

Finding Information Again Using An Individual's Web History

Roy A. Ruddle

School of Computing, University of Leeds, Leeds, UK.
Human Perception, Cognition & Action Dept, Max Planck Institute for Biological Cybernetics, Tübingen,
Germany.

r.a.ruddle@leeds.ac.uk

ABSTRACT

In a lifetime, an “average” person will visit approximately a million webpages. Sometimes a person finds they want to return to a given page at some future date but, having no recollection of where it was (URL, host, etc.) and so has to look for it again from scratch. This paper assesses how a person's memory could be assisted by the presentation of a “map” of their web browsing activity. Three map organisation approaches were investigated: (i) time-based, (ii) place-based, and (iii) topic-based. Time-based organisation is the least suitable, because the temporal specificity of human memory is generally poor. Place-based approaches lack scalability, and are not helped by the fact that there is little repetition in the paths a person follows between places. Topic-based organisation is more promising, with topics derived from both the web content that is accessed and the search queries that are executed, which provide snapshots into a person's cognitive processes by explicitly capturing the terminology of “what” they were looking for at that moment in time. In terms of presentation, a map that combines aspects of network connectivity with a space filling approach is likely to be most effective.

Keywords

Navigation; Web history; Information retrieval.

1. INTRODUCTION

One of the key usability challenges facing the Web is making it navigable. Navigable as a whole (where technologies such as semantic search have great potential), as an individual (so people can remember where they have been) and in a collaborative sense (so we can benefit from each other's navigational efforts).

In a lifetime, an “average” person will visit approximately a million webpages (Weinreich et al., 2006). A handful will be frequented often, and others will have no future relevance. However, in a significant minority of cases the person finds they do want to return to a given page at some future date but, having no recollection of where it was (URL, host, etc.), has to search for it again from scratch, often with frustratingly little success. No existing method is an effective aid for this, be it improvised (e.g., emailing oneself a URL), part of web browser functionality (e.g., bookmarks or history list) or a networked service (e.g., Google's Web History). See (Bruce et al., 2004).

2. An individual's history

The present paper assesses how a person's memory for their web browsing activity could be assisted so that any piece of information can be easily found again. The general approach is to generate a “map” of the activity, so retrieval takes advantage

of our everyday spatial memory processes. For example, we represent physical places in a partly hierarchical manner so that even if one doesn't know exact location of a place, step-wise localisation leads to approximately the correct position, from where a recognition-based strategy can take over. Similarly, on web we rarely teleport direct to the webpage we're looking for, and instead tend to “orienteer”, narrowing in on the target in sequence of short steps (Teevan et al., 2004).

The raw data were year-long personal browsing details, recorded using Google's Web History, which captures the terms used for each (Google) search as well as every URL that was visited (note: secure (https) requests are not recorded, and a person may turn off recording if they wish). The remainder of this paper uses the data from one person as an illustrative example.

This person's history contained 5936 items, which included 985 searches (a few of performed more than once), and visits 3243 different URLs. The main approaches of organising these items into a space are based on:

- time
- place
- topic

2.1 Time-based organisation

Time-based organisation is the approach used by Web History and one of the approaches offered by history lists (e.g., in Firefox). However, people's coding specificity for “when” events happened is limited to a resolution of several months (Wagenaar, 1986), so time-based organisation is poorly suited to remembering occasional visits that took place a substantial time ago. Generally, time is only suitable as a secondary filter, to place limits on the space to be searched in information retrieval.

2.2 Place-based organisation

Place-based organisation may be expressed at levels of detail that range from individual URLs to hosts (e.g., google.com), and displayed in ways such as an alphabetical list (e.g., as with browser history lists) or a network. The lists are unusable - unless one can remember the part of the page's name - and frequency information provides little additional benefit because most pages/hosts are rarely visited, including (probably) the one you've forgotten and are seeking.

A network can derive links from the order in which URLs were visited (see Figure 1). Although this provides more structure than a list, the result is unlikely to be a useful aid to navigation because the graph covers a very large area if displayed so that the labels are readable, related information topics may be

widely separated, and the number of nodes increases rapidly with time (studies show that the majority of URLs/hosts people visit are new; (Weinreich et al., 2006)).

Many attempts have been made to analyse Web navigation in terms of paths. However, string matching showed that there was little repetition in the paths followed, with the longest repeated path having only seven URLs (this echoes analyses of server weblogs, which showed that the median path length was approximately three clicks; (Pirolli and Pitkow, 1999)). As a result, techniques based on “trails” are unlikely to be a useful tool in aiding navigation with web histories.

