LogCanvas: Visualizing Search History Using Knowledge Graphs

Luyan Xu
DEKE key lab(MOE)
Renmin University of China
Beijing, China
xuluyan@ruc.edu.cn

Zeon Trevor Fernando
L3S Research Center
Leibniz Universität Hannover
Hannover, Germany
fernando@L3S.de

Xuan Zhou
School of Data Science & Engineering
East China Normal University
Shanghai, China
zhou.xuan@outlook.com

Wolfgang Nejdl
L3S Research Center
Leibniz Universität Hannover
Hannover, Germany
nejdl@L3S.de

ABSTRACT

In this demo paper, we introduce LogCanvas, a platform for user search history visualization.Different from the existing visualization tools, LogCanvas focuses on helping users re-construct the semantic relationship among their search activities. LogCanvas segments a user's search history into different sessions and generates a knowledge graph to represent the information exploration process in each session.A knowledge graph is composed of the most important concepts or entities discovered by each search query as well as their relationships. It thus captures the semantic relationship among the queries.LogCanvas offers a session timeline viewer and a snippets viewer to enable users to re-find their previous search results efficiently. LogCanvas also provides a collaborative perspective to support a group of users in sharing search results and experience.

CCS CONCEPTS

• **Human-centered computing → Collaborative and social computing systems and tools**; **Graph drawings**; **Information visualization**;

KEYWORDS

search history visualization; information-refinding; collaborative search;

ACM Reference Format:
Luyan Xu, Zeon Trevor Fernando, Xuan Zhou, and Wolfgang Nejdl. 2018. LogCanvas: Visualizing Search History Using Knowledge Graphs. In *SIGIR '18: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, July 8–12, 2018, Ann Arbor, MI, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3209978.3210169

SIGIR '18, July 8–12, 2018, Ann Arbor, MI, USA

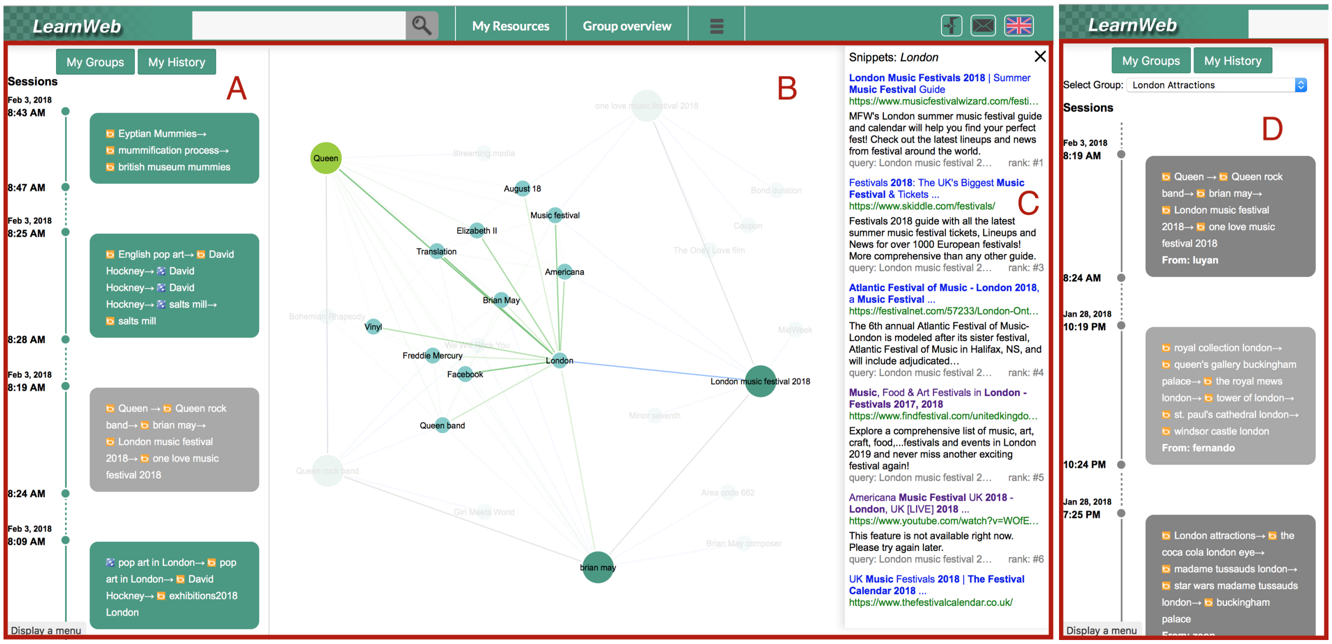
ACM ISBN 978-1-4503-5657-2/18/07.
https://doi.org/10.1145/3209978.3210169

