personalization of queries [2, 3]. Click entropy, measured using the query history and documents clicked by the users, is one such measure [2]. This method was recently improved by a topic model-based extension [3] and is referred to as topic entropy. In this article, we improve topic entropy by measuring how each user's topical user profile differentiates from the query words' topics. Using the topic distributions of clicked documents for each user as a feature, the potential for personalization is modeled on a fine-grained level. Through experiments, we show that the proposed method can process queries without any history and is more effective for queries with low frequencies. This allows the system to alleviate the cold-start problem and allows determining the ambiguity of new queries.

Topic models are also used for reranking the search results [4, 5]. User profiles can suffer from data sparsity problems if a user does not have sufficient history. To resolve this problem, the users are first grouped

*Correspondence: samirakarimi@hacettepe.edu.tr

using the latent topics modeled using latent Dirichlet allocation (LDA). Our proposed ranking model reranks the search results with respect to user group profiles instead of individual user profiles. The proposed group profiles improve the retrieval effectiveness when using both long-term and short-term query histories.

In summary, this article proposes two novel algorithms concerning two subtasks of Web search personalization. The first algorithm focuses on estimating the potential for personalization using our proposed metric, which we will refer to as unified topic entropy. The second is a reranking method for personalization using grouped topic profiles. When these two contributions are used in combination, a clear improvement over the baseline methods is achieved.

The rest of the paper is organized as follows. Section 2 discusses the related work on personalization potential measures and personalized search approaches. We present our new metric of unified topic entropy and its relation with other potential personalization metrics in Section 3. Section 4 presents our new proposed personalization search method based on grouped user profiles. The evaluation methodology is given in Section 5 and evaluation results are presented in Section 6. Section 7 contains the concluding remarks.

2. Related work

Estimating the potential for personalization and ranking for personalization are two related tasks that can be employed simultaneously in a search engine. This section presents the relevant work for these two tasks in separate subsections.

2.1. Potential for personalization

A search engine typically has access to only the query issued by the user. Given this query, predicting its potential for personalization is an important task with a direct impact on the final effectiveness of the system. Teevan et al. [6, 7] evaluated different metrics to predict the ambiguity of a query and its potential for personalization. They evaluated intrinsic features like query length and click entropy introduced by Duo et al. [2]; clarity measure, which compares the language model of the retrieved result set to a background language model [8]; and result entropy for predicting the potential for personalization. Wang et al. [9] proposed user entropy, which averages the click entropy by each user, and discussed that user entropy is useful for low-frequency queries. They reported click entropy as a reliable method for predicting the potential when a history for the query is available. Click-entropy models the ambiguity using only the user interactions, ignoring the contents of the documents.

Instead of just relying on the click information, augmentation of click-entropy with the content of the documents was also investigated [3, 10]. Song et al. [10] discussed the relationship between query ambiguity and topic distributions. They used the latent topic model variable to model the clicked documents' content and improve the click-entropy model for predicting the ambiguity of queries. The topic model-based approach proposed in this research is motivated in a similar way but extends the model proposed by Yano et al. [3] so that the newly proposed metric can handle new queries and perform better for low-frequency queries.

2.2. Personalized search

As opposed to estimating the potential for personalization, a large body of research exists for the personalization process. A personalization system first models the user profile and reranks the results using this profile. In the process of personalization, user profiles are modeled by user behavior, content, and the context of users. A natural source for building a user profile is the user's browsing history. Matthijs and Radlinski [11] used the words in titles, full text, and metadata of browsed web pages to construct a user profile composed of

terms. External sources like the Open Directory Project (ODP) are also used as external knowledge sources for modeling user profiles [12–14].

Similar to our method, topic model-based personalization methods exist [4, 15, 25]. Harvey et al. [4] used latent Dirichlet allocation (LDA) and built latent topic models to represent the document sets. The users are modeled by the topic distributions of the documents that they have clicked on. Vu et al. [25] used a time-aware topic model for personalization with the motivation of capturing the dynamic nature of users' interests. Since user interests and search intentions are changing during a search session, long-term and short-term profiles were also discussed in some papers such as [16, 17]. Vu et al. [16] created a temporal user profile using the user's clicked documents and used these profiles for ranking the results. Bennet et al. [17] split the user profile into three based on different temporal periods and built a long-term profile, a daily profile, and a session profile. In their experiments, they showed that using these profiles is more effective than click entropy and query position in a search session.

