Revisiting Information Retrieval Tasks with User Behavior Models

Yiqun Liu
Tsinghua University
yiqunliu@tsinghua.edu.cn
and
Jiaxin Mao
Tsinghua University
maojiaxin@gmail.com

Analyzing and modeling user behavior in the Information Retrieval (IR) process is important for IR research. In this paper, we propose a new research schema in IR research. In this schema, we first investigate humans' cognitive process when completing a specific IR task and build cognitive models for that task. The findings in the investigation and the proposed cognitive models are then utilized to improve machines' performance in the IR task. Through this research schema, we revisit three IR tasks. In the first study, we carefully analyze users' clicking behavior in mobile search and propose a mobile click model to extract relevance feedback from the mobile search logs. In the second study, we conduct an eye-tracking study to investigate how human assessors read a document during relevance judgment task and adopt the findings in building a novel retrieval model that can better approximate humans' relevance judgment. In the last study, we conduct another eye-tracking study to investigate humans' reading behavior when completing the reading comprehension task. We build a prediction model for user attention and leverage the predicted attention signals to improve the machine reading comprehension model. By successfully adopting this research schema to three different IR tasks, we demonstrate its effectiveness and generalizability.

DOI: 10.1145/3352683.3352686 http://doi.acm.org/10.1145/3352683.3352686

1. INTRODUCTION

During the last three decades, the Web has enabled billions of users to access a huge amount of diverse information, including hypertext documents, images, video, audio, and social media. However, it also causes information overload for the individual user with a limited cognitive capacity. Therefore, Information Retrieval (IR) applications, especially the Web search engines, have become one of the most important and most popular applications on the Web because they can help users locate relevant information within a second.

The information retrieval process is essentially a "human-in-the-loop" process. It begins when a user has an information need in mind and submit a *query* to the IR system (e.g. a

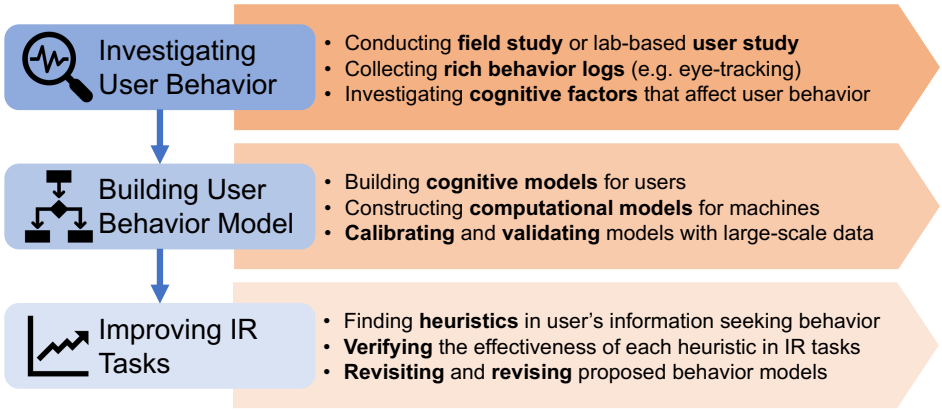

Fig. 1. Steps in constructing behavior-oriented models for IR tasks.

Web search engine) accordingly. Receiving the query, the search engine will try to infer the user's information need, retrieve a set of information objects (e.g. web pages, images, video, and etc., also called *documents*), and rank them by their relevance to the user's information need. The user will then examine and interact with the ranked list of information objects. If her information need cannot be fulfilled by the returned documents, the user can reformulate her query and restart the loop. Due to the "human-in-the-loop" nature, modeling user behavior in the IR process is vital for the IR research and the design of effective IR systems. Previous studies have proposed a variety of *user behavior models* and leveraged them in different IR tasks including deriving an optimal ranking principle [Robertson 1977] [Fuhr 2008], designing plausible evaluation metrics for IR systems [Moffat et al. 2013], and extracting relevance feedback from user behavior logs [Joachims et al. 2005] [Chuklin et al. 2015].

