Towards Query Log based Personalization using Topic Models

Mark J. Carman and Fabio Crestani
University of Lugano
Faculty of Informatics
Lugano, Switzerland
{mark.carman,fabio.crestani}@usi.ch

Morgan A. Harvey and Mark Baillie
University of Strathclyde
Department Computer and Information Sciences
Glasgow, UK
{morgan.harvey,mb}@cis.strath.ac.uk

ABSTRACT

We investigate the utility of topic models for the task of personalizing search results based on information present in a large query log. We define generative models that take both the user and the clicked document into account when estimating the probability of query terms. These models can then be used to rank documents by their likelihood given a particular query and user pair.

Categories and Subject Descriptors

H.3.3 [**Information Storage & Retrieval**]: Information Search & Retrieval

General Terms

Design, Measurement, Experimentation

Keywords

Personalized Search, Topic Models, Query Log Analysis

1. INTRODUCTION

Click-through data, in the form of query logs, is an abundant and important source of information for improving search engine retrieval performance. This data contains the search history of individual users and can therefore be utilized to personalize search results. Many studies have suggested that while difficult, if done correctly, personalization can indeed improve the quality of search results [**?**].

The aim of this paper is to investigate the applicability of topic modeling [**?**] approaches to the problem of query-log based personalization. The main contributions of this work are as follows:

- We describe two different topic models capable of factorizing the query log into a set of parameter matrices, and define document ranking functions based on the learned parameters.
- We train the models and evaluate their personalization performance on a real query log dataset.

2. RELATED WORK

Click-through data has been used extensively in the past as pairwise preference judgements for training learning-to-rank algorithms [**?**]. The aim of these rank learning techniques is to use the query log as a proxy for human relevance judgements. Each clicked-document for a particular query is assumed to be either a vote confirming its relevance or a preference for that document over other documents present higher in the ranked list that were not clicked on. These pseudo-judgements are then used to calculate optimal weights for combining relevance features for each query-document pair.

In our work, we follow the same intuition; that each click on a URL represents an implicit vote for the relevance of the document to the query. Our work departs from that of learning-to-rank systems in that we are not attempting to learn a query independent weighting scheme for features defined over document-query pairs. Rather we are interested in a query dependent and personalized ranking function using only the data present in the log itself. The score given by such a function could then be used alongside relevance information (such as content-based retrieval functions) as features for a learning-to-rank system.

There have been a number of recent studies on the viability of query log based personalization [**?**, **?**]. Dou et al. [**?**] investigated a number of heuristics for creating user profiles and generating personalized rankings. In their work they used a preset set of interest categories and a K-nearest neighbor (KNN) approach for grouping users. They then used a borda-fuse approach to merge personalized and baseline rankings. The techniques we outline in this paper generalize these more heuristic techniques for clustering and smoothing user profiles by estimating the latter as parameters of a generative process. We also offer personalized ranking formulas which are principled in the sense that they are derived directly from the generative process.

In [**?**] the authors perform a large-scale study of query logs obtained from the Yahoo search engine to investigate user activity and determine if such data could be utilized to generate user profiles for personalization. The results of the study indicate that users' queries exhibit a fairly consistent set of topics with each user varying in their specific topical interests. Furthermore the work showed that a user's interests tended to converge to a stable distribution, but only after a large number of queries. These results inform the

work in this paper and lend credence to the notion that hidden topic models might provide a workable solution to the problem of generating reliable and generalizable user profiles from query log data.

3. TOPIC MODELS

In this section we introduce and compare different latent variable generative processes for modeling the data found in a query click-through log. Before introducing these models however, we first briefly discuss Latent Dirichlet Allocation (LDA) [**?**, **?**] which is a simple latent topic model that we extend to build more complicated models for query log analysis. We will also be using LDA as a non-personalized baseline for our personalization experiments later on.

