Topics, Tasks & Beyond: Learning Representations for Personalization

Rishabh Mehrotra
Dept of Computer Science,
University College London
London, UK
r.mehrotra@cs.ucl.ac.uk

ABSTRACT

Accurate understanding of a user's interests, preferences and behaviours is possibly one of the most critical research challenges faced while developing personalized systems for behavior targeting and information access. We intend to develop comprehensive latent variable models for web search personalization which jointly models user's topical interests along with user's click based relevance preferences while at the same time taking into account user's intended search tasks along with information about other similar users. We further augment this model by incorporating topic-level relevance parameters, which, to the best of our knowledge, is the first attempt at modeling result ranking preferences at the topic level. Additionally, we intend to explore the possibility of modeling users in terms of the search tasks they perform thereby coupling users' topical interests with their search task behavior to learn user representations. Finally, we wish to evaluate the proposition of extending user representations to hierarchical structures as an alternative to existing flat representations. The evaluation of these alternative approaches for user modeling is based on their performance on a variety of tasks such as collaborative query recommendations, user cohort modeling and search result personalization. This proposal provides the motivation to pursue these research directions, summarizes key research problems being targeted, glances through potential ways of tackling these research challenges and highlights some initial results obtained.

Categories and Subject Descriptors

H.3.3 [**Information Storage And Retrieval**]: Information Search and Retrieval—*User Modelling*

Keywords

User modelling; Personalization; Latent Variable Models; Search Tasks

WSDM'15, February 2–6, 2015, Shanghai, China.
Copyright 2015 ACM 978-1-4503-3317-7/15/02 ...$15.00.
http://dx.doi.org/10.1145/2684822.2697037.

1. INTRODUCTION

Personalization has become an important research topic in machine learning fuelled in part by its major significance in e-commerce and other businesses/services that try to tailor to user-specific needs or preferences. Online products, news, search, media, advertisement, user interfaces, and to a lesser extent healthcare, are several of the areas that have depended on some form of personalization to improve satisfaction or business goals in general. In order to address personalization problems machine learning has long relied on tools such as collaborative filtering (matrix factorization) and models originally developed not necessarily for personalization. However, even though the data available for personalization has grown in richness and size, and the available processing power has also increased, the basic tenet for the methods used has not changed in a major way.

As a consumer of the informational content, different users have distinct preferences of information for decision making; thus accurately understanding their respective information needs and decision preferences is crucial for providing effective decision support. While human behaviours are largely determined by their own goals and preferences, the mined knowledge reveals users' underlying intentions and behaviour patterns, which provide unique signals for human centric optimization and personalization. Web search personalization has recently received a lot of attention by the research community. Personalized search leverages information about an individual to identify the most relevant recommendations for that person. A challenge for personalization is in collecting user profiles that are sufficiently rich to be useful in settings such as result ranking, query recommendations, etc, while balancing privacy concerns.

Accurate understanding of a user's search interests and preferences is arguably one of the most critical research challenges in web search. Existing work on user level personalization has made use of ODP related topical categories or has relied heavily on term based representation of user profiles. There are several challenges with topical analysis of user actions. Most of existing work on analyses user's topical interests and his/her associated click patterns in an isolated manner. We should note that both queries and click-throughs convey useful clues about user's search intent. Capitalizing on both 1) user's topical interests and 2) user's unique result ranking preference should provide a comprehensive understanding of user's search intent. While personalization techniques rely heavily on using ODP-like topic categories, such a method restricts topic-coverage as search logs offer much richer content in terms of the number of topics spanned as well as the granularity level of the topics. Personalization models which learns latent topics in an unsupervised manner from search logs would help capture topical interests embedded in search logs. Efficient user profiles can then be generated by modeling users by a mixture of interests (/topics) and identifying distinct ranking preferences of individual users based on his click patterns.

Furthermore, it should be noted that users interact with search engines to accomplish some task such as *arrange a trip*, *plan a wedding* etc. Such broad requirements prompts the use of multiple queries, sometimes spanning multiple sessions. We define search tasks as the group of queries a user issues to accomplish such overall intended task and advocate the use of such search tasks to build individual user models. We postulate that in a web search setting, a user representation based on the search tasks users' perform would better capture user actions, interests and preferences.

