Enhancing the Personal Web Revisitation using the Context & the Content Keywords Along with Recommendations

Mrs. Sreemol V. S.
M.Tech Student
Department of Computer Science & Engineering
Thejus Engineering College, Thrissur, India

Abstract— Among the various activities of the users in the web, revisiting the previously viewed web pages has become a dominant one. As the web is massive and dynamic this process is not an easy task .There are chances for the users to forget the URL and title of the viewed page after many days when they try to re-find it. This paper utilizes the humans recalling process of the past events by using the episodic memory and the semantic memory. Thus a web revisitation technique has been proposed here to recollect information from the past by using the context and the content keywords. When a user accesses a web page more than a time period, it is understood that the page has a tendency to be revisited later by the user. So by using the unigram extraction the context information like time, location and activity of the user related to the focused page and the highlighted text from the focused web page are captured and stored in the database. Later when the user wants to revisit the web page the system searches the context and the content terms stored in the database and returns a ranked list of web pages to the user. Also as an enhancement to the proposed system a recommendation of related web pages viewed by another users of similar interests are and shown.

Key words: Web Revisitation, Web Mining, Context & Content Extraction, Natural Language Processing

I. INTRODUCTION

With the advent of Computer technology and World Wide Web it has been easy for people to get the desired information that they want merely at their finger-tips. Among the various activities of the user's in the web, revisiting a web page has been found to be a dominant one. Web revisitation refers to getting back to a previously viewed web page. This activity is a common procedure yet not an easy task due to the large amount of information accessed by the user on the web. The working of the human memory in facilitating the recall process is considered here for web revisitation. Human memory usually uses the episodic and the semantic memory to recall facts or events from the past. The episodic memory here is related to the contextual cues and the semantic memory is related to the content cues .Both these cues are considered in this proposed method as a best recall cue to get back to the previously viewed web page.

Consider if a person wants to revisit a web page that he has viewed in the last month. It is possible that the person might forget the URL of the web page, more likely there are chances for the person to remember the context information like the time or location or the concurrent activity like the music that was playing in the background while focusing the web page. So using any of this context or content cue to retrieve the required page makes Revisitation a faster and efficient method.

Studies of web access have shown that the users frequently revisit pages that they have seen in the past. Re-finding information on the web is a common yet a time consuming and challenging task. Remembering the exact URL of the web page is not possible for a long term visit, yet like the episodic memory of brain which remembers the episode of events, there is a possibility of people to remember the context and the content terms related to the accessed page. Therefore revisitation has been enabled here by means of the contextual and content cues.

Re-visiting previously seen information on the web is a significant task that users often perform and is the area that researchers continue to improve. Re-finding information on the web deals with two steps that is firstly locating the web page secondly finding desired information from that page. Based on the pattern of revisitation of users it can be termed as either short term or long term revisitation. Despite various improvements on revisitation techniques still there are substantial shortcomings. This paper has been proposed to remove the difficulty in retrieving the required page that had been viewed in the past.

II. RELATED WORK

Computers and Internet have become the part of our day today life without which life seems to be difficult now a days. People surf on the net to obtain wide range of information related to any field on just a mouse click. The study by McKenzie, B. and Cockburn [1] suggests that revisitation constitutes a major part of the entire web activity. It refers to visiting a web page that has already been visited previously. Though revisitation is a common task yet it is a tedious one due to large amount of personally accessed information by user on the web. There are various techniques already in use to help achieve revisitation which have been discussed here.

Greenberg, S. and Cockburn [2] states that the majority of revisitation tasks are accomplished via the Back and Forward button in-built in the web browser. It is one of the effective methods used for revisitation because of its simplicity in the mode of usage and its rapidness in returning to very recently viewed web pages. It needs only simple forward and backward clicks to get back to the previously and currently visited web pages during a session. For short term revisitation it is worthwhile and easy to use the Back button as it can handle the current session well but is not advisable for long term revisitation. According to Greenberg and Kaasten the Back and Forward button has a stack-based nature [3] where the newly accessed pages by the user will be added upon the top of the stack and when the user revisits a page the stack pointer moves backwards till the previously visited page has been reached. Perhaps the Back button which is session based could be improved further by basing it on recency list [5] rather than a stack model. The working of.

