Semantic History Map: Graphs Aiding Web Revisitation Support

Jakub Šimko
Slovak University of Technology
Bratislava, Slovakia
Email: kubo.simko@gmail.com

Michal Tvarožek
Slovak University of Technology
Bratislava, Slovakia
Email: tvarozek@fiit.stuba.sk

Mária Bieliková
Slovak University of Technology
Bratislava, Slovakia
Email: bielik@fiit.stuba.sk

***Abstract*—We present a novel approach intended to reduce user effort required to retrieve and/or revisit previously discovered information by exploiting web search and navigation history. In our approach, we collect streams of user actions during search and navigation sessions, identify individual user goals and construct and persistently store visual trees representing session history. We provide users with a History Map – a scrutable graph of semantic terms and web resources with full-text search capability over individual history entries, constructed by merging individual session history trees and the associated web resources. The Map semantically organizes a user's browsing history (with the help of the Delicious folksonomy) and enables him to quickly recall information distributed over several documents and/or sessions. We present experimental results of session identification and also evaluate our prototype over generic web pages and as well in conjunction with our personalized faceted semantic browser Factic with promising initial results.**

***Index Terms*—history; exploratory; search; graphs;**

I. Introduction

Several studies have shown that web page revisitation accounts for about 50% to 80% of browsing behaviour on the Web [1]. While about 50% of revisits occur quickly (i.e., within 3 minutes), many revisits occur after much longer time with 15% being long-term revisitations (i.e., revisitations after more than a week) [2]. Existing browser aids such as differently colored links for recent visits disappear after a few days, history lists are unusable, manual bookmarks often become difficult to manage as their number and age grow, and require users to evaluate future page relevance at the time of browsing, putting extra effort on users [3]. Revisitation strategies such as re-searching or re-tracing pages by guessing the original query or the sequence of links that lead to a specific page require considerable user effort and do not ensure success as search engines return different results, even for the same query, or the user cannot remember the correct trail (i.e., sequence of links) from a start to the target page [2].

From the end-user's perspective, Semantic Web applications (e.g., query builders, browsers, search engines) provide working but complex user interfaces that induce a high cognitive load on users. This, in conjunction with new trends in web usage such as tabbed browsing, interactive (asynchronous) web applications or *exploratory search*, which stresses learning and investigative tasks that often span several web resources and/or browsing sessions [4], makes effective navigation and orientation support crucial for users.

We propose *History Map* – a novel method for history (e.g., search and browsing history) acquisition, semantic analysis, searching and browsing intended to reduce user effort required to revisit previously discovered information specifically aimed at preserving the context and semantic relations between (distributed) web resources. By analyzing users' web search logs, History Map identifies sessions (groups of actions representing search goals) and organizes them into a single graph of terms and web resources, which is thereafter used to provide user guidance and revisitation support.

In this paper, we describe our Semantic History Map approach with particular focus on experimental validation of our search session identification method, which we performed using web search logs originally released by AOL.

II. Related Work

A broad survey describing existing (semantic) web history and revisitation approaches, along with open problems including acquisition, search and visualization of history entries and metadata was presented in [5]. Term extraction algorithms and online API's like OpenCalais (www.opencalais.com) or Alchemy (www.alchemyapi.com) have been employed for metadata acquisition with good results in the English language. Multilingual approaches have yet to be perfected [6].

The acquisition of implicit user interests is a key concept of Google Personalized Search that provides history browsing adapted to estimated user preferences. Other projects use clickstream analysis and data mining techniques to estimate user search patterns [7] and detect completion of online tasks [8]. Consequently, current browser and search engine extensions support features such as full-text search in history (e.g., the Firefox plugin WebMynd) or tree-based history visualization more suited to the recursive nature of web navigation (e.g., the Firefox plugin HistoryTree, Pad Tree or WebView [5]). Still, users often encounter typical search/navigation related problems such as keyword guessing, information overload, irrelevant results, dead links or disorientation.