We intend to evaluate the usefulness of deviating from the traditional way of representing users based on a flat d-dimensional representation to user modeling over hierarchies. One possibility is to develop models which learn hierarchical models of user's intended tasks and learn user representations on top of such hierarchies. Overall, we intend to develop comprehensive latent variable models for user modelling and personalization which is able to capture all these goals in a fully unsupervised fashion. We review existing work in Section 2 while the specific research challenges along with possible solutions are discussed in Section 3. Section 4 discusses the results obtained so far while Section 5 highlights key issues for discussion at the consortium.

2. RELATED WORK

Personalizing web search has received a lot of attention by the research community. Prior work has primarily focused on mining general search behaviours but has considerably ignored the importance of identifying individual user's search preferences as well as user variability. A prominent line of prior research uses long-term histories to directly improve retrieval effectiveness. Various authors have considered topic-based representations for personalization [5][6][7] making use of hand-picked ODP topical categories. While such topics are easily specified, much human effort is required in labelling queries for each topic. ODP categories based methods restricts topic coverage in a major way as search logs offer much richer content both in terms of the number of topics involved as well as the granularity level of each topic.

Another line of research in web search personalization has focused on using term based representations for user profiles. Authors in [8] build user interests profiles using terms extracted from user's browsing history following which the term weights are generated using different weighing schemes. While query terms are representative of user interests, they often limit the scope of personalization as different users inherently follow different distributions over words and queries belonging to the same topic/interest might not contain any over-lapping terms. Finding similar users and building user cohorts becomes difficult in such settings.

Finally, very recently, authors in [9] have proposed a generative model which models users as a mixture over latent user groups wherein each group shares a common distribution over queries and a common click preference pattern. While it is tempting to group users into categories, this grouping tends to be rather limited when it comes to large number of users and large amount of behavioural data. An inflation in the number of user clusters decreases the interpretability of these user groups. Also, the proposed model is limiting in the sense that it does not capture the idea that a user might share topical interests with other users in the group but have a different click preferences in terms of what kind of documents the user ideally prefers and clicks.

3. RESEARCH CHALLENGES

Overall, the goal of this research is to develop latent variable models which succinctly represent user information and help in building user profiles for use in personalization based services. We next describe the different research questions we intend to answer with this research.

3.1 Joint model of user interests with click preferences

We intend to propose a latent variable model which captures user's interests in an unsupervised manner wherein a user is characterized by a mixture of interests (/topics) and distinct ranking preferences of individual users are identified based on his click patterns. Our model is able to learn these latent topical interests from search logs by attempting to describe the available data instead of manually fixing the topics beforehand. Such a topical mixture carries considerably more information than user clusters with users sharing topical interests. The click preference of each user is portrayed by the corresponding relative importance of the ranking features, which leads to distinct click patterns over the returned documents. For example, for the same query, some user may want high authority websites (i.e., larger weight on the pagerank score), while the other user may prefer the documents better matched with their queries (e.g., larger weight on the relevance features such as BM25).

Our strategy is thus to describe users as a mixture of topics and to assume that each search task is motivated by choosing a topic of interest first and subsequently a query to describe that search task from the catalogue of words consistent with that particular interest(/topic). Once the query is issues, based on the documents clicked by the user, our model updates the parameters associated with user's click preferences.

3.1.1 *Incorporating Topic-level preference*

As a next step, we augment the framework proposed above by incorporating topic-level click patterns shared by queries belonging to each topic. We postulate that similar to how different users might prefer different ranking features, topics themselves can have an inherent preference over click patterns. Queries belonging to topics like education might warrant documents which better match the query to rank higher while some other topic like News might warrant high authority websites to rank higher than others. We aim to capture this topic-level feature ranking preference by incorporating a parameter β_t for each topic t which would correspond to the relative importance of ranking features for queries arising from the particular topic t.

