A Framework to Infer Webpage Relevancy for a User

Saniya Zahoor, Mangesh Bedekar and Varad Vishwarupe

Abstract The Web is a vast pool of resources which comprises of a lot of web pages covering all aspects of life. Understanding a user's interests is one of the major research areas towards understanding the web today. Identifying the relevance of the surfed web pages for the user is a tedious job. Many systems and approaches have been proposed in literature, to try and get information about the user's interests by user profiling. This paper proposes an improvement in determining the relevance of the webpage to the user, which is an extension to the relevance formula that was proposed earlier. The current work aims to create user profiles automatically and implicitly depending on the various web pages a user browses over a period of time and the user's interaction with them. This automatically generated user profile assigns weights to web pages proportional to the user interactions on the webpage and thus indicates relevancy of web pages to the user based on these weights.

Keywords User profiling · Web personalization · Implicit user behavior modeling · Client side analysis

1 Introduction

In recent years, there has been a tremendous growth in the web and its usage, so much so that today many users find it difficult to get information that is relevant to them. Moreover, the behavior of the user is dynamic which makes it difficult to

S. Zahoor · M. Bedekar (✉)
Computer Engineering Department, MAEER's MIT Kothrud, Pune, India
e-mail: mangesh.bedekar@gmail.com

S. Zahoor
e-mail: saniya.zahoor@yahoo.com

V. Vishwarupe
IT Engineering Department, MIT College of Engineering, Kothrud, Pune, India
e-mail: varad44@gmail.com

S.C. Satapathy and S. Das (eds.), *Proceedings of First International Conference on Information and Communication Technology for Intelligent Systems: Volume 1*, Smart Innovation, Systems and Technologies 50, DOI 10.1007/978-3-319-30933-0_16

track his current interests and changes in his interests. If the user's interests are asked explicitly, most users tend to either ignore giving information or fill in wrong/incomplete information. For example, when asked for rating webpage's, users at times tend to either give wrong information, or incomplete information which is also learned as it is stated explicitly. So, there is a need to learn the user implicitly, whereby the system learns about the user and his interests in a transparent manner. The user does his usual web activities and no questions are asked. Various activities that the user does on his browser are tracked to infer interest. All this is done to obtain knowledge about user's behavior of web and thus learn about his interests. The User Profile, thus built contains entities found interesting by the