These problems are also prevalent in the Semantic Web environment, where most applications (i.e., query builders or browsers) provide very limited if any support for revisitation and history tracking, which, in conjunction with the complexity of semantic information spaces and query construction, seriously hinders widespread acceptance of these solutions.

The context of a visited document (i.e., associated queries or web resources) often helps users to determine its relevance and/or navigate to related pages during exploratory search tasks. In such cases, the rediscovery of distributed information located in multiple web resources is paramount and the overhead of having to search for each document separately is huge. E.g., SearchBar—a search-centric web history—defines context by matching query terms against document titles, and provides users with a hierarchic list of topic, queries and visited results [9]. Another project, HyperHistory by Nagel and Sander [10] implied semantic relationships between visited documents based on their common occurrence in web navigation clickstreams.

However, defining context using only terms in page titles is insufficient as titles cannot describe entire page content, and provide only limited information for user modeling required for personalized history browsing. With respect to open-ended exploratory search tasks, a major drawback of current approaches is the focus on specific web document retrieval as opposed to support for distributed information recall via revisitation of multiple related resources. Here, any metadata available in the Semantic Web (such as social bookmarking portals) can provide valuable information about individual visited resources and the true intent of users ultimately facilitating future information revisitation. Due to these obvious advantages, the lack of history based solutions specifically aimed at Semantic Web applications is surprising.

III. Enhancing Personal Browsing History

The goal of our approach is to provide revisitation support for previously discovered (distributed) information during exploratory search sessions. To do so, we devised two interconnected approaches to revisitation support:

- *Search History Tree* (SHT) – an in-session tree-based history visualization,
- *Semantic History Map* (SHM) – an interactive, semantically organized, graph-based visualization of longer-term browsing history that shows the original context of individual history entries.

Connections among queries and web resources in the trees/map preserve original context of the search actions. SHT also provides guidance for complex search sessions via full-text search and exploits implicitly or explicitly discovered item relevance.

Our method records user sessions (i.e., queries and visited web resources), identifies and separates individual user goals (i.e., coherent user sessions with similar terms), preserves their context by persistently storing history trees corresponding to relations between queries and visited web resources, and ultimately synthesizes navigable graphs from extracted terms, visited resources and user goals. Beside automated tree construction, we also enable users to create bookmarks and derive implicit feedback by clickstream analysis.

A. Search History Tree

We describe SHT by showing how Alice, a new resident of London, can find a restaurant serving Chinese crispy duck and preferably also fried ice cream for dessert (see Fig. 2). Alice starts with the query “Chinese food” and immediately visits two websites about Chinese cuisine creating two *web document* nodes with thumbnails. As this was not what she was looking for, she adds “London” to her query creating a new *query node*, which results in sites referring to restaurants. She now adds “crispy duck” and later simplifies the query as her husband does not like “crispy duck”. Next, she searches for fried ice cream by substituting “fried ice cream” for “duck” creating a new *query node* connected to the common ancestor. As the results are irrelevant, Alice examines the SHT, finds the query that returned the best results—“Chinese food London”—and clicks that node in the SHT to bring up those results again for closer examination.

Fig. 1. Search history tree as shown in our faceted semantic browser Factic during exploration of a photo collection (left).

SHT continuously records user activity in a browser (e.g., queries, back button use, result visits) and constructs a tree-based representation of query modifications. Queries are defined as either full text query changes or faceted restriction changes, when a semantically rich corpus is explored (as shown in Fig. 1). During sessions (i.e., at creation time) the purpose is to provide orientation support within recent queries and results, and streamline revisitation of results or queries. We also store history trees for future reference and processing.

We define sessions based on goals that users want to achieve rather than instances of web search applications. A goal is defined as a set of weighted terms related to individual sessions. Crucial is the correct recognition of different goals i. e. the correct grouping of individual web search log entries. This is a non-trivial task as users seldom work on single task in a single browser instance consequently requiring the analysis of the semantics of the performed user actions. Prior to session identification, we determine: 1. what search logs to cluster, 2. how to compute their term and URI vectors and 3. how to evaluate vector similarity.