2.3 Topic-based organisation

Analysis of the Web History data showed that 17% of the recorded items were searches, no doubt due to the utility of search in information retrieval and lack, so far, of effective techniques for aiding people’s memory for where they have previously been in an information space. Search queries provide snapshots into a person’s cognitive processes, explicitly capturing the terminology of “what” they were looking for at that moment in time, which may be complemented by data

(e.g., title, full text or metadata) from the items that were subsequently retrieved.

Topic-based organisation combines search queries with the item data (in the present study, these data were the caption provided for item by Google’s Web History) to create a map of the topics covered in a person’s history. String edit distance (to account for mistypes and spelling variations) and word overlap metrics were used to consolidate the queries/data into 570 topics that each contained an average of nine URLs, and the growth rate is substantially lower than for time- and placed-organisation (see Figure 2).

For presentation, a topic network indicates the key similarities between topics, whereas a tree map is a more compact way of showing the textual details of the topics themselves (see Figures 3 & 4). However both of these forms of presentation have disadvantages, with the network still occupying a large amount of space so navigation within it is time consuming, and the tree map obscuring structural information about how topics relate to each other.

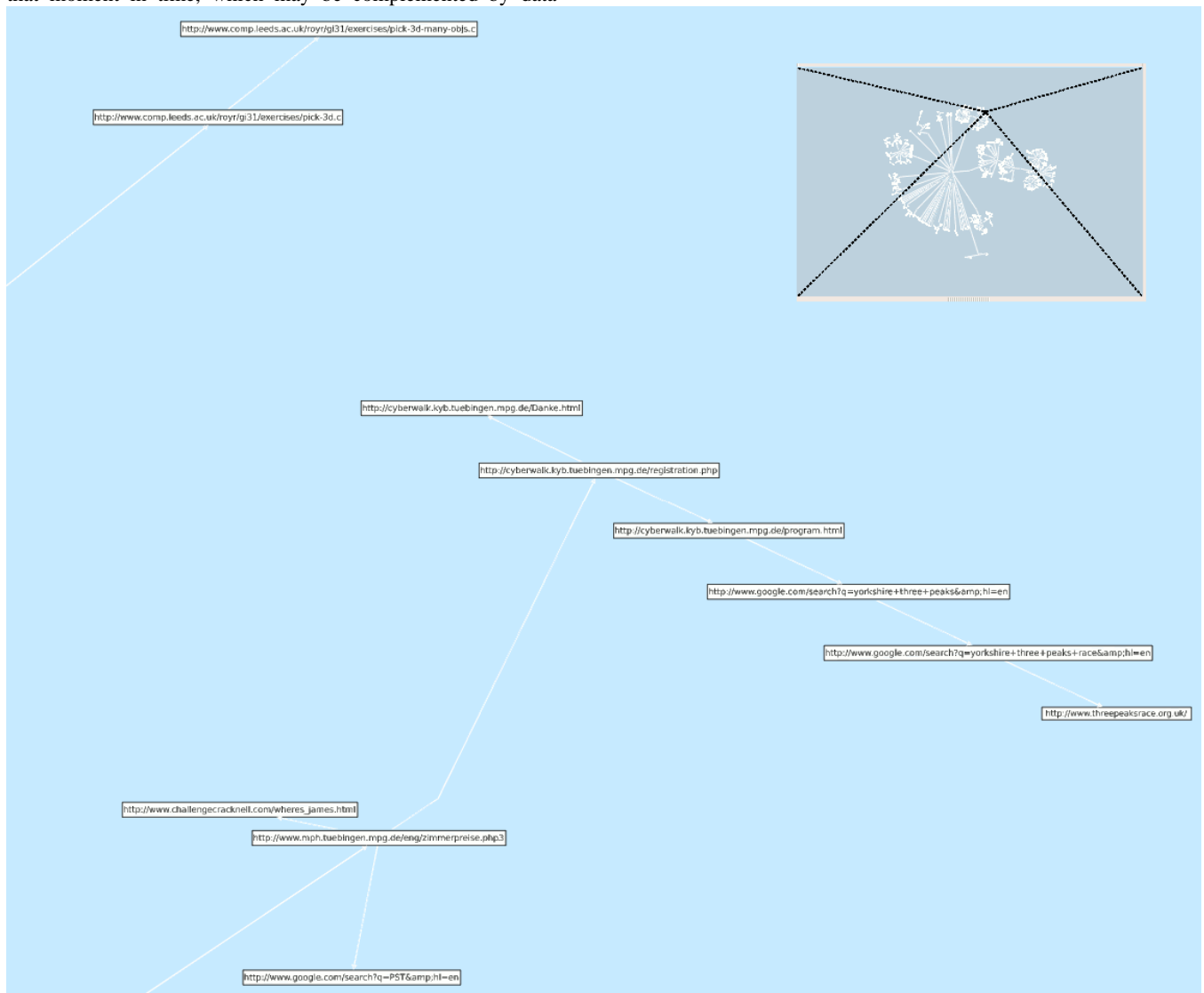


Figure 1. Overview (inset) and close up of places-based layout of a 5396 item history.

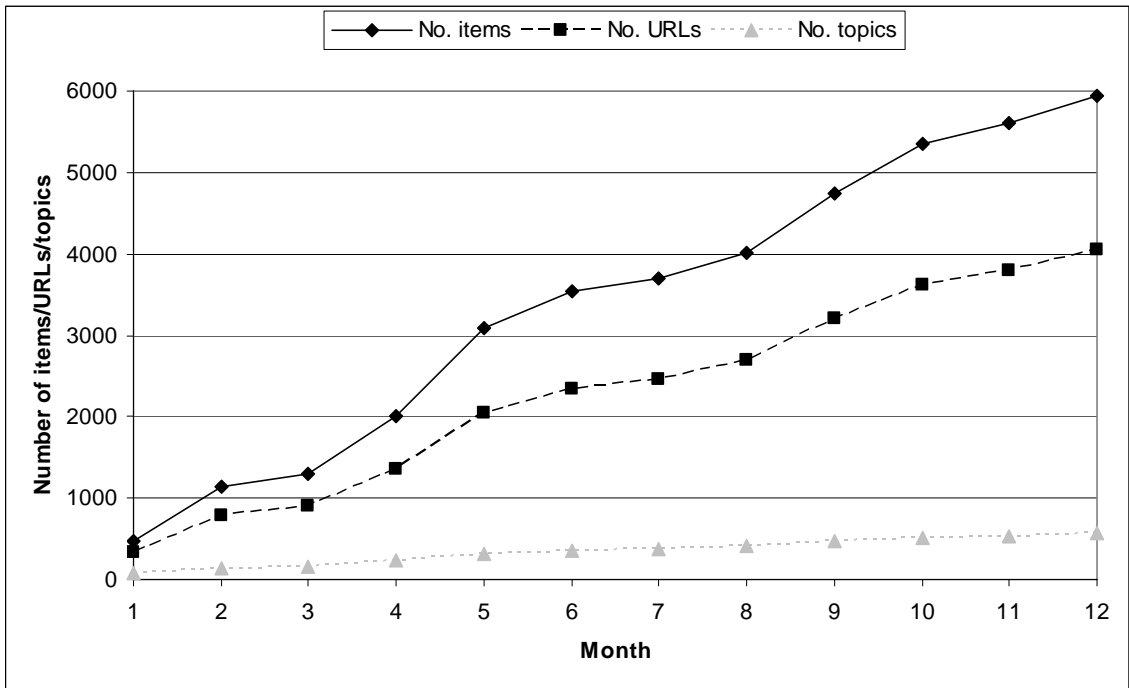


Figure 2. Growth of items, URLs and topics over 12 months of an individual's web history.