3.2 Modelling users in terms of search tasks

Existing user modelling methods for web search rely heavily on per user topical interests and hence, fail to differentiate between users which share similar topical interests. We postulate that in web search setting, search logs contain information about various actions that users perform and profiling users based on search tasks would better capture the heterogeneity in user information. Our goal here is to use search log data to create a list of global search tasks. One possible way is to follow the approach of task discovery as proposed in [?] wherein a task is defined as the maximal subsequence of possibly non-consecutive queries in referring to the same latent user need which makes the set of all user tasks a partitioning of the set of all user queries. We formulate the task discovery problem as follows: given a query log QL and a user u, let T_u be the set of user tasks discovered by a query partitioning scheme π; the user task discovery problem can then be described as finding the best query partitioning strategy π^* that approximates the actual set of user tasks Θ such that:

$$\pi^* = argmax_{\pi}\xi(\Theta, T, \pi) \tag{1}$$

where function $\xi(.)$ is an accuracy measure which evaluates how well the query partitioning strategy π approximates the actual user tasks Θ. We use cosine similarity to measure this accuracy. This step is followed by clustering the user tasks identified to obtain universal tasks across all users. The final set of user tasks obtained are represented by a set of query terms and henceforth define the set of tasks used for experiments.

Based on the extracted search tasks, we construct a user-task association matrix which represents the search tasks users have been involved with. For each user u_i, we create a bag-of-queries representation from the list of queries issued by the user and compare each user with each of these search tasks t_j obtained above. For each user-task $<u_i, t_j>$ pair, we populate the corresponding value in the user-task association matrix (R) with the cosine similarity score (r_{ij}) we obtain for the pair. For tasks in which users do not have any matching queries, we assign a score of 0 to the corresponding pair. We model the user-task association in terms of probabilistic matrix factorization problem and learn latent vector representation for each user from the user-task association matrix by fitting a probabilistic model.

We further augment our task based user profiles by incorporating user's topical interest profiles and next describe our tensor based approach for the joint model.

3.3 Coupling Tasks & Topics

Traditional approach to user modelling have heavily relied on building user's topical profiles and developing personalized systems based on the learnt topical profiles. We augment our task based user representation model with user's topical information by coupling the topical interest with task based information in the form of a tensor and learning user profiles based on the decomposition of the $< user, topic, task >$ tensor. We next describe the overall system in detail.

3.3.1 Learning Topical Interest Profiles

Given user's history of search queries, we aim to develop a topic interest model which captures user's interest distribution over different topics. While most of existing techniques rely heavily on using ODP-like topic categories for modeling user's topical interests, such methods severely restrict topic-coverage as search logs offer much richer content in terms of the number of topics spanned as well as the granularity level of the topics - models which learns latent topics in an unsupervised manner from search logs would help capture topical interests embedded in search logs. We make use of the Latent Dirichlet Allocation model to learn the latent set of topics embedded in the search log. We hypothesize that each search query is motivated by choosing a topic of interest first and subsequently a query is issued to describe that search need from the catalogue of words consistent with that particular topic. Based on this intuition, we learn a LDA based topic model and use the learnt model to do topical inference for each user to obtain a topic-distribution for the user over the set of learnt topics. We refer to this distribution as a user's ***topical profile***.

3.3.2 Coupling Topics & Tasks

Our main intuition behind leveraging both the topical profile as well as the search task profile of users is to better differentiate between users who share similar topical profiles. Consider the task of personalizing search results for a query like "building algorithm for user models". This query broadly belongs to the computer science topic and a software engineer would expect a different set of results than a PhD researcher. Indeed, a software engineer performs different search tasks than a researcher and by leveraging their search task information, the user model could help in ranking results accordingly. We formulate this intuition in our model by coupling task information with topical information on a per-user basis. We construct a 3-mode tensor $< user, topic, task >$ to jointly capture user's topical as well as search task based information. Next, we briefly describe the tensor formulation:

