Develop, Implement, and Improve a Web Session Detection Model

Chaoyu Ye
Mixed Reality Lab
University of Nottingham, UK
psxcy1@nottingham.ac.uk

Supervisors: Dr. Max L.Wilson & Prof. Tom Rodden
Mixed Reality Lab
University of Nottingham, UK
firstname.lastname@nottingham.ac.uk

ABSTRACT

More research in web and Information Retrieval is turning towards session-based retrieval rather than single item or query investigation. However, most of the session detection attempts only used simplistic rules (e.g. "30 mins inactivity creates a new session"). Up to this point, there are various fuzzy definitions of session, but no general consensus about it in the literature [3]. Whilst comparably little work has involved the mental model about the "web session" from real users. In response to these, my research focuses on web session detection involving real users with a comprehensive set of factors identified by them rather than the "simple fixed timeout". My objective is to develop a session detection model with corresponding rules for each factor, and then embedded them into a Chrome Extension to automatically detect more accurate web sessions from log data.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval - *search process.*

Keywords

Sessions, Web History, Log Analysis, Qualitative

1. MOTIVATION

As the information needs of users on the web vary from straight-forward information seeking to complex topic exploration, recent research has moved beyond trying to provide optimal results for a current query, towards trying to model and support a "search session" [8]. The notion of session can be dated back to decades ago. Early work on DIALOG [14] kept track of searchers' local queries and allowed them to reuse them, especially for information re-finding. The research on web session started in mid 90s [1]. Despite this history, there are only various fuzzy definitions of session [11, 5], rather than a general consensus of it [3]. Silverstein [11], for example, defined session as "a series of queries by a single user made within a small range of time; a session is meant to capture a single user's attempt to fill a single information need", and Jansen et al [5] stated that "A session is a series of interactions by the user towards addressing a single information need". Starting from last decade, there has been increasing focus on web sessions, where search engines are keen to better support searchers who continue to search for more than a few queries or minutes (e.g. [16]) such as query reformulation (e.g. [7]). While, the absence of a general consensus of session definition hinders the practical way for grouping relevant queries. For the present, most of them have used the "fixed timeout" [1, 4]. For example, Catledge and Pitkow suggested a 25.5 minute inactivity between two adjacent activities was best to divide logs into sessions in 1995 [1], and similarlity, Google employed 30 minutes interval as one of their session conditions [15]. In addition, some others utilised "query change" in the search activities to detect the "topical shifting" to detect the session boundary [7, 5]. Gayo-Avello provided a comprehensive summary of previous session detection methods including both temporal and lexical clues.

Besides the gap between the concepts and practice, all of the previous session study didn't really involve the end users' concern and their mental model about it. Additionally, most of them mainly focused on the consecutive search activities without considering other behaviours such as interleaving. Some research, for example, has shown that people typically engage in a range of interleaving activities [6, 10, 12]. Moreover, most of the previous research in session generally only reinforce the study on search and query log. Browse and search, however, are highly related to each other. Cheng et al, [2] for example, stated that some searches were actually triggered by what people have browsed. The involvement of browse behaviour would benefit the fully understanding of search session.

2. RESEARCH QUESTIONS

My primary PhD research objective is to develop a web session detection model within a set of factors considered by people and implement it as a Chrome Extension to automatically detect users' web session boundaries covering both Search and Browse. Particularly, I have the following goals which are partially completed.

RQ1) What factors are considered by people in session detection

RQ2) What are the practical rules for each factor

RQ3) How to work these practical rules for different factors together

IIiX '14 Aug 26 - 29 2014, Regensburg, Germany
Copyright is held by the owner/author(s).
ACM 978-1-4503-2976-7/14/08.
http://dx.doi.org/10.1145/2637002.2637057.

In RQ1, it is about understanding the factors considered by people when they are determining the sessions, which includes these situations for each factor: “when it divides the session?”, “when it joins two seemingly separate sessions together?”, “when it dominate other factors?”, and “when it is overriden by other factors?”. The answers to these deliver a comprehensive conceptual understanding of the factors rather than directly applied into practical implementation. It will help to model user’s highly dynamic behaviour in terms of factors, and bring some implicit factors that have effect but haven’t been paid attention to.

In RQ2, it is about the practical implementation within web data (e.g. clickthrough data) to simulate each factor from the answer to RQ1. The feasibility and barrier should be carefully considered.

In RQ3, it is about combining the rules from RQ2 together. As these factors are highly interrelated to each other, working them together require refinements on the conflicts, especially the domination in different situations.

To validate the rules in RQ2 and RQ3, a Chrome Extension will be developed and rules will be embedded to be evaluated.

3. METHODOLOGY

My overview approach is to 1) first understand factors considered by users in the session detection; 2) develop a tool to practically implement them based on their web data; 3) conduct iterative experiments to validate and improve it.