Probabilistic topic models are also used for personalization [18]. Some authors used pLSI [19] and Kullback–Leibler divergence to estimate a query model. In a similar method [20], a text similarity algorithm using latent Dirichlet allocation was proposed for personalization. The authors used topic model and word cooccurrence analysis to calculate topics in the text. More recently, topic models were used for query suggestions [21]. The authors used topic models to model the semantic relationships on an AOL query log. They reported unseen queries as an important shortcoming for their method. Amer et al. [22] used word embeddings as opposed to topic models for user profiles; however, their model failed to improve search effectiveness. On the other hand, Vu et al. [23] learned LDA-based vector embeddings for personalization and achieved clear improvement.

3. Predicting potential for personalization

Personalization is not appropriate for all user queries and may even yield worse results than generic ranking methods. The ranking for a navigational, specific, and unambiguous query is usually stable and its ranking does not depend on user preferences. Better rankings can be obtained for those queries without personalization. For example, the query "myspace" is usually a navigational query for the social networking website regardless of the user issuing this query. For such a query, trying to personalize can produce an inferior ranking. Although there are some metrics such as click entropy and topic entropy to identify queries as to whether they require personalization, they have limitations, especially for queries without history. To overcome these limitations, we present a new metric called unified topic entropy that estimates the potential for personalization using topic distributions of individual query words.

First, we summarize click entropy and topic entropy metrics since we compare their performances with the performance of our new metric called unified topic entropy. Click entropy measures the query's personalization potential using the clicked documents for the same query. If the click entropy for a query is high, it means that different users click on different documents and the query is ambiguous. Click entropy [7] is defined in Equation 1 as the entropy of the documents' click probability distribution for the query:

$$ClickEntropy(q, D_q) = \sum_{d \in D_q} -P(d|q)log(P(d|q)), \tag{1}$$

where D_q is the set of documents clicked for query q and $P(d|q)$ is the number of clicks for a document d divided by the total number of clicks for query q. For an unambiguous query, relevant documents are clicked on with a higher frequency by different users, creating a probability distribution with less uncertainty.

As can be seen from Equation 1, click entropy is purely based on documents and not their contents. When different documents with similar contents are clicked on by users for a query q, click entropy will be high, signaling a false ambiguous query. Topic entropy [3], on the other hand, models $P(d|q)$ using the topic model distribution of the documents, able to account for documents with similar contents:

$$TopicEntropy(q, D_q) = \sum_{d \in D_q} P(d|q) KL(P(z|d)||P(z|q)) \tag{2}$$

$$= \sum_{d \in D_q} P(d|q) \sum_{z \in Z} P(z|d) log(\frac{P(z|d)}{P(z|q)}), \tag{3}$$

where $P(z|d)$ is the probability of the topic z for the given document d. The topic set Z is obtained using latent Dirichlet allocation (LDA). $P(z|q)$ is the probability of the topic z for the given query q and it is estimated using the documents clicked for a query q as in Equation 4:

$$P(z|q) = \sum_{d \in D_q} P(z|d) P(d|q). \tag{4}$$

Topic entropy is the weighted sum of Kullback–Leibler divergences of query and document topic distributions and Yano et al. [3] modeled the topic entropy as the center of gravity for the topic distribution divergences. While this measure incorporates document similarities, the users' behavioral differences are only modeled through the $P(d|q)$ component. Topic entropy is still not defined (its value is zero) for the new queries, the same as click entropy. We try to address these problems with our new metric. Although Yano et al. [3] also proposed topic user entropy (TUE) as in Equation 6 to incorporate the users' behavioral differences, in their experiments the correlation of topic user entropy results with human judgments is low compared to topic entropy.

$$TUE(q, U_q, D_{u,q}) = \sum_{u \in U_q} \frac{1}{|U_q|} \sum_{d \in D_{u,q}} P(d|u,q) KL(P(z|d)||P(z|q)) \tag{5}$$

$$= \sum_{u \in U_q} \frac{1}{|U_q|} \sum_{d \in D_{u,q}} P(d|u,q) \sum_{z \in Z} P(z|d) log(\frac{P(z|d)}{P(z|q)}), \tag{6}$$

where $D_{u,q}$ represents the documents clicked on by user u for query q, and U_q is the user set issuing query q. It is assumed that the probability of each user issuing the query is equally likely.