Back and Forward based on the recency-based list uses the buttons which moves up and down the history list [2] [4].

In the World Wide Web there are lots of web sites available but among these web sites people feel only some of them as useful and interesting[10]. So such websites have a tendency to be revisited again by users. The difficulty of people in remembering the address of the websites leads to inability to access the previously viewed webpage. David Abrams et al [1] presented an option to get rid of this problem by the usage of bookmarks. Now a day's all modern web browsers include bookmark features. Bookmarks [9] are saved shortcuts that direct the browser to a specific web page. It stores the title, url and icon of the corresponding page. With bookmarks it is possible to access the frequently visited websites and useful references since it is not needed to remember the URLs. A bookmark stores only the location of a webpage and not the contents of the webpage. Each browser has a built-in tool for managing the list of bookmarks [8]. But if a person uses other browser than the personal browser, then it is not possible to access the bookmarks saved in the personal PC. As an extension to bookmark, social bookmarking methodology has come into action. Social bookmarking [10] [2]means saving the bookmarked pages in some sites which supports social bookmarking so that it can be accessed from any browsers. According to Tsubasa Takahashi et al.[6] it is an information sharing service that allows individuals to bookmark and annotate web pages of interest or those that impress them. Its use can be further expanded like sharing of bookmarked pages of one person in a social bookmarking site with other people of similar field or interest. The various bookmarking sites available are delicious, dzone, clip marks, twine. Despite the benefits of bookmarks some major problems also existed with them. With bookmarks it is only possible to get back to the website and not the exact information what the user have viewed previously.

History tools of web browsers stores the accessed title and url of the web pages including the time at which they were accessed. It also divides and stores the accessed page in the sections like today, yesterday, last week etc. Users can revisit the required page by simply searching the title or keywords of title. Potentially any web page visited earlier could be revisited using these lists [21]. However, the entries expire after several days or weeks and long-term revisits are not supported beyond a certain date. Various advancements had been made in the web browser history to improve revisitation by combining the use of web site thumbnails and content snippets to assist users to easily browse or search their histories by time. Saurabh Kumar [18], have suggested that the storage of accessed web pages in the history helps to mine and rank the web pages according to frequently accessed ones thereby making revisitations easy and also supports recommendation system [20] by association rule mining.

A web search engine [26] is a software system that is designed to search for information on the World Wide Web. The search results are generally presented in a line of results often referred to as search engine results pages. A search engines main process includes web crawling, indexing and searching. Web search engines get their information by web crawling from site to site. Indexing means associating words and other definable tokens found on web pages to their domain names and HTML-based fields. A query from a user can be a single word. The index helps find information relating to the query as quickly as possible. Tyler and Teevan [28], studied how search engines are used for re-finding previously found search results. It explored the differences between queries that had substantial 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 may be a way to bridge a task between two different sessions

The revisitation then took an improvement when Li Jin et.al [1], proposed a method of revisitation by using any of the context or the content keyword to get back to the previously viewed web page. The concept of humans natural recall process [33] using the episodic memory and semantic memory is used here. Both the context information like time, location and activity regarding the browsed page and the content terms related to the browsed page were considered as a cue that helped the users to recall about the browsed web page. Web mining techniques and natural language processing methods were used to build this system.

III. METHODOLOGY

A. Problem Definition

A good revisitation approach should retrieve the relevant pages to the user in a faster and efficient manner. The retrieval of the apt content is a tedious process and it is even more tedious to do it in a faster way. When long term revisitation occurs there are chances for users to forget the title and URL of the page so it becomes difficult for them to retrieve the previously focused web page. In order to overcome this situation it is necessary to build a system that will help the users in the recalling procedure. A framework has been proposed here to make the revisitation process easier and faster by the use of any of the context or the content cues for retrieving the required web pages.