An LDA model consists of two parameter matrices Φ and Θ containing estimates for the probability of a word given a topic $P(w|z)$ and a topic given a document $P(z|d)$. Each column of the respective matrices contains a probability distribution over words for a particular topic and over topics for a particular document (denoted ϕ_z and θ_d respectively). In order to prevent overfitting, LDA places a symmetric Dirichlet prior on both these distributions, resulting in the following expectations for the parameter values under the respective posterior distributions $P(\phi_z|\mathbf{w}, \mathbf{z})$ and $P(\theta_d|\mathbf{z}, \mathbf{d})$, where $\mathbf{w}$ is the vector of words occurrences w_i in the corpus, $\mathbf{z}$ is an assignment of topics to each word position z_i and $\mathbf{d}$ is the vector of documents d_i associated with each word position:

$$\hat{\phi}_{w|z} \triangleq \mathbf{E}_{P(\phi_z|\mathbf{w},\mathbf{z})}[P(w|z)] = \frac{N_{wz} + \beta\frac{1}{W}}{N_z + \beta} \tag{1}$$

$$\hat{\theta}_{z|d} \triangleq \mathbf{E}_{P(\theta_d|\mathbf{z},\mathbf{d})}[P(z|d)] = \frac{N_{zd} + \alpha\frac{1}{Z}}{N_d + \alpha} \tag{2}$$

Here N_{wz}, N_{zd} and N_z are counts denoting the number of times the topic z appears together with the word w, or with the document d or in total respectfully. W is the vocabulary and Z is the number of topics. Finally α and β are hyperparameters, which determine the number of "pseudo-counts" from a uniform distribution over the vocabulary/topics that are added to the observed counts of words/topics.

The Gibbs sampling procedure for LDA involves iteratively updating the assignment of each topic z_i in the topic vector $\mathbf{z}$ by sampling a value from the conditional distribution $P(z_i|\mathbf{w}, \mathbf{z}_{-i}, \mathbf{d})$, which is conditioned on the current assignment to all topic variables except z_i, (denoted $\mathbf{z}_{-i}$). In LDA the word assignment is conditionally independent of the document given the topic assignment, so:

$$P(z_i|w_i, d_i) = \frac{P(z_i, w_i|d_i)}{P(w_i|d_i)} \propto P(w_i|z_i)P(z_i|d_i) \tag{3}$$

Thus the expected value for the conditional distribution is simply:[1]

$$\mathbf{E}[P(z_i|\mathbf{w}, \mathbf{z}_{-i}, \mathbf{d})] \propto \hat{\phi}_{w_i|z_i}\hat{\theta}_{z_i|d_i} \tag{4}$$

Where the estimates $\hat{\phi}_{w|z}$ and $\hat{\theta}_{z|d}$ are calculated over $\mathbf{z}_{-i}$ rather than $\mathbf{z}$. After sufficient iterations of the sampler, the Markov chain converges (as seen by minimal change in the model likelihood $\prod_i \sum_z \hat{\phi}_{w_i|z}\hat{\theta}_{z|d_i}$) and the parameters of the LDA model can be estimated from $\mathbf{z}$. For increased accuracy, the parameter estimates are averaged over consecutive samples (iterations) $\{\mathbf{z}^{(n-k)}, \ldots, \mathbf{z}^{(n)}\}$ from the end of the chain.

[1]Since the expectation of the product of two independently distributed random variables is the product of their expectations.

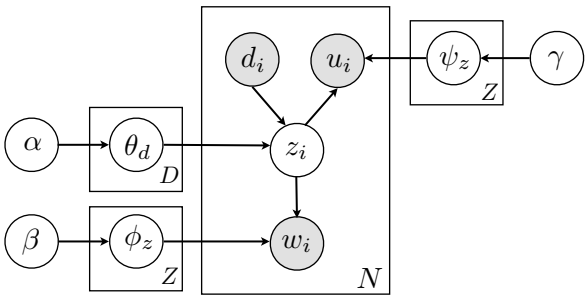

Figure 1: Personalization Topic Model 1 (PTM1) has 3 observed variables and one latent variable per word position. The difference with LDA being the addition of an observed user variable u_i, which like the word w_i is dependent on the topic z_i.

We now investigate Topic Models which contain three observed variables: the document d_i, the user u_i and the word w_i (as apposed to the two observed variables of LDA). The new user variable corresponds to the unique identifier for the user in the log who submitted the corresponding query.