1 INTRODUCTION

When people use search engines to retrieve information, acquire knowledge or solve daily-life problems, they often are not satisfied by a single-shot query. Instead, they will issue a series of queries and have multiple rounds of interaction with the search engines. This is known as an information exploration process, in which each round of interaction is a stepping stone for a user to achieve his / her final goal. Users' interaction with search engines is usually recorded in search history logs. If used wisely, these search history logs can help users preserve and recall the process of their information exploration, so that they can re-find forgotten information or knowledge quickly. A survey of experienced Web users found that people would like to use search engines to re-find online information, but often have difficulty remembering the sequence of queries they had used when they originally discovered the content in question [1]. In addition, studies have shown that as many as 40% of users search queries are attempts to re-find previously encountered results [18].

Besides being helpful in information re-finding, search histories can also benefit collaborative search. In collaborative search, a group of users undertake various search subtasks, aiming to accomplish a complex collaborative task, e.g., planning for a trip. By seeing each other's search histories, group members from different backgrounds can learn from each other, as they will be looking at different aspects of the same topic / task. This helps them form a more complete view of a certain topic or detect fake information more effectively.

To make the best of search histories, researchers have worked on tools that can track users' search history and visualize it in an understandable and in-depth presentation [3, 14]. Search logs record a searcher's explicit activities, including the queries submitted and the answers (search results) clicked. In reality, such explicit activities provide only partial information about an information exploration process. More intellectual activities are carried out in the searcher's mind. Studies have shown that a good visualization of a user search history should not only present the explicit activities, represented by search queries and answers, but also depict the latent information exploration process in the searcher's mind [7]. Such a visualization can help users quickly re-construct the knowledge acquired in the process of information exploration.



(a) Personal View

(b) Group View: group session viewer of "London Attractions"

Figure 1: Overview of LogCanvas. (A): personal session viewer in which queries are clustered into sessions in a most recent ordering, session "Queen →...→ one love music festival 2018" is selected; (B): knowledge-graph viewer, presenting the knowledge graph of the selected session; subgraph of node "London" is highlighted by moving the mouse over it; (C): snippets viewer, showing previous search results of a selected node "London"; (D): group session viewer, in which a session from another user (fernando) is selected;

In this paper, we present LogCanvas, a platform for graph-based search history visualisation. The visualization interface is illustrated in Figure 1. Users' search activities are clustered into sessions (on the left). Queries of each session are embedded into a knowledge graph to help users understand what the queries returned as results and how they are related. This knowledge graph is constructed based on the concepts extracted from the search result snippets. A snippets viewer on the right helps users re-find information of previous searches. Additional features enable users to interact with the search histories. For instance, users can focus on a subgraph of a certain node (e.g. "London") through a mouse hover (Figure 1(a)). All nodes float in the workspace, allowing users to drag and move nodes to positions they like.

As an example, suppose an art enthusiast, who is a member of a "London Attractions" searching group, has conducted several rounds of searches about London. In the end, he wants to quickly review his search history, in order to plan the trip to London. Through the session viewer, the user can identify the sessions about activities in London as shown in Figure 1(a)-A. By selecting a certain session which includes queries such as "London music festivals 2018" or "Brian May", the user can view its corresponding knowledge graph in the knowledge graph viewer(Figure 1(a)-B). All concepts in this graph are extracted from the user's search results, enabling him to quickly grasp the knowledge structure of his search history. Through concepts such as "Music festival", "Americana", and "Brian May", the user remembers that in this search session he mainly searched about music festivals in London, specifically for American-style music and the Queen Band. To get details about music festivals in London in 2018, the user clicks on "London", through the snippets viewer (Figure 1(a)-C), he reviews all filtered search result snippets that are related to music festivals in London. In order to collect more information about attractions in London, the user can also turn to other searchers in the collaborative group "London Attractions", to view group members' search histories(Figure 1(b)-D). His group members with their different backgrounds have found interesting things he has never thought of. He gains insights from knowledge graphs of other group members, acknowledging and using the suggestions from them. Meanwhile, he contributes his perspective to the group.