To accomplish RQ1, I have conducted a study within the similar interview methods to Sellen et al. [10]. 20 participants were employed to engage in a 90-120 minutes interview. They were asked to markout their own defined sessions in the printout of their own web histories, and discussed the factors triggerring the session boundary. After that, the card sorting technique [9] was applied to probe their mental models of session. Three types of data were collected and analysed: logs, interview data, and card sorts. The logs and card sorts were quantitatively summarised, and the interview data was transcribed from the audio recordings and was analysed using an open inductive form of Grounded Theory [13] by two researchers. Disagreements were discussed carefully and codes were merged or divided as their definitions, and the definitions of the categories and themes, developed.

To accomplish RQ2 and RQ3, I am working on a Chrome Extension capturing user’s web data including “mouse & keyboard event, timestamp & viewing time, URL, page title, query (if it is a search), top 10 keywords of page content, tab Id, window Id, openerTabId”. An initial session-detection rule based on the 6 factors in our taxonomy will be embedded. In addition, other findings from last experiment will also be considered. Then, this Extension will be installed in participants’ laptop to collect their web data for 7 days, and it will automatically generate a list of system-defined sessions based on the rules. After that, the participants will be invited to join an interview to markout their own-defined session on partial of their own web histories, and discuss the comparison between the “system-defined” and “user-defined” sessions. Both qualitative (e.g. Grounded Theory) and quantitative data analysis will be applied.

4. PROGRESS

Up to this point, we have developed a preliminary taxonomy of 6 factors in the session detections: Topic Change, Task Change, Phase Change, Group of People involved, Time Gap, and Multitasking. The first three of them refer to the lexical clues about the activities grouped into sessions. These 6 factors are interrelated, and they represent common themes discussed by participants, rather than rules that can directly applied. It is not, however, a matter of how often each factor applies, but instead how much each factor applies at different times. Besides the taxonomy, the feature study of session itself was also covered and there were some interesting findings: 1) the tolerance of time gap for dividing sessions varies from types and scale of web activitiy, e.g. search with more queries input or browse with more pageviews have higher tolerance of time gap; 2) people have higher tolerance of time gap of real-life interruption than other web activities, e.g. people may still resume from the previous session after hours gap caused by cook or sleep; 3) the query number, pageviews and length have some effects on the judge of activity scale, which may even lead to mis-estimation of the activity length.

After working on RQ1 and getting it acceptted to one conference, I am working on RQ2 about determing the initial rules for each factor. As telling differences of the scale between Topic, Task, and Phase is still challenging, we decided to group these three as one factor “whether they are content-relevant to each other” temporarily in the initial stage. The relevance will be determined based on the web data as shown in Table 1. For the Group of People involved, it will only work on social network and mailboxes so that the activities from one social network or mailbox will be grouped into one session. Then for the Time Gap, we will build an “intelligent time gap” calculation within the consideration of the query number, pageviews, and length, and it will also check whether the interruption is from real-life or not. The corresponding weight for them need further consideration. The Multitasking factor is still under construction.

Table 1: Factors and Web Data

Factors	Web data involved
Topic, Task, and Phase	Title, Domain in URL, Top 10 keywords, openerTabId, Query (if it is a search), tabId, windowId
Group of People	Domain in URL, a dataset of mailbox and social network domain
Time Gap	query number (if search), pageviews, viewing time, timestamp, title, domain in URL, openerTabId, tabId, windowId, Top 10 keywords
Multitasking	tbd

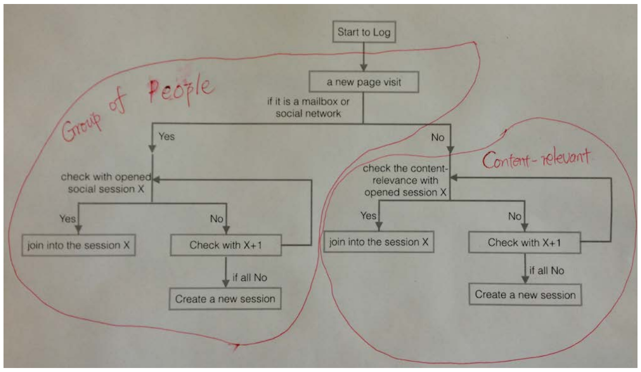

Figure 1: Decision Making of New Page Visit

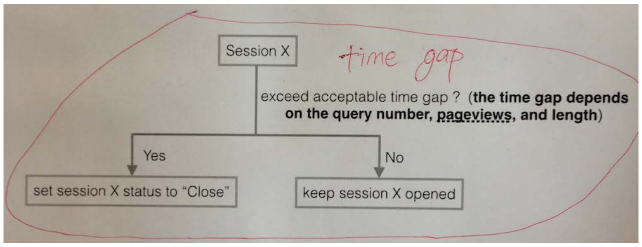
Session X
time gap
exceed acceptable time gap ? (the time gap depends on the query number, pageviews, and length)
Yes
No
set session X status to "Close"
keep session X opened

Figure 2: Session Status

Currently, we are developing the Chrome Extension. After finishing data capture, I am working on the practical rules. There are two main threads in the system: decision making of new page visit (T1) in Figure 1 and session status check (T2) in Figure 2. The T1 shows the procedure for checking whether a new page visit belongs to an existing "opened" sessions or not. It will firstly check whether it is mail or social network activities, and then check the relevant with some other "opened" sessions based on the transition and content. T2 is about checking the status of existing sessions. There are two status of a session: closed (i.e. when the inactivity time gap exceeds the "calculated timeout") and opened (i.e. when it does not). The "closed" session will not be involved in the iterative check in T1 with new page visit. The calculated timeout is related to the query number, pageviews, length, and also the type of interruption (i.e. whether it is from real-life or other web activity).