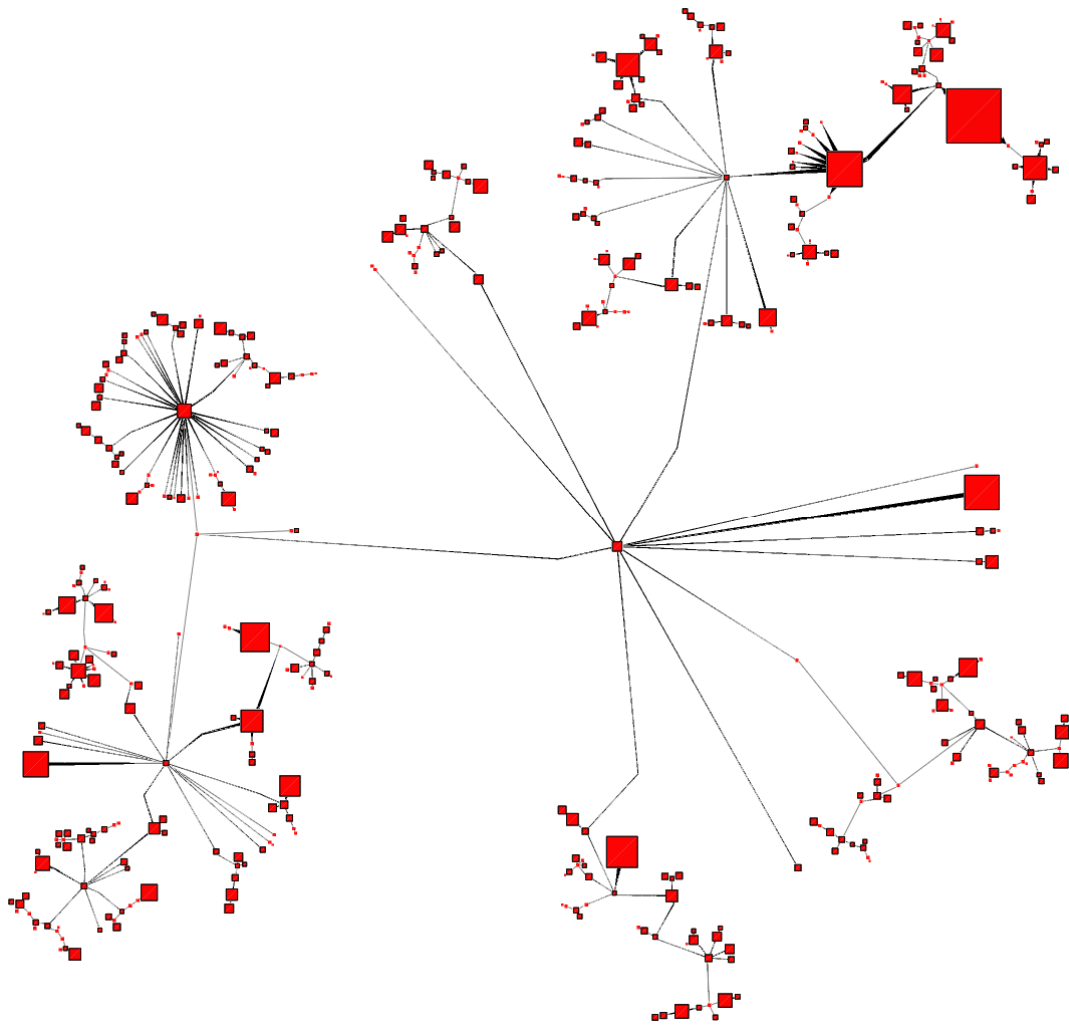


Figure 3. Network showing main associations of 570 topics. Node size indicates relative number of URLs in each topic.