To realize this user interface, we applied several techniques. First, query logs are clustered into sessions according to time intervals. Second, for queries of each session, related concepts and entities are extracted from the search result snippets using the Yahoo Fast Entity Linker [2]. Third, correlations between the concepts are computed using a method based on the entity co-occurrence frequency in wikipedia.

The visualisation platform can be used in different platforms, and is currently integrated into LearnWeb[1], an online environment

[1]https://learnweb.l3s.uni-hannover.de/

that supports collaborative sensemaking by allowing users to share and collaboratively work on content retrieved from a variety of web sources [10–12].

In the remainder of this paper, we briefly introduce the enabling technologies of the user interface and how we are going to demonstrate it in the conference.

2 EXISTING SYSTEMS

Research on archived data visualization and information re-finding is relevant to LogCanvas, as it concerns preserving and visualizing users' search histories.

Systems such as popHistory [4] and Warcbase[8, 9] save users' visit data, based on which they can extract and display the most visited websites to users. History Viewer [16] tracks processes of exploratory search and present users with interaction data to enable them to revisit the steps that led to certain insights.

Information re-finding tools such as SearchBar [13] provide a hierarchical history of recent search topics, queries, results and users' notes to help users quickly re-find the information they have searched. Some other tools, such as SIS (Stuff I've Seen), collect users' personal data, such as email and docs, and offer a diary list [6] to help users quickly locate past events or visited web-pages based on dates. Some recent work [5, 15] has investigated how to combine context analysis and information re-finding frameworks to remind users about historical events according to users' current context.

In collaborative search systems such as Coagmento [17] and SearchTogether [14], visualization of search history usually involves multiple users' search logs including their search queries, bookmarks, etc. Interfaces of this kind display search histories separately according to datatypes or categories and support notepad functions which allow group members to share experience.

Most of the previous visualisation tools focus primarily on the selection of suitable data to present on the user interfaces. They leave it to the users to re-construct short term memory and semantic relationships. By contrast, the visualisation of LogCanvas not only provides a detailed overview of search history and efficient ways to re-find information, but also introduces a knowledge graph that helps users connect their search activities into coherent processes of semantic information exploration.

3 OUR SYSTEM

3.1 Overview

The overview of the visualization platform is shown in Figure 2. Given a users' search history log, the following steps of data preprocessing are used to prepare the data for final visualisation:

1) Session Segmentation – the queries in the log are split into different sessions according to the time interval between searches(e.g. "epidemiology", "anorexia", etc.);

2) Search Result Acquisition – the top 10 snippets of each query are fetched from the archived search results;

3) Entity and Concept Extraction – the most relevant entities and concepts are extracted from the search results and form a knowledge graph; we use Yahoo's Fast Entity Linking toolkit[2](yahooFEL) [2] to extract the entities and concepts from the snippets of the top-10

[2]https://github.com/yahoo/FEL

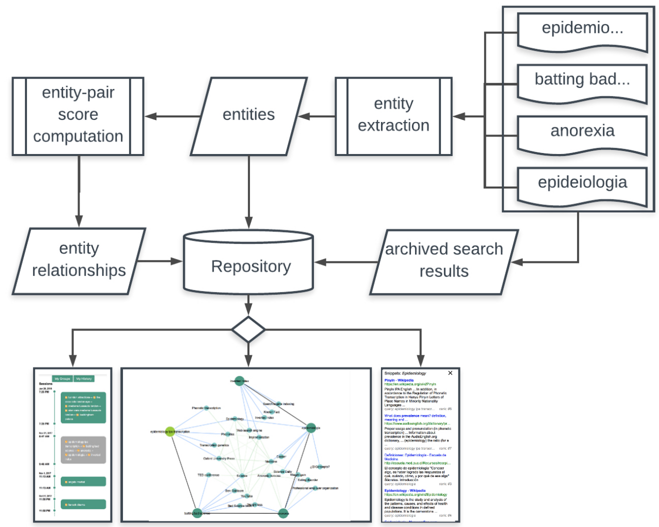


Figure 2: Overview of the visualization module

search results and select the top 5 entities / concepts to add to the knowledge graph; the selection is based on the scoring method described in section 3.2.1;

4) Edge Score Measuring – the entity and concept nodes in the knowledge graph are connected by edges, which represent the semantic relationships; the edge score is computed based on the co-occurrence of the entities and concepts in Wikipedia, as described in section 3.2.2;

5) Group clustering – a user's query session is added to the collaborative group he/she belongs to, if any search result of the session is tagged as useful to that group.

