A Personalized Search Using User's- Profile

Naeem Naik, L.M R.J. Lobo

Abstract-Users' interest is an important area in the field of IR that attempts to adapt ranking algorithms so that the results returned are tuned towards the searcher's interests. In this work we use user data to build personalized ranking models in which user profiles are constructed based on the user's tagging data over a topic space. However, instead of employing a human-generated ontology, we use novel latent topic models to determine these topics. This means that the topic space used is determined based purely on the tagging data itself and therefore does not require human involvement to determine the topics.

Our experiments show that by introducing user profiles as part of the ranking algorithm, rather than by re-ranking an existing list, we can provide personalized ranked lists of documents which improve significantly over a non-personalized baseline. Further examination shows that the performance of the personalized system is particularly good in cases where prior knowledge of the search query is limited. This is especially useful as these are the cases where we are unable to determine when same tag has totally different intention.

Index Terms— image search, metadata, optimization,

I. INTRODUCTION

The vocabulary problem, where people use the same terms to describe different needs, is a well-known issue affecting Information Retrieval systems and was identified early on in the field's development. Despite this, most IR systems treat all users equally and, given a specific query, will attempt to return an optimal ranked link for the average user". Recent years have seen a gradual increase in the use of personalization to improve search results, corresponding with an equally gradual increase in our understanding of how to tackle the problem [1]. Personalization can be seen as a special case of augmentation to search systems where additional context, beyond merely the search query issued, is used to enhance rankings. The key idea is that by understanding something about the user issuing a query, we can tailor the ranked list presented such that the likelihood of highly ranked documents being relevant is increased. Much early work into such techniques was unsuccessful and many studies have subsequently shown that great care must be taken when applying personalization so as to avoid damaging an already near optimally ranked list [2, 3].

In order to construct a personalized ranking for a given user's query, it is first necessary to have some knowledge about that user's previous tagging behavior.

This prior evidence is often referred to as a user's profile" and should, in some sense, represent the topical interests of the user.This profile can be built by considering tags made by the user one and, for each of these searches, the query terms used.

Manuscript received October, 2013.

Mr. Naeem Naik, Walchand Institute of Technology, Solapur, India.
Prof. L.M R.J. Lobo, Walchand Institute of Technology, Solapur, India.

In early work these profiles were constructed using the raw terms of prior queries or the content of the clicked documents, usually in the form of language models, however this often proves to be ineffective, perhaps because such a representation of interests too fine-grained given the limited amount of data available [4]. An approach for dealing with this sparsity is no instead base the profile on the main topics discussed in each document, where the topical description of the document is obtained from a human-generated online ontology, such as the Open Directory Project [5, 6, 7].

This approach introduces the problem, however, that many documents may not be present in the online categorization scheme and requires that people are involved in determining the correct categories for each document, a process that is both expensive and error-prone.

Click-through data, in the form of query logs, is a potentially abundant - since any search engine could generate them - and important source of information regarding search behavior and can therefore be utilized for personalized search tasks. Query logs generally take the form of triples, consisting of a user ID, a search query and a clicked document. 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 presented higher in the ranked list that were not clicked on [8].

In this work, we follow the intuition that each click on a URL represents an implicit vote for the relevance of the document to the query and that the query words used to search for a document represent that document's content. This framework allows us to construct representations of documents, to build personalized search models and to fairly evaluate the performance of these models, since the query logs represent user-specific relevance judgments in context.

II. RELATED WORK

The idea of using tagging data of a user with a search system to construct a user profile has been around for some time, and there is significant variation in ways that the problem has been tackled. The approaches differ based on what length of profile data is used and how the data chosen are then turned into a suitable user profile. In some cases researchers have considered only the information from the small corpus in order to build short-term profiles, whereas other work has attempted to identify longer-term user interests. Recent work has even shown how these short and long-term profiles can be combined [9]. In general short-term data is often too sparse to allow for robust personalization performance and only delivers solid improvements late in long search sessions, which are relatively rare. In this work we focus on long-term tagging data to build user profiles as it provides a richer source of information about the user's interests and preferences.

Once prior tagging data has been chosen, it must then be converted into a user profile which should form a representation of the user's interests, be they long-term or

simply with regard to the current corpus. These profiles can be generated in a number of different ways. Some approaches use vectors of the original terms, often weighted in some fashion. Others attempt to map the user's interests onto a set of topics, which are either defined by the users themselves or extracted from social websites.

Furthermore once users' corpus is getting created we apply the LDA algorithm to model user specific topic for each user. Once each user's specific topic is generated, we use this topic to map user's query. The module is placed in online stage and called as user specific query mapping. After the user's query is mapped to one or more user specific topic, we rank the topic by using topic sensitive user preferences. These topic sensitive user preferences are calculated based upon how frequently used tags a single word from one of the topics. After calculating topic sensitive user preferences and ranking the mapped topics, we then rank the images from these topics and these images are then displayed to the user in a systematic manner.