What search logs to cluster – queries bundled with subsequent search results. Usually, we recognize two types of web search logs - query entries and the corresponding visited results. However, from a user agenda point of view,

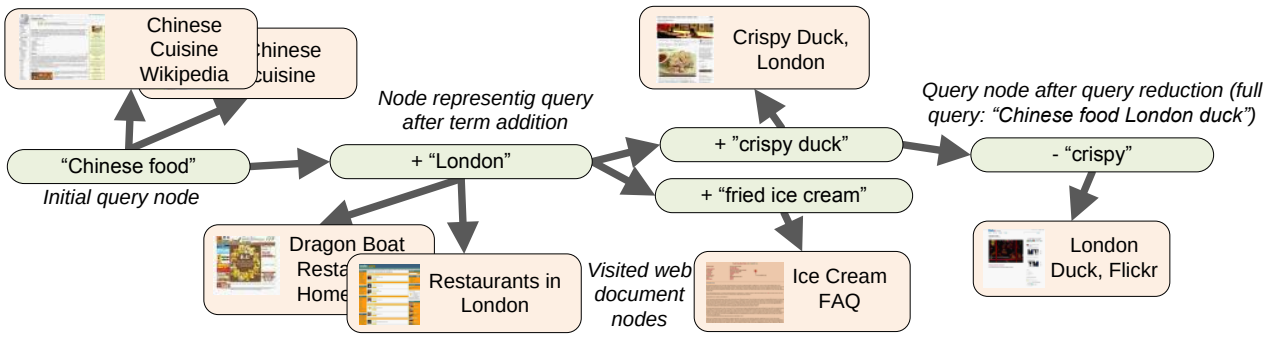

Fig. 2. A search session as shown with Search History Tree (normally shown vertically). Query nodes (shown in the middle layer) display information about query modification and form the core of the session. Web document nodes with thumbnails are attached to queries from which they were accessed.

a single query with subsequent result visits serves the same goal so we consider it to be a single element for clustering, represented by the aggregated vector and time span. We preserve the inner structure of the element for later stages, but that is transparent to session identification process.

Computing term vector of query-results group. SHT parses queries into words and using WordNet.net uses lemmatization to create weighted term vectors (excluding stopwords). Using external term extraction services (TagTheNet, OpenCalais), it retrieves term vectors of the visited results. Existing metadata and other related resources are added to vectors as URIs. The combined vector of the group is afterwards computed as the sum of the query vector and normalized sum of all visit vectors. The normalization is required to suppress "overrun" by general terms in the aggregated vector produced by term extraction services.

Comparing query-result groups. When users create a new query-result group (by entering a new query), the group's aggregated vector is compared with recently identified sessions. Each session is characterized by an aggregated term vector of its members (i.e., the normalized sum of the group's aggregated vectors). When resolving similarity, two criteria are commonly considered: term vector cosine similarity and time distance [11], [12]. We adopted this approach and combine criteria using the $N \times N \rightarrow N$ fuzzy function. The output of the fuzzy function is the final decision whether to continue in an existing session or start a new session: *certain continuation, weak continuation, uncertain, weak split, certain split*. If there are multiple candidates for a session continuation (more than one session is similar to the actual query), the query is attached to the one with the best score.

B. Semantic History Map

Individual Search History Trees are synthesized into a Semantic History Map – graph of terms and web resources. Let us consider Alice's SHM comprising two sessions, one dealing with Chinese food, another performed to find cheap local lunch facilities (see Fig. 3). Both sessions deal with similar topics and are bound closely together by merging identical results (restaurant portals) and by word proximity (food – lunch). Alice can navigate the map in order to revisit or reconstruct information distributed among several documents or sessions in the past.

In order to provide full-text search capability, we create two term indexes. The *item index* reflects characteristics of individual history entries, the *goal index* lists whole session trees and their overall term properties. Therefore, the results of history search are twofold:

- **Past goal summary** representing a whole session from the user goal index, ideal as a starting point for revisitation of distributed information.
- **Past query or web search result** corresponding to an individual history entry from the item index. Tooltips show the original context of the item, i.e. the neighboring elements in its original Search History Tree. The context serves as a cue for users, in addition to the document's text snippet or related terms, to recall whether it was the desired target document or not.