B. Module Description

The proposed personal web revisitation technique makes use of the context and the content keywords to recall the previously viewed web page. Also the related topics viewed by other users are shown as recommendations which may be helpful for them to get more information regarding that topic from other websites which were unknown to them. The modules involved with this project include the following:

- Context Extraction Module
- Content Extraction Module
- Revisitation & Recommendation Module

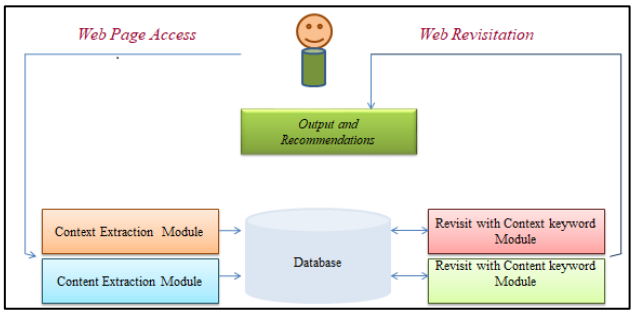

Fig. 1: The Web Revisitation Framework

The chrome extensions are used for the extraction process which are indeed small programs that add functionality to the chrome browsers. Any modifications can be done to the appearance of the web page using the chrome extensions. The chrome extensions contain a content script, a background script and a manifest file. The first step in creating a chrome extension is to write the manifest file. It is a json formatted file that contains information about the extension. The content script works in the background of a web page. Consider an event such as loading of the web page including the text, images and links etc. After fully loading the page, a timer is set to run. The timer here is set to around 25000 millisecond which is equivalent to 25 second. After the given time stamp a message passing takes place between the content script and the background script. A greeting variable representing hello and the three data including the highlighted text, url and title of the focused page are passed to the background script.

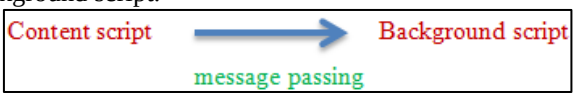

Fig. 2: Message Passing 1

In the background script an event that the message has been received is considered. It checks the greeting variable to identify from where the data has come and then stores the three data in a local variable. After this an Ajax call is performed to pass the data from the background script to the java program and this is made possible by the java servlets. Servlets are java files used for communicating data using http protocol. Servlets use the GET method or POST method to pass data, here the POST method have been used. Thus the data are sent to the java program and hence this is the working of the chrome extension.

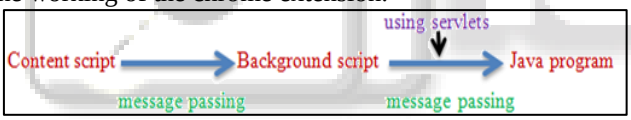

Fig. 3: Message Passing 2

1) Context Extraction Module

This module captures the context related information during the browsing of a web page which has a tendency to be revisited by the user. The location information is captured using the functionality called as geo-location in the html5 which is supported by majority of the browsers including chrome. The activity is given manually by the user. On the browser action of clicking the Go button the location, activity and time for the particular session of log in and log out will be stored in the database automatically.

2) Content Extraction Module

The content extraction module will extract the topmost relevant keywords from the highlighted text and title of the browsed page and stores it in the database. These terms are used as a recall cue to retrieve the previously browsed web page. The terms are pre-processed, ranked and then stored in the database.

a) Pre-Processing

The entire highlighted content terms of the web page are not stored in the database as it may contain unimportant words which will utilize unnecessary space. A process called as cleaning takes place so as to remove the not relevant terms before it is being stored in the database. There are various steps involved in the pre-processing process as tokenisation, segmentation and stemming

b) Tokenisation & Segmentation

This is a natural language processing method. This is done in this framework by using the Stanford nlp library in java. It is process in which the paragraph texts are broken into small pieces of text or single units called as the tokens. Tokens are indeed the smallest individual units of a program which may contain words, punctuations, numbers etc. In the second stage the tokenized words are segmented into sentences rather than paragraphs. Here all the tokenized words are converted to lower case then the punctuations are removed also the stop words. Stop words are unimportant words like am, is, are, was, were etc. whose list can be created by us so that if such words happen to come in the highlighted text it will be removed.