5. FUTURE PLAN

In the rest of my second year, I will concentrate on determining the rules and validating their practical implementation in my tool with couple of studies. In the following months, some further debugging on the Chrome Extension will be done and also some more work on the initial practical session detection rules, e.g. deciding the initial parameters in the time gap calculation. After that, we will conduct an experiment with the main objective to study the differences between user-defined and system-generated sessions. The data on the difference between them will be applied to refine the collaboration of rules in our system. The participants will be asked to have the Chrome Extension installed in their laptops to collect their web data for 7 days, and all of the data will be submitted and stored in our server securly and anonymously. There are two phases in total. Phase one is to ask participants to markout their own session. Phase two is the mainly focusing on the comparison between the user-defined and system-generated sessions. They will be asked to have a look at them to mark out: 1) "incorrect" session in our system-generated sessions. 2) Vote the "better" session in the group of different sessions between their own-defined and system-defined sessions. After that, there will be a discussion about their own rules for the session determine. The feedback from this will contribute to reality of our system-generated session afterwards. There will be an iterative process within one or two more similar experiments. Hopefully, by the end of my second year, I can accomplish RQ2 and achieve the Chrome Extension with some updated rules from the experiment above.

In my third year, I will keep working on the rules, especially the conflict between rules, for example, one factor may be dominated by others in some situations. I will run a couple of studies to further exam it following the similar structure of experiment above. I may also expand my study to collect a longer period history from each participant like 1 month instead of 7 days. In addition, I am planning to look for one relevant internship to get bigger datasets involved to finalise the rules and tool.

The expected final outcome will be a comprehensive list of practical rules for session detection and a Chrome Extension embedded with these rules.

6. REFERENCES

[1] L. D. Catledge and J. E. Pitkow. Characterizing browsing strategies in the World-Wide web. *Computer Networks and ISDN Systems*, 27(6):1065–1073, 1995.

[2] Z. Cheng, B. Gao, and T.-Y. Liu. Actively predicting diverse search intent from user browsing behaviors. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 221–230, New York, NY, USA, 2010. ACM.

[3] D. Gayo-Avello. A survey on session detection methods in query logs and a proposal for future evaluation. *Information Sciences*, 179(12):1822 – 1843, 2009. Special Section: Web Search.

[4] D. He and A. Göker. Detecting session boundaries from Web user logs. *In Proc. 22nd BCS-IRSG*, pages 57–66, 2000.

[5] B. J. Jansen, A. Spink, C. Blakely, and S. Koshman. Defining a session on Web search engines. *JASIST*, 58(6):862–871, 2007.

[6] B. Mackay and C. Watters. Exploring multi-session web tasks. *In Proc. CHI2008*, pages 1187–1196, 2008.

[7] S. Ozmutlu. Automatic new topic identification using multiple linear regression. *Information Processing & Management*, 42(4):934–950, 2006.

[8] K. Raman, P. N. Bennett, and K. Collins-Thompson. Toward whole-session relevance: exploring intrinsic diversity in web search. Proc. SIGIR '13, pages 463–472, New York, NY, USA, 2013. ACM.

[9] G. Rugg and P. McGeorge. The sorting techniques: a tutorial paper on card sorts, picture sorts and item sorts. *Expert Systems*, 14(2):80–93, 1997.

[10] A. J. Sellen, R. Murphy, and K. L. Shaw. How knowledge workers use the web. In *Proc. CHI2002*, pages 227–234. ACM Press.

[11] C. Silverstein, H. Marais, M. Henzinger, and M. Moricz. Analysis of a very large web search engine query log. *SIGIR Forum*, 33(1):6–12, Sept. 1999.

[12] A. Spink, M. Park, B. J. Jansen, and J. Pedersen. Multitasking during Web search sessions. *IP&M*, 42(1):264–275, 2006.

[13] A. Strauss and J. Corbin. Grounded theory methodology. *Handbook of qualitative research*, pages 273–285, 1994.

[14] R. K. Summit. Dialog: An operational on-line reference retrieval system. *ACM Annual Conference/Annual Meeting*, 1967.

[15] G. A. T. Trevor Claiborne. Update to sessions in google analytics, 2011.

[16] R. W. White and S. M. Drucker. Investigating behavioral variability in web search. In *Proc. WWW2007*, pages 21–30. ACM, 2007.