The SHM is constructed via merging stored Search History Trees into a single graph via matching identical terms and resources from different history trees. Since this alone may produce too few connections, we also connect terms by exploiting the existing folksonomy of Delicious (http://del.icio.us). We address dense (sub)graphs or too many irrelevant connections via term filters, item relevance ratings and successive filtering. The creation of Semantic History Maps follows these steps:

1) Copy each SHT into the SHM and transform query nodes from history trees into term nodes of the SHM.
2) Merge multiple identical web search results or queries into single nodes.
3) Preserve original multi-term queries as term nodes. For each particular term create a new term and attach it to the original query as a predecessor.
4) If any of the SHM's terms is also present in the external folksonomy, add all its folksonomy neighbours to the SHM (this will load directly related parts of the folksonomy into the SHM).
5) Connect multi-term queries with their subqueries. If the term set of query *A* is a subset of the term set of query *B* then **A** is a subquery of **B**.

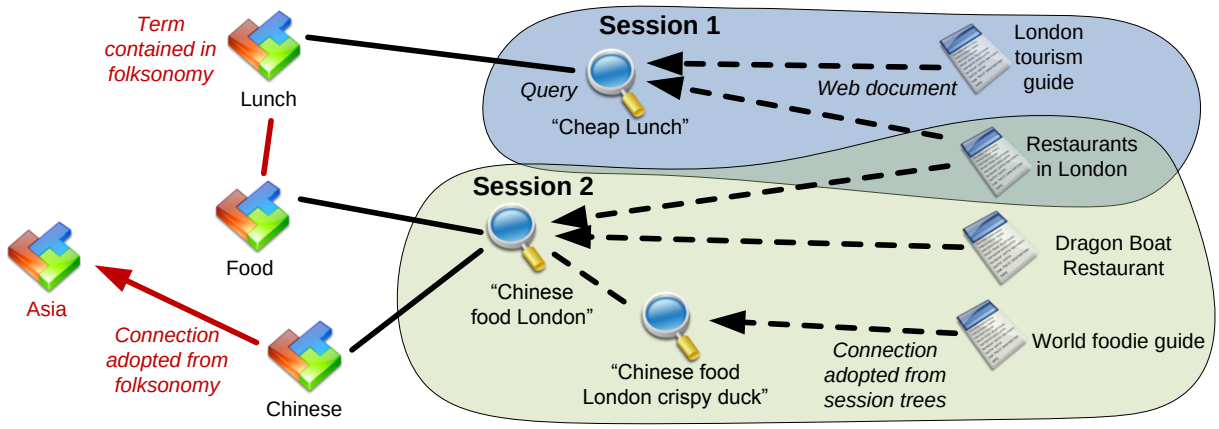

Fig. 3. Example of a Semantic History Map. Bubbles denote original session trees (right). Folksonomy terms (left) are linked to query terms (center) and web resources (right) based on data derived from browsing history.

IV. Validation

As proper identification of user goals is crucial for the success of our approach, we evaluated the success rate of our session identification method by experiments with the released AOL web search log corpus (made public in 2006 [13]) and a small group of expert users. Subsequently, we evaluated Semantic History Map during search/browsing sessions and subsequent resource revisitation via explicit and implicit user satisfaction measurement in live user experiments.

A. Experiment with AOL corpus

1) Data: As the AOL corpus is large (2 GB of logs), we picked 100 users with the highest number of performed actions (around 2000-5000 query entries – queries or result visits) and thus high potential for creating meaningful SHTs and SHMs. We identified hundreds of sessions for each of these users. However, we were unable to use term extraction over the visited results since the AOL data set contained only domain URLs. We define a certainty index for each processed query indicating the strength of the classifier's decision (certainty is the minimum of absolute values of all fuzzy matchings against previous existing sessions). We defined 3 possible values of certainty (certain, weak and uncertain) and randomly chose 20 instances of each. We decided not to choose randomly from the entire set, as we wanted to observe the behavior of our method in ambiguous cases. For each chosen query, we took 10 neighboring queries (5 before and 5 after it) and created a list of 11 queries and their respective result visits (total of 60 lists).