Note that TUE weights the divergence of the document model from the query model by $P(d|u,q)$, which is the number of times user u clicks on document d for query q divided by the total number of clicks of u for q. For a user who did not issue q previously, TUE is not defined since no document is clicked. In order to solve this cold-start problem, we try to benefit from extracted topics of topical user model $P(u|q)$ in our new metric. We define $P(u|q)$ in Equation 8 and it is the probability distribution of the query on the users using the LDA topic model:

$$P(u|q) \propto P(u)P(q|u) = P(u) \prod_{w \in q} P(w|u) \tag{7}$$

$$= P(u) \prod_{w \in q} \sum_{z \in Z} P(w|z)P(z|u), \tag{8}$$

where $P(u)$ is the probability of user u and it is estimated by the proportion of queries submitted by user u to the total number of queries. $P(w|z)$ is the probability of the word w of the query for topic z and $P(z|u)$ is the probability of the topic z for the given user u. $P(z|u)$ is also estimated using all documents D_u clicked by user u as in Equation 9, and it is used to weight the contribution of each topic for the query:

$$P(z|u) = \sum_{d \in D_u} P(z|d)P(d|u). \tag{9}$$

Using $P(u|q)$ as the weighting factor instead of $P(d|q,u)$, we define our new metric, called unified topic user entropy (UTUE), as in Equation 10. This metric unifies all users who have or have not issued the query in the past:

$$UTUE(q, U_q, D_u) = \frac{1}{|U_q|} \sum_{u \in U_q} P(u) \sum_{d \in D_u} \prod_{w \in q} \sum_{z \in Z} P(z|u)P(w|z)P(z|d) log(\frac{P(z|d)}{P(z|w)}). \tag{10}$$

As a new query will only be submitted by a single user and will not have any clicked documents, $D_{u,q}$ will be an empty set. As a result, $TUE(q, U_q, D_{u,q})$ will be equal to zero. Instead of depending on the clicked documents for the specific query q, the documents clicked on by user D_u for all queries are used to compare the user profile with the query. Furthermore, instead of using $P(z|q)$, which depends on the clicked document set for query q, the topic distribution of words in the query is used. With these two approximations, the proposed method can estimate the potential for personalization for a query without any history.

4. Ranking process for personalized search

Personalization is the task of reranking the retrieved document set concerning the user profile. The list of documents produced by the search engine for the query is reordered using the user profile. While this task on its own is independent of the potential for personalization tasks, we argue that selective reranking can yield better results. Three ranking methods are used in our evaluations, where the first one uses a generic document scoring function based on topic models without any personalization. The second model uses the personalization factor for the user profile built using the documents clicked on by the user. Although user profiles are indicative of the user's interests, they can be incomplete and misleading due to data sparsity. For user profiles with less browsing history, the data available might not be sufficient. To resolve the data sparsity, the history of similar users can be grouped and used for personalization. As a final ranking model, we propose a new group-based personalization method.

4.1. Generic ranking without personalization

Given the LDA topic models, documents and words are associated with topics in the document set. Building on the same framework introduced by Harvey et al. [4], documents are ranked with respect to the LDA model

$P(d|q)$ called the NonPTM (nonpersonalized topic model) here. The $P(d|q)$ is estimated using the Bayes rule and the LDA generative model as follows:

$$NonPTM(d,q) = P(d|q) \propto P(d)P(q|d) = P(d)\prod_{w\in q} P(w|d) = P(d)\prod_{w\in q}\sum_{z} P(w|z)P(z|d), \tag{11}$$

where $P(d)$ is the prior document probability and z is the topic latent variable estimated using LDA. $P(w|z)$ and $P(z|d)$ are obtained from the LDA topic model. Since NonPTM is a method without any personalization, comparisons with this baseline method will reveal the improvement of personalization over generic ranking with topic models.

4.2. User profile-based personalization

The personalization method of Vu et al. [15, 25] is reproduced for completeness. A user's topical profile is modeled by the set of documents D_u that the user clicked on. Using the topic distributions of the user's documents that are associated with topics, the user profile can be considered as the vector of posterior probabilities of topics given the user. The user profile is calculated as in Equation 12:

$$P(u|z) \propto P(u)P(z|u) = P(u)\frac{1}{K}\sum_{d_i\in D_u} \alpha^{t_{d_i}-1}P(z|d_i). \tag{12}$$

Similar to the approach of Vu et al. [25], an exponentially decaying function is used in the user profile to give more weight to recently clicked documents and t_{d_i} is equal to 1 for the most recent relevant document for the exponential decay function penalizing older clicks. The accumulated evidence is transformed into a probability using the K normalization function calculated as the sum of document biases $K = \sum_{d_i} \alpha^{t_{d_i}-1}$ and the α parameter of the decaying function is set to 0.95, the same as in the work of Vu et al. [25]. Then the personalization-based ranking function is defined as in Equation 13, which will be referred to as the personalized topic model (PTM):

$$PTM(d,q,u) = P(d|q,u) \propto P(d)\prod_{w\in q} P(w,u|d) = P(d)\prod_{w\in q}\sum_{z} P(w|z)P(u|z)^{\lambda}p(z|d). \tag{13}$$

The λ parameter weighs the effect of the user's topical profile on the ranking process and it is equal to 0.175, similar to Harvey at al. [4].

