Activity metadata for enhancing Web document retrieval

James Gheel
Institute of Lifelong Learning, University of Ulster
Newtownabbey
Co. Antrim, Northern Ireland
+44 (0)28 90 368658
j.gheel@ulster.ac.uk

Prof. Terry Anderson
School of Computing and Maths, University of Ulster
Newtownabbey
Co. Antrim, Northern Ireland
+44 (0)28 90 366903
tj.anderson@ulster.ac.uk

ABSTRACT
Web users frequently revisit pages that are of particular relevance to them. They also tend to have these documents on-screen and interact with them for longer periods than other pages. By recording document access frequency and activity metadata, which is based on transient user interactions within the browser, it is possible to infer the importance the user attaches to a given page. Activity metadata, access history metadata and document content can be stored in a locally held repository. This repository will help the user remember and quickly retrieve high interest documents they have accessed in the past.

This paper discusses the nature of activity metadata generated and recorded during Web document use, how it relates to the document, and how it can be gathered, stored, represented and visualized for subsequent retrieval. Selected prototype implementation issues are also presented.

Keywords
Metadata, user activity, document retrieval, document representation, document visualization.

1. INTRODUCTION
The number of documents available on the Web continues to increase enormously each year. As Lyman and Varian state [1], it has become "the information medium of first resort for its users". The great benefit of information availability through the Web is reduced as users struggle with information overload and loss of potential productivity.

Traditional mechanisms such as databases have been able to cope with vast increases in information due to their scalable design and techniques such as clustering. Several improvements have been made to assist users in searching for unknown information in unstructured repositories by learning from their searching habits and by cross-referencing the context of a user's search from existing documents [9].

The arguably unparalleled success of Web search engines such as Google and Yahoo, is largely based on content indexing and sophisticated use of information in page links. Highly effective algorithms have been devised to assess the level of importance the Web collectively attaches to a particular site or page. However, comparatively little research has been focused on the importance a particular site or page has for an individual user, often only to predict future pages of interest [3, 14], or to allow keyword searching of previously view pages [15]. In fact, there is strong evidence that Web page revisitation is a prevalent behaviour [12].

While bookmarks are simple and highly effective, they can be somewhat cumbersome to manage and keep up-to-date [11, 12]. Address-bar histories and auto-complete functions perform a similar function, but have the advantage (and sometimes disadvantage) of being automatically maintained by the browser. They make effective use of the metadata of past browsing behaviour. However, we will argue that there is much more metadata that can be recorded and used to help users revisit documents, a process that often involves recalling the context of use when the document was last seen [2].

This paper takes the concept of Web document metadata further and suggests that documents should be linked with metadata derived from the user's interactions. It is human nature to repetitively organise our life and carry out our lives in much the same way each day, week, month, etc. This also applies to our use of Web documents. How we have previously used the documents can support their future retrieval. If we solely rely on a particular search engine to find the same information, often we won't enter the same keywords and may have a tedious and frustrating task in re-finding a document.

2. DOCUMENT STORAGE AND RETRIEVAL
A user selects and views a document because of its relevance to their task. Our local file storage typically contains documents we have already seen. But in the case of Web-based documents, if no bookmark or entry in favourites has been set, or if it is a significant period since it was visited, there is likely to be no trace either in the browser history or cache. The only trace is in the user's memory as a page location (URL), title or some fragment of content. But low frequency of access does not imply that it is unimportant. Such access may be vital, for example, to lower-priority research topics, annual financial guidance or work-related manual.

There are several problems with bookmarks and browser histories [11]. Firstly, bookmarking mechanisms revert to the same hierarchical file and folder metaphor that exists for our local storage, implying that we will organise information from different

Supporting Human Memory with Interactve Systems, workshop at the HCI 2007 (British HCI conference 2007), September 4th, 2007, Lancaster, UK

locations into different categories suitably structured. Gemmell et al. for example [6] argue several valid points in their justification for a non-hierarchically organised (and unlimited capacity storage system). They point to several studies that show filing objects into a single hierarchy is too restraining, that items usually belong in several categories anyway and that users would rather not have to categorise items at all. Secondly, an historical list of our previously accessed documents doesn't incorporate any contextual information about the document we viewed. That is, how long did the user spend reading the document, what was its key topic, author, etc? Thirdly, the document the user once viewed may have been deleted, replaced or modified by the time they revisit a bookmark - to a large extent resulting in lost information and knowledge for the user.