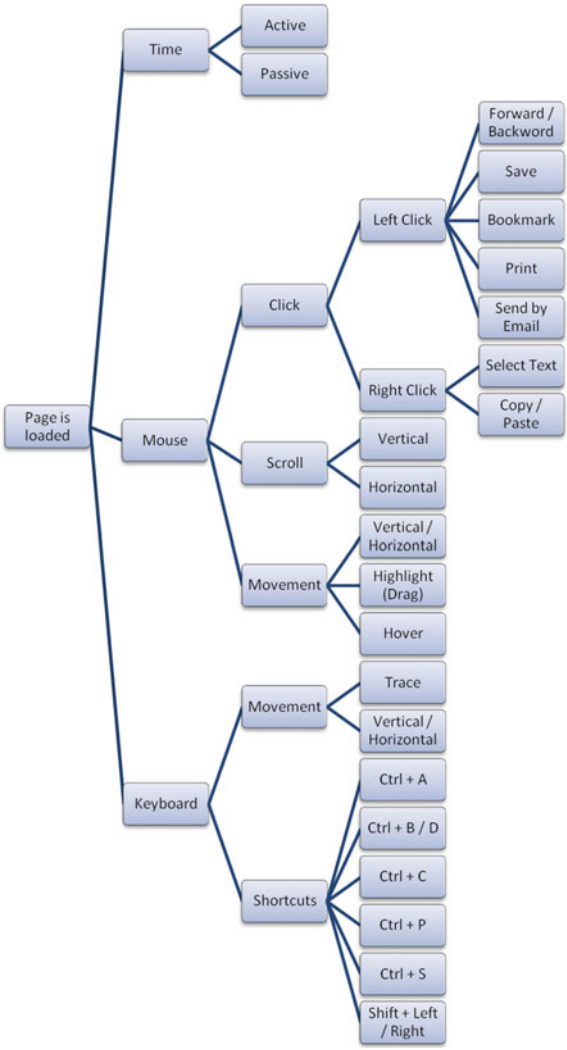

Fig. 1 Various implicit interest indicators

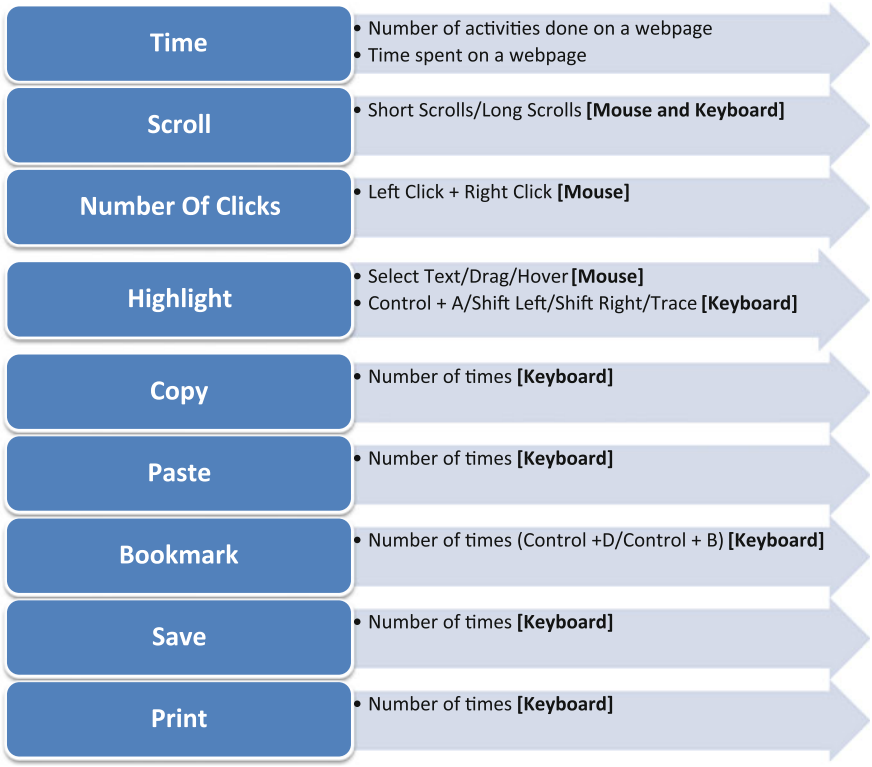

Fig. 2 Implicit interest indicators

user [1–4]. Zahoor et al. [5] presented an comparative survey of various implicit interest indicators used for User Profiling as shown in Fig. 1.

Similarly, select text, control all, drag and so on do the highlighting of text and should have the same weight. The regrouped interest indicators are as indicated in Fig. 2.

In this paper, the focus is on the above set of implicit interest indicators (9) to propose a measure that will generate user profiles automatically depending on the various web pages a user browses over a period of time and thus indicates relevancy of web pages for a particular user.

2 Related Work

Researchers have done work of allotting weights to the pages visited. This was primarily done on relating the number of pages a user visits and number of pages he finds relevant for a given search. The method proposed by Teevan et al. [6] ranks

documents by summing over terms of interest the product of the weight of the term and the frequency with which that term appears in the document. Rating any web page explicitly after each search is a tedious task and is not in line with the usual behavior of the user.

White et al. [7] have compared web retrieval systems with explicit v/s implicit feedback. They go on to show that the implicit feedback systems can indeed replace explicit feedback systems with little or no effect to the users search behavior or task completion. Shapira et al. [8] proposed to combine some well known interest indicators to get better results for relevancy of web pages, they also propose several more implicit interest indicators. Li et al. [9] proposed a system to identify users' interest based on his behavior in the browser he uses to access the Internet. They also go on to show that change in users interests can be handled by the system. Papers not complying with the LNCS style will be reformatted. This can lead to an increase in the overall number of pages. We would therefore urge you not to squash your paper.

3 Modifications Proposed

The paper proposes a User Relevance Factor to determine if a web page is relevant or not, and if relevant how much. The visited web pages are sorted in order of their relevant factors and only the top few pages (based on a threshold, to be taken from the user) are said to be relevant. The activity by the user on the webpage can be inferred by the active time spent by the user, mouse movement, scrolling behavior, save, bookmark and print on the webpage as indicated in [10].

The relevance factor, Rf is, calculated by the following formula,

$$Rf = \log\left(\frac{(time) \times (movement)}{scroll}\right) * Save * Print * Bookmark$$

The formula included 6 factors in it and it was based on the concept that increase in the implicit interest indicators increases the relevancy of the search results for a particular user but there were following issues with the formula,

SCROLL: If a user is scrolling a lot, then it nullifies the effect of spending time on the webpage and performing actions on it. Placing Scroll at the denominator changes the effect on the other activities on the webpage. A concept of ratio of short scrolls and long scrolls (S/L) is introduced to handle this problem. Short scrolls means that user is reading something on the webpage, and the small scrolls done more number of times indicates the user is positively interested in the webpage while as long scrolls means that user is seeing the page quickly and the long scrolls done less number of times indicates less interest of the user on the webpage. The relevance factor R will be directly proportional to short scrolls (S) and inversely proportional to long scrolls (L).