4.3. Group profile-based personalization

The function $PTM(d,q,u)$, which is reproduced from Vu et al. [1], depends on users and their topic distributions estimated using the documents clicked on by the users. One disadvantage of $PTM(d,q,u)$ is that it is built considering only the documents that user u has clicked on and the set of documents clicked on might be sparse for some users. Data sparsity can be resolved by backing off to the group of users with behavior similar to that of user u. We propose a group profile-based personalization method, which first groups users with respect to their topic distributions, and we use group profiles in the ranking process.

Users are clustered using their $P(z|u)$ topic probability distributions. Topic probability distributions depend on the documents that are clicked on by users and they are estimated with Equation 9. The K-means

clustering algorithm is used to partition the users into $|C|$ groups. The number of clusters $|C|$ is a parameter and C_u is the cluster of user u. Our proposed group profile-based personalization method called the grouped personalized topic model (GPTM) determines the ranking score with respect to the group profiles and it is defined as in Equation 15:

$$GPTM(d,q,u) = P(d|q,C_u) \propto P(d) \prod_{w \in q} P(w,C_u|d) \tag{14}$$

$$= P(d) \prod_{w \in q} \sum_{z} P(w|z) P(C_u|z)^{\lambda} P(z|d). \tag{15}$$

Equation 15 generalizes the user ranking to the clusters of users, resolving the sparsity problem. The λ parameter weighs the effect of group profile on the ranking process and group profiles are computed as follows:

$$P(C_u|z) \propto P(C_u)P(z|C_u) = P(C_u)\frac{1}{K} \sum_{d_i \in D_{C_u}} \alpha^{t_{d_i}-1} P(z|d_i). \tag{16}$$

The computation of group profiles is similar to the computation of user profiles except that it depends on the documents that are clicked on by all users in the cluster.

5. Evaluation methodology

5.1. Datasets

In order to investigate the effectiveness of the proposed methods, a dataset of web search engine logs is used. To the best of our knowledge, the AOL Query Log dataset is the only available dataset with anonymized user information. The log contains three-month query logs starting from March 2006 and it contains 657,426 anonymous users. As done by Harvey at al. [4], we cleaned the dataset by only retaining queries that resulted in a click on a URL. Then the data are filtered by removing URLs clicked on less than 100 times.

As a second dataset, the TREC 2014 Session Track[1] data are used for the experiments. Session Track consists of 1021 query sessions for 60 different topics along with the clicked documents and user ids. The URLs are manually annotated by judges for the topics as spam (–2), not relevant (0), relevant (1), highly relevant (2), key (3), and navigational (4). We use the content of the clicked URL to create topic models of user profiles. The extracted dataset is shown in Table 1. To evaluate the personalized model, we divided the dataset into 95% for training and the last 5% of queries for testing.

Table 1. Extracted data from AOL and TREC 2014 Session data set for experimentation.

	#Queries	#Users	#URLs
AOL Dataset	1,452,012	4,217	11,209
TREC 2014 Session Dataset	2550	148	1097

5.2. Evaluation metrics

The potential for the personalization metric is evaluated using a methodology similar to that of Yano et al. [3]. The correlations between human judgments for query ambiguity and the automatic measures are reported.

[1]TREC 2014 Session Track (2014). TREC Session Dataset [online]. Website https://trec.nist.gov/data/session2014.html [accessed 23 Feb 2017].

First, for three different frequency levels, queries are randomly sampled and 200 queries are selected for each frequency level. A total of 600 queries are annotated by human judges as "clear", "broad", and "ambiguous". Five human annotators are used for the annotations and the interrater agreement is estimated using Fleiss' kappa as 0.436.

For evaluation, labeled queries are assigned weights; for ambiguous, broad, and clear the weights are defined as 2, 1, and 0 values. Then, for each query label, scores are calculated as the sum of human-assigned labels. The rank correlations between the human scores and the potential personalization metrics are calculated using Kendall's τ.