The work presented in this paper based on LDA for the problem of personalized search. These include a user-topic distribution directly in the model, thereby considering the user as part of the generative process. When evaluating these models using tagging data it was found that they had an overall no negative effect on the ranked lists produced and were therefore able to improve upon the personalized LDA baseline. We now present an approach to tag based search for personalization using sets of latent topics derived directly from the tag data itself where the user is specifically included as part of the generative process but rather is subtly introduced as part of the ranking formula.

By means of a large-scale experiment we are able to demonstrate performance improvements over an un-personalized baseline and show that this new model is particularly effective in cases of sparse prior data where tagging frequencies can be utilized to generate good ranked lists.

III. PROPOSED SYSTEM

Module Description:

1) Corpus Creation:

DA has the assumption that each topic, or a language model, is a multinomial distribution over all words. For a document, multiple occurrence of the same word is a draw from this multinomial distribution. So, in theory, word counts matter. However, in terms of inference, if you use Gibbs sampling, you probably need to re-sample topic assignments for multiple occurrence. But if you use variational inference, you just do it the same across multiple occurrences.

Here in our proposed systems corpus for each user is created by using his tagging data corresponding to each tagged image by him. Each document is created using the collection of tags (by removing duplicate tags) given by a specific user for a specific image.

The collection of such document constitutes a corpus for a specific user. Furthermore each user's corpus contains a user's vocabulary, a file containing all the tags removing duplicates. Each tag or word in this file has an index. These indices are then placed at the specific place in each document. Now, with the documents for each tagged image and a vocabulary our corpus is ready for LDA.

2) User specific topic modeling:

LDA algorithm topic modeling:

In natural language processing, latent Dirichlet allocation (LDA) is a generative model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's creation is attributable to one of the document's topics. LDA is an example of a topic model and was first presented as a graphical model for topic discovery byDavidBlei, Andrew Ng, and Michael Jordan in 2003.

With plate notation, the dependencies among the many variables can be captured concisely. The boxes are "plates" representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document. M denotes the number of documents, N the number of words in a document. Thus:

A is the parameter of the Dirichlet prior on the per-document topic distributions,

β is the parameter of the Dirichlet prior on the per-topic word distribution,

θ_i is the topic distribution for document i,

ϕ_k is the word distribution for topic k,

z_{ij} is the topic for the jth word in document i, and

w_{ij} is the specific word.

The w_{ij} are the only observable variables, and the other variables are latent variables. Mostly, the basic LDA model will be extended to a smoothed version to gain better results. The plate notation is shown on the right, where K denotes the number of topics considered in the model and:

ϕ is a K*V (V is the dimension of the vocabulary) Markov matrix each row of which denotes the word distribution of a topic.

The generative process behind is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words.

3) User Specific Query Mapping:

In the online stage, when user u submits a query q, 9 we first perform user-specific query mapping - estimate the conditional probability that q belongs to user u-specific topics:

P (topicj|q, u) =p (topicj|u) ·p (q|topicj, u) / p (q)

The user query is mapped against one or more user specific topic. Once the query is mapped, we rank the mapped topics based the topic sensitive user preferences.

4) The rank of the user u-specific topics is decided by p (topicj|u), the probability that user u is interested in topicj. This can be calculated by aggregating user u's preference over all the images.

Once we are done with maps topics ranking, we then go for image ranking.

The rank of image i can be obtained as:

Rank (i |q; u) ∝ 1 / Σ p (topicj|q; u) p (topicj|i; u).

Using the distributions obtained from this model we should be able to construct a ranking formula which, given a query, will consider the probability of each document given the word in the query and the interests of the user who submitted it. However, as outlined earlier in the paper, in order for personalization to work it must be applied very subtly.

By directly including the user in the topic model we are saying that his/her topical interests are equally important when describing a document he/she has clicked as the words assigned to that document to describe it. The work of Carman demonstrated that this assumption is clearly far too strong as they were unable to obtain successful results from similar models. Instead we consider a different model which does not explicitly include the user in the Markov chain topic sampling but instead calculates an interest distribution for each user after the sampler has converged.

System Architecrure:

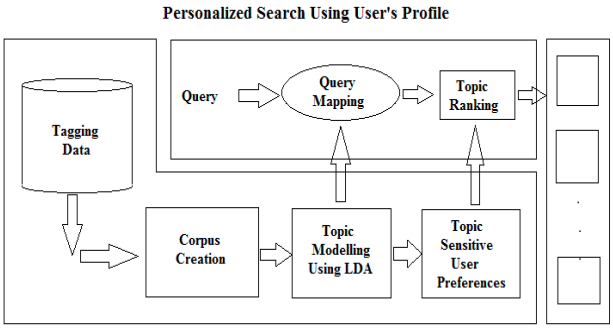

Figure 1

IV. OVERVIEW AND DIAGRAM

The overview of proposed systems can be described in two stages, offline stage and online stage as shown in Fig 2.

