# *B-hist*: Entity-Centric Search over Personal Web Browsing History

## http://mem0r1es.io

Michele Catasta[1], Alberto Tonon[2], Vincent Pasquier[3], Gianluca Demartini[2], Philippe Cudré-Mauroux[2], and Karl Aberer[1]

[1] EPFL, Lausanne—Switzerland
{firstname.lastname}@epfl.ch
[2] eXascale Infolab, University of Fribourg—Switzerland
{firstname.lastname}@unifr.ch
[3] University of Applied Sciences, Western Switzerland
{firstname.lastname}@gmail.com

**Abstract.** Web Search is increasingly entity-centric; as many common queries target specific entities, search results are progressively augmented with semi-structured and multimedia information about entities. However, search over personal Web browsing history still revolves around keyword-search mostly. *B-hist* aims at providing Web users with an effective tool for searching and accessing information previously looked up on the Web by providing multiple ways to filter results using temporal ranges, session-based clustering, and entity-centric search.

## 1 Motivation

Web Search today has very much become powered by semantic data: Search Engine result pages (SERPs) are rich of structured content including pictures, maps, factual data, in addition to the standard links pointing to Web pages. This is possible thanks to structured knowledge bases and LOD datasets such as Freebase and thanks to semantic annotations of Web pages using, for instance, schema.org.

A related task, which in our opinion has not yet received the full benefits of the Semantic Web, is search over personal Web browsing history. Most browsers provide a very limited keyword-based search over previously visited pages. The goal of our system, called *B-hist*(standing for *'Better history'*), is to bring entity-centric access to personal browsing activities thanks to semantic technologies such as the ones we developed in our recent research work [7, 6].

*Related Systems.* The Mozilla foundation is working on a related system called Pancake[4] whose goal is to integrate search results from browsing history, social streams, and Web search. While their focus is on integrating content from different sources, the system we propose rather aims at semantically enriching the

[4] https://wiki.mozilla.org/Pancake


**Fig. 1.** Welcome screen of the *B-hist* dashboard. The user has access to his/her browsing activities in the previous two weeks aggregated over time, entities, and sessions.

search experience over personal browsing history easing the access and recall of previously seen information.

A commercial product related to *B-hist* is being developed by CottonTracks[5] and provides a clustered access to personal browsing history. However, *B-hist* provides a much richer set of information access functionalities thanks to the semantic enrichment of Web browsing history, which is its core competitive advantage.

## 2 System Description

Our system provides a multi-dimensional access to one's personal Web history by letting users select the desired pieces of information by means of several filters: temporal, entity-centric, and session-based. In the following we describe the main components of *B-hist* and its backend data processing architecture.

### 2.1 System Components

**Chrome Browser Extension.** The initial data collection is handled by a Web browser extension[6], which is responsible both to gather raw data from the user browsing activity as well as to let the user set preferences and to access the search dashboard of *B-hist*. Specifically, the settings of the extension allow the user to filter-out some domains as well as to allow/disallow https domains from being stored, indexed, and searched by the system. The extension also opens a new browser tab displaying the welcome screen of *B-hist* where the user can start looking for information in her browsing history (see Figure 1).

[5] `http://cottontracks.com/`

[6] At this point, we provide an extension for the Chrome browser.

***B-hist* Data Processing.** Once the raw HTML data is gathered from an accessed page, it goes trough our TRank [6] processing pipeline (see Figure 2) where the main textual content is kept (using approaches from [2]), entities are extracted (using a Conditional Random Field approach trained on a news corpus [1]), and entity types are selected (using approaches from [6]). Such metadata on the Web pages is stored and indexed in *B-hist* (see Figure 2).

To store and index data (e.g., timestamps, sessions, and cluster information) we use both an inverted index (i.e., Apache Lucene[7]) and a lightweight DBMS (i.e., SQLite). The raw HTML coming from the browser extension is however not stored in *B-hist*, as this would require too much storage on the long term.

In parallel to the TRank pipeline, a batch process of session discovery and categorization is accessing the data from the browser and creating additional metadata grouping pages in coherent sessions with a common user intent.

