Beyond Sessions: Exploiting Hybrid Contextual Information for Web Search

Jia Chen*
Department of Computer Science and Technology, Institute for Articial Intelligence, Beijing National Research Center for Information Science and Technology, Tsinghua University
Beijing 100084, China
chenjia0831@gmail.com

ABSTRACT

It is essential to fully understand user intents for the optimization of downstream tasks such as document ranking and query suggestion in web search. As users tend to submit ambiguous queries, numerous studies utilize contextual information such as query sequence and user clicks for the auxiliary of user intent modeling. Most of these work adopted Recurrent Neural Network (RNN) based frameworks to encode sequential information within a session, which is hard to realize parallel computation. To this end, we plan to adopt attention-based units to generate context-aware representations for elements in sessions. As intra-session contexts are deficient for handling the data sparsity and cold-start problems in session search, we would also attempt to integrate cross-session dependencies by constructing session graphs on the whole corpus to enrich the representation of queries and documents.

KEYWORDS

session search, cross-session context, self-attention

ACM Reference Format:
Jia Chen. 2020. Beyond Sessions: Exploiting Hybrid Contextual Information for Web Search. In *The Thirteenth ACM International Conference on Web Search and Data Mining (WSDM '20), February 3–7, 2020, Houston, TX, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3336191.3372179

1 INTRODUCTION

Despite the achievements that modern search engines have made, we still have a long way to go before fully understanding users' search intents. To fulfill their complex information needs, users will issue a sequence of queries, examine and interact with some of the results. Such behavior is regarded as search tasks or sessions. Queries in sessions can be ambiguous, which may cause difficulties for Ad-hoc retrieval methods, e.g., traditional IR approaches like BM25 and deep learning-to-rank models such as DRMM [4]. Numerous studies have shown the effectiveness of exploiting user interactions to enrich their retrieval models. Most of these methods regard both the query/document content and user interactions as sequential information, thus utilizing Recurrent Neural Network (RNN) based frameworks for session context modeling.

However, for lack of appropriate sequence transduction tools, most existing approaches employ RNN-based frameworks (e.g., the encoder-decoder architecture) or convolutions for session modeling. Previous studies have pointed out the constraints of these architectures. For example, the inherently sequential nature impedes parallelization within training examples, which may further limit batching operations when encountering long input sequences [7]. Thus, we first plan to find a better way of modeling session-level contexts. This motivates our first research question:

RQ1: How to represent the contextual information within a session effectively and efficiently?

Previous log-based user behavior investigations [2] find that in real-world search scenarios, over 70% sessions contain only two queries, while long sessions (with 5+ queries) are scarce and noisy. On one hand, limited context information, i.e., short query sequence or sparse click signals, may cause difficulties for models that rely merely on intra-session context to accurately capture user intents. For example, the famous *cold-start* problem for initial queries within a session. On the other hand, the data sparsity problem for long-tailed or out-of-vocabulary queries remains to be handled. It is hard to understand these queries if we ignore the semantic connections between them and frequent ones. To this end, we aim at using cross-session contextual information, which motivates our second research question:

RQ2: How to exploit the inter-session context information as the wisdom of crowds for a specific query?

Due to the diversity of user intents in different search conditions, it is crucial to find how to utilize the contextual information from two levels. Suppose we represent the cross-session context information of a query/document node u as the aggregation of its neighboring nodes N_u in a constructed session-flow graph, then given its intra-session and inter-session context-aware representation $\mathbf{e}_u$ and $\mathbf{e}_{N_u}$, we would like to study appropriate techniques of combining two components together to formulate the hybrid context-aware representation. It worths investigating when to rely more on intra-session contexts and when to refer more the wisdom of crowds. All of the above challenges motivates our third research question:

RQ3: How to combine both the intra-session and inter-session contextual information for task optimization?

In the next, we will retrospect the related work and introduce the possible solutions for the proposed research questions.

2 RELATED WORK

2.1 Search session modeling

As user interaction behavior can reflect their search intents, numerous studies aim at incorporating users' session-level contextual information into the optimization for various IR tasks. Some researches exploit the query sequence and the user clicks within a session for context-aware query suggestion [6]. Since the core objective of search engines is to optimize document ranking, contextual information is also considered for session-based retrieval [1],

*Second year Ph.D. student.

WSDM '20, February 3–7, 2020, Houston, TX, USA

ACM ISBN 978-1-4503-6822-3/20/02.
https://doi.org/10.1145/3336191.3372179

e.g., the session search or exploratory search. These studies have achieved great improvements in understanding user intent, but may still be confronted with the data sparsity problem on long-tailed or unseen queries.

2.2 Self-attention mechanism

RNN-based architectures are widely used for sequence modeling due to their effective power of representing sequential information. However, RNN-based architectures have their limitations: low parallelization, hard to perceive distant inputs, etc. To this end, Vaswani et al. [7] propose a new framework, namely *Transformer*, which abandons the recurrence and relies entirely on an attention mechanism. Extend experiments show both the effectiveness and efficiency of *Transformer*. Inspired by this work, many follow-up work adopt similar frameworks with only attention mechanism for context-aware dependency modeling and also achieve considerable success in corresponding domains.

