DUMPLING: A Novel Dynamic Search Engine

Andrew Jie Zhou, Jiyun Luo, Hui Yang
Department of Computer Science, Georgetown University
{jz398,jl1749}@georgetown.edu, huiyang@cs.georgetown.edu

ABSTRACT

In this demo paper, we introduce a new search engine that supports Information Retrieval (IR) in a dynamic setting. A dynamic search engine distinguishes itself by handling rich interactions and temporal dependency among the queries in a session or for a task. The proposed search engine is called Dumpling, named after the development team's favorite food. It implements state-of-the-art dynamic search algorithms and provides: (i) a dynamic search toolkit by integrating the Query Change Retrieval Model (QCM) and the Win-win search algorithm; (ii) a user-friendly interface supporting side-by-side comparison of search results given by a state-of-the-art static search algorithm and the proposed dynamic search algorithms; (iii) and APIs for developers to apply the dynamic search algorithms to index and search over custom datasets. Dumpling is developed under the umbrella of a bigger project in the DARPA Memex program to crawl and search the dark web to support law enforcement and national security.

1. INTRODUCTION

Dynamic search is an emerging topic in Information Retrieval (IR) research [4]. In dynamic search, we model dynamic systems which change or adapt over time or a sequence of events using a range of techniques from artificial intelligence and reinforcement learning. Many of the open problems in current IR research can be described as dynamic systems, for instance, information retrieval in a session.

In a dynamic search setting, a user issues multiple queries during a session to accomplish a search task. The common process is that the user sends a query, gets the top ranked documents, changes her query and sends it again. As a result, a series of queries, a series of retrieved documents, and rich user interactions will be generated, until the session stops when the user finishes her search task. During a session, the temporal dependencies between queries are presented in that way that the previous queries and the previously obtained search results will influence how the user issues the current query and how the search result rankings could be optimized for the current query.

The dynamic search algorithms implemented in Dumpling include the Query Change Retrieval Model (QCM) [1] and the Win-win search algorithm [3]. Not only does it implement a dynamic search engine, Dumpling also provides a convenient user interface for a user to compare the results from the dynamic search engine and the static search engine. By comparing the retrieved documents, the user can easily evaluate the performance of different search engines. We have deployed Dumpling for DARPA's Memex program. The evaluations of the systems are done by obtaining real user feedback for the search results. They show that the dynamic search algorithms dramatically improve the search accuracy over the static search algorithms. To further test the performance of the dynamic search engine, Dumpling also provides simple APIs to update the dataset. Dumpling can automatically build an index and apply the dynamic search algorithm to a new dataset. Dumpling is not only a dynamic search system but also provides potential to further the research of dynamic search.

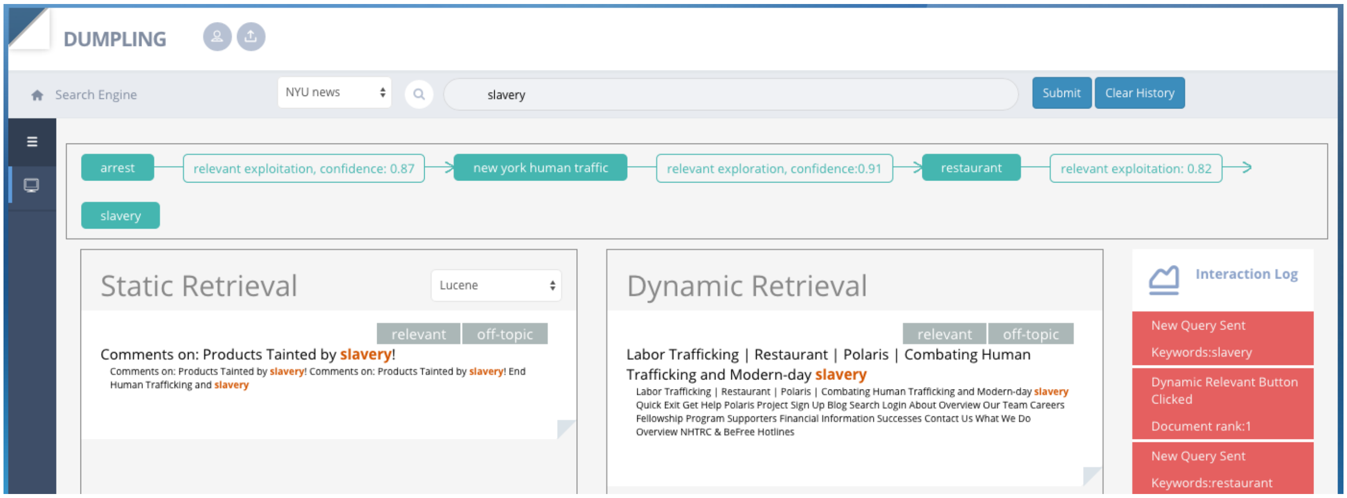
2. QUERY CHANGE RETRIEVAL MODEL

The Query Change Retrieval Model (QCM) analyzes the evolution of queries as well as previously retrieved documents to enhance the dynamic search. QCM's document ranking function is based on the previous query q_{i-1}, the current query q_i, the previously retrieved documents D_{i-1}, and the discount factor $\gamma \in (0, 1)$: $Score(q_i, d) = P(q_i|d) + \gamma \sum_a P(q_i|q_{i-1}, D_{i-1}, a) \max_{D_{i-1}} P(q_{i-1}|D_{i-1})$. It measures the document relevance to the i^{th} query based on the current reward (current document relevance) and the discounted sum of past accumulated rewards/relevance in a session. The most important feature of this formula is that the past queries and retrieved documents have less influence on the user's current search. The QCM model is highly effective and consistently won the top positions in the recent TREC Session Track evaluations [2].