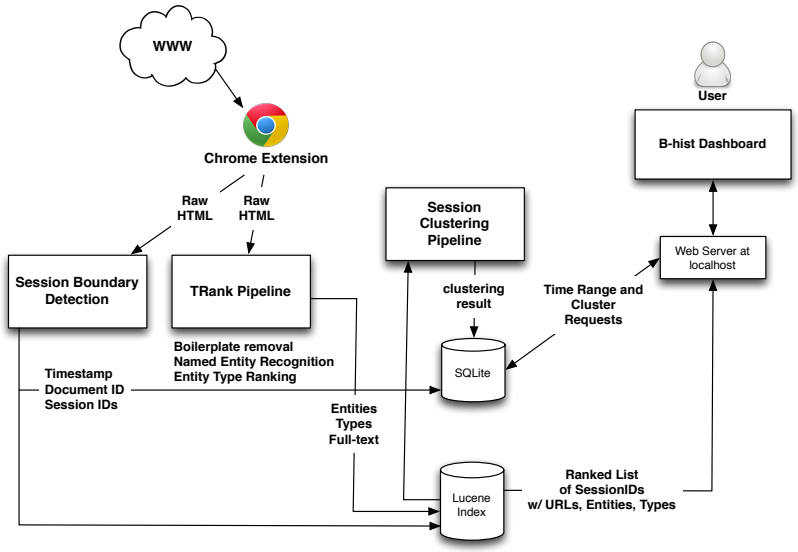


**Fig. 2.** *B-hist* data processing architecture: First the raw HTML of a Web page visited by the user is provided by the browser plugin. Next, the page is processed through boilerplate removal, named-entity recognition, entity type selection, and clustered in an existing or new session. Then, the generated metadata is stored and indexed. Finally, sessions are clustered together in semantically coherent groups. The user's access to information happens via the *B-hist* dashboard.

[7] `http://lucene.apache.org/`

*Browsing Session Detection and Clustering.* Following the large amount of literature available for Web search session detection, we apply existing approaches to set the boundaries of browsing sessions in *B-hist*. Specifically, the system looks for idle browsing times to identify the end of a browsing session and the beginning of the next one. After experimental validation of different thresholds, we picked an idle time interval of 26 minutes, which has also been shown to be effective to detect the end of a Web search session in [5]. Each Web browsing session $s$ is then identified by the list $\tau(s)$ composed by the top-$n$ most frequent types associated to entities contained in the web pages composing $s$. In order to do this, *B-hist* exploits TRank [6] to recognize named entities and to assign a unique entity type to each of them (e.g., Tom Cruise $\rightarrow$ American Actor). Thanks to this, we can define the distance $\delta$ between two browsing sessions as

$$\delta(s_0, s_1) = \left( \sum_{(t,t') \in \tau(s) \times \tau(s')} dist(t, t') \right) \cdot \frac{1}{|\tau(s) \times \tau(s')|},$$

where $dist(t, t')$ is the distance between two types $t$ and $t'$ in the TRank type hierarchy, and is defined as the sum of the number of steps in the hierarchy needed to reach their least common ancestor starting from each one of them. We finally use $\delta$ to cluster the sessions by using a variant of $k$-means clustering algorithm in which the centroid of each cluster is a list composed by the $n$ most frequent types identifying its sessions. The main property of the variation of $k$-means we use is that there is no need to specify the number $k$ of clusters. Rather, we specify a threshold $\Delta$ and, each time a session we want to cluster is further than $\Delta$ from every existing cluster, a new cluster of browsing sessions is created. With such approach we group together browsing sessions about similar entities creating thematic clusters for the user to browse.

***B-hist* Search Dashboard.** After the process described above, the web pages' contents and generated metadata are available for search via the *B-hist* dashboard (see Figure 1). The user is first presented with a summary of the latest two weeks of browsing activities. Each element of the dashboard serves both for filtering and for providing information as the information displayed in each component is updated dynamically after each click.

User interactions are handled by four different components:

- A search box powered by entity suggestion
- A time-based focus with interval selection
- An entity-centric filter
- Groups of semantic sessions.

The main point of entry to search through one's personal browsing history is the familiar search box. The *B-hist* search box is powered by a query autocompletion feature that suggests entities appearing in the user's browsing history based on the query he/she is typing in the box. Such a functionality can be used by users as a way to self-select the sessions they are most interested in.

Thus, user-initiated session clustering becomes an alternative to the algorithmic clustering that *B-hist* precomputes and proposes on its middle panel. A second possibility to filter results is based on the time dimension (left panel): the default view is on the previous two weeks but the user can change it by selecting a different interval in the calendar (with a minimum granularity of one day). The third option to filter results is to select an entity or an entity type in the left panel (below the calendar). Thanks to this panel, the user can specify which entity (or entity type) he/she is interested in and see the clusters, time periods, and URLs most relevant to it. The forth option to interact with the user history is the session clusters in the middle panel: first, the user is presented with a set of clusters which are relevant to the current filters. Then, if the user clicks on a cluster, he/she will be presented with the set of entities belonging to the pages in that cluster. The right panel of the dashboard contains a list of URLs ordered by access time which reflects the currently selected filters.

Each update to the filters will automatically update the results in the other components of our user interface. We expect the user interaction with *B-hist* to finish either when the intended URL has been found and clicked (i.e., re-finding activity) or simply when the user identifies an entity or entity type she was trying to recall using *B-hist*.

*On-line availability.* The system is accessible on-line at `http://mem0r1es.io` where we provide a screencast demonstrating the end-user *B-hist* dashboard. Moreover, for the purpose of judging our system at the Semantic Web Challenge, we provide access to an online deploy of the *B-hist* dashboard which allows to search over a fictitious browsing history.

## 3 Conclusions and Next Steps

In this document, we described *B-hist*: The first semantically-enriched Web browsing activity search end re-finding tool.

