ECNU at CLEF PIR 2018 : Evaluation of Personalized Information Retrieval

Qingchun Bai[1], Jiayi Chen[1], Qinmin Hu[2] and Liang He[1]

[1] School of Computer Science and Software Engineering, East China Normal University, Shanghai, China
{qchbai, jychen}@ica.stc.sh.cn, lhe@cs.ecnu.edu.cn
[2] Department of Computer Science, Ryerson University, Toronto, Canada
vivian@ryerson.ca

Abstract. Personalized Information Retrieval (PIR) is an effective solution when purposes of queries are issued but users receive the same results. The PIR-CLEF 2018 task aims to explore the methods and evaluations of PIR. By analyzing the provided data we generate query level and session level baselines. We compare baselines and extended models we propose, and experiment results show that insufficient relevance information has a negative impact on the performance of models and evaluation process. Since personalization ranking based on typical users interests is not effective in reality, especially when the results of relevance feedback is not satisfactory, we consider that the PIR task should not only relate to context, but to the various search intentions. We propose several suggestions about data and evaluation process.

Keywords: Personalized Information Retrieval; Query Expansion; Data Analysis

1 Introduction

The PIR-CLEF 2018 task aims to explore the methods and evaluations of Personalized Information Retrieval (PIR). PIR has drawn great attention to help understand user's behaviors in the interaction with IR systems. Personalized search is an effective solution when queries purposes are issued but users received the same results.

Existing works [3,5,4] have proved that personalized ranking can be considered as a good solution for PIR task. The foundation and key of personalized ranking service is how to obtain persons attempt to frame user's interest model. Various penalization strategies are carried out, for example, In [7], a method is discussed to identify user's interest automatically, which based on the assumption that a user's general preference may help the search engine disambiguate the true intention of a query. The approach described in [8] considered a user's prior interactions with a wide variety of content to personalize that user's current web search. More recently, in [1], a dynamic personalized ranking model is proposed to recommend the most relevant information which combined different sources of information.

For another piece of research, the research focus on understanding the user's intent of search session information[2,6,9,10], results show that it is possible to understand the user's intent, since people all have intentions in the process of seeking information and also have reasons to believe these information seeking intentions.

These prior works on personalized information retrieval have focused on independent issues with independent data. A few of them also have focused on the analysis of required data and evaluation of personalized ranking. In fact, we consider that personalized ranking based on typical user's interests is not effective in reality, especially when the results of relevance feedback are not good, the re-ranking model cannot achieve the desired result.

Therefore, in this study, we aim to explore the potential of the task and understand both the data and evaluation of the personalized search and ask the following research questions:

- If the provided PIR data can satisfied the task?
- How to achieve personalization, and what kind of data is needed to support this research?
- How to evaluate it?

To achieve this aims, this paper is organized as follows. In section 2, we briefly review and analysis the current PIR data. In section 3, we describes our baselines method to the data in detail. In section 4, experiments and results are presented. Finally, we discuss the task and evaluation in section 5.

2 Data Review and Analysis

We present the statistics about dataset in this section, the review of current dataset described in Section 2.1, and explores the potenital of the dataset. Then we analysis the query session given by the data in Section 2.2.

Statistics about Dataset In this section, we briefly review current dataset of PIR task and provide a comprehensive analysis on this task. In PIR-CLEF 2018, data are provided with six csv files including information below:

- the search tasks (sessions) of ten users;
- the queries submitted by all users and all documents returned by ClueWeb API;
- relevance scores labeled by users and original ranks of documents;
- personal information like gender and job;
- remarks written by users;
- statistical information of terms in queries.

Statistics about Sessions A user can submit several queries in a query session. These queries are aiming at different objectives. To find the objective of the user, we gather all queries as one query to represent the objective. We need to submit this query to the API and evalutate the performance.

3 Methods

The users submit their queries to the ClueWeb API[3] and annotate whether the returned documents are relevant. The users divide relevance into four grades: relevant, somewhat relevant, not relevant and off topic with scores ranging from four to one. According to the description above, we define that documents are relevant to the query only when those scores are four. Figure 1 shows the framework of the personalized ranking part.