TIME ISSUE: Time is taken directly proportional in the formula. There are some issues in the following cases:

- user spends more time on the webpage, does less activity;
- user spends small time and does a lot of activity.

The formula gives higher priority to the former which should not occur. The time factor in the equation was modified to ratio f, where f equals the number (N) of activity done on the webpage divided by the time (T) spent on the webpage which shows the user's engagement on the webpage. There can be four cases for f which are shown in Table 1.

SAVE, PRINT, BOOKMARK: what if a user performs these actions n number of times that too should have an effect on the relevancy for a user. If a user prints a page n number of times, that means the page has higher relevancy for him as compared to the case if he prints the page once. Thus, capturing actions and capturing the number of times the action can be included in the formula as well.

Factors included are Highlight, Copy, Paste, Bookmark, Save, Print, Number of Clicks on a Webpage (Left Click & Right Click), Number of Activity done on the Webpage divided by time spent on the Webpage (N/T) and Number of Short Scrolls divided by Number of Long Scrolls (S/L) as the implicit interest indicators in relevance factor as mentioned below,

$$R = \frac{N}{T} + (Sa * n^{s} + B * n^{b} + Pr * n^{Pr} + C * n^{c} + Pa * n^{Pa} + H * n^{h} + Cl * n^{Cl}) + \frac{S}{L}$$

where

N Number of activities on a particular webpage during a particular session
T Time spent on the webpage during a particular session
n number of times the activity is performed during a particular session
Cl clicks = leftclicks + rightclicks
Sa Save,
B bookmark,
Pr Print,
Pa Paste
H Highlight
C Copy,
L Long scrolls,
S Short scrolls.

Table 1 Activity V/S time

Activity	Time	
Cases	More time	Less time
More activity	Medium value	High value
Less activity	Less value	Medium value

Table 2 Initial weights of interest indicators

Click	Highlight	Copy	Paste	Bookmark	Save	Print
0.33	0.20	0.15	0.12	0.10	0.07	0.03

After various iterations, the values mentioned in table gave best results

Among the Nine implicit interest indicators that are in the formula, Time (seconds) and Scroll (distance covered) are taken directly as value and the rest of the seven interest indicators have the ordering as,

$$\text{Click} > \text{Highlight} > \text{Copy} > \text{Paste} > \text{Bookmark} > \text{Save} > \text{Print}$$

Click is placed at the initial position, as it is the most frequent action that a user does on the webpage. It also indicates about which part of the webpage (DOM) a user is interested. To highlight text on a webpage, the user needs to first click then highlight. In order to paste 'copied' text, the user first needs to copy the text then highlight and then paste. Click, highlight, copy and paste tells us which part of the webpage a user is interested and that will have higher weight as compared to bookmark which tells us about the URL the user is interested in. The order followed for Bookmark, save and print is the same as in [11]. The weights for these implicit activities are as indicated in Table 2.

4 Implementation Details

The implementation starts with installation of XAMPP server on the client's machine and creation of database with tables containing specific columns that represent a lot of useful information. One of the tables contains the two columns, one for the URL and the other for the time spent. Likewise there are other tables which contain information about user's interest on the webpage. Once the server has been installed with the required database and tables, every time the user uses a browser he needs to switch on the server after which the data starts getting stored in the database.

JavaScript (JS) is an interpreted computer programming language. Mozilla Firefox Browser is free and open source. One of the main reasons why Mozilla Firefox was used was because of its unique add-on Greasemonkey. Once all these user initiated events are captured, this data along with the URL of the web page is stored into the database. The relevance factor script is then invoked using Greasemonkey which monitors the users' behavior on the web page. In this way all the necessary data values required for the method are captured and stored. Since the databases are stored on XAMPP server.