c) Stemming

The stemming process is done based on certain rules for each case. It checks each word and if possible to remove items from the word, removes it and returns its root .For eg. Consider the word fishes, from this word es can be removed and made to fish as it signifies the same meaning similarly the word sailing can be made to sail after stemming process. The Porter Stemmer algorithm is used to perform stemming in this system. It is one of the best stemming algorithms that is publically available but its accuracy cannot be perfect always as some exception cases like leaves and leaf where after stemming leaves can be made only to leav and not leaf.

d) Ranking

Ranking is performed here to find out the most relevant keywords which have a tendency to be used by the user for the recall process. The ranking is done here by using the HITS algorithm.

– HITS algorithm

HITS refer to hyperlink induced topic it is used to rank the web pages and that concept adopted here to rank the keywords in order to find out the influential words in the paragraph. In this algorithm two factors are considered

– Hub
– Authority

Hub means a web page containing many out links and Authority means it is a web page to which many other links point to or in links. In the case of a sentence for example I am fine, am is the keyword considered here and the word that comes before am that is ‘ I ‘is considered as the in link of am and the word that comes after am that is ’fine’ is considered as out link. Thus we take the in link and out link count of the keyword am in the paragraph which is also called as hub score or authority score, we can calculate the total score by taking the sum of this hub score and authority score. Based on this score the keywords are ranked and selected as the top words. These top words are then stored in the database.

In the next step the retrieved relevant keywords are given as input to calculate the final score which is got by calculating the sum of the term score and the highlighted score. This score indicates how relevant the key terms are in the documents. Term score is calculated as:

Term score = tf * idf ———— 1

tf = Term frequency

idf=Inverse document frequency

Term frequency refers to how many times a word has come in the paragraph and inverse document frequency is the inverse of how many number of paragraphs that contains the word. Heading score is a score given to the keyword if it comes in the title of the web page. Based on the sum of these scores the final score is calculated .Thus a hash map will be returned to the database containing the key value pair that is the keyword and its score. Finally from the java program the data are inserted to the database using the Java database connectivity (JDBC).

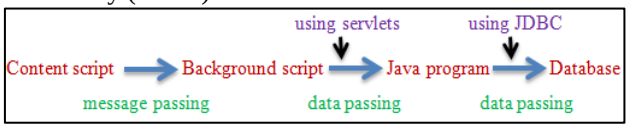

Fig. 4: Message Passing 3

3) Revisitation and Recommendation Module

The revisitation phase deals with the returning of the output related to the users search with the stored content and context keywords from the database. After the desired results are got they are ranked and returned. A recommendation of the related pages viewed by other users of similar interests are listed and shown in this system. Content based recommendations are used to achieve this. The concepts of term frequency and inverse document frequency are used here to retrieve the information. They are used to determine the relative importance of a keyword. A vector space model is used to find the related words. Each item is stored as a vector in an n-dimensional space and the angles between the vectors are calculated to find similarity between the vectors. The cosine of the angle is taken between the user vector and document vector and predictions are made based on that.

C. Performance Evaluation

An evaluation of the performance of the proposed system is done in comparison with the existing history search listings. A short study was conducted for a week in both the systems and performance evaluated on the basis of the following performance metrics.

1) Page Finding Rate

It is the property of finding the required page. If the required page is found then the finding rate is set to one and then necessarily incremented else set to zero.

2) Precision

It corresponds to the number of browsed pages before getting the desired result.

3) Recall

It is the ratio of the relevant pages to the total retrieved pages.