3.2 Key Methodologies

3.2.1 Entity and Concept Extraction. We use the Yahoo entity linking toolkit to get the candidate entities and concepts from the top-10 search results. To select the five most relevant entities / concepts, we compute a quality score $qScore(e)$ for each candidate e. In general, the larger $qScore(e)$, the more relevant e to the query. $qScore(e)$ is defined as follows:

$$qScore_e = Freq_e * \frac{1}{|avgFel_e|} \tag{1}$$

where

$$avgFel_e = \frac{\sum_{i=1}^{n} fel\_score_{e_i}}{n} \tag{2}$$

In eq. 2, e_i refers to the ith entity e extracted from the top-10 search result snippets. $fel\_score_{e_i}$ is a negative value returned by yahooFEL. It represents the confidence of yahooFEL in that e_i is a relevant entity to the query. n is the count of occurrence of an entity e in the top-10 search result snippets. We sum the $fel\_score_{e_i}$ from different snippets and get an average $avgFel_e$, which represents the confidence of yahooFEL in that e is relevant to the entire session. In eq. 1, $Freq_e$ is the frequency of e occurring in the top-10 result snippets. We include this to favor more frequently occurring entities in the top-10 results. In addition, we remove candidate entities that have a word length less than 4, as they are likely to be stop words.

3.2.2 Edge Score Computation. An edge in the knowledge graph represents the semantic association between two entities / concepts. For entities in one session, we enumerate all possible entity pairs and count the co-occurrence of each pair in Wikipedia. The co-occurrence count is the number of documents returned from the Wikipedia index built from Solr[3], using a boolean query with the two entities as phrase terms (e.g., "computer science" AND "information retrieval"). We assume that the more frequently two entities co-occur in Wikipedia articles, the more correlated they are.

Some strongly correlated entities can still get a low co-occurrence count when one of the entities is not common in wikipedia. On the other hand, some entity pairs with high co-occurrence count (i.e., "human" and "animal") are common sense correlations uninteresting to users. Moreover, correlation of the same entity pair can have different meanings in different search sessions. For example, the relationship between "apple" and "toolkit" in the session of "computer" is different from that of "fruit cultivation".

Therefore, to make the entity relationships meaningful targeting a certain session, we normalize the co-occurrence counts using a non-linear function. For each entity pair (e_i, e_j) of a session, the $eScore_{(e_i,e_j)}$is defined as follows:

$$eScore_{(e_i,e_j)} = \begin{cases} 1 - \frac{\lambda}{\lambda + \max\{C_{(e_i,e_j)}\}} & \text{if } \max\{C_{(e_i,e_j)}\} > 1000 \\ \frac{C_{(e_i,e_j)}}{\max\{C_{(e_i,e_j)}\}} & \text{otherwise} \end{cases}$$

where $C_{(e_i,e_j)}$ means how many times two entities co-occur in wikipedia and $\max\{C_{(e_i,e_j)}\}$ is the entity pair with the largest co-occurrence count in a session. With the non-linear function, even when there is an entity pair with an extremely large co-occurrence count(>1000), the edge scores of the other entity pairs can still be significant. This makes sure that all relevant edges are visible in the graph visualisation. We empirically set λ to 50 in the normalization.

3.2.3 Data Collection. To collect users' search histories, whenever a user submits a query to the platform (i.e., LearnWeb), we record the query, the search objective (text, image, video) and the search service provider (bing, flickr, youtube, etc.) in the history log, and annotate them with a timestamp. All top search results viewed by the user are also stored based on the click or save information (when a result is saved to a group in LearnWeb). These preprocessing steps are then performed offline once a certain amount of log has been accumulated - we run the edge score computation script at the end of each day. All results are stored in a relational database format in a MariaDB repository, so they can be quickly retrieved during visualization.

4 DEMONSTRATION

In the demonstration, we will mainly show how LogCanvas can help users re-find information and how it can benefit collaborative search.

