

Analysing Web Search Logs to Determine Session Boundaries for User-Oriented Learning

Ayşe Göker and Daqing He

School of Computer and Mathematical Sciences,
The Robert Gordon University
Aberdeen AB25 1HG, Scotland
{`asga`, `dqh`}@scms.rgu.ac.uk

Abstract. Incremental learning approaches based on user search activities provide a means of building adaptive information retrieval systems. To develop more effective user-oriented learning techniques for the Web, we need to be able to identify a meaningful session unit from which we can learn. Without this, we run a high risk of grouping together activities that are unrelated or perhaps not from the same user. We are interested in detecting boundaries of sequences between related activities (sessions) that would group the activities for a learning purpose. Session boundaries, in Reuters transaction logs, were detected automatically. The generated boundaries were compared with human judgements. The comparison confirmed that a meaningful session threshold for establishing these session boundaries was confined to a 11-15 minute range.

1 Introduction

Given the increased use of the Web, the amount of information available, and greater variety of regular users, it is imperative to have adaptive techniques for Web-based Information Retrieval Systems (IRSs) which meet individual users' needs more effectively. To this end, research has included work in user profiles, automated browsing and suggesting hyperlinks [1, 5].

Recurring patterns in users' search *activities* (queries, judgements and navigation) can be exploited with learning techniques to enable user-adaptability. This paper focuses on the temporal ordering of activities clustered according to close proximity in time. Although, other forms of activity clustering (i.e., topicality, browsing patterns) are possible, initially we use *time* information, and investigate the extent to which this alone is effective. We group activities and refer to the resulting unit as a **session**. If we view a user with an interest in a specific topic as acting in a particular **role**, then it is not unreasonable to assume that the activities in the same session are likely to correspond to one role. We argue that there are contextual connections between activities if we view the retrieval process as an interactive problem solving task with a goal [3]. Hence, our aim is to specify a session so that it contains data pertaining to one role.

In Web Transaction Log (TL) analyses, studies often group all activities for one user or IP number into a unit referred to as a session [4]. The appropriateness of this grouping is debatable [3], particularly where the time span is large.

Additionally, the final cut-off point for the TLs is usually arbitrary. This presents us with the risk of grouping together activities that are unrelated. Researchers focusing on Web navigation activities have used the time between two adjacent page accesses to help cut sessions [2]. However, their work focuses on users' navigation behaviour, and does not include activities of using Web search engines. This paper describes an automatic method for detecting session boundaries and then presents the results of the comparison with human judgements.

2 The Method and Data

Due to a lack of adequate information about Web users, our empirical method of detecting session boundaries currently uses only time information. Our aim is to examine the effectiveness of using reliable and easily obtainable information, like time, in detecting session boundaries. A time span called **session interval** could be defined in advance to be used as a threshold. Two adjacent activities are assigned to two different sessions if the time between them exceeds this threshold. The identification of session boundaries then becomes a process of examining the time gap between activities and comparing with the set session interval. Each session has a number of activities in a sequence and within the context of the experiments, we refer to this number as the **iteration** of the session,¹ *e.g.* if a session has three activities, its iteration number is three.

The experiments were based on a set of transaction records of searches by Reuters Intranet users, referred to as the *Reuters logs* (Reuters Ltd.). The search engine used is a local version of *AltaVista*. The time range of the logs extends seven days from 30th March 1999. There are 9,534 activities from 1,440 unique IP addresses. Each record contains: *Time stamp*, *IP address*, and *CGI command*. This command includes information about the query terms, the search method (simple/advanced) and the subsequent page numbers for the same search.

3 The Experiments

Our experiments consisted of two stages: automatically detecting session boundaries then comparing them with human judgements.

3.1 The Automatic Detection of Session Boundaries

We first cut the logs with a large session interval and grouped the sessions with the same iteration number together in order to see the distribution of various sessions. Then, gradually we decreased the session interval and obtained the corresponding distributions, which show the percentage of sessions with a particular iteration in relation to the total number of sessions.

Ideally, a session should contain only those activities from one role. An **optimal session interval** that enables this should not be too large in order to

¹ We have chosen this terminology to emphasise the sequence in the activities within the session and their likelihood of being related to the same role.

avoid the risk of grouping activities from different roles together. Also, it should not be too small as there would be less information available on the role.