In the case of Web documents, navigation is much less structured than for a locally held file. After using a search engine, even if exactly the same keywords are entered as were originally used, a user might follow a different set of links, taking them further from the document they were looking for and perhaps even away from the topic they were browsing [4].

The next section discusses the potential of using the document contents, structure and implicit information in how we browse, retrieve and use documents to assist us further in the retrieval and use of these and related documents.

3. DOCUMENT-INFLUENCED USER ACTIVITY

The number of times we view or open a document is a valuable indicator of its importance to a user – perhaps its perceived authority on some topic, or as a highly reliable source of information. However, if the time spent on a page is usually very brief, then it is probably only a link to a more useful page [12]. Recalling even approximately the day or time we last accessed a document is often a major part of how we remember and relocate a document.

Some browser history facilities now track the number of times the user accesses a Web page document. However, they tend to be quite short term, such as a matter of days, and cannot assist in refinding material over a long period.

The size of a document will influence the amount of time required to read it, in the same way that more text in a document and fewer images will require the user to read more. On a library shelf this is obvious perhaps by the size of a book in conjunction with the size of the characters used on the pages.

When we revisit a Web page, being able to tell automatically if it has changed and even the changes that were made is beneficial for a user in the time taken to review a document. Subsequently, this can be used as an indicator of past change frequency and quantity.

A user may scroll repeatedly in a Web document, indicating to the user's attentiveness to a document [7]. Similarly, in a browser with a tabbed user interface, repeatedly flicking to a certain tab indicates a high level of relevance to a task or subject of interest. With a physical book, research paper or magazine, this would show as marks on the pages, a change in the colour at the edge of the pages, or generally a degradation of paper quality increasing over time.

The duration of a document being open, taking into account whether it is in focus or not, directly reflects the importance it has to our task and perhaps the quality of the content – in the same way that a document on a user's desk is important, but not as important as one in the centre of the desk.

Additionally, if a user noticeably takes information from a document, i.e. copying and pasting elsewhere, this points towards another level of the document's relevance. Conversely, if a user is required to enter information into a Web form, for example in an information request or on a forum, being able to recall this text and interaction with the Web page could help relocate it.

Finally, usage of hyperlinks is of key importance. For example, the main value of a 'hub' page is a set of pointers to a chosen topic. The number of times links are clicked in a document therefore indicates something of that page's worth to the user. The short duration on screen of a sequence of documents may suggest relevance to a target document in that succession of links. Being able to recreate the steps made in a browsing trail and visually showing this at another point in time can mimic the path in a user's long-term memory, thereby rekindling their ability to remember and find a particular document and related documents [5].

3.1 Activity Metadata

In addition to the metadata that occurs as part of the document itself and its contents [10], the user's interaction with a Web document can provide guidance to usage as important as that with printed documents and books in their aging and signs of usage.

This document metadata can be further clarified and classified to assist in collecting, analysing and representing to a user for document retrieval.

Table 1. Web Document Activity Metadata

Higher level user-document activity	**Lower level user-document activity**
Number of accesses	Mouse movement over document
Date/Time of last access	Scrolling required to view document
Duration of document onscreen – focused	Link selection/click count on document
Duration of document onscreen – unfocused	Quantity of data entry
	Text selected/copied

Table 1 shows the two different types of Web document activity metadata. In the first instance, the document itself as a whole influences or initiates the activity performed i.e. higher level user-document activity. Secondly, the user influences the activity performed on the document or parts of it, i.e. lower level user-document activity.

4. REPRESENTING WEB DOCUMENT ACTIVITY

The resulting goal of gathering and generating activity metadata is to assist the user in subsequent retrieval of documents they have already encountered – a time consuming task, particularly as the number of documents a person has accessed increases over time and as the length of time since the document was last seen reduces memory retrieval capability.

To this end, we have undertaken to derive information from users' browsing habits when using the Firefox Web browser and the design of a system, called MetaReminder [13], that leads to a way of visualising the results.

Custom browser code is invoked in a Firefox extension when a page is loaded in the Browser. This code stores the currently viewed Web page document exactly as it has been downloaded to the browser. The HTML document is checked for malformed HTML and then re-formatted to allow for Document Object Model (DOM) parsing. Non-activity metadata that is relevant to the document, is extracted such as: title, description, number of links and size [10] and creates (or initiates creation of) the higher level user-document activity metadata as defined in column one of Table 1. Activity metadata is stored as the result of user interaction with the page.