2) Methodology: We created an application that displayed the lists and asked participants identify individual search sessions (i.e., identify queries that addressed the same goals). Users were able to see what queries were in the list as well as visits and time gaps between the logged actions. Participants were familiarized with the concept of a session goal and asked to group together those queries, that fulfilled the same need for the user (fact retrieval, entertainment, learning...). Each list was evaluated by at least 3 participants.

Category	Sure	Weak	Unsure	**Total**
Succeed	6	10	2	**18**
Failed	12	7	15	**34**
Controversial	2	3	1	**6**
Total	20	20	18	**58**

TABLE I
NUMBER OF SUCCESSFULLY EVALUATED FRAGMENTS OF AOL SET.

Category	Sure	Weak	Unsure	**Total**
Success rate	0.76	0.75	0.53	**0.68**
Controversial couples	34	37	24	**95**
Total couples	196	190	178	**564**

TABLE II
RATIO OF SUCCESSFULLY EVALUATED BINARY PAIRS.

After evaluation, we examined differences between participants' opinions. Each evaluated list represented a set of 10 binary relationships (pairs) between a "central query" and other 10 queries in its vicinity. The pair value represented whether those two queries should be in the same session or not. We aggregated values by vote and excluded those which did not receive at least 3/4 of votes, considering them as controversial.

3) Results and Lessons Learned: Afterward, we compared the participants' knowledge with the results of our session identification method, computing the total number of successful and failed lists (table I) and pairs (table II). A list was considered failed if at least one of the 10 pairs mismatched in analytical and participants' evaluation. A list was controversial if at least half of its pairs were controversial.

The overall success rate of 68% is not satisfying but still encouraging since the AOL data set was huge in quantity but provided only minimum information on individual entries. In practice, we would have access to proper web resources metadata, term vectors or even explicit user feedback. Another issue, that might have had negative impact on the success rate was the disabled lemmatization due to misspelling of the query.

B. Live Experiments

We evaluate SHT and SHM over generic web documents and also in conjunction with our personalized faceted semantic browser Factic [14] (see Fig. 1), which facilitates exploratory search over a collection of semantically annotated photographs or scientific publications respectively. Initial experiments with our prototype SHT integrated with the faceted browser indicate promising results in terms of improved user orientation in the already explored part of the information space.

In our controlled experiment we evaluate the quantitative benefit of our history approach with a given set of exploratory and query answering tasks given a time limit against a set of baseline approaches. As a controlled experiment's time span is not long enough to cover the evaluation of long-term evolution and use of a user's personal SHM, its primary goal is to evaluate orientation support during complex search sessions. We implicitly measure task success rate and the time spent on individual tasks, and gather explicit user feedback via post-experiment questionnaires.

In our uncontrolled user study, volunteers with experience with existing baseline history tools (standard browsers with mature history extensions) will be offered to use our approach (application) as their primary search tool. The focus of this longer-term study is to gather real-world usage data and also (qualitative) feedback from users via questionnaires with the primary goal being SHM validation. The key idea is to confront the overall number of revisitations with the number of those where SHM was used. Within this scope, we distinguish cases when users use SHM directly or tried to search with the regular search function first.

Our secondary goal is to measure user affinity for new alternative widgets (graphs) instead of classic approaches (graph navigation vs. back button usage), or gather user feedback on the proposed approach/application in general.

V. Conclusion and Future Work

We described our ongoing research dealing with web search history. Its purpose is to provide a comprehensible and effective way to deal with web revisitation. We devised *Search History Tree* to provide visual in-session orientation support for complex (exploratory) search sessions. We take advantage of both traditional term extraction and matching approaches and semantics enabled approaches for SHT construction and matching against the fulltext/faceted queries and resources. SHT reveals the original context of history items in order to increase success rate of exploratory search tasks of re-learning and re-investigating content, while also providing standard browser history functionality like full-text search and personalized item relevance rating.

