Personal Web Revisitation by Context and Content Keywords with Relevance Feedback through Mining Usage History

[1]Priti Chorade,[2]Mr.S.M.Shinde
[1] M.E.Student ,SVERI's College of Engineering ,Pandharpur,
[2] Assistant Professor ,SVERI's College of Engineering ,Pandharpur,

***Abstract*-Design a web revisitation framework by considering user behavior and ambiguity of the web contents. The web is playing a significant role in delivering information to users' fingertips. A web page can be localized by a fixed url, and displays the page content as time-varying snapshot. In web crawling phenomena web behaviors like web revisitation is to refind the previously viewed web pages, not only the page url, but also the page snapshot at that access timestamp.Work introduced human's natural recall process of using episodic and semantic memory cues to facilitate a personal web revisitation technique through context and content keywords. Technique proposedis for context and content memories' acquisition, storage, decay, and utilization for page re-finding.**

Index Terms*: -*Web revisitation, access context, page content, relevance.

I. INTRODUCTION

Re-back to previously viewed web pages is a common yet hard task for users due to the large amount of personally accessed information on the web. This work leverages human's natural recall process of using episodic and semantic memory cues to facilitate recall, and presents a personal web revisitation technique called WebPagePrev through context and content keywords.Techniquefor context and content memories' acquisition, storage, decay, and utilization for page re-finding. A relevance feedback idea is involved to tailor individual's memory strength and revisitation habits.

- The system used to compare with the various web revisitation tool as Memento, History List method, and Search Engines, the proposed system delivers the best re-finding quality.
- With relevance feedback, the finding rate increases and average rank error minimizes compared to stable memory management strategy.Among time, location, and activity context factors and context+content based re-finding deliver the best performance, compared to ontext-based re-finding and content-based re-finding.

II. OBJECTIVES

Some methods and tools like bookmarks, history tools, search engines, metadata annotation and exploitation, and contextual recall systems have been used to support personal web revisitation. The context information considered in this work includes access time, location and concurrent activities automatically inferred.
The main Objectives of our work are:

- Design a personal web revisitation technique that allows users to back to their previously traversed pages by access context and page content keywords.
- Designing dynamic tuning strategies to tailor to individual's memorization strength and recall habits based on relevance for performance improvement.
- Designing mechanism to the prediction of users' revisitation.
- Design and implement technique to support users' ambiguous re-finding requests.
- Evaluate the effectiveness of the proposed technique *system*, and report the findings in web revisitation users.

III. RELEVANCE

Relevance feedback is an interactive procedure that has been shown to work particularly well in classical information retrieval and more recently in web search domain. When a user interacts with the *system*during web revisitation phase, s/he can either manually enter some context information, or pick up suggested values from contextual hierarchies by clicking the leftside buttons of time, location, and activity. Each contextual hierarchy is dynamically maintained by analyzing the user's clicking behaviors and the statistical frequencies of captured context instances. Frequently accessed context items are top listed in the corresponding contextual hierarchy. User's types in re-finding requests are automatically corrected by the system based on its indexed content and context keywords. The user can scroll the page up and down with the mouse wheel to inspect all the result pages. If the user double-clicks and dwells on a page by printing,downloading, or reading for a while, we treat the page query relevant. With this feedback information, the web revisitation engine gets to know the system performance, and tune related necessary parameters to improve it gradually. Meanwhile, to keep pace with the user's context memorization strength, the engine tunes the leveled decay rates for probabilistic context memory according to the located levels of typed context keywords.

IV. LITERATUREREVIEW

To support personal web revisitation, some techniques and tools are developed, including bookmarks, history tools, search engines, metadata annotation and exploitation, and contextual recall systems.

A.Bookmarks

Apart from *back=forward* buttons, manually/ automatically bookmarking favorite web pages in web browsers enables users to get back to the earlier accessed pages. According to the user is every visited web page and browsing preferences built bookmarks automatically and organized them into a regency list or layered structure respectively.

Gamez *et al.* Further used classifiers to determine a few of the bookmarks that are more probably to be hit later and showed them in the browser bookmarks personal toolbar, so that the user can access the desired web page through a single mouse click.

Kawase *et al.* recommended visited pages relevant to the currently viewed pages, and presented them with a dynamic browser toolbar. Besides, the *search bar* tool allowed users to organize their essential search keywords and click pages under diverse topics. Users can make entries on the topics for easy navigation. With the *Landmark* tool, users can also mark a specific part of the page.

B.History Tools.

History tools of web browsers maintain the user has accessed URLs chronologically according to visit time (e.g., today, yesterday, last week, etc.), and accessed page titles and contents.