While the IR community has paid considerable attention to modeling user behavior, we notice that some existing user behavior models in IR are based on oversimplified assumptions, and therefore, may fail to accurately model real user's behavior in the complex and constantly evolving search environment. For example, in the traditional Web search environment where "ten blue links" will be returned for each query, we can assume that the user will examine the search results linearly and click all relevant results until her information need is satisfied. Based on this assumption, a simple yet effective model can be constructed to infer relevance information from click logs [Chapelle and Zhang 2009]. However, modern Web search engines often federate highly heterogeneous vertical results into the search engine result pages (SERPs), thus this assumption may not always hold and the corresponding user behavior model may not be valid. On the one hand, some vertical search results with rich media content are more visually attractive than other results, which may alter the user's examination order on them [Liu et al. 2015]. On the other hand, in order to reduce the interaction cost of users, some vertical results directly present useful information in the snippets. The user does not need to click these results even if they are relevant, which also violates the above assumption.

Having noticed the weakness of existing user behavior models in IR, we proposed a new research schema and revisited three IR tasks that heavily rely on user behavior models. In this research schema, we regard users' search process as a *cognitive* process which means that we not only care about modeling users' behavior but also want to investigate the latent cognitive process behind their explicit behavior patterns.

As shown in Figure 1, the first step of this research schema is to carefully investigate user behavior in the IR process. In this step, we build our own experimental search platform and conduct field studies and lab-based user studies for different IR tasks. While field studies let us investigate user behavior in a more natural settings, in the lab-based user study, we can collect rich user behavior information, such as the eye-tracking data, in a more controlled environment.

Based on the findings in the first step, we build user behavior models in the second step. The user behavior models can categorized as two broad classes: 1) cognitive models for users to characterize, predict, and explain their behavior in a specific IR task; 2) computational models that enable machines to automatically complete the IR task. It is also important to empirically calibrate and validate the proposed models. Therefore, after defining the model, we will train and test it with large scale datasets collected from the field studies or the query logs from commercial search engines.

In the last step of the research schema, we leverage the findings and user behavior models to improve the performance of the IR tasks. We first get can inspirations and heuristics from users' information seeking behavior. Then we can verify their effectiveness in improving the performance of IR tasks. Finally, we can revisit and revise the proposed models with effective heuristics inspired by human behavior.

In the rest of this paper, we will give examples on how we adopt this research schema by briefly summarizing how we revisit three IR tasks with this schema, respectively in Section 2–4. After that, we will conclude this paper and discuss future directions in Section 5.

2. EXAMPLE STUDY 1: USER BEHAVIOR MODELING FOR MOBILE SEARCH RANKING

By modeling users' click behavior as a stochastic process, click models can be used to extract unbiased relevance feedback from the biased click logs. However, users' search behavior in the mobile environment is fundamentally different from that in the traditional desktop search environment, which may render the existing click models less effective in the mobile environment.

2.1 Investigating User Behavior in Mobile Search

To address this problem, we first conducted a user study to investigate users' search behavior on mobile devices [Mao et al. 2018] and found that some vertical results in mobile search have a low *click necessity*, which means that they directly present rich information on the SERP, and therefore, can be useful for users even without being clicked. A further analysis based on click logs from a mobile search engine also showed that on average the vertical results in mobile search have a relatively low click-through rate, which could be attributed to that many vertical results on mobile SERPs can directly satisfy users' needs

without being clicked.

2.2 Constructing Click Models for Mobile Search

Failing to take the click necessity into consideration, a click model may mistakenly interpret a low click-through rate on the results with a low click necessity as a negative relevance feedback from users. Therefore, we need to construct a new click model for mobile search. In a follow-up study [Mao et al. 2018], we proposed two behavioral biases that are prevalent in mobile search and incorporated them into a novel click model named Mobile Click Model (MCM).