Table 3 URL's and corresponding values

Case	URL	Cl	H	C	P	B	Sa	P	S	L	T
1	URL1	3	2	1	1	1	2	1	44	19	189
2	URL2	4	4	1	1	1	1	1	103	51	96

5 Mathematical Proof of the Formula

Consider the case of a user who visits two URLs with the following values (Table 3).
Case 1:

$$f = (3+2+1+1+1+2+1)/189 = 11/189 = 0.0582.$$
$$Ne = 63.\ \text{Average Scroll(i)} = 4805.134/63 = 76.27.$$
$$S = 44, L = 19, S/L = 44/19 = 2.3158.$$
$$R(1) = (0.0582) + (0.33*3 + 0.20*2 + 0.15*1 + 0.12*1 + 0.10*1 + 0.07*2 + 0.03*1) + (2.3158) = 4.304.$$

Case 2:

$$f = (4+4+1+1+1+1+1)/96 = 13/96 = 0.1354.$$
$$Ne = 154.\ \text{Average Scroll (i)} = 37785.53/154 = 245.360.$$
$$S = 103, L = 51, S/L = 103/51 = 2.019608.$$
$$R(2) = (0.1358) + (0.33*4 + 0.20*4 + 0.15*1 + 0.12*1 + 0.10*1 + 0.07*1 + 0.03*1) + (2.01960) = 4.745.$$

As can be observed, R(2) has higher value than R(1) i.e.; Case 2 URL is more relevant to the user as compared to the Case 1 URL. Subjectively, the user was asked which URL is more relevant to him—Case 1 or Case 2, and the answer was similar.

6 Conclusion

User's behavior on a webpage can reveal a lot about his interests on the web. The actions that the user performs and his usage on the web can be captured and can help in understanding the user very well. Almost all actions of the user, which are done in the browser, can be captured via the Mouse and the Keyboard. Only Keyboard and Mouse actions are not sufficient to identify the user's interests. The proposed framework handles the problem well by considering the ratio of short to long scrolls to get a proper interest measure. Time spent, considered as a single entity is a misnomer to indicate relevancy which has to be handled too.

The framework ranks all the visited web pages according to its relevancy to the user hence this will be used in giving relevant search results to the user. For each search term the pages browsed by the user are recorded and ranked according to the users profile. Once a concrete database gets created over a period of time, as soon as the user searches any term he will get a list of web pages visited by him for similar search done earlier which would be ranked according to the user's personal relevance.

References

1. Faucher, J., McLoughlin, B., Wunschel, J.: Implicit web user interest, Technical Report MQP-CEW-1101, Worcester Polytechnic Institute, Spring (2011)
2. Hauger, D., Paramythis, A., Weibelzahl, S.: Using browser interaction data to determine page reading behavior. In: UMAP'11, Proceedings of the 19th International Conference on User Modeling, Adaption, and Personalization, pp. 147–158. Girona, Spain, 11–15 July 2011
3. Kríž, J.: Keyword extraction based on implicit feedback. Inf. Sci. Technol. Bull. ACM Slovakia, **4**(2), 43–47
4. Leiva Torres, L.A., Hernando, R.V.: A gesture inference methodology for user evaluation based on mouse activity tracking. In: IHCI 2008, Proceedings of the IADIS International Conference on Interfaces and Human Computer Interaction, Amsterdam, The Netherlands, 25–27 July 2008
5. Zahoor, S., Bedekar, M., Kosamkar, P.: User implicit interest indicators learned from the browser on the client side. In: International Conference on Information and Communication Technology for Competitive Strategies, Udaipur, Rajasthan, India, 14–16 Nov 2014
6. Teevan, J., Dumais, S., Horvitz, E.: Personalizing search via automated analysis of interests and activities. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '05), pp. 449–456. ACM, New York
7. White, R., Ruthven, I., Jose, J.M.: The use of implicit evidence for relevance feedback in web retrieval. In: Proceedings of the Twenty-Fourth European Colloquium on Information Retrieval Research (ECIR '02). Lecture Notes in Computer Science, pp. 93–109. Glasgow (2002)
8. Shapira, B., Taieb-Maimon, M., Moskowitz, A.: Study of the usefulness of known and new implicit indicators and their optimal combination for accurate inference of users interests. Proceedings of SAC '06, pp. 1118–1119
9. Li, F., Li, Y., Wu, Y., Zhou, K., Li, F., Wang, X.: Discovery of a user interests on the internet. In: Proceedings of the IEEE/WIC/ACM, International Conference on Web Intelligence and Intelligent Agent Technology, pp. 359–362 (2008)
10. Zahoor, S., Dr. Bedekar, M.: Implicit client side user profiling for improving relevancy if search results, CCSEIT-2014. In: Proceedings of Fourth International Conference on Computational Science, Engineering and Information, Technology, Army Institute of Technology, Pune, India, 8–9 Aug 2014
11. Zahoor, S., Rajput, D., Bedekar, M., Kosamkar, P.: Capturing, understanding and interpreting user interactions with the browser as implicit interest indicators. In: ICPC 2015, International Conference on Pervasive Computing, Sinhgad College of Engineering, Pune, 8–10 Jan 2015