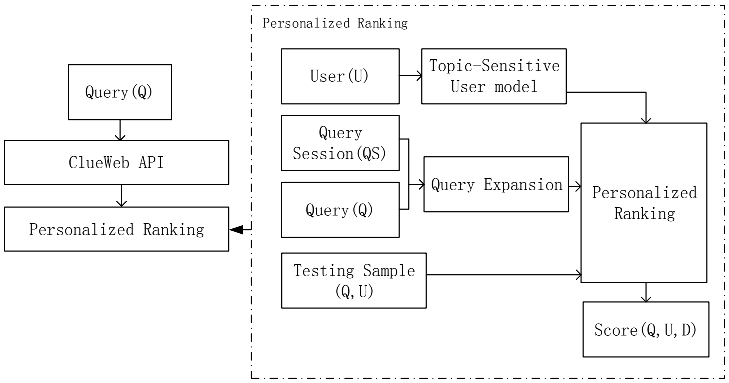

Fig. 1: Personalized Ranking Framework.

Baselines We propose two baselines: query-level baseline and session-level baseline. In query-level baseline, we evaluate each query independently. While in session-level baseline, we collect all relevant documents of queries in the session and consider them as the relevant documents of the search task. We evaluate the performance of each query on its search task. We also assume that queries belonging to one session represent different aspects of user's need. So we sum up all queries in a session to one and submit it to the ClueWeb API. We evaluate and display the performance of each session baseline in Table 1.

Query Expansion We first take user's feedback into account. After the user labels the documents with relevance score, we choose ten words with highest frequency in relevant documents as expansion words and add them to the original session-level query. We submit the new queries to the API and evaluate the performance of this method. Then we assume top-20 documents returned from API are relevant and select ten most frequent words in documents as expansion words. The first method is listed in Table 2 with suffix "RF" while the second one is named with suffix "PRF".

[3] http://clueweb.adaptcentre.ie/WebSearcher/search?query=queryString &page=pagenumber

Topic-Sensitive User model We propose a language modeling approach to personalized search based on users' search behavior and preference. To capture the user's searching interesting and implicit purpose, we propose to use LDA-based approach for simulating users which does not merely focus on simulating the search behaviours but also considers search sessions of the task.

4 Experiments and Results

4.1 Performance of Each Query Session

Table 1 shows the performance results of each query session, there are total 14 automatic session IDs and we obtain a list as follows. "Sum Up" means the performance of the new query generated by all queries in a session and "Single Query" is mean performance of all queries in a session. From this table, we observe that different session have the wide variations performance. The best result is session 156, the category of the query is "Travel". The user gives a higher relevance score which denotes the relevance of the document to the topic (1 off-topic, 2 not relevant, 3 somewhat relevant, 4 relevant). the description of the users is "The relevant documents are documents that list as many historical and popular places in venice. I don't want to see other documents that talk about other related places. In addition to that I am not interested about accomodation during my search."

Table 1: Performance of Each Query Session

Session ID	Sum Up				Single Query			
	MAP	NDCG	P@5	P@10	MAP	NDCG	P@5	P@10
107	0.1581	0.5546	0.2000	0.3000	0.1069	0.2241	0.4222	0.3666
154	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
156	0.7588	0.9294	1.0000	1.0000	0.3878	0.5295	0.5000	0.5000
161	0.1321	0.4625	0.2000	0.3000	0.0839	0.1985	0.2800	0.2600
162	0.1621	0.5695	0.2000	0.3000	0.0702	0.2152	0.3333	0.3667
172	1.0000	1.0000	0.2000	0.1000	0.2000	0.2000	0.0400	0.0200
173	0.0890	0.3940	0.4000	0.4000	0.0329	0.1035	0.2375	0.1625
175	0.0120	0.1947	0.0000	0.0000	0.0523	0.1751	0.1666	0.1666
176	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
201	0.0990	0.3851	0.2000	0.1000	0.0893	0.2334	0.2000	0.1400
202	0.3711	0.6418	0.2000	0.1000	0.3711	0.6418	0.2000	0.1000
203	0.4052	0.6232	0.6000	0.4000	0.4052	0.6232	0.6000	0.4000
204	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
205	0.0060	0.1706	0.0000	0.0000	0.1145	0.1448	0.0800	0.1000
mean	0.2281	0.4232	0.2285	0.2142	0.1367	0.2349	0.2185	0.1844

4.2 Performance Comparison

We further make a comparison between baselines and our methods in Table 2. In query-level evaluation, our methods are worse than baselines. In session-level evaluation, relevance feedback method performs better than baselines because we can obtain users' interests from relevant documents. However pseudo relevance feedback and LDA methods get worse performance than baseline.