—**Click Necessity Bias**: *Some types of search results (e.g. the knowledge graph and direct answer results) have low click necessity because they can satisfy users' information needs without requiring any clicks, which will lower the click probabilities of these results.*

—**Examination Satisfaction Bias**: *A user can feel satisfied and leave the SERP after examining a search result that is both attractive and with low click necessity.*

To incorporate the **Click Necessity Bias**, we extended the examination hypothesis [Craswell et al. 2008] by assuming that a user will click a search result if and only if: 1) she has examined it; 2) it appears to be relevant and attractive; 3) she needs to click it to get useful information. We use C_i, A_i, and E_i to denote whether the user click result i, whether the result is attractive, and whether the user has examined the result. We further introduce a binary variable N_i to represent the click necessity of the result. $N_i = 1$ indicates that the user needs to click the result to get useful information and $N_i = 0$ indicates that the user can be satisfied without clicking it. Therefore, our assumption can be formally written as:

$$C_i = 1 \iff E_i = 1 \wedge A_i = 1 \wedge N_i = 1 \tag{1}$$

For the **Examination Satisfaction Bias**, we use a binary variable S_i^E to denote whether the user is satisfied just by *examining* result d_i (examination satisfaction), S_i^C to denote whether the user is satisfied after *clicking* it (click satisfaction). We assume that: 1) a user will be satisfied once she encounters either an examination satisfaction event ($S_i^E = 1$) or a click satisfaction event ($S_i^C = 1$); 2) once the user is satisfied, she will not examine follow-up results. We further use S_i to denote user's *state of satisfaction* after position i. Therefore, we have:

$$S_i = 1 \iff S_{i-1} = 1 \vee (S_i^E = 1 \vee S_i^C = 1) \tag{2}$$

$$P(E_i = 1 | S_{i-1} = 1) = 0 \tag{3}$$

Theoretically, MCM extends the examination hypothesis and can be seen as a unified generalization of two most widely-adopted click models, DBN [Chapelle and Zhang 2009] and UBM [Dupret and Piwowarski 2008]. Empirically, experiments on large-scale mobile search logs show that MCM achieves substantial performance gains over the baseline models in predicting user clicks in mobile search.

2.3 Search Result Ranking with MCM

After training MCM on mobile click logs, we could use its learnt parameters to compute a relevance score for each mobile search result in the logs. This score could be used to rank the search results according to users' implicit relevance feedbacks. We defined the relevance score as the probability of becoming satisfied when a user examines a result d_i of type v_i:

$$\begin{aligned} score(q, d_i) &= P(S_i = 1 | E_i = 1) \\ &= P(A_i = 1)[P(N_i = 1)P(S_i^C = 1) + (1 - P(N_i = 1)P(S_i^E = 1)] \end{aligned} \tag{4}$$

Experiments on real mobile search logs demonstrated that we could significantly improve the ranking performance of mobile search with the relevance score estimated by MCM.

3. EXAMPLE STUDY 2: READING BEHAVIOR INSPIRED RELEVANCE ESTIMATION

Retrieval models aim to estimate the relevance between a document and a query. Ideally, this relevance estimation should approximate the relevance judgment made by users. However, existing retrieval models work in a rather different manner from how humans read the document and make relevance judgment. By investigating how humans read during the relevance judgment and re-examining the existing retrieval models, we may fill this gap and make retrieval models' relevance estimation more consistent with humans' relevance judgment.

3.1 Investigating Reading Behavior in Relevance Judgment

In the second study, we first conducted an eye-tracking study [Li et al. 2018] with 29 participants to investigate human assessors' reading behavior, especially how they allocate their attention on different parts of the document, during a specific information retrieval task. In the eye-tracking study, we required each participant to make 4-level relevance judgment for 60 documents from 15 topics and recorded her eye-movement during the whole relevance judgment process as signals for her reading attention. With the collected data, we investigated how users' attention was affected by different factors including the position bias, linguistic features, search task types, and query terms.

3.2 Building Reading Models for Relevance Judgment Task

Through a systematical investigation, we further found that human assessors' reading process can be characterized by a two-stage reading model. In the first stage, users tend to allocate a high level of attention in reading the beginning of a document to form a preliminary relevance judgment. Then in the second stage, users will adopt different reading strategies based on the preliminary relevance judgment. They will either read the document to acquire relevant information or skim the document to validate the preliminary judgment.