We would like to estimate a model in which the topic assignment depends on both the document and the user. More precisely, we would like to directly estimate the probablity $P(z|d, u)$ as a parameter of our model. Unfortunately that would require estimating $(Z-1)DU$ different parameter values, where D is the number of documents and U is the number of users. Obviously, for any reasonable size dataset (containing many thousands of documents and users) there will be insufficient data to estimate all of the parameters. In other words, we are unlikely to see enough queries for each pair of documents and users to be able to estimate an entire distribution over topics. We thus need to investigate models that factorize the conditional distribution by assuming conditional independence between u_i and d_i given z_i.

Figure **??** shows the plate notation for our first model, a slight variation on LDA, which we will refer to as Personalization Topic Model 1 (PTM1). In this model, the user assignment u_i is chosen according to a multinomial distribution ψ_z, which depends on the topic z_i, in the same way as the word assignment w_i is chosen from a topic specific multinomial ϕ_z. The user distributions are smoothed using a symmetric Dirichlet prior with concentration parameter γ, resulting in the following estimate for the probability of a user given a topic:

$$\hat{\psi}_{u|z} = \frac{N_{uz} + \gamma\frac{1}{U}}{N_z + \gamma} \tag{5}$$

Thus the PTM1 model has the same parameters as LDA (Φ and Θ) plus an additional user probability matrix of size ZU which we denote Ψ. For the Gibbs sampling routine, the probability of choosing a topic assignment under the model can be factorized as follows:

$$\begin{aligned} P(z_i|w_i, d_i, u_i) &= \frac{P(z_i, w_i, u_i|d_i)}{P(w_i, u_i|d_i)} \\ &\propto P(w_i|z_i)P(u_i|z_i)P(z_i|d_i) \end{aligned} \tag{6}$$

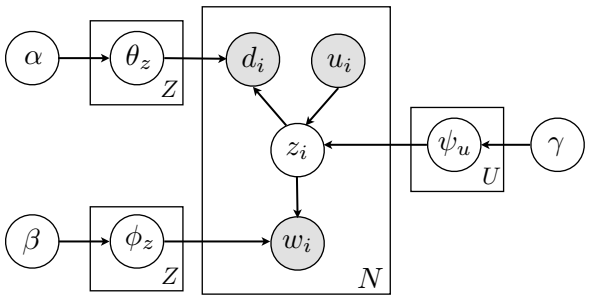

Figure 2: PTM2 is the same as PTM1 except that the topic z_i is now dependent on the user u_i rather than the clicked document d_i which is instead itself now dependent on the topic z_i.

Thus we have the following update formula:

$$\mathbf{E}[P(z_i|\mathbf{w}, \mathbf{z}_{-i}, \mathbf{d}, \mathbf{u})] \propto \hat{\phi}_{w_i|z_i}\hat{\psi}_{u_i|z_i}\hat{\theta}_{z_i|d_i} \quad (7)$$

The PTM1 generative model is a little counterintuitive in that it supposes that a document chooses a topic and then the topic chooses a word and a user, making the user a byproduct of the process rather than the initiator of it. We can also consider a model in which the user first chooses the topic and then the topic chooses the document. We refer to this model as PTM2 and show a graphical representation of it in Figure **??**. According to this model we need to estimate the parameters $\hat{\theta}_{d_i|z_i}$ and $\hat{\psi}_{z_i|u_i}$ instead of $\hat{\psi}_{u_i|z_i}$ and $\hat{\theta}_{z_i|d_i}$. The Gibbs sampling update for the second model is then practically the same as before:

$$\mathbf{E}[P(z_i|\mathbf{w}, \mathbf{z}_{-i}, \mathbf{d}, \mathbf{u})] \propto \hat{\phi}_{w_i|z_i}\hat{\theta}_{d_i|z_i}\hat{\psi}_{z_i|u_i} \quad (8)$$