Tauscher and Greenberg analyzed 10 weeks of usage data collected from 30 participants when using a commercial browser chrome, and discovered that people tend to revisit pages just visited, access only a few pages frequently, browse in tiny clusters of related pages and generate only short sequences of reproduced URL paths, which can be used to develop guidelines for the design of history mechanism.

Google Web History keeps user's search keywords and clicked pages, and categorizes them into image, news, ordinary page, etc. Users can navigate or search the history by page title/content keywords.

Contextual Web Historyimproved the visual appearance of the web browser history by combining website thumbnails and content snippets to assist users to browse or search their histories by time quickly.

Visual History Toolencoded four features of a visited web page, which consists of user's page interests measured by dwell time, the frequency and recency of the visit, and navigational associations between pages. List- and graph-based forms are then adapted to provide navigation histories.

xMem improved history mechanisms by intermixing semantic aspects with the temporal dimension of the accessed pages. It organized the pages into groups and presented a navigational history instead of merely exploiting time sort history.
Search Panel connected web page and method metadata into an interactive design of the retrieved texts that can be done for sense-making, navigation, and re-finding texts.

C.Search Engines

Tyler and Teevan studied how search engines are used for re-finding previously found search results. It explored the differences between queries that had substantial/minimal changes between the previous query and the revisit query. Through observing the differences between re-finding behavior occurring within the same session and across multiple sessions, the results showed that cross-session re-finding might be a way to bridge a task between two different sessions.

Re: Search supported simultaneous judgment and re-finding on the web. Past queries were indexed to recognize repeated searches, and the most recently viewed results were stored in a result cache. When a user's query was similar to a preceding query, *Re: Search* obtained the current results from a current search engine, and fetched relevant previously viewed results from its cache. The newly available results were then merged with the previously viewed results to create a list that supported intuitive re-finding and contained new information. Adar *et al*. analyzed 5-week web interaction logs from over 612,000 users, and interview studies from 20 participants who installed software to log web page visits for one to two months. They identified twelve different types of revisitation curves corresponding to four groups (i.e., fast, medium, slow, and hybrid revisits), and regarded each of them as a signature of user behavior in accessing a given web page. The analysis of revisitation behaviors for web users in various contexts could empower search engines to support fast better, fresh, and useful finding and re-finding.

D. Metadata Annotation and Exploitation

Haystack stored arbitrary objects of interest to a user and recorded arbitrary (predefined or user-defined) properties of and relationships between the stored information. It coined a uniform resource identifier (URI) to name anything of interest, including a document, a person, a task, a command/menu operation, or an idea. Once named, the object can be annotated, related to other objects, viewed, and retrieved through arbitrary properties, which served as useful query arguments, as facets

for metadata-based browsing, or as relational links to support the associative web browsing. Bearing the similarity to *Haystack*, a SQL-based *MyLifeBits* platform was built for designating, storing, and obtaining a personal lifetime archive. It stored content and metadata for a variety of item types, including contacts, documents, email, events, photos, music and video, which were linked together implicitly using "*time*", or explicitly linked with typed links such as a "*person in photo*" connection between a contact and a photo, or a "*comment*" link between a voice comment and a document. By connecting, the traditional folder (directory) tree was replaced by a more general "*collections*" function using a directed acyclic graph (DAG).

E. Leveraging Access Context and Page Content.

Stuff I have Seen built a unified index of information that a person has seen on the computer, including emails, web pages, documents, media files, calendar appointments, etc., and allowed the use of such contextual cues as time, author, thumbnails, and previews to search for information. Deng *et al*. allowed users to re-find web pages and local files through previous access context, including time, location, and concurrent activities. It clustered and organized context instances in context memory, and dynamically degraded the context instances to mimic user's memory decay feature. A query-by-context model for information recall was presented upon the context memory. *You Pivot* leveraged human user's natural method of recall by allowing a user to search through their digital history (e.g., files, URLs, physical location, meetings, and events) for the context they do remember. The user can then Pivot, or see everything that was going on while that context was active. Further, *YouPivot* displayed a visualization of the user's activity, providing another method for finding context. *Memento* provided users with detailed topic-phrases extracted from access context and page content to aid web revisitation.