3.3.3 Constructing $< user, topic, task >$ Affinity Tensor

To jointly model the user's topical and task preferences, we construct a 3-mode tensor - users, topics and tasks. A third order tensor can be represented as $T \in \Re^{I_1 \times I_2 \times I_3}$ with each element of the tensor denoted as $t_{i,j,k}$ with $i \in (1, I_1)$, $j \in (1, I_2)$ and $k \in (1, I_3)$. The symbol $\circ$ represents the vector outer product. Each element of our tensor ($T \in \Re^{I_1 \times I_2 \times I_3}$), $t_{i,j,k}$ defines user i's combined task based and topical preference - a user's participation in a certain task gets weighted by his topical affinity, thereby coupling his task based and topical affinity. More formally, we define each tensor-component value as follows:

$$t_{i,j,k} = U_{i_{topic_j}} \times U_{i_{task_k}} \tag{2}$$

where $U_{i_{topic_j}}$ is user U_i's topical affinity for topic j obtained from the LDA model learnt before while $U_{i_{task_k}}$ represents the task affinity for user U_i's for search task k obtained in earlier the user-task association phase. I_1, I_2, I_3 are the different dimensions of the different modes of the tensor - in our case, these represent the number of users, number of topics and the number of search tasks extracted respectively. Thus, for each user we construct his coupled task-topic affinity value and populate the corresponding component in the tensor T.

3.3.4 Tensor Decomposition

Tensor decomposition methods are regarded as higher-order equivalents to matrix decompositions. The PARAFAC tensor decomposition [?] allows us to leverage connections between the different users across different topics and different search tasks. By PARAFAC, the input tensors are transformed into Kruskal tensors, a sum of rank-one-tensors. Formally, the tensor $T \in \Re^{I_1 \times I_2 \times I_3}$ is decomposed into com-

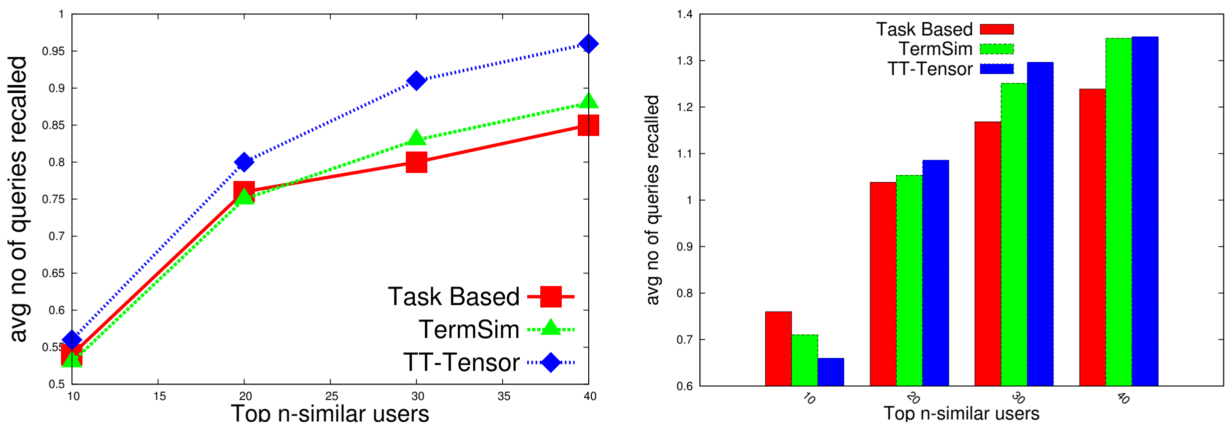

Figure 1: Performance on Collaborative Query Recommendation (left figure: top-10 recommended queries & right figure: top-20 recommended queries). The average number of query matches between the recommended set of queries and user's own test set of (unseen) queries is plotted against the number of similar users considered n where n refers to the number of similar users considered.

ponent matrices $U \in \Re^{I_1 \times d}$, $T \in \Re^{I_2 \times d}$ and $S \in \Re^{I_3 \times d}$ and d principal factors λ_i in descending order. Via these, tensor T can be written as a Kruskal tensor by:

$$T \approx \Sigma_{k=1}^{d} \lambda_k \cdot U^k \circ T^k \circ S^k \tag{3}$$

where λ_k denotes the k-th principal factor. The goal is to compute a decomposition with d-components that best approximates our tensor T, i.e., to find

$$min_{\widetilde{T}} \|T - \widetilde{T}\| \tag{4}$$

such that

$$\widetilde{T} = \Sigma_{k=1}^{d} \lambda_k \cdot U^k \circ T^k \circ S^k \tag{5}$$

We make use of the Alternating Least Squares (ALS) approach to solve the above objective - having fixed all but one matrix, the problem reduces to a linear least-squares problem. In our experiments, we pre-assume a value of $d = 20$ thereby making each user representation a 20-dimensional vector which we use for our experiments.