The personalization is evaluated using success at rank k ($S@k$), the mean reciprocal rank (MRR) up to rank 10, and normalized discounted cumulative gain (nDCG@k). Success at rank k is the proportion of recommended items in the top-k (here $k = 1, 10$) set that are relevant. MRR is calculated as in Equation 17:

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i}, \tag{17}$$

where Q denotes the number of queries and $rank_i$ is the rank of document d for query q obtained from the ranking model. $NormalizedDCG$ is a measure of ranking quality discussed in [24] and measures the usefulness, or gain, of a document based on its position in the result list. Queries are sorted according to the potential personalization metrics and personalization is selectively applied to queries above a threshold.

5.3. Number of topics

The number of topics used for LDA is an important parameter. The relationship between MRR and this parameter is investigated in a small development set. Parameters of the LDA model are trained using the training corpus.[2] Figure 1 shows the MRR for different topic numbers ranging from 10 topics to 100. The results indicate that using 40 topics yields the best results.

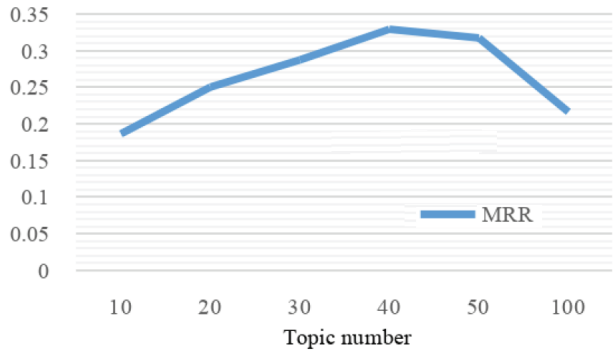

Figure 1. The changes in MRR with different topic numbers using the LDA model.

6. Evaluation results

Using the evaluation methodology defined in Section 5, the performances of selective personalization and group profiles are evaluated. This section is organized to highlight the key findings for these metrics and compare the performance of the proposed method to that of state-of-the-art algorithms.

[2]GENSIM (2019). Gensim library for the LDA estimation [online]. Website https://radimrehurek.com/gensim/ [accessed 01 Nov 2019].

6.1. Quantifying query ambiguity

The performance of the three potential personalization metrics is investigated using the 600-query ambiguity dataset built in this research. Table 2 shows the correlation between human judgments and three metrics, namely click entropy [2], topic entropy [3], and the proposed UTUE metric. The results are grouped into three, highlighting the effectiveness of the metrics for each frequency category. There are 200 queries at each frequency level. The low-frequency group is also separated into three subgroups to reveal the performance gains of the proposed UTUE metric for low frequencies. If a query's frequency is equal to one, it means that the query is new and its click entropy and topic entropy are equal to zero. Furthermore, 34% of the queries are new without any history where only UTUE can be used for calculating the potential for personalization.

The results in Table 2 show that UTUE outperforms the topic entropy and the click entropy for the queries without history and the first column indicates that the improvement is doubled for those queries. UTUE's performance for the queries whose frequency is less than 10 is given in the second column, where UTUE outperforms the other metrics. However, for queries with frequencies higher than 10, topic entropy and click entropy is able to estimate the potential more accurately and it outperforms the proposed UTUE. Figure 2 shows the plot of frequency and rank correlation for both topic entropy and UTUE. While topic entropy is more accurate for identifying query ambiguity at higher frequencies, UTUE is more effective for lower frequencies. This means that the topic entropy measure can successfully determine the ambiguity of a query if there are enough previous clicks for that query, but UTUE performs better for low-frequency queries as it tries to determine the ambiguity of a query from the clicked documents using the individual words forming the query. This result confirms our intuition that UTUE can be used for queries where the other metrics fall short, in queries without history and for low-frequency queries, and topic entropy should be preferred for the other queries.

Table 2. Kendall's τ between methods and ambiguity levels at different query frequencies.

	$\tau for LowFreq$			$\tau for MidFreq$	$\tau for HighFreq$
	$F = 1$	$2 \leq F < 10$	$10 \leq F < 50$	$50 \leq F < 150$	$150 \leq F < 400$
Click entropy	0.149	0.181	0.252	0.226	0.214
Topic entropy	0.149	0.268	0.340	0.301	0.283
UTUE	0.297	0.273	0.182	0.207	0.185

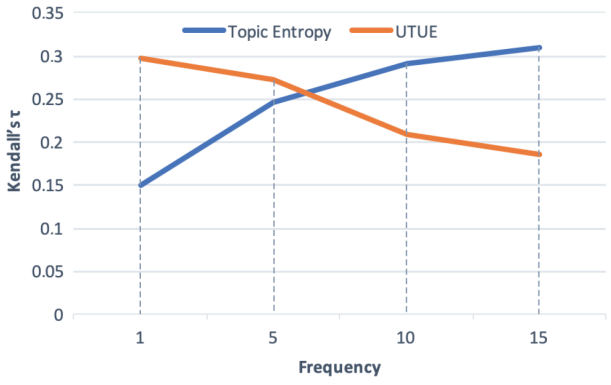

Figure 2. The changes in Kendall's τ in topic entropy and UTUE metric for low-frequency queries.