The performance of the proposed system with the existing system showed that the personal web revisitation technique showed better results. The finding rate of personal web revisitation was 0.85% compared to the History listings with 0.82%.The average precision was found to be more in the case of history and average recall good for our proposed system.

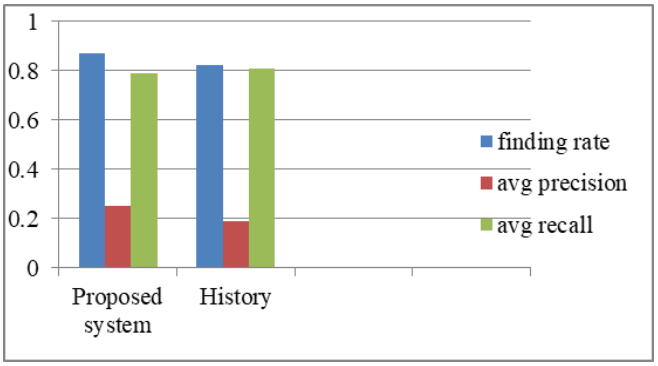

Fig. 5: Performance Evaluation Graph

IV. CONCLUSION

Web revisitations have been made easier and faster by the use of any of the context or the content cues to retrieve the previously focused web page. The human's natural method of recollecting the past events using the episodic and the semantic memory have been used in this system. The enhancement for this system is a recommendation system which shows a list of related topics searched by other users of similar interest. This has indeed helped users to get more information regarding their desired topic.

REFERENCES

[1] Li Jin, Gangli Liu, Chaokun Wang and Ling Feng: (2017), Personal Web Revisitation by Context and Content Keywords with Relevance Feedback, IEEE Transactions on Knowledge and Data Engineering

[2] McKenzie B. and Cockburn A: (2001), an Empirical Analysis of Web Page Revisitation, the Proceedings of the 34th Annual Hawaii International Conference on System Sciences (HICSS-34), Hawaii.

[3] A. Cockburn, S. Greenberg, S. Jones, B. Mckenzie, and M. Moyle: (2003), improving web page revisitation: analysis, design and evaluation, IT & Society, 1(3):159–183.

[4] Greenberg, S., Ho, G. and Kaasten. S: (2002), Contrasting Stack-Based and Recency Based Back Buttons on Web Browsers, Report 2000-666-18, Department of Computer Science, University of Calgary, Alberta, Canada.

[5] Greenberg, S. and Cockburn A: (1999), Getting Back to Back: Alternate Behaviors for a Web browser's Back Button, Proceedings of the 5th Annual Human Factors and the Web Conference, NIST, Gaithersburg, Maryland, USA.

[6] https://stackoverflow.com/questions/1313788/how-does-the-back-button-in-a-web browser-work.

[7] Kaasten, Sh., and Greenberg, S: (2000), Integrating Back, History and Bookmarks in Web browsers, In Extended Abstracts of the ACM Conference of Human Factors in Computing Systems (CHI'01), ACM Press.

[8] David Abrams, Ron Baecker, Mark Chignell's: (1998), Information Archieving with Bookmarks Personal Web Space Construction and Organization, the Conference paper processing of the Conference on Human Factors in Computing Systems Los Angeles,U.S.A HI 98 18-23 .