3. WIN-WIN SEARCH

A user tends to query several times in a row during one search session. She changes keywords in the queries hoping to find more information. The intension of a user differs from one session to another. Sometimes, the user finds some relevant information by the first query, and then wants to find more detailed information about subtopics by the next query (We term this behavior as exploitation.) Other times, the user gets what they need by the first query, and then

SIGIR'15, August 09-13, 2015, Santiago, Chile.
ACM 978-1-4503-3621-5/15/08.
DOI: http://dx.doi.org/10.1145/2766462.2767873 .



slavery

Comments on: Products Tainted by slavery!

Comments on: Products Tainted by slavery! Comments on: Products Tainted by slavery! End Human Trafficking and slavery

Labor Trafficking | Restaurant | Polaris | Combating Human Trafficking and Modern-day slavery

Labor Trafficking | Restaurant | Polaris | Combating Human Trafficking and Modern-day slavery Quick Exit Get Help Polaris Project Sign Up Blog Search Login About Overview Our Team Careers Fellowship Program Supporters Financial Information Successes Contact Us What We Do Overview NHTRC & BeFree Hotlines

New Query Sent
Keywords:slavery

Dynamic Relevant Button Clicked
Document rank:1

New Query Sent
Keywords:restaurant

Figure 1: Dumpling User Interface

wants to change to a brand-new topic by the next (We term this behavior as exploration). In the model of the Win-win search algorithm, we categorize the user into 1 of 4 states: $\{RT$: find something relevant and explore a new topic; RR: find something relevant and exploit the same topic; NRT: find nothing relevant and explore a new topic; NRR: find nothing relevant and exploit the same topic.$\}$ The user's state moves from one state to the next along with the evolution of queries. The belief of the user's state can be presented as $b_t = S_m$ $(t = 1, 2, ..., n)$ (n stands for the total number of iterations) and m= {RT, RR, NRT, NRR} [3]. Dumpling infers the user's current state by analyzing previous returned documents, the user's queries and the user's interactions during the session. Every time the user sends a new query, Dumpling updates the user's state using $b_{t+1}(s_j) = Pr(s_j|o_t, a_t, b_t) = \frac{O(s_j,a_t,o_t)\sum_{s_i \in S} T(s_i,a_t,s_j)b_t(s_i)}{P(o_t|a_t,b_t)}$. where t indicates the search iteration; s_i and s_j are any two states; $b_{t+1}(s_j)$ is the probability of the user being believed at state j in iteration $t+1$; o_t is the observation of the user's behavior; a_t is the search engine's action. Dumpling thus infers the user's state, and chooses appropriate retrieving algorithm and arguments.

4. USER INTERFACE & EVALUATION

Dumpling supports dynamic search to different datasets which can be chosen by a user. Currently we use datasets provided by the DARPA Memex program. The documents are crawled from the dark web including forum posts, online sex ads, child labor ads, etc. The search engine is located at `dumplingproject.org`. Due to the sensitivity of the data, our search results are password protected and not open to the public. Figure 1 shows the interface of Dumpling. It supports the following functionalities. Dumpling provides very convenient APIs to maintain the dataset. With Dumpling's API, the user can add, update and delete index based on new crawling results. The extensibility of Dumpling is enhanced by the easy-to-change dataset. The search engine provides side-by-side comparison between the dynamic search algorithm and static search algorithm (elastic search in our case). We also have a decision state pane to show our estimated decision making states of the user during a search.

algorithm	dynamic search	static search
Precision@10	0.52	0.32
# of Queries	362	
# of Sessions	178	

Table 1: Search Accuracy in Precision at 10.

Dumpling provides a convenient user interface for a user to compare the dynamic search results and the static search results. Dumpling gathers explicit feedback from real users in US agencies working on national security. The user can provide both positive and negative feedback of the search results and the feedback is used in the dynamic search algorithms. In Figure 1, in the first query, a user uses "arrest"; in the second query, the user uses "new york human traffic". From the user's perspective, she initially wants to know who was arrested recently. Then she wants to further exploit who is recently arrested by New York police because of human trafficking. For the static search algorithm, the retrieved documents only include New York human trafficking news, the keyword "arrest" seldom appears in the documents. For the dynamic search algorithm, the retrieved documents contain much more information about arresting, which better suits the user's needs.

5. ACKNOWLEDGEMENT

The research is supported by DARPA FA8750-14-2-0226, NSF IIS-1453721, and NSF CNS-1223825. Any opinions, findings, conclusions, or recommendations expressed in this paper are of the authors, and do not necessarily reflect those of the sponsor.

References

[1] D. Guan, S. Zhang, and H. Yang. Utilizing query change for session search. In *SIGIR'13*.

[2] E. Kanoulas, M. M. Hall, P. Clough, B. Carterette, and M. Sanderson. Overview of the trec 2011 session track. In *TREC'11*.

[3] J. Luo, S. Zhang, and H. Yang. Win-win search: Dual-agent stochastic game in session search. In *SIGIR'14*.

[4] H. Yang, M. Sloan, and J. Wang. Dynamic information retrieval modeling. SIGIR '14.