V. PROPOSED WORK

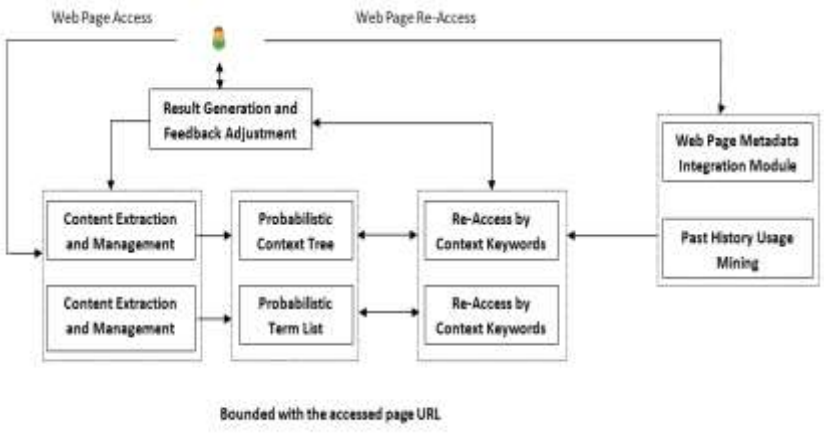

Figure 1. Block diagram of the proposed system

It consists of two main phases.

Phase 1 - Preparation for web revisitation

When a user obtains a web page, which is of potential to be revisited later by the user (i.e., page access time is over a threshold), the *context acquisition and supervision module* captures the current access context (i.e., time, location, activities inferred from the currently running computer programs) into a probabilistic context tree. Until the *content extraction and management* module performs the unigrambased extraction from the displayed page fragments and obtains a list of probabilistic content terms. The probabilities of acquired context occurrences and quoted content terms indicate how fit the user will refer to them as visual cues to get back to the earlier focused page.

Phase 2 - Web revisitation

When a user demands to get back to a earlier focused page through context and content keywords, the *re-access by context keywords* module and *re-access by content keywords* module search the probabilistic context tree repository and probabilistic term list repository, respectively. The *result generation and feedback adjustment* module combines the two search results and returns the user a ranked list of visited page URLs. The relevance feedback mechanism dynamically tunes essential parameters (including memories' decay rates, page reading time threshold, interleaved window size threshold, weight vectors in computing the association and impression scores), which are critical to the construction and management of context and content memories for personal web revisitation. This section describes the acquisition and management of user's previous access context and content-related information to prepare for user's web revisitation.

Context Acquisition and Management Module

Context Acquisition

Three kinds of user's access context, i.e., access time, access location, and concurrent activities, are captured. While access time is determinate, access location can be derived from the IP address of user's computing device. By calling the public IP localization API, can map the IP address (e.g., "166.111.71.131") to a region (e.g., "Beijing, Tsinghua University"). To get a high-precision location, we further build an IP region geocoding database, which could translate a static IP address to a definite place like "Lab Building, Room 216". If the user's GPS information is available, a free GPS localization application could also help localize the

user to the point of Interest (POI) in the region. User's concurrent activities are inferred from his/her computer programs, running before and after the page access.

Construction of Probabilistic Context Trees

Access context (i.e., time, location, and concurrent computer programming activities) is organized in a **probabilistic context tree** to support generalized revisit queries due to human user's cognitive understanding and progressive decay during learning and recalling in [0,1], stating the likelihood that this context node is used as a contextual cue. In the activity subtree, leaf nodes' scores are the association scores defined. As time and location are deterministic, leaf nodes in the time and location sub-trees.

Decay and Reinforcement of Probabilistic Context Trees

The obtained probabilistic context trees will evolve dynamically in life cycles to reflect the gradual debasement of human's episodic memorization as well as the context keywords that users will use for recall. That is, for each connection in the probabilistic context tree, its associationscore will progressively decay with time.

Content Extraction and Management Module

Apart from access context, users may also get back to the previously viewed pages through some content keywords. Instead of extracting content terms from the full web page, only consider the page segments shown on the screen. There are many term weighting schemes in the information retrieval field. The most generic one is to calculate term frequency-inverse document frequency *(tf-idf*). For personalized web revisitation, merely counting the occurrence of a term in the offered page segment is not sufficient. Also, the user's web page browsing behaviors (e.g., visitation time length and highlighting or not), as well as page's subject headings, are counted as user's opinion and potential interest notices for succeeding recall. In a similar practice as access context, we attach an impression score to each quoted content term *d*, determining how likely the user will point to it for recall based on the four normalized features.

VI. RESULT AND DISCUSSION

Relevance feedback is an interactive procedure that hasbeen shown to work exceptionally well in classical information retrieval and more recently in web searchdomain. Fig. 2 shows *top*-4 previously visited web pages below the re-finding context keywords *{"busy", "programming", "at lab", "in April"}*, and content keywords *{"retarget", "project"}*. The user can scroll up & down with themouse wheel to see all the result pages. If the userdouble-clicks and dwells on a page by printing,downloading, or reading for a while, we treat the pagequery relevant.