6.2. Effect of selective personalization

In order to investigate the importance of selective personalization, different potential personalization metrics are used to predict the query's potential and they are normalized using the maximum value. Then, for a threshold ξ, if the potential is below this value it is ranked with the topic model-based ranking algorithm $NonPTM(d,q)$; otherwise, it is ranked with personalized $PTM(d,q,u)$. A more accurate personalization metric is expected to yield better performance gains with selective personalization as it can identify queries more suitable for personalization.

Tables 3 and 4 report the MRR, S@1, S@10, and nDCG@10 scores for the three potential personalization metrics in the AOL and Session Track 2014 datasets. The first row represents the ranking score when using only $NonPTM(d,q)$, which is no personalization. The last row shows the result when all queries are reranked using $PTM(d,q,u)$. Naturally, these two cases are independent of the potential metric used and are common for all three metrics. When we consider the results of UTUE, it is evident that it achieves a higher score for all different thresholds. This indicates that it assigns a more accurate prediction for personalization, and the queries with lower UTUE scores do not benefit from personalization. A similar result is observed between topic entropy and click entropy, confirming the experiments of Yano et al. [3]. Topic entropy performs better than click entropy.

The results of UTUE for $\xi > 0.6$ reveal the highest ranking scores for all measures. This indicates that using personalization only for queries with potential higher than 0.6 is a better strategy than using other thresholds. When considering the difference between applying personalization to all of the queries and selective personalization with $\xi > 0.6$, the performance gain for MRR is as high as 0.264 in the AOL dataset and 0.228 in TREC 2014.

Table 3. Comparison of selective personalization using different potential personalization metrics in the AOL dataset. Personalization PTM is applied only to queries with potential $> \xi$.

ξ	Click entropy			Topic entropy			UTUE		
	S@1	S@10	MRR	S@1	S@10	MRR	S@1	S@10	MRR
None	0.205	0.371	0.267	0.205	0.371	0.267	0.205	0.371	0.267
$\xi > 0.8$	0.331	0.477	0.382	0.328	0.496	0.385	0.384	0.560	0.445
$\xi > 0.6$	0.353	0.498	0.416	0.413	0.572	0.481	**0.462**	**0.620**	**0.536**
$\xi > 0.4$	0.309	0.454	0.354	0.359	0.542	0.420	0.408	0.601	0.478
$\xi > 0.2$	0.227	0.413	0.298	0.298	0.482	0.378	0.357	0.542	0.431
All	0.206	0.387	0.272	0.206	0.387	0.272	0.206	0.387	0.272

6.3. User topical profile versus group topical profile

In order to test our hypothesis that group profiles resolve the sparsity problem, we compare the selective personalization effectiveness of user-based $PTM(d,q,u)$ and group-based $GPTM(d,q,u)$. Using the threshold $\xi = 0.6$, the two ranking methods are compared. An important parameter defined by Vu et al. [16] is the temporal decaying model for the documents clicked on by the users. In order to take into account the difference between a user or group profile formed by short- and long-term user interactions, two separate experiments are performed. The first short-term experiment uses the document clicks performed in a month by the users, while the long-term uses all the clicked documents. To investigate the relationship between the number of clusters

Table 4. Comparison of selective personalization using different potential personalization metrics in Session TREC2014 dataset.

ξ	Click entropy				Topic entropy				UTUE			
	S@1	S@10	MRR	nDCG@10	S@1	S@10	MRR	nDCG@10	S@1	S@10	MRR	nDCG@10
None	0.163	0.328	0.231	0.259	0.163	0.328	0.231	0.259	0.163	0.328	0.231	0.259
$\xi > 0.8$	0.268	0.418	0.327	0.340	0.275	0.461	0.357	0.403	0.322	0.517	0.426	0.452
$\xi > 0.6$	0.307	0.465	0.372	0.389	0.376	0.539	0.440	0.491	**0.419**	**0.588**	**0.493**	**0.527**
$\xi > 0.4$	0.272	0.420	0.339	0.351	0.332	0.507	0.384	0.438	0.356	0.540	0.442	0.469
$\xi > 0.2$	0.203	0.383	0.285	0.302	0.256	0.451	0.359	0.407	0.304	0.507	0.418	0.440
All	0.171	0.364	0.265	0.298	0.171	0.364	0.265	0.298	0.171	0.364	0.265	0.298

and MRR, different numbers of clusters are evaluated ($k \in \{10, 20, 30, 40, 50, 100\}$) in Figure 3. As shown in Figure 3, the best result is obtained with $k = 30$.