The activity metadata is combined in one complete XML document and maps as a one-to-one relationship to the original HTML document. Keeping both parts separate ensures flexible access to information, i.e. quick display of the original HTML Web page document and manipulation of the metadata, for example in a search.

When a user wishes to retrieve a previously viewed Web page, they activate a button from the browser toolbar and a locally stored Web page containing Java applets is displayed within the browser. This shows visual representations of document history navigation, based on the activity of the user when engaged with the Web page document, in addition to embedded document metadata and browser generated metadata.

4.1 Visual Representation

In order to properly support the task of retrieving previously viewed Web page documents, using a user's activity metadata on those documents as previously described, a suitable visual representation of the metadata needs to be provided to the user.

Our approach to this in MetaReminder is 3-fold. Firstly, a time-based arrangement of documents is employed (Figure 1), such as that in Lifestreams [8], where the ordering of a list of Web page documents by time allows other temporal related metadata (such as link succession and document usage duration) to be incorporated in the visual arrangement of those documents. Secondly, a flexible query mechanism allows users to dynamically determine the metadata values of a desired document, either individually or in a compound manner, to eventually narrow down the pool of available documents that match the selected metadata criteria. Lastly, a free-text search allows the user to search for any keyword in their repository of browsed documents, either in the Web document itself or in the annotated comments they have added during Web browsing.

5. RESULTS AND CONCLUSION

Firstly, it has been found that a large number of the Web pages that users read are visited very infrequently – in this instance, 72% were visited only once and 96% visited 5 times or less. This has confirmed and justified the purpose of this research and the necessity in helping users remember and retrieve previously viewed Web page documents – particularly those that are less likely to remain in a user's long term memory either by lack of relevance at the time, or due to a long period of time since last seen. In addition, the results have shown that as the number of times the user visits a page, the activity performed on that Web page document as a result increases proportionately. Also, users tend to visit a large number of pages with relatively small values of metadata, for example, with a small size and a small number of links, thereby supporting the theory that users need support in remembering a combination of metadata items in order to distinguish documents from each other when finding and reminding.

The overall use of the MetaReminder tool has been successful, with an 84% success rate from a group of evaluators in finding documents last used over a range of increasing time spans and that have been used infrequently.

The evaluation users have shown in the use of MetaReminder that they prefer the flexibility in being able to match the metadata they remember to their Web document management tasks, without being restricted for example by a narrowly focused tool or facility such as a Web browser history or bookmark function. This was explicitly evident in the use of the Metadata Dynamic Query facility of MetaReminder. This flexibility was confirmed in the feedback gathered from open answered questions.

An important finding was that the MetaReminder tool had a significant impact on what users felt was important in terms of metadata in helping them in their Web document management tasks once exposed to new types of metadata. A comparison of the rating of each of the metadata items used in MetaReminder by the evaluation users and a group of people who didn't use it, showed that those who weren't exposed to MetaReminder were more likely to find traditional items useful such as title of page, location of page, subject and description.

This is in contrast to the evaluation users of MetaReminder who were more likely to find the activity metadata of date and time of access and number of accesses (non-transient metadata), followed by subject, description and annotated comments as useful. In addition it has been discovered that that the users had formed a particular mental process of what they found useful in helping them in their task, i.e. they look for significant overall document activity first (date, time, access count), followed by an overall summary of the document from document metadata (subject, description and comments). This was evident in both the results from the use of the Browsed Document Stream facility in MetaReminder and by the rating of the metadata items in the questionnaire.