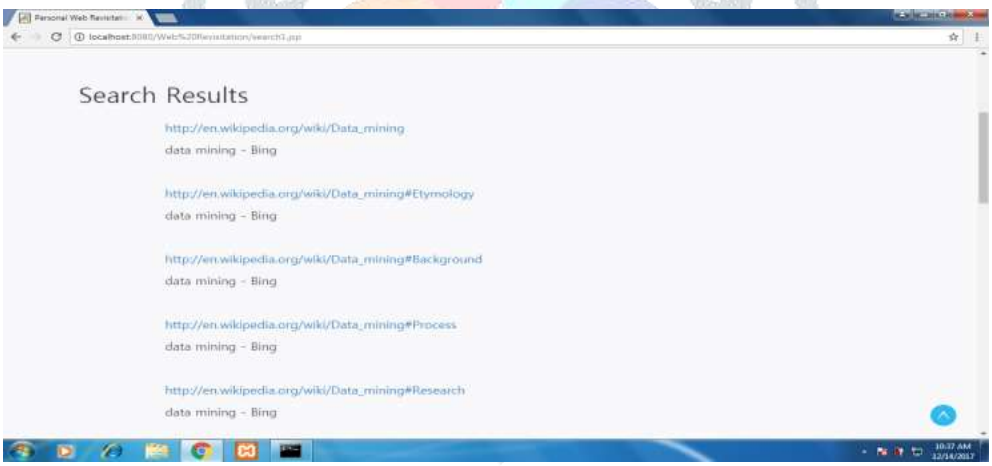

Figure 2:Search Results

History of web browsers maintaina user has accessed URLs chronologically according to visittime (e.g., today, yesterday, last week, etc.), and accessedpage titles and contents as shown in fig3.

Figure 3: History

Figure 4: Performance metrics used to result analytics.

The web revisitation performance metrics include pages'finding rate, average precision, average recall and averagerank error for a set of re-finding requests.

VII. Conclusion

Personal web revisitation technique based on context and content keywords introduced. Context instances and page content are respectively organized as probabilistic context trees and probabilistic term lists, which dynamically evolve by degradation and reinforcement with relevance feedback. Work introduced human's natural recall process of using episodic and semantic memory cues to facilitate a personal web revisitation technique through context and content keywords. Technique proposed is effective for context and content memories' acquisition, storage, decay, and utilization for page re-finding.

VIII. References

[1] A. Cockburn, S. Greenberg, S. Jones, B. Mckenzie, and M. Moyle. Improving web page revisitation: analysis, design, and evaluation.IT & Society, 1(3):159–183, 2003.
[2] L.Tauscher and S. Greenberg. How people revisit web pages: empirical findings and implications for the design of history systems. International Journal of Human-Computer Studies, 47(1):97–137, 1997.
[3] J. Teevan, E. Adar, R. Jones, and M. Potts. Information re-retrieval: repeat queries in yahoo's logs. In SIGIR, pages 151–158, 2007.
[4] M. Mayer. Web history tools and revisitation support: a survey of existing approaches and directions. Foundations and Trends in HCI, 2(3):173–278, 2009.
[5] L. C. Wiggs, J. Weisberg, and A. Martin. Neural correlates of semantic and episodic memory retrieval. Neuropsychologia pages 103–118, 1999.
[6] M. Lamming and M. Flynn. "forget-me-not": intimate computing in support of human memory. In FRIEND21 Intl. Symposium on Next Generation Human Interface, 1994.
[7] E. Tulving. What is an episodic memory? Current Directions in Psychological Science, 2(3):67–70, 1993.
[8] C. E. Kulkarni, S. Raju, and R. Udupa. Memento: unifying content and context to aid webpage re-visitation. In UIST, pages 435–436,2010.
[9] J. Hailpern, N. Jitkoff, A. Warr, K. Karahalios, R. Sesek, and N. Shkrob. Youpivot: improving recall with contextual search.

In CHI, pages 1521–1530, 2011.
[10] T. Deng, L. Zhao, H. Wang, Q. Liu, and L. Feng. Refinder: a context-based information re-finding system. IEEE TKDE,25(9):2119–2132, 2013.
[11] T. Deng, L. Zhao, and L. Feng. Enhancing web revisitation by contextual keywords. In ICWE, pages 323–337, 2013.
[12] H. Takano and T. Winograd. Dynamic bookmarks for the WWW.In HYPERTEXT, pages 297–298, 1998.
[13] S. Kaasten and S. Greenberg. Integrating back, history and bookmarks in web browsers. In HCI, pages 379–380, 2001.
[14] J. A. Gamez, J. L. Mateo, and J. M. Puerta. Improving revisitation browsers capability by using a dynamic bookmark personal toolbar. In WISE, pages 643–652, 2007.
[15] R. Kawase, G. Papadakis, E. Herder, and W. Nejdl. Beyond the usual suspects: context-aware revisitation support. In HT, pages 27–36, 2011.