4. RANKING DOCUMENTS

For the non-personalized case (using the LDA model) we rank documents according to their likelihood given the query, which can be estimated as follows:

$$\begin{aligned} P(d|q) \propto P(d)P(q|d) &= P(d)\prod_{w\in q} P(w|d) \\ &= P(d)\prod_{w\in q}\sum_z \hat{\phi}_{w|z}\hat{\theta}_{z|d} \end{aligned} \quad (9)$$

In order to personalize the ranking we can rank documents according to their likelihood given both the query and the user:

$$\begin{aligned} P(d|q,u) \propto P(q,d|u) &= P(d|u)P(q|d,u) \\ &= P(d|u)\prod_{w\in q} P(w|d,u) \end{aligned} \quad (10)$$

Thus our personalized ranking formula consists of a *user specific* "document prior" $P(d|u)$ and a query-term likelihood $P(w|d,u)$. The first term estimates the user's probability to choose the document *a priori*, and the second models their likelihood to choose word w to describe it.

For PTM1 we can calculate these two estimates as follows:

$$P_1(d|u) \propto P(u|d)P(d) = P(d)\sum_z \hat{\psi}_{u|z}\hat{\theta}_{z|d} \quad (11)$$

$$P_1(w|d,u) = \frac{P(w,u|d)}{P(u|d)} = \frac{\sum_z \hat{\phi}_{w|z}\hat{\psi}_{u|z}\hat{\theta}_{z|d}}{\sum_z \hat{\psi}_{u|z}\hat{\theta}_{z|d}} \quad (12)$$

Model	S@1	S@10	MRR@10
LDA	**0.2283**	**0.4693**	**0.3063**
PTM1	0.1809	0.4162	0.2560
PTM2	0.1927	0.4178	0.2640

Table 1: Initial retrieval performance on the three month dataset with 160 topics. The not personalized LDA baseline performs best overall.

Similarly for PTM2:

$$P_2(d|u) = \sum_z \hat{\theta}_{d|z}\hat{\psi}_{z|u} \quad (13)$$

$$P_2(w|d,u) = \frac{P(w,d|u)}{P(d|u)} = \frac{\sum_z \hat{\phi}_{w|z}\hat{\theta}_{d|z}\hat{\psi}_{z|u}}{\sum_z \hat{\theta}_{d|z}\hat{\psi}_{z|u}} \quad (14)$$

5. EXPERIMENTS

In order to evaluate our models on real data we made use of the AOL Query Log dataset. The log contains the queries of 657,426 anonymous users over 3 months from March to May, 2006. It is, as far as we know, the only publicly available dataset of sufficient size to perform our analysis. We protected user privacy by analyzing results only over aggregate data.

To clean the data we first discarded all queries which didn't result in a click on a URL. We then selected only those URLs which more than 100 users had clicked on, and selected from the remaining users only those with more than 200 queries. In order to parse the queries we first separated the words according to whitespace. All punctuation was removed and Porter's algorithm was used for stemming. We did not remove any stopwords but did remove words that appeared less than 3 times in the dataset.

The resulting dataset contained 2152721 queries, 6581 users, 15996 documents and had a vocabulary of 53132 words. We note that the user profiles in the dataset should be considered "long-term" since they are over a 3 month period, (in contrast to "short-term" profiles that are built using click data from only the previous week, day or even session).

We separated the dataset into a training and testing set by retaining the last 5% of queries by each user for testing. To ensure the significance of the results all retrieval metrics were computed over the test queries for 1000 users from the log.

To evaluate retrieval performance we calculated the success at rank k (S@k) and the mean reciprocal rank (MRR):

$$S@k = \frac{1}{Q}\sum_i^Q \mathtt{I}(r(d_i,q_i) \le k) \quad (15)$$

$$MRR = \frac{1}{Q}\sum_i^Q \frac{1}{r(d_i,q_i)} \quad (16)$$

Here Q denotes the number of queries, $r(d,q)$ is the rank of document d for query q and $\mathtt{I}()$ is an indicator function returning 1 whenever its argument is true and zero otherwise.

With regard to parameter settings, for LDA we set the concentration parameters α and β to be 50.0 and $0.1W$ respectively, which are common settings used in the literature [**?**]. (Here W is the size of the vocabulary.) For the personalization models, we set α to 50.0 or $0.1D$ (depending on the model), β to $0.1W$, γ to 50.0 or $0.1U$.