Overall, the above formulation helps us to couple user's topical interests with their search task associations and learn a user representation based on this coupled tensor. This tensor decomposition based user modelling approach allows us to use multi-modal user information and leverage insights from each of them while learning user representations. An important characteristic of the proposed tensor based approach is that this method is generic enough and allows us to plug-in other sources of user information - click models, data from advertisement responses, etc.

3.4 Extending user representations to hierarchies

The user modeling techniques described above represent users in a flat d-dimensional representation so far. We postulate that hierarchical representations for users would better capture the heterogeneity in user information and aid in personalization. The challenges in moving from a flat representation to a hierarchical are many including coming up with algorithms to find similarity functions which make use of non-flat hierarchical structure based representations. We intend to extend our work to include more structure based user representations and compare them against traditional ways of representing users.

4. PROGRESS SO FAR

The research so far has focused on evaluating the benefits of learning user profiles based on the search tasks users are involved with. We postulate that in a web search setting, a user representation based on the search tasks users' perform would better capture user actions, interests and preferences. Given a search log, we extract search tasks performed by users and find user representations based on these tasks. More specifically, we construct a user-task association matrix and borrow insights from *Collaborative Filtering* to learn a low-dimensional factor model wherein the actions/interests/preferences of a user are determined by a small number of latent factors. By applying probabilistic matrix factorization to the user-task association matrix, we learn *task-based user representations* for each user.

A good user profile for query recommendation should capture a user's specific interests & informational needs. Based on this intuition, we evaluate performance of the the task-based user modeling approach on *Collaborative Query Recommendation* where the goal is to recommend queries to a user based on queries issued by similar users. We calculate the weighted frequency of a candidate query for 10 most similar users of the target user u, and selected the top n queries as recommendation. We make use of the AOL log dataset which consists of ∼20M web queries collected over three months and use data for about ∼1200 users who have issued more than 550 queries. We run our Task Discovery algorithm on the set of queries for each of these users which results in a total of ∼0.12M tasks which we cluster using cosine similarity score to obtain a set of 1529 search tasks using which we create the user-task association matrix. Our baseline(*TermSim*) is a method that only uses bag-of-words based representation for each user where the terms are extracted from user queries & similar users found using cosine similarity between each user's bag-of-word based representations. We consider the test-set of queries in the target user as relevant, and computed average number of relevant queries matched in the recommendation query set as the performance metric.

We plot the average number of query matches between the recommended set of queries and user's own test set of (unseen) queries against the number of similar users considered n where n refers to the top-10 (left) and top-20 (right) query suggestions from n-most similar users. Our initial results [14] (Figure 1) show that the proposed Topic-Task Tensor based user modelling approach(*TT-Tensor*) performs better than *TermSim* as well as *TaskBased* which demonstrates that combining search task information with user's topical interests thus help us better capture different aspects of user profiles and can serve as potent user modelling tools. Since *TermSim* relies strictly on term matching for measuring user similarities, its coverage is limited: it might not capture insights for the users with too few queries or those who shared the same search interest but issued different queries or performed different tasks. Task based user modelling can help in better differentiating between users which have similar topical interests but perform different tasks. The proposed tensor based approach combines the best of both the worlds and hence was able to leverage the topical user profile information with the task aspect.