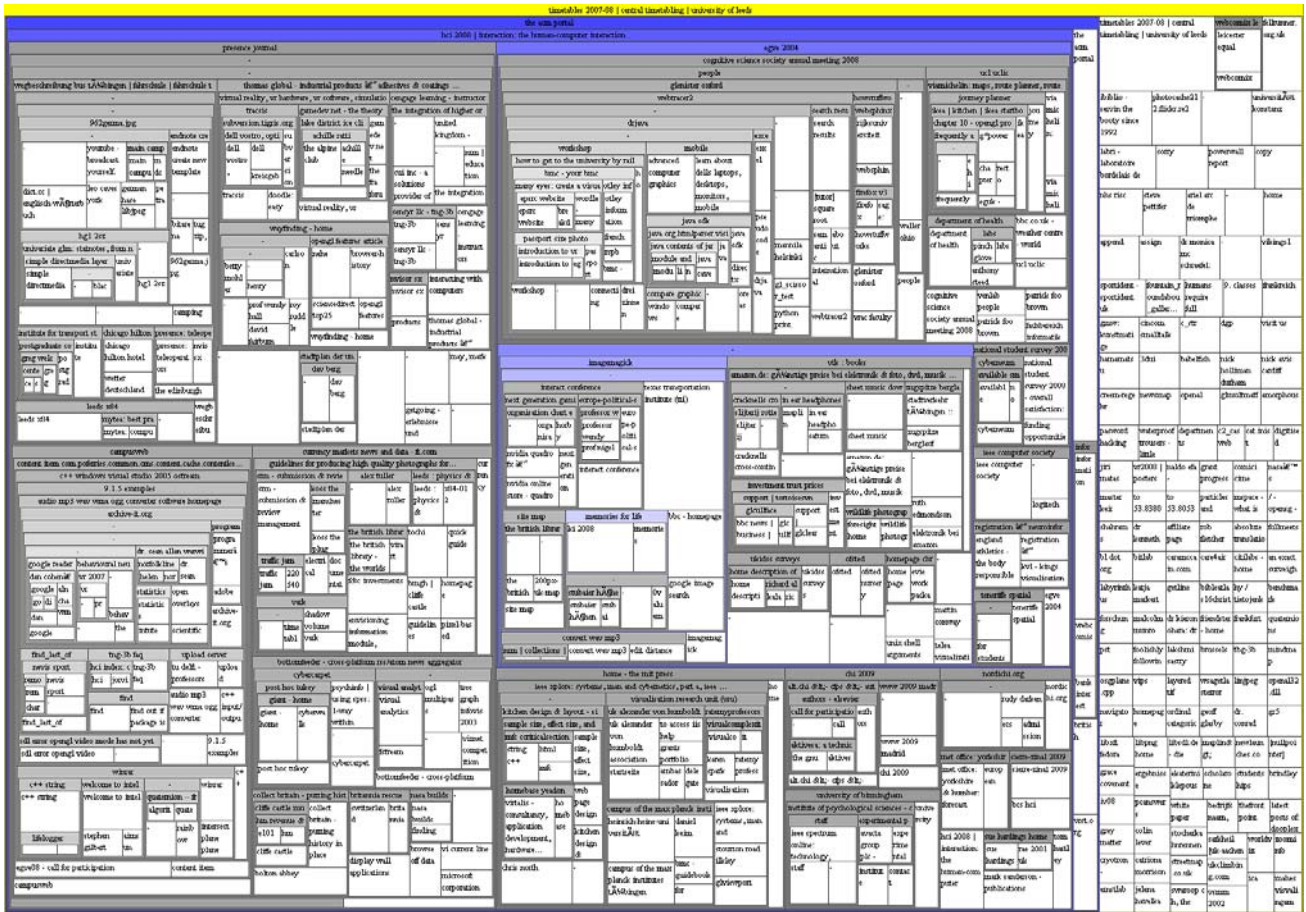


Figure 4. Tree map showing the same data as Figure 3.

3. Future work

The overall challenge is to develop aids that make information spaces as easy to navigate as the physical world, despite the fact that the former are structurally much more complex and contain less salient navigational cues.

It is technically feasible for the location and content of every item a person retrieves from the Web to be stored in a personal history, together with the actions the person performed (path, search queries, etc.). However there are many unknowns, and in particular:

- a) What data that would most benefit the finding and retrieval an item in the future?
- b) How should those data be processed?
- c) How should the resultant information be presented?
- d) What algorithms and architecture would allow the above to be implemented on a large scale?

4. ACKNOWLEDGMENTS

This research was supported by Study Leave and an Alexander von Humboldt Fellowship for Experienced Researchers awarded to Dr. Roy Ruddle, hosted by Prof. Heinrich Bülthoff, Max Planck Institute for Biological Cybernetics, Tübingen, Germany.

5. REFERENCES

[1] Bruce, H., Jones, W., Dumais, S., 2004. Information behaviour that keeps found things found. Information

Research 10, paper 207. <http://InformationR.net/ir/10-1/paper207.html>. [Last accessed 13 July 2006].

- [2] Pirolli, P., Pitkow, J., 1999. Distribution of surfers' paths through the World Wide Web: Empirical characterizations. *World Wide Web 2*, 29-45.
- [3] Teevan, J., Alvarado, C., Ackerman, M., Karger, D., 2004. The perfect search engine is not enough: A study of orienteering behavior in directed search. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, New York, pp. 415-422.
- [4] Wagenaar, W., 1986. My memory: A study of autobiographical memory over six years. *Cognitive Psychology* 18, 225-252.
- [5] Weinreich, H., Obendorf, H., Herder, E., Mayer, M., 2006. Off the beaten tracks: Exploring three aspects of web navigation. In: *Proceedings of the 15th International Conference on World Wide Web*, ACM, New York, pp. 133-142.