Next, we devised *Semantic History Map* – a way to construct a navigable semantic graph of terms and related web documents out of browsing history, which interactively visualizes a user's history improving orientation and re-exploration success rates specifically for longer term revisitation patterns.

We see two primary directions for future work – a more thorough validation of our approach via a long-term live user study, and further statistical evaluation of our session identification approach using web search logs, e.g., from AOL web search log.

Acknowledgement

This work was partially supported by the grants VEGA 1/0508/09, KEGA 028-025STU-4/2010 and it is the partial result of the Research & Development Operational Programme for the project Support of Center of Excellence for Smart Technologies, Systems and Services, ITMS 26240120029, co-funded by the ERDF.

References

[1] E. Adar, J. Teevan, and S. T. Dumais, "Large scale analysis of web revisitation patterns," in *CHI '08: Proc. of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*. New York, NY, USA: ACM, 2008, pp. 1197–1206.

[2] H. Obendorf, H. Weinreich, E. Herder, and M. Mayer, "Web page revisitation revisited: implications of a long-term click-stream study of browser usage," in *CHI '07: Proc. of the SIGCHI conf. on Human factors in computing systems*. New York, NY, USA: ACM, 2007, pp. 597–606.

[3] S. Kaasten and S. Greenberg, "Integrating back, history and bookmarks in web browsers," in *CHI '01: CHI '01 extended abstracts on Human factors in computing systems*. New York, NY, USA: ACM, 2001, pp. 379–380.

[4] G. Marchionini, "Exploratory search: from finding to understanding," *Communications of the ACM*, vol. 49, pp. 41–46, 2006.

[5] M. Mayer, "Web history tools and revisitation support: A survey of existing approaches and directions," *Foundations and Trends in HCI*, vol. 2, no. 3, pp. 173–278, 2009.

[6] C. Roussey, S. Calabretto, and F. Harrathi, "Multilingual extraction of semantic indexes," in *SADPI '07: Proc. of the 2007 international workshop on Semantically aware document processing and indexing*. New York, NY, USA: ACM, 2007, pp. 1–6.

[7] N. Sadagopan and J. Li, "Characterizing typical and atypical user sessions in clickstreams," in *WWW '08: Proceeding of the 17th international conference on World Wide Web*. New York, NY, USA: ACM, 2008, pp. 885–894.

[8] P. J. Kalczynski, S. Senecal, and J. Nantel, "Predicting on-line task completion with clickstream complexity measures: A graph-based approach," *Int. J. Electron. Commerce*, vol. 10, no. 3, pp. 121–141, 2006.

[9] D. Morris, M. Ringel Morris, and G. Venolia, "Searchbar: a search-centric web history for task resumption and information re-finding," in *CHI '08: Proc. of 26th SIGCHI conf. on Human factors in comp. systems*. NY, USA: ACM, 2008, pp. 1207–1216.

[10] T. Nagel and R. Sander, "Hyperhistory," in *HYPERTEXT '05: Proceedings of the sixteenth ACM conference on Hypertext and hypermedia*. New York, NY, USA: ACM, 2005, pp. 276–277.

[11] Z. Zhang and O. Nasraoui, "Mining search engine query logs for query recommendation," in *WWW '06: Proceedings of the 15th international conference on World Wide Web*. New York, NY, USA: ACM, 2006, pp. 1039–1040.

[12] J. Huang and E. N. Efthimiadis, "Analyzing and evaluating query reformulation strategies in web search logs," in *CIKM '09: Proceeding of the 18th ACM conference on Information and knowledge management*. New York, NY, USA: ACM, 2009, pp. 77–86.

[13] G. Pass, A. Chowdhury, and C. Torgeson, "A picture of search," in *InfoScale '06: Proceedings of the 1st international conference on Scalable information systems*. New York, NY, USA: ACM, 2006, p. 1.

[14] M. Tvarožek and M. Bieliková, "Adaptive faceted browser for navigation in open information spaces," *WWW 2007*, pp. 1311–1312, 2007.