[9] Takumi Yoshida and Ushio Inoue: (2013), A Bookmark Recommender System based on Social Bookmarking Services using Wikipedia, 14th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing.
[10] Andriy Shepitsen, Jonathan Gemmell, Bamshad Mobasher, and Robin Burke: (2008), Personalized Recommendation in Social Tagging Systems Using Hierarchical Clustering, in RecSys'08 Proceedings of ACM Conference on Recommender System Pages 259-266 Lausanne, Switzerland.
[11] Nagehan Ihan, Sule Gunduz Oguducu: (2009), A Recommender model for social bookmarking sites, Fifth International Conference on Soft Computing, Computing with Words and Perceptions in System Analysis, Decision and Control, ICSCCW.
[12] Tsubasa Takahashi and Hiroyuki Kitagawa: (2008), S-BITS:Social-Bookmarking Induced Topic Search, Ninth IEEE International Conference on Web-Age Information Management.
[13] Nisar Muhammad, Saeed Mahfooz, Shah Khusro and Azhar Rauf: (2013), A Literature Survey on Ranking Tagged Web Documents in Social Bookmarking Systems, IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 11, Issue 5, PP 56-69.
[14] https://en.wikipedia.org/wiki/Bookmark_World_Wide_Web.
[15] https://techterms.com/definition/bookmark[fevicon
[16] https://www.youtube.com/watch?v=HeBmvDpVbWc.
[17] Bonnie MacKay, Melanie Kellar and Carolyn Watters: (2005), An evaluation of landmarks for re-finding information on the web, proceedings of Extended Abstracts on Human Factors in Computing Systems in CHI, pages 1609–1612.
[18] Saurabh Kumar: (2013), Mining User Interests from Web History, the IEEE/WIC/ACM International Conferences on Web Intelligence (WI) and Intelligent Agent Technology (IAT).
[19] B. Tan, Y. Lv, and C. Zhai: (2012), Mining long-lasting exploratory user interests from search history, Proceedings of the 21st ACM international conference on Information and knowledge management. ACM, pp. 1477–1481.
[20] Xiaobin Fu, J. Budzik, and K. J. Hammond: (2000), Mining navigation history for recommendation, Proceedings of the 5th international conference on Intelligent user interfaces. ACM, pp. 106–112.
[21] Matthias Mayer : (2009),Web History Tools and Revisitation Support: A Survey of Existing Approaches and Directions, Foundations and Trends R in Human–Computer Interaction Vol. 2, No. 3 ,173–278.
[22] Tauscher, L. and Greenberg, S: (1997), 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
[23] S. S. Won, J. Jin, and J. I. Hong: (2009), Contextual web history: using visual and contextual cues to improve web browser history, In CHI, pages 1457–1466
[24] Allison Woodruff, Andrew Faulring, Ruth Rosenholtz, Julie Morrison, Peter Pirolli: (2001), Using Thumbnails to Search the Web, Proc. CHI2001, 2001, pp.198-205.
[25] Guodong Si,Hongzhi Song,Jiale He,Yi Fu and Yu Zhao: (2012),A Graphical web revisitation tool, IEEE International Conference on Computer Science and Automation Engineering (CSAE).
[26] https://en.wikipedia.org/wiki/Web_search_engine
[27] https://en.wikipedia.org/wiki/Web_crawler
[28] S. Tyler and J. Teevan: (2010), Large scale query log analysis of re-finding, Proceedings of the third ACM international conference on Web search and data mining pages 191-200
[29] Joshua Hailpern, Nicholas Jitkoff, Andrew Warr, Karrie Karahalios, Robert Sesek, Nik Shkrob: (2011), YouPivot:Improving recall with contextual search, CHI 2011, ACM Press, pp. 1521-1530
[30] Ricardo Kawase, George Papadakis, Eelco Herder: (2011), Supporting revisitation with contextual suggestions,Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries, Pages 227-230.
[31] Jakub Simko, Michal Tvarozek and Maria Bielikova: (2010), Semantic History Map: Graphs Aiding Web Revisitation Support, IEEE Workshops on Database and Expert Systems Applications.
[32] T. Deng, L. Zhao, H. Wang, Q. Liu, and L. Feng: (2013), Refinder: A Context-based Information Re-finding System", IEEE Transactions on Knowledge and Data Engineering,Vol. 25,No.9.
[33] E. Tulving: (1993), What is episodic memory?, Current Directions in Psychological Science, 2(3):67–70.
[34] R. Kawase, G. Papadakis, E. Herder, and W. Nejdl: (2011), beyond the usual suspects: context-aware revisitation support, In HT, pages27–36.
[35] https://www.analyticsvidhya.com/blog/2015/08/beginners-guide-learn-content-based-recommender-systems/