Table 2: Performance Comparison

Methods	MAP	NDCG	P@5	P@10	P@20
Query-Baseline	0.2216	0.3239	0.1922	0.1688	0.0929
Query-RF	0.1084	0.1908	0.1013	0.0000	0.1908
Query-PRF	0.1460	0.2300	0.1169	0.0935	0.0623
Session-Baseline	0.2283	0.4232	0.2286	0.2143	0.1619
Session-RF	0.2743	0.4251	0.4000	0.2714	0.2190
Session-PRF	0.0695	0.1549	0.1286	0.1071	0.0929
Session-LDA	0.1549	0.3005	0.1714	0.1786	0.1357

5 Discussion

It is within our expectation that all query-level evaluations are worse than that of baseline. We think this phenomenon is caused by the lack of relevant documents. In the provided data, each user only labels about twenty documents whose original ranks range from 0 to 100 and few of them are relevant. Assuming the users get 100 documents returned per query, eighty percent of relevance information is lost. In this scenario, any document not occurring in this list is considered as irrelevant which means a relevant document can be annotated as irrelevant from the perspective of user. Insufficient relevance information even make us hard to evaluate on certain queries. In session 154, 176 and 204, all documents are irrelevant so that even though we find the potentially relevant documents, we cannot know whether they are relevant from the angle of users.

We also think the evaluation process should be upgraded. Unlike existing search task like TREC tracks, personalized information retrieval focuses more on individual differences. In PIR-CLEF 2018, some users receive the same task but the queries they submit are different. For example, user 8,11 and 12 receive the same task of traveling, but their queries are about Dublin, Tokyo and Barcelona. The individual differences are expressed by queries so we think this task is still an ad-hoc retrieval task. So if we want to focus on individual differences, we need more users to join in the data collection.

We suggest that the complete logs of users can be provided. By analyzing the relevant documents annotated by users, we get an improvement in session aspect as listed in Table 2. However our method still can be improved. In this task we are provided with users' actions such as opening document and submitting a query. But we think these data is not sufficient enough because only part of actions are provided so that we cannot analyze users' preference by their actions.

In conclusion, we put up with three suggestions. The first one is that more complete relevance labels should be provided. Then we think more participants can join in the data collection to provide more personalized data. The last one is that we think detailed user actions can help improve the performance.

6 Conclusions

We have proposed a view of PIR task that implies that personalization should be with respect not only to context, but to the various information that people have during the course of an information search session. We focus on taking user's feedback into account and propose two extend models : Query Expansion method and Topic-Sensitive user model. We first conduct experiments on each query session, results show that different session have the wide variations performance. Then we compared baselines with extended models. Noting that topic-sensitive strategy does not work very well, insufficient relevance information has a negative impact on the performance of models and evaluation process. We will extract more useful features and focus on the learning to rank approaches in the future.

References

1. E. Ali. Dynamic personalized ranking of facets for exploratory search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1379–1379. ACM, 2017.
2. Z. Carevic, M. Lusky, W. van Hoek, and P. Mayr. Investigating exploratory search activities based on the stratagem level in digital libraries. *International Journal on Digital Libraries*, pages 1–21, 2017.
3. Z. Dou, R. Song, and J. R. Wen. A large-scale evaluation and analysis of personalized search strategies. In *International Conference on World Wide Web*, pages 581–590, 2007.
4. W. Guo, S. Wu, L. Wang, and T. Tan. *Multiple Attribute Aware Personalized Ranking*. Springer International Publishing, 2015.
5. W. Guo, S. Wu, L. Wang, and T. Tan. Personalized ranking with pairwise factorization machines. *Neurocomputing*, 214(C):191–200, 2016.
6. M. Mitsui, J. Liu, N. J. Belkin, and C. Shah. Predicting information seeking intentions from search behaviors. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1121–1124. ACM, 2017.
7. F. Qiu and J. Cho. Automatic identification of user interest for personalized search. In *Proceedings of the 15th international conference on World Wide Web*, pages 727–736. ACM, 2006.
8. J. Teevan, S. T. Dumais, and E. Horvitz. Personalizing search via automated analysis of interests and activities. In *ACM SIGIR Forum*, volume 51, pages 10–17. ACM, 2018.
9. W. van Hoek and Z. Carevic. Building user groups based on a structural representation of user search sessions. In *International Conference on Theory and Practice of Digital Libraries*, pages 459–470. Springer, 2017.
10. G. H. Yang, X. Dong, J. Luo, and S. Zhang. Session search modeling by partially observable markov decision process. *Information Retrieval Journal*, 21(1):56–80, 2018.