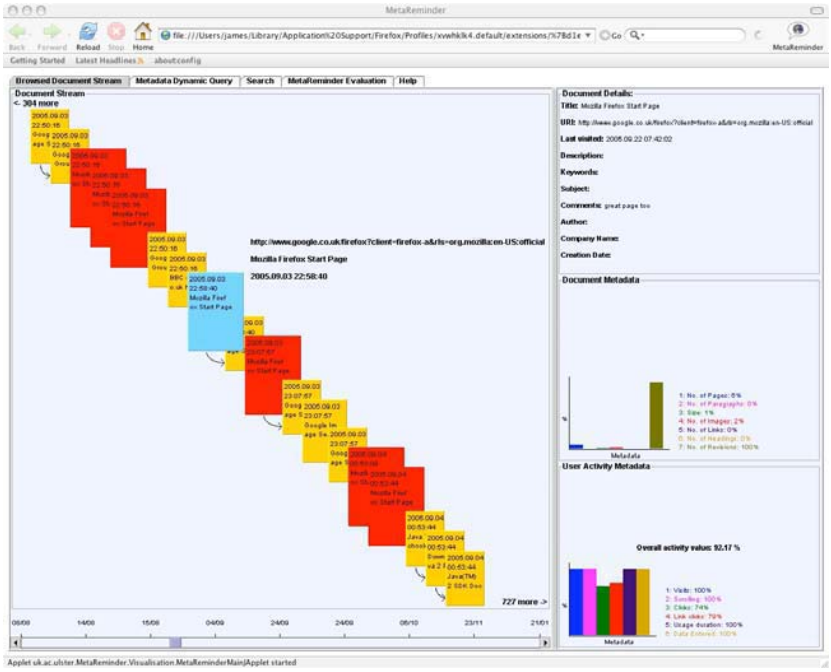

Figure 1. The MetaReminder Browsed Document Stream.

6. ACKNOWLEDGMENTS

This work was partly funded by the Centre for Software Process Technologies at the University of Ulster which is supported by the EU Programme For Peace And Reconciliation in Northern Ireland and The Border Region Of Ireland (PEACE II).

7. REFERENCES

[1] Lyman, Peter and Hal R. Varian, *"How Much Information"*, 2003. Available from: http://www.sims.berkeley.edu/how-much-info-2003.

[2] Villa, Robert and Chalmers, Matthew, *"A framework for implicitly tracking data"*, Proceedings of the Second DELOS Network of Excellence Workshop on Personalisation and Recommender Systems in Digital Libraries, Dublin City University, Ireland, 18 - 20 June 2001.

[3] Shavlik, Jude and Goecks, Jeremy, *"Estimating Users' Interest in Web Pages by Unobtrusively Monitoring Users' Normal Behavior"*, Proceedings of the 2000 AAAI Spring Symposium on Adaptive User Interfaces.

[4] Levene M. and Loizou G., *"Web interaction and the navigation problem in hypertext"*, Encyclopedia of Microcomputers, 28(7):381-398, Marcel Dekker, NY, 2001.

[5] Chalmers, M., Rodden, K. and Brodbeck, D.: *"The order of things: activity-centred information access"*; Computer Networks and ISDN Systems, 30 (1998), 359–367.

[6] Gemmell, J., Bell, G. and Lueder, R., *"MyLifeBits: a personal database for everything"*, in Communications of the ACM, vol. 49, Issue 1 (Jan 2006), pp. 88.95.

[7] Claypool, Mark, Le, Phong, Waseda, Makoto and Brown, David, *"Implicit Interest Indicators"*, In Proceedings of ACM Intelligent User Interfaces Conference (IUI), Santa Fe, New Mexico, USA, January 14-17, 2001.

[8] Freeman, E. and Fertig, S., *"Lifestreams: Organizing your Electronic Life"*, AAAI Fall Symposium: AI Applications in Knowledge Navigation and Retrieval, November 1995, Cambridge, MA.

[9] Lawrence, S., *"Context in Web Search"*. IEEE Data Engineering Bulletin. Volume 23, Number 3, pp. 25-32, 2000.

[10] Gheel, J., and Anderson, T., *"Data and Metadata for Finding and Reminding"*, In Proceedings of 1999 IEEE International Conference on Information Visualisation, July 14 - 16, 1999, London, England.

[11] Abrams, David, Baecker, Ron and Chignell, Mark, *"Information Archiving with Bookmarks: Personal Web Space Construction and Organization"*, In Proceedings of CHI '98.

[12] Cockburn, Andy and McKenzie, Bruce, *"What do Web users do? An empirical analysis of Web use"*, in International Journal of Human-Computer Studies, 54(6), 903-922, 2001.

[13] Gheel, J., *"MetaReminder"*, Available from: http://homepage.mac.com/james.gheel/MetaReminder/

[14] Goecks, Jeremy, Shavlik, Jude, *"Automatically Labeling Web Pages Based on Normal User Actions"*, Proceedings of the Workshop on Machine Learning for Information Filtering, 1999 International Joint Conference on Artificial Intelligence (IJCAI), Stockholm, Sweden, 1999.

[15] Google Inc., *"Google Web History"*, Available from: http://www.google.com/psearch