The current version of *B-hist* runs on the user machine: In order to preserve his/her privacy, no data is ever sent to any third party. However, we envision a server-side version of the system using scalable storage, indexing and processing techniques (e.g., Apache Solr and Hadoop as described in [6]). In such a setting, users would be sharing their browsing activities (as they already do using any of the commercial Web browsers) and would obtain additional functionalities. For example, one could provide personal analytics functionalities (e.g., 'How do I spend my time online?') and recommendations using, for instance, collaborative filtering approaches that correlate data across similar *B-hist* users.

We also envision a 'forget' functionality as not all information accessed online stays pertinent on the long term. By analyzing user interaction with *B-hist*, the system would learn which type of information the user is most interested in and would consider other types of information as less important (i.e., similarly to the way in which the human memory works).

## 4 Acknowledgements

This work was supported by the Swiss National Science Foundation under grant number PP00P2_128459, and by the Haslerstiftung in the context of the Smart World 11005 (Mem0r1es) project. We also thank for their help and feedback Martin Grund, Eugenia Martin, and Ruslan Mavlyutov.

## References

1. Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.
2. Christian Kohlschütter, Peter Fankhauser, and Wolfgang Nejdl. Boilerplate detection using shallow text features. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pages 441–450, New York, NY, USA, 2010. ACM.
3. Ravi Kumar and Andrew Tomkins. A characterization of online browsing behavior. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 561–570, New York, NY, USA, 2010. ACM.
4. Claudio Lucchese, Salvatore Orlando, Raffaele Perego, Fabrizio Silvestri, and Gabriele Tolomei. Identifying task-based sessions in search engine query logs. In *WSDM*, pages 277–286, 2011.
5. Claudio Lucchese, Salvatore Orlando, Raffaele Perego, Fabrizio Silvestri, and Gabriele Tolomei. Discovering tasks from search engine query logs. *ACM Trans. Inf. Syst.*, 31(3):14:1–14:43, August 2013.
6. Alberto Tonon, Michele Catasta, Gianluca Demartini, Philippe Cudré-Mauroux, and Karl Aberer. TRank: Ranking Entity Types Using the Web of Data. In *International Semantic Web Conference*, 2013.
7. Alberto Tonon, Gianluca Demartini, and Philippe Cudré-Mauroux. Combining inverted indices and structured search for ad-hoc object retrieval. In *SIGIR*, pages 125–134, 2012.

## A Appendix. Addressed Evaluation Criteria

### A.1 Open Track Mandatory Requirements

*The application has to be an end-user application, i.e. an application that provides a practical value to general Web users or, if this is not the case, at least to domain experts.*

The *B-hist* system provides the end-users with a Web application that runs on their local machines (This is done to preserve their privacy) and let them navigate their browsing history in a better way following new possible access paths more than the standard ranked list of visited pages: entities, sessions, and time.

*The information sources used*
- *should be under diverse ownership or control*
- *should be heterogeneous (syntactically, structurally, and semantically), and*
- *should contain substantial quantities of real world data (i.e. not toy examples).*

The information indexed by *B-hist* starts with Web pages the user has visited additionally enriched with Linked Dataset such as DBPedia entities and YAGO entity types. The underlying data which is used are Web pages together with real-world entities from popular LOD repositories.

*The meaning of data has to play a central role.*
- *Meaning must be represented using Semantic Web technologies.*
- *Data must be manipulated/processed in interesting ways to derive useful information and*
- *this semantic information processing has to play a central role in achieving things that alternative technologies cannot do as well, or at all;*

The semantic enrichment of Web pages is done using Named Entity Recognition tools (see [1]) together with Entity Type Selection tools (see [6]). We choose not to use already embedded schema.org annotations as their are originally thought to enable search engines to provide better search functionalities and not to be consumed by end users as we use them in *B-hist*.

The *B-hist* system provides more than just search over one's browsing history: We expect users not necessarily to click on a URL in the list to re-find the information they are looking for. A successful interaction may end, for example, by reading the name of an entity which the user could not remember before. Thus, *B-hist* goes beyond pure document search by adding metadata about entities, sessions, and time.

### A.2 Open Track - Additional Desirable Features

The *B-hist* Web interface is an attractive and functional search dashboard which allows different types of interaction: controlled keyword queries, time filtering, entity navigation, and session selection.

Concerning *B-hist* scalability, the incremental approaches we used (e.g., for clustering, indexing) allow to extend the currently indexed history dataset as new pages are visited by the user with no need to recompute the metadata over the entire collection. Moreover, a previous study found out that the median number of pageviews per day is 59 with a maximum below 1000 [3]. Such a workload allows our system to process data without consuming much resources from the user local machine.

Experimental evaluation of the scalability and effectiveness of the entity type ranking component has been presented in [6] while the session detection approaches are taken from IR literature (e.g., [4, 5]).

The *B-hist* system is novel as it is the first system in applying semantic technologies to Web History Search. It goes beyond pure information retrieval systems by adding the time, entities, and session dimensions to Web history search and re-finding. For such reasons, it also has a clear commercial potential. Moreover, by extending it to a server-side approach with, for example, entity recommendations functionalities, it may also be used for targeted advertising.

While the semantic processing happens at the textual level, the *B-hist* user interface presents the user with more familiar images of the browsed entities.