3.3 Improving Relevance Estimation with Reading Inspired Model

Based on the analysis and findings on humans' reading behavior during relevance judgment, we re-examined existing retrieval models and proposed new ones in a follow-up study [Li et al. 2019]. We first summarized six *reading heuristics* from users' reading patterns:

—**Sequential reading**: Reading direction is from top to bottom.
—**Vertical decaying attention**: Reading attention is decaying vertically.
—**Query centric guidance**: Reading attention is higher in the contexts around query terms.
—**Context-aware reading**: Reading behavior is influenced by the relevance perception from previously read text.
—**Selective attention**: Users will skip some seemingly irrelevant text during relevance judgment.
—**Early stop reading**: Users will stop reading once the read text is enough to make relevance judgment.

Then, we reviewed existing retrieval models with these reading heuristics and found that none of them satisfies all the heuristics. Therefore, we integrated all the effective reading heuristics into a new retrieval model called Reading Inspired Model (RIM). Experiment results on a publicly available datasets show that by integrating more reading heuristics, the RIM outperforms most existing retrieval models. A deeper investigation into the trained RIM further suggests that the RIM can capture users' selective attention and early stop reading patterns.

4. EXAMPLE STUDY 3: HUMAN BEHAVIOR INSPIRED MACHINE READING COMPREHENSION

Machine Reading Comprehension (MRC) is an emerging task in NLP and IR research. Recently, a number of deep neural models have been proposed to solve the MRC tasks, and have achieved a performance comparable to humans in some simplified MRC task settings such as SQuAD [Rajpurkar et al. 2016]. However, there is still a large gap between the performance of humans and machines in more practical and challenging MRC settings.

4.1 Investigating Reading Behavior in Reading Comprehension Task

Following the rationale and methodology we had adopted in Example Study 2, we believed that a better understanding of how human reads and allocates their attention during reading comprehension process could help to improve the performance of MRC tasks. Therefore, we conducted another eye-tracking study to investigate how humans read a document during the reading comprehension task. By analyzing the data collected from 32 participants, we found that after finding some possible answer candidates, the user would revisit the whole document before submitting the answer. For the document that contains an answer, the attention distribution in the last reading period concentrates on the beginning of the document, indicating that users are more likely to reread the document to verify the final

answer. We also found that human's attention distribution was affected by both question-dependent factors and question-independent factors.

4.2 Building Reading Models for Reading Comprehension Task

Such reading behavior in reading comprehension task could also be modeled as a two-stage process in which the first stage is to search for possible answer candidates and the second stage is to generate the final answer through a comparison and verification process. Based on the two-stage reading models, we could extract features to predict human's attention signals on the read document.

4.3 Improving MRC Performance with Predicted Attention

As we also found that the user tended to pay more attention on the answer text and by using the attention signal as features we could improve the performance of MRC tasks, we also tried to leverage the predicted attention signals as features in building the answer sentence retrieval model for the MRC task. Results showed that although there was a large gap between the models based on real and predicted attention, using the predicted attention signals as features could indeed improve the performance of answer sentence retrieval.

5. CONCLUSION

Because the user plays an important role in the IR process, modeling user behavior is vital for the IR research and the design of IR systems. In our recent work, we propose a new research schema in which we throughly investigate humans' cognitive process when completing some IR tasks and adopt the findings in building an improved computational model for the corresponding task. Through this research schema, we revisit three IR tasks: relevance feedback extraction, relevance estimation, and machine reading comprehension. These studies not only enhance our understanding of how human complete such tasks, but also guide the design of novel computational models for each task and substantially improve the task performance, which in turn demonstrate the effectiveness and generalizability of the proposed research schema. In future work, we will continue to explore how to adopt this research schema to other IR tasks, such as modeling users' session-level information needs and modeling relevance at the passage-level.

ACKNOWLEDGMENTS

This work is supported by Natural Science Foundation of China (Grant No. 61622208, 61732008, 61472206) and National Key Basic Research Program (2015CB358700). We thank Sogou.com for the anonymized mobile search log used in this work.