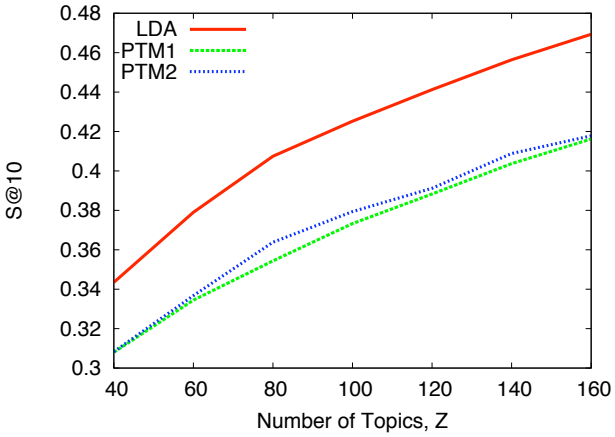

Figure 3: Retrieval performance on the three month dataset versus topic count. LDA consistently outperforms the PTMs across all topic count values.

For all models we ran the Gibbs sampler for 400 iterations, discarding the first 300 iterations as burn-in and averaging parameter estimates over the last 100. These settings appeared to give consistently good convergence in terms of model likelihood.

Table **??** gives the results for models with a topic count of 160. Surprisingly, the personalization topic models with their additional parameters and ability to leverage information about the user do not appear to improve performance over the baseline ranking. Across all metrics the not-personalized LDA-based ranking is preferred. We believe that there may be three reasons for this. Firstly, it may be the case that the learnt models have not converged on the best possible parameter settings since the generative models all assume that the document label and the user label are equally important, when in reality the document is far more important for determining the query terms. Secondly, by blindly applying personalized ranking algorithms to all queries, we may be degrading performance on those queries whose meaning was not initially ambiguous. Thirdly, the long term profiles of the users may themselves be so heterogenous as to not be useful for predicting the future relevance of documents to the user. These conjectures motivate our future work where we plan to investigate other convergence criteria, user profiles over shorter time periods and also query ambiguity measures.

We then investigated the effect of changes in the topic count on the relative performance of the personalization models. Figure **??** shows plots of performance against the number of topics for the long-term dataset. We see from the first plot that as the number of topics increases so does the performance of the different models. The LDA consistently outperforms the personalization models across the topic counts.

6. CONCLUSIONS

In this paper we have shown that it is possible to factorize a sizable search engine query log using topic modeling techniques. The results of our analysis applying two different topic modeling based personalized ranking formulas to a query log dataset were largely negative, indicating the complexity of the personalization problem.

Results on a short term user-profile dataset were far more positive, however.

that there are a large number of occasions where the personalised models do outperform LDA and in future work we want to look into this in more detail.

and it's probably because not all queries are equally amenable to personalisation, we need some way to identify a priori which ones are amenable...

NEED TO ADD SOMETHING HERE?

This research represents a first step in applying topic modeling techniques to the difficult task of query log based search engine personalization. There are many opportunities for future work, including: (1) investigating methods for "folding in" documents and users who were not in the training set, (2) integrating other forms of evidence into the ranking including the content of the documents, (3) extending the topic models in order to account for and make use of the temporal information about users, documents and terms in the log or applying n-gram based topic models to allow for better analysis of concepts within queries, (4) improving the cleaning of the log data in particular taking into account click bias in the models, (5) experimenting with the PTMs on other tripartite data sources such as Twitter streams.

7. REFERENCES

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

[2] Z. Dou, R. Song, and J.-R. Wen. A large-scale evaluation and analysis of personalized search strategies. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 581–590, New York, NY, USA, 2007. ACM.

[3] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Science*, 101:5228–5235, 2004.

[4] T. Joachims. Optimizing search engines using clickthrough data. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 133–142, 2002.

[5] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101, 2003.

[6] S. Wedig and O. Madani. A large-scale analysis of query logs for assessing personalization opportunities. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'06)*, pages 742–747, 2006.