2.3 Graph-based Information Retrieval

Graph-structured information has been widely discussed in the IR literature. Jiang et al. [5] use click-through bipartite graphs to enrich vector representations for the reduction of the lexical gap between query and documents. With the emergence of deep learning, many efforts have been made to employ graph neural networks for various IR tasks. For example, graph convolutional neural networks (GCNs) have been adopted in text classification and web-scale recommendation systems. We would like to use similar structures to mine the wisdom of crowds buried in the whole corpus and alleviate the data sparsity problem.

3 METHODOLOGY

3.1 Session-level context modeling

To better capture user intent in complex search scenarios, researchers usually encode query history and user interactions into context-attentive session representations [3]. Previous studies usually adopt hierarchical RNN-based frameworks for session modeling, which shows less superiority in handling short sequences and parallel computing. Inspired by previous work, we will eschew RNN architectures and design new transformer units to encode session contexts for a specific query/document into context-aware vectors. The goal of designing the corresponding units is to model intra-session contexts with both high effectiveness and efficiency.

3.2 Exploiting cross-session dependencies

Utilizing session-level context information helps for better user intent understanding. However, there also remains the data sparsity problem for long-tailed queries as well as the cold-start problem in search scenarios where less contextual information can be inferred. Query-session graph and co-occurrence commodity graph have been proved not only effective for overcoming the query-item sparsity problem but also generalize well for unseen or long-tailed queries [8]. Therefore, we attempt to exploit the wisdom of crowds buried in cross-session dependencies by constructing a graph on the session data. As shown in Figure 1, we try to model both the co-clicked and co-semantic dependencies between queries and further enrich the node embeddings by leveraging these cross-session relationships. Approaches such as Graph Neural Network (GNN) or direct node embedding aggregation can be employed here for cross-session context representation.

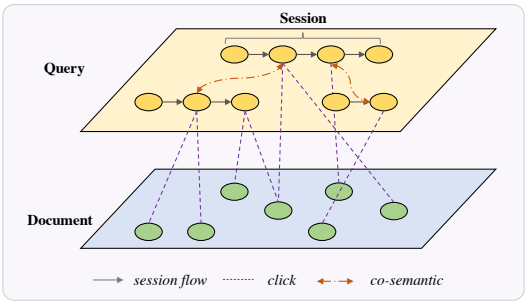

Figure 1: Graph structure constructed on session data.

Here we can simply represent the cross-session contexts of a specific query/document node as the aggregation of its neighboring nodes, for example. Let N_u be the set of neighboring nodes of node u in co-clicked/co-semantic level, a possible attention-based aggregation can be formulated as follows:

$$\mathbf{e}_{N_u} = \sum_{v \in N_u} \alpha_v \mathbf{e}_v, \tag{1}$$

$$\beta_v = tanh(\mathbf{e}_u^T W^{c/s} \mathbf{e}_v), \tag{2}$$

$$\alpha_v = \frac{exp(\beta_v)}{\sum_{v' \in N_u} exp(\beta_{v'})}; \tag{3}$$

where $\mathbf{e}_v$ denotes the embedding of node v in clicking space or semantic space, W^c and W^s are the bilinear weight matrices in click-level and semantic-level, respectively.

3.3 Integrating hybrid context information

To enrich the representation of query or document with both intra-session and inter-session contexts, we should find a good way of integrating them. A simple method of combining two contextual representations is to directly concatenate them. The specific calculation method for the weight matrices of intra-session and cross-session contextual information remains to be investigated.

REFERENCES

[1] Wasi Uddin Ahmad, Kai-Wei Chang, and Hongning Wang. 2019. Context Attentive Document Ranking and Query Suggestion. *arXiv preprint arXiv:1906.02329* (2019).
[2] Jia Chen, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2019. TianGong-ST: A New Dataset with Large-scale Refined Real-world Web Search Sessions. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. ACM, 2485–2488.
[3] Jia Chen, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2020. A Context-Aware Click Model for Web Search. In *Proceedings of the Thirteenth ACM International Conference on Web Search and Data Mining*. ACM.
[4] Jiafeng Guo, Yixing Fan, Qingyao Ai, and W Bruce Croft. 2016. A deep relevance matching model for ad-hoc retrieval. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, 55–64.
[5] Shan Jiang, Yuening Hu, Changsung Kang, Tim Daly Jr, Dawei Yin, Yi Chang, and Chengxiang Zhai. 2016. Learning query and document relevance from a web-scale click graph. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 185–194.
[6] Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie. 2015. A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, 553–562.
[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.
[8] Yuan Zhang, Dong Wang, and Yan Zhang. 2019. Neural IR Meets Graph Embedding: A Ranking Model for Product Search. In *The World Wide Web Conference*. ACM, 2390–2400.