Figure 4 shows the MRR-based performance comparison for the two models PTM and GPTM. As can be observed, using group profiles improves MRR by 0.09 when using long-term user profiles while short-term user profiles are improved by 0.08. Thus, using group profiles instead of user profiles improves both cases. As expected, the short-term user profile is more effective than the long-term.

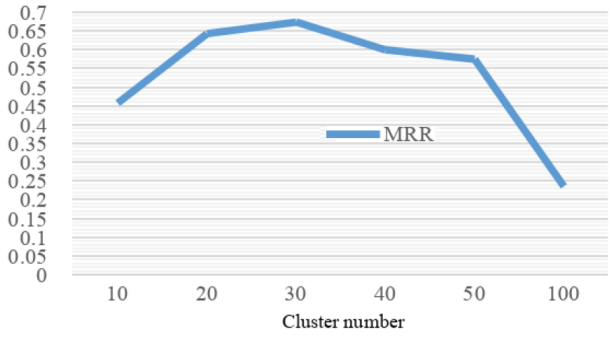

Figure 3. The changes in MRR in GPTM model in different cluster numbers.

7. Conclusions and future work

In this paper, a selective personalization strategy is proposed. As one important stumbling block, the potential for personalization of new or low-frequency terms is not handled by the state-of-the-art metrics. For this reason, we propose a new metric able to handle such new queries, which make up an important portion of the queries in a search engine's log. When evaluated for the task of characterizing queries with respect to ambiguity, the proposed metric improved the state-of-the-art for such new and low-frequency queries.

When compared to the method proposed by Yano et al. [3], our proposed potential personalization metric is defined in terms of the latent topic models rather than relying solely on the query history directly. This allows the UTUE to generalize better to rare queries as well as new queries that are not issued previously as it is. Using the topic models, these queries are modeled using similar queries in a more flexible way.

Furthermore, we show that selective personalization using a combination of UTUE and topic entropy improves personalization effectiveness. Handling low-frequency queries with UTUE and reverting to topic entropy for the other queries, a better selective personalization strategy is proposed. Our results indicate a

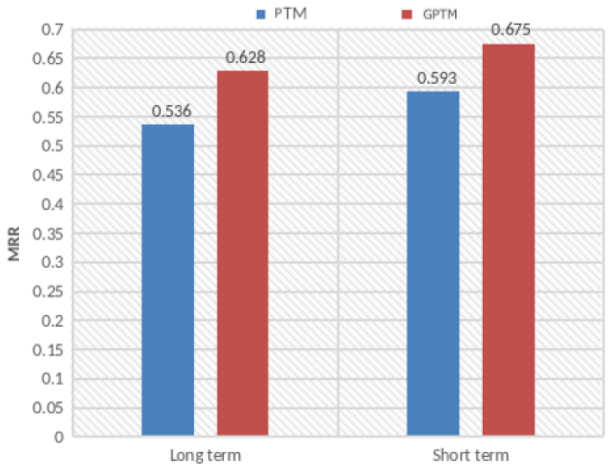

Figure 4. Ranking performance of the two models, PTM and GPTM, on queries.

4%-5% improvement with just this strategy. This, we believe, proves that handling low-frequency queries better is an important subtask for such selective personalization systems.

Finally, noticing a similar sparsity problem in user profiles based on topic models, rather than depending solely on the user performing the query, consolidating profiles of similar users improves personalization. The proposed group profiles improve the MRR of the queries by 9% for the long-term profiles and improve the short-term profile by 8%. The topic model-based search system achieves a 67% MRR score. To the best of our knowledge, grouping users by their profiles built using topic models is a novel method.

References

[1] Teevan J, Dumais S, Horvitz E. Beyond the commons: investigating the value of personalizing web search. In: Proceedings of the Workshop on New Technologies for Personalized Information Access; Edinburgh, UK; 2005. pp. 84-92.

[2] Dou Z, Song R, Wen J. A large-scale evaluation and analysis of personalized search strategies. In: Proceedings of the 16th International Conference on World Wide Web; Banff, Canada; 2007. pp. 581-590.

[3] Yano Y, Tagami Y, Tajima A. Quantifying query ambiguity with topic distributions. In: Proceedings of the 25th ACM International Conference on Information and Knowledge Management; Indianapolis, IN, USA; 2016. pp. 1877-1880.