In offline stage, first we acquire data in U x I x T relationship. Then to calculate User Annotation Prediction we apply Ranking Based Multi-Correlation Tensor Factorization. After calculating RMTF we create Corpus which contains the tagging data of all the users for LDA. LDA gives the User Specific Topic for each user. Once we get User Specific Topics, we can determine Topic sensitive user preferences. All these modules are executed in offline mode.

In Online stage, when user fires a query, the user query is mapped to one or more User Specific Topics. Once a given query is mapped, then we consider the mapped topics and Topic Sensitive User Preferences to rank the Images which are retrieved. The ranked results are then displayed to the user, which are obviously Personalized Ranked List.

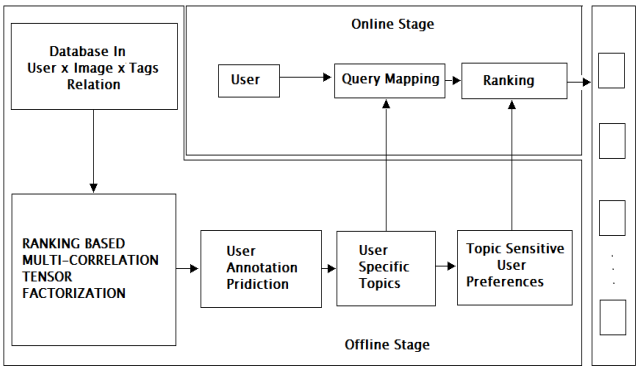

Figure 2. Overview and diagram

V. INPUT OUTPUT

In our proposed system, during offline mode the input should be ternary relationship of Users, Images and Tags. In offline mode we calculate some of the required parameters for the personalized image search. Fig 3 shows the input output of proposed system

During the online mode we have a user name and a single word textual query as an input; the query is normally from the user's corpus. The output of our proposed systems contains a ranked list of images which is derived from Query mapping and Ranking modules.

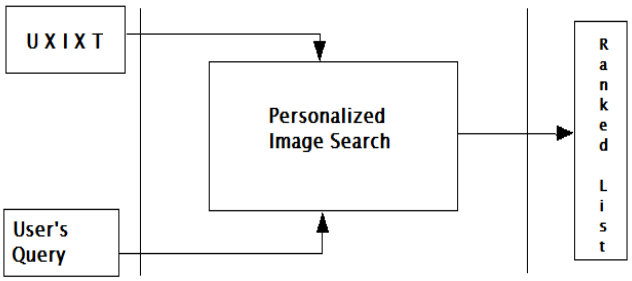

Figure 3 Input Output

VI. CONCLUSIONS

The results of our analysis indicate that it is possible to improve performance through personalisation by making use of topic-model based user profiles. While in theory, personalisation can offer a path to achieving substantial gains in retrieval performance, in practice performance improvements over all queries will be quite small with respect to the performance of the un-personalised retrieval system. Thus personalisation needs to be introduced with great care in order to obtain gains without adversely affecting average performance.

During the user-specific topic modeling process, the obtained user-specific topics represent the user's distribution on the topic space and can be considered as a user's interest profile. Therefore, this framework can be extended to any applications based on interest profiles.

REFERENCES

[1] Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. In Bocca, J. B., Jarke, M. & Zaniolo, C. (Eds.), *Proceedings of the 20th Int. Conf. Very Large Data Bases, VLDB* (pp. 487— 499).

[2] J. Teevan, S. T. Dumais, and E. Horvitz. Potential for personalization. ACM Trans. Comput.-Hum. Interact. 17(1):4:1{4:31, Apr. 2010.

[3] Z. Dou, R. Song, and J.-R. Wen. A large-scale evaluation and analysis of personalized search strategies. Proceedings of the 16th international conference on World Wide Web, pages 581{590, 2007.

[4] Morgan Harvey et al, "Building User Profiles from Topic Models for Personalised Search"

[5] P. A. Chirita, W. Nejdl, R. Paiu, and C. Kohlschutter. Using odp metadata to personalize search. In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '05, pages 178{185, New York, NY, USA, 2005. ACM.

[6] S. Gauch, J. Cha_ee, and A. Pretschner. Ontology-based user pro_les for search and browsing. WIAS, pages 219{234, 2003.

[7] Sieg, B. Mobasher, and R. Burke. Web search personalization with ontological user pro_les. In Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, CIKM '07, pages 525{534, New York, NY, USA, 2007. ACM.

[8] S. Wedig. A large-scale analysis of query logs for assessing personalization opportunities. In Proceedings of KDD '06, pages 742{747, 2006.

[9] Wood, K. R., Richardson, T., Bennett, F., Harter, A., & Hopper, A. (1997). Global teleporting with Java: toward ubiquitous personalized computing. Computer, 30(2), 53-59.