REFERENCES

CHAPELLE, O. AND ZHANG, Y. 2009. A dynamic bayesian network click model for web search ranking. In *Proceedings of the 18th international conference on World wide web*. ACM, 1–10.

CHUKLIN, A., MARKOV, I., AND RIJKE, M. D. 2015. Click models for web search. *Synthesis Lectures on Information Concepts, Retrieval, and Services 7,* 3, 1–115.

CRASWELL, N., ZOETER, O., TAYLOR, M., AND RAMSEY, B. 2008. An experimental comparison of click position-bias models. In *Proceedings of the 2008 international conference on web search and data mining*. ACM, 87–94.

DUPRET, G. E. AND PIWOWARSKI, B. 2008. A user browsing model to predict search engine click data from past observations. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 331–338.

FANG, H., TAO, T., AND ZHAI, C. 2004. A formal study of information retrieval heuristics. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 49–56.

FUHR, N. 2008. A probability ranking principle for interactive information retrieval. *Information Retrieval 11,* 3, 251–265.

JOACHIMS, T., GRANKA, L. A., PAN, B., HEMBROOKE, H., AND GAY, G. 2005. Accurately interpreting clickthrough data as implicit feedback. In *Sigir*. Vol. 5. 154–161.

LI, X., LIU, Y., MAO, J., HE, Z., ZHANG, M., AND MA, S. 2018. Understanding reading attention distribution during relevance judgement. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. ACM, 733–742.

LI, X., MAO, J., WANG, C., LIU, Y., ZHANG, M., AND MA, S. 2019. Teach machine how to read: Reading behavior inspired relevance estimation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM.

LIU, Z., LIU, Y., ZHOU, K., ZHANG, M., AND MA, S. 2015. Influence of vertical result in web search examination. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 193–202.

MAO, J., LIU, Y., KANDO, N., LUO, C., ZHANG, M., AND MA, S. 2018. Investigating result usefulness in mobile search. In *European Conference on Information Retrieval*. Springer, 223–236.

MAO, J., LUO, C., ZHANG, M., AND MA, S. 2018. Constructing click models for mobile search. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. SIGIR '18. ACM, New York, NY, USA, 775–784.

MOFFAT, A., THOMAS, P., AND SCHOLER, F. 2013. Users versus models: What observation tells us about effectiveness metrics. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. ACM, 659–668.

RAJPURKAR, P., ZHANG, J., LOPYREV, K., AND LIANG, P. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

ROBERTSON, S. E. 1977. The probability ranking principle in ir. *Journal of documentation 33,* 4, 294–304.

Yiqun Liu works as professor and co-chair at the Department of Computer Science and Technology in Tsinghua University, Beijing, China. He obtained his Ph.D. and B.Eng in Computer Science at Tsinghua University. His major research interests include Web Search, User Behavior Analysis, and Web Data Mining. He serves as the co-Editor-in-Chief of Foundations and Trends in Information Retrieval (FnTIR) and also on the editorial boards of Journal of the Association for Information Science and Technology (JASIST) and Information Retrieval Journal (IRJ). He serves as Program Co-chair of SIGIR2018, area chair of ACL2018 and WWW2020, Program Co-chair of NTCIR-13/14, Program Co-chair of ICTIR2020, General Co-chair of AIRS2016 as well as (senior) program committee members of several important international academic conferences including SIGIR, WWW, KDD, IJCAI, AAAI, CIKM and WSDM.

Jiaxin Mao is a post-doctoral research fellow at the Department of Computer Science and Technology in Tsinghua University, Beijing, China. He obtained his Ph.D. and B.Eng in Computer Science at Tsinghua University. His research interests lie in the area of Information Retrieval, Web Search, and User Behavior Analysis and Modeling. He has published over 10 papers in top-tier conferences and journals in IR, including SIGIR, WWW, TOIS, and CIKM. He served as reviewer and PC member for TOIS, TKDE, JASIS, SIGIR, WWW, and CIKM. He also serves as the ACM SIGIR Student Affairs Co-Chair since 2018.