[4] Harvey M, Crestani F, Carman M. Building user profiles from topic models for personalised search. In: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management; San Francisco, CA, USA; 2013. pp. 2309-2314.

[5] Carman M, Crestani F, Harvey M, Baillie M. Towards query log based personalization using topic models. In: Proceedings of the 19th ACM International conference on Information and Knowledge Management; Toronto, Canada; 2010. pp. 1849-1852.

[6] Teevan J, Dumais S, Horvitz E. Potential for personalization. In: ACM Transactions on Computer-Human Interaction; 2010. pp. 1-31.

[7] Teevan J, Dumais S, Liebling D. To personalize or not to personalize: modeling queries with variation in user intent. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval; Singapore; 2008. pp. 163-170.

[8] Cronen-Townsend S, Zhou Y, Croft W. Predicting query performance. In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval; Tampere, Finland; 2002. pp. 299-306.

[9] Wang Y, Agichtein E. Query ambiguity revisited: clickthrough measures for distinguishing informational and ambiguous queries. In: HLT '10 Human Language Technologies: 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics; 2010. pp. 361-364.

[10] Song R, Luo Z, Wen J, Yu Y, Hon H. Identifying ambiguous queries in web search. In: Proceedings of the 16th International Conference on World Wide Web; Banff, Canada; 2007. pp. 1169-1170.

[11] Matthijs N, Radlinski F. Personalizing web search using long term browsing history. In: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining; Hong Kong; 2011. pp. 25-34.

[12] Sieg A, Mobasher B, Burke R. Web search personalization with ontological user profiles. In: Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management; Lisbon, Portugal; 2007. pp. 525-534.

[13] Chirita A, Nejdl W, Paiu R, Kohlschütter C. Using ODP metadata to personalize search. In: Proceedings of the 28th Annual International ACM Sigir Conference on Research And Development in Information Retrieval; Salvador, Brazil; 2005. pp. 178-185.

[14] Karimi S, Abri R. Improvement of semantic search results with providing an updatable dynamic user model. International Journal of Computer Applications; 2016; 155 (4): 7-14. doi: 10.5120/ijca2016912285

[15] Vu T, Willis A, Tran S, Song D. Temporal latent topic user profiles for search personalisation. In: ECIR 37th European Conference on IR Research; Vienna, Austria; 2015. pp. 605-616.

[16] Vu T, Willis A, Kruschwitz U, Song D. Personalised query suggestion for intranet search with temporal user profiling. In: Proceedings of the 2017 Conference Human Information Interaction and Retrieval; Oslo, Norway; 2017. pp. 265-268.

[17] Bennett P, White R, Chu W, Dumais S, Bailey P et al. Modeling the impact of short- and long-term behavior on search personalization. In: Proceedings of the 35th international ACM SIGIR Conference on Research and Development in Information Retrieval; Portland, OR, USA; 2012. pp. 185-194.

[18] Hofmann T. Probabilistic latent semantic analysis. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters; 2010. pp. 1167-1175.

[19] Wei S, Yu Z, Ting L, Sheng L. Bridging topic modeling and personalized search. In: Proceedings of the Fifteenth conference on Uncertainty in Artificial Intelligence: Posters; Stockholm, Sweden; 1999. pp. 289-296.

[20] Shao M, Qin L. Text similarity computing based on LDA topic model and word co-occurrence. In: Proceedings of 2nd International Conference on Software Engineering, Knowledge Engineering and Information Engineering: Posters; Singapore; 2014.

[21] Momtazi S, Lindenberg F. Generating query suggestions by exploiting latent semantics in query logs. Journal of Information Science; 2016; 42 (4): 437-448. doi: 10.1177/0165551515594723

[22] Amer N, Mulhem P, Géry M. Toward word embedding for personalized information retrieval. Neu-IR: SIGIR 2016 Workshop on Neural Information Retrieval; Pisa, Italy; 2016.

[23] Vu T, Nguyen D, Dat Q, John M, Song D et al. Search personalization with embeddings. European Conference on Information Retrieval; Aberdeen, UK; 2017. pp. 598-604.

[24] Manning C, Raghavan P, Schütze H. Introduction to Information Retrieval. Cambridge, MA, USA: Cambridge University Press, 2008.

[25] Vu T, Willis A, Song D. Modelling time-aware search tasks for search personalisation. In: Proceedings of the 24th International Conference on World Wide Web; Florence, Italy; 2015. pp. 131-132.