5. ISSUES FOR DISCUSSION

The major issue for discussion revolves around the fallacies in existing approaches to personalization and understanding realistic expectations from personalization systems. The question of what makes a good personalization system needs to be addressed which would in itself set some basic standards for future personalization systems and help streamline the research in personalization. Another important issue worth deliberation is the role of evaluation in personalization. While industrial researchers have access to a massive user centric data, the amount of research possible from outside the industry in an academic setting is rather limited. How can we negate this factor and ways of coming up with possible remedies is worth discussing. Additionally, while the current research plan is majorly motivated for web search based personalization, an important question is how well these personalization approaches generalize to other domains. Can we come up with domain-agnostic personalization techniques? Would transfer learning play a role in this regard? Should we start looking at cross-domain personalization techniques to harness the rich multi-domain user data we currently have.

6. CONCLUSION

Understanding preferences and informational needs of users is a complex task. In this research, we intend to propose techniques of user modeling aimed at capturing the vast heterogeneity among users thereby helping in developing better personalized systems. The proposed research aims to develop varied user models based on their informational preferences, their topical interests as well as their search task behaviours alongside extending traditional flat representations to richer hierarchies.

7. ACKNOWLEDGEMENTS

This work was supported in part by a Google Faculty Research Award.

8. REFERENCES

1 Andrew Postlewaite. Social norms and social assets. Annual Review of Economics, 3:239-259, 2011.

2 Ehsan Abbasnejad, Scott Sanner, Edwin V. Bonilla, Pascal Poupart: Learning Community-Based Preferences via Dirichlet Process Mixtures of Gaussian Processes. IJCAI 2013.

3 White, R. W., Chu, W., Hassan, A., He, X., Song, Y., Wang, H. (2013, May). Enhancing personalized search by mining and modeling task behavior. In Proceedings of the 22nd international conference on World Wide Web (pp. 1411-1420). International World Wide Web Conferences Steering Committee.

4 Blundell, C., Teh, Y. W. (2013). Bayesian Hierarchical Community Discovery. In Advances in Neural Information Processing Systems (NIPS 2013) (pp. 1601-1609).

5 Kim, Jin Young, Kevyn Collins-Thompson, Paul N. Bennett, and Susan T. Dumais. "Characterizing web content, user interests, and search behavior by reading level and topic." In Proceedings of the fifth ACM international conference on Web search and data mining, pp. 213-222. ACM, 2012.

6 White, Ryen W., Wei Chu, Ahmed Hassan, Xiaodong He, Yang Song, and Hongning Wang. "Enhancing personalized search by mining and modeling task behavior." In Proceedings of the 22nd international conference on World Wide Web, pp. 1411-1420. International World Wide Web Conferences Steering Committee, 2013.

7 Bennett, Paul N., Ryen W. White, Wei Chu, Susan T. Dumais, Peter Bailey, Fedor Borisyuk, and Xiaoyuan Cui. "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, pp. 185-194. ACM, 2012.

8 Matthijs, Nicolaas, and Filip Radlinski. "Personalizing web search using long term browsing history." In Proceedings of the fourth ACM international conference on Web search and data mining, pp. 25-34. ACM, 2011.

9 Wang, Hongning, ChengXiang Zhai, Feng Liang, Anlei Dong, and Yi Chang. "User Modeling in Search Logs via a Nonparametric Bayesian Approach." (WSDM 2014).

10 Teh, Yee Whye, Michael I. Jordan, Matthew J. Beal, and David M. Blei. "Sharing Clusters among Related Groups: Hierarchical Dirichlet Processes." In NIPS.

11 Wang, Chong, John W. Paisley, and David M. Blei. "Online variational inference for the hierarchical Dirichlet process." In International Conference on Artificial Intelligence and Statistics, (AISTATS 2011) pp. 752-760. 2011.

12 Ahmed, Amr, Yucheng Low, Mohamed Aly, Vanja Josifovski, and Alexander J. Smola. "Scalable distributed inference of dynamic user interests for behavioral targeting." In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 114-122. ACM, 2011.

13 D. Blackwell and J. MacQueen. Ferguson distributions via polya urn schemes. The Annals of Statistics, 1973.

14 R. Mehrotra, E. Yilmaz and M. Verma. "Task-Based User Modelling for Personalization via Probabilistic Matrix Factorization." In Proceedings of the 8th ACM Conference on Recommender Systems (RecSys), 2014.