On our LearnWeb platform, we have collected a large number of user search histories. The histories include those of individuals who used LearnWeb to search and explore learning resources on the Web. They also include histories of collaborative search processes, in which a group of users studied a topic together and shared their findings. In the demonstration, we will visualize these search histories in LogCanvas. We will let the audience interact with the search histories to show how our visualization can help them quickly understand the search processes.

We will also provide a number of test scenarios for the audience to try out our system. They will first perform some search tasks using LearnWeb. During this process, audience can create or join a collaboratively searching group. We will generate the visualization of their search histories on the fly, and demonstrate how accurately LogCanvas can visualize their information exploration processes.

At this moment, an online demonstration of LogCanvas[4] is accessible using the demo account (username: *luyan*, password: *test*).

ACKNOWLEDGMENTS

This research has been supported in part by project ALEXANDRIA which is funded by European Research Council under the EU 7th Framework Programme (FP7/2007-2013) / ERC 339233.

REFERENCES

[1] Anne Aula, Natalie Jhaveri, and Mika Käki. 2005. Information search and re-access strategies of experienced web users. In *WWW 2005*. 583–592.
[2] Roi Blanco, Giuseppe Ottaviano, and Edgar Meij. 2015. Fast and space-efficient entity linking for queries. In *WSDM 2015*. 179–188.
[3] Susan E Brennan, Xin Chen, Christopher A Dickinson, Mark B Neider, and Gregory J Zelinsky. 2008. Coordinating cognition: The costs and benefits of shared gaze during collaborative search. *Cognition* 106, 3 (2008), 1465–1477.
[4] Matthew Carrasco, Eunyee Koh, and Sana Malik. 2017. popHistory: Animated Visualization of Personal Web Browsing History. In *CHI EA 2017*. 2429–2436.
[5] Tangjian Deng, Liang Zhao, Ling Feng, and Wenwei Xue. 2011. Information re-finding by context: a brain memory inspired approach. In *CIKM 2011*. 1553–1558.
[6] Susan Dumais, Edward Cutrell, Jonathan J Cadiz, Gavin Jancke, Raman Sarin, and Daniel C Robbins. 2003. Stuff I've seen: a system for personal information retrieval and re-use. In *SIGIR 2003*. 72–79.
[7] Marti Hearst. 2009. *Search user interfaces*. Cambridge University Press.
[8] Jimmy Lin. 2015. Scaling down distributed infrastructure on wimpy machines for personal Web archiving. In *WWW 2016*. 1351–1355.
[9] Jimmy Lin, Milad Gholami, and Jinfeng Rao. 2014. Infrastructure for supporting exploration and discovery in web archives. In *WWW 2014*. 851–856.
[10] Ivana Marenzi. 2014. *Multiliteracies and e-learning2.0*. Vol. 28. Peter Lang, Bern, Switzerland.
[11] Ivana Marenzi and Wolfgang Nejdl. 2012. *I search therefore I learn - Active and collaborative learning in language teaching: Two case studies*. IGI Global, 103–125.
[12] Ivana Marenzi and Sergej Zerr. 2012. Multiliteracies and active learning in CLIL - the development of LearnWeb2.0. In *IEEE Transactions on Learning Technologies*.
[13] Dan Morris, Meredith Ringel Morris, and Gina Venolia. 2008. SearchBar: a search-centric web history for task resumption and information re-finding. In *CHI 2008*. 1207–1216.
[14] Meredith Ringel Morris and Eric Horvitz. 2007. SearchTogether: an interface for collaborative web search. In *UIST 2007*. 3–12.
[15] Maya Sappelli, Suzan Verberne, and Wessel Kraaij. 2017. Evaluation of context-aware recommendation systems for information re-finding. *Journal of the Association for Information Science and Technology* 68, 4 (2017), 895–910.
[16] Vinícius CVB Segura and Simone DJ Barbosa. 2016. History viewer: displaying user interaction history in visual analytics applications. In *HCI 2016*. 223–233.
[17] Chirag Shah and Roberto González-Ibáñez. 2010. Exploring information seeking processes in collaborative search tasks. *ASIST* 47, 1 (2010), 1–7.
[18] Jaime Teevan, Eytan Adar, Rosie Jones, and Michael AS Potts. 2007. Information re-retrieval: repeat queries in Yahoo's logs. In *SIGIR 2007*. 151–158.

[3]http://lucene.apache.org/solr/

[4]http://learnweb.l3s.uni-hannover.de/lw/searchHistory/entityRelationship.jsf