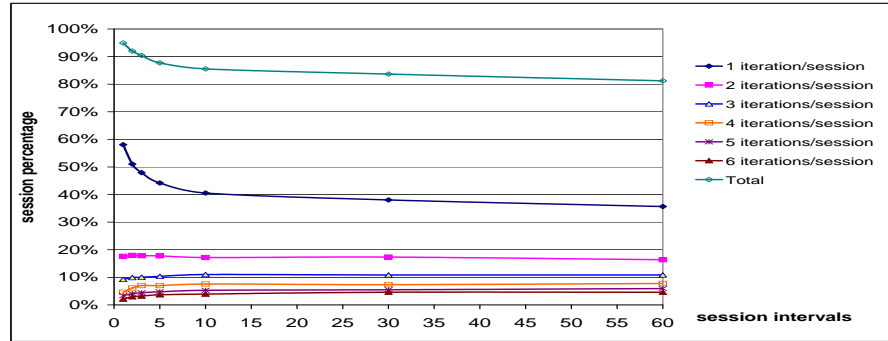


Fig. 1. The frequency of sessions given different session intervals

We monitored distributions of sessions with 6 iterations or less as their total covers the majority (81%) of sessions [3]. The results (Fig. 1) show that most short sessions are not affected when the session interval is larger than 15 mins. When the session interval is shorter than 10 mins, the percentage of sessions with 1 iteration increases dramatically, whereas the percentages of sessions with 3-6 iterations decrease. So, the optimal session interval with regard to the likelihood of grouping activities from the same role together is within 10-15 mins.

3.2 The Human Identification of Session Boundaries

The automatic detection process may result in the following errors: **Type A errors** occur when two adjacent activities for related search statements are allocated into different sessions; **Type B errors** occur when unrelated activities are allocated into the same session. The former is the result of selecting a too tight session interval, whereas the latter is the result of a too loose interval. We view Type B errors as potentially the most damaging to our learning purpose as it could make the role prediction invalid. Hence, we give it a higher weight (nominally twice that of Type A) in this experiment.

Two experts in query formulation worked through the logs and marked the places of a context/role change. Their judgements were compared and anomalies due to oversight or lack of knowledge of domain-specific vocabulary were reduced to a minimum. These judgements were taken as the basis of comparison with the automatic detection method. Resulting types of errors are shown in Fig. 2, which shows that Type A errors decrease sharply until about 15 mins, then continue dropping slowly. Type B errors, on the other hand, increase steadily between 5-15 mins, but do so at a slower rate thereafter. The total percentage of errors (Type A and B) decrease dramatically until about 15 mins. For our learning purpose, we prefer low percentages of total and Type B errors. Hence, the results indicate

that the optimal session interval for the logs in this experiment is between 11-15 mins. This confirms the results of Sect. 3.1.

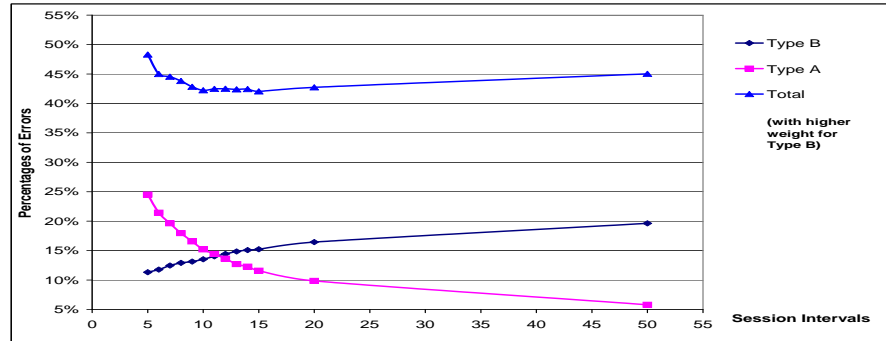


Fig. 2. The frequency of session cut errors given different session intervals

4 Conclusion and Future Work

In this paper, we have presented a method for detecting session boundaries by using a minimal amount of user information that is typically available in Web logs. Our results, after comparing with human judgements, indicate that an optimal session interval should be within the range of 11-15 minutes. In future work, we intend to explore methods of improving automatic session boundary detection by reducing both the percentages of total errors and Type B errors. We envisage using topic information and a statistical model of activity sequence to adjust the automatically generated session boundaries. Additionally, it may be possible to refine the session boundaries by examining the statistical distributions of the intervals between activities within a session.

Acknowledgements We thank Reuters Ltd. for the logs; Jim Jansen, David Harper (especially on future work) and Robin Boswell for their helpful comments.

References

1. Balabanovic M., Shoham Y., and Yun Y.: An Adaptive Agent for Automated Web Browsing. Tech. Rep. CS-TN-97-52, Dept. of Comp. Sci., Stanford University (1997)
2. Catledge L. and Pitkow J.: Characterizing Browsing Strategies in the World-Wide Web. In *3rd International World-Wide Web Conference* (1995) http://www.igd.fhg.de/archive/1995_www95/papers/
3. He D. and Goker A.: Detecting session boundaries from Web user logs. In *22nd Annual Colloquium on IR Research IRSG 2000*, Cambridge, UK (2000) 57–66
4. Jansen J., Spink A., Bateman J., and Saracevic T.: Real Life Information Retrieval: A Study of User Queries on the Web. *SIGIR Forum*, **32(1)**(1998) 5–17
5. Joachims T., Freitag D., and Mitchell T.: WebWatcher: A Tour Guide for the World Wide Web. In *Proceedings of IJCAI97* (1997) 770–775