

Identification of User Sessions with Hierarchical Agglomerative Clustering

G. Craig Murray¹, Jimmy Lin¹, Abdur Chowdhury²

¹ College of Information Studies, University of Maryland
{gcraigm,jimmylin}@umd.edu

² AOL Search, AOL Inc., Sterling, VA
cabdur@aol.com

Abstract

We introduce a novel approach to identifying Web search user sessions based on the burstiness of users' activity. Our method is user-centered rather than population-centered or system-centered and can be deployed in situations in which users choose to withhold personal content information. We adopt a hierarchical agglomerative clustering approach with a stopping criterion that is statistically motivated by users' activities. An evaluation based on extracts from AOL Search™ logs reveals that our algorithm achieves 98% accuracy in identifying session boundaries compared to human judgments.

1. Introduction

Studies of users' Web search behavior based on log analysis begin with an operational assumption that individual users' activities can be grouped into distinct sessions. A common approach is to set an arbitrary timeout threshold and assume that any gap in activity from a particular IP address that exceeds the threshold is actually a break between an individual user's sessions. This approach has a number of limitations. The universal application of a single time threshold (e.g., 20 minutes without activity) to all users as a criterion for new sessions has not been vetted and we cannot say that there is some singular criterion that is appropriate for all users.

Other approaches to session identification look at query similarity, identifying breaks between sessions by changes in query terms. We believe these conflate two separate issues—shifts in topics of interest and delays between query activities. However, some information needs and interests are persistent, as reflected by identical queries separated by very long intervals of time. Meanwhile, users often have momentary shifts in topical focus while actively engaged in searching. This results in searches for one topic that are often embedded in longer efforts to find information on completely different topics (e.g., a hobby enthusiast taking a “break” from a work-related search task). Definitions of session boundaries that rely on semantics of query terms are dangerously circular and conceal persistence and recurrence of users' long-term information needs. Additionally, characterizing individual users' search behavior by query terms may not be viable as users become more cautious about privacy concerns.

We believe that users' search activity can be modeled on three different layers: a layer of activity, a layer of topicality, and a layer of semantic expression. On the activity layer, we can measure the amount of time users engage in different types of search activities; the frequency of queries, time spent browsing, etc. On the topical layer we can model aspects of users' interests; how focused or diffuse they are, how they develop, shift, and decay. On the semantic expression layer, we can analyze means of expressing topical interests, digesting search results, learning new terms and refining queries. Although these three layers interact, we see them as independent axes of analysis, as evidenced by multiple topical shifts within a single session and persistent information needs across multiple sessions. Understanding the relationships between these layers is critical to understanding user behavior.

To explore the validity of a three-layered approach, we begin by modeling users' search sessions at the activity layer. We propose a purely time-based approach to session identification that does not rely on a semantic layer, but is able to identify sessions at the activity layer with a high degree of accuracy. This approach allows

us to isolate the different layers of interest and study user behavior without conflating relevant factors. Our model is capable of describing user search behavior—individually and in aggregate—without exposing user’s personal interests and without relying on query term analysis.

2. Related work

Other studies have analyzed query logs to identify patterns at separate layers of activity and topicality (e.g., Lau & Horvitz, 1999). Clustering techniques have been used to group queries semantically (Beeferman & Berger, 2000) and to group users by activity patterns (Wang & Zaiane, 2002). We apply a clustering technique to identify session boundaries at the activity layer.

The concept of “session” has many definitions in the literature. Arlitt (2000) looked for a fixed global time threshold to identify user sessions by assuming that any gap of time between queries (from a single user) that exceeds a threshold of minutes is also a break between user sessions. Arlitt swept this timeout threshold from a high value to a low value and observed the tradeoffs between number of sessions and number of active sessions. We believe there is merit in the statistical post hoc analysis of large data sets, but fixed thresholds that strike a good balance for the system are really uninformative about the individual users.

In other research, Spink and colleagues defined a session as “the entire set of queries by the same user over time” (Spink, Jansen, & Ozmutlu, 2000; see also Jansen & Spink, 2003). In fact their “sessions” are a function of the short time span of their data and the volume of use by their users. Meanwhile, He and colleagues (He & Göker, 2000; He, Göker, & Harper, 2002) have taken a mixed approach to identifying sessions in Web search logs. They used an interval thresholding technique in which multiple different time intervals were evaluated against a training set. He, et al. also combined the time interval information with an analysis of users’ query terms from one query to the next. They find that the probability of complete changes in query terms is related to the gap of time between queries, i.e. long gaps of time predict complete changes in query terms. However, in our own analysis of 8 million queries we found that sequential identical queries (from a single user) were also quite likely to be separated by long time intervals, highlighting the interplay between a layer of activity and a layer of topical interest.

We believe that a mixed approach to modeling search behavior has merit but that any single time threshold adds unnecessary noise to the data. In our study, we isolated the issues of timing and semantics in order to explore the validity of time-based session boundaries. We analyzed frequency and burstiness of activity from users and identified a user-centered time threshold that is highly accurate and independent of query semantics.

3. HAC

We performed our analysis of user behavior using data from usage logs of AOL Search, a large popular search engine. Three months of data were collected from 216,000 users in November and December of 2004 and January of 2005. The logs contain an anonymous user ID number, the query search terms submitted and the submission time of the search. Unlike IP addresses or cookies, the user ID’s in our data remain distinct regardless of shared network caching, network address translation, or user mobility. A total of 8,269,030 queries were issued by these users in the three months that were logged. The log entries were resorted by user ID and then by query time. This allows us to perform user-by-user analysis on very large groups of users over an extended period of time.

We implemented a variant of hierarchical agglomerative clustering (HAC) (Willett, 1988) to identify individuals’ session breaks. Our algorithm has two parts. First, we loop through the gaps between each user’s queries, shortest to longest, to identify time intervals that significantly increase the variance. Then we take a user-specific criterion based on variance, and group each user’s queries into sequential sessions based on their individual criterion. Figure 1 and Figure 2 give an overview of the clustering algorithm.

Our approach has a driving central principle—find intervals of time that are significantly longer than the average time between queries *within* a session. The result of taking the argmax of the ratio described is that we find the interval which had the most significant effect on the variance when added to smaller intervals. We find that for the majority of users, this approach produces a singular spike in the rate of increase of the variance of time intervals. That is, for most users there are small steps up in variance, followed by a large step up in variance at some interval, and then subsequently very small steps up in variance with larger intervals. The user-based threshold lies just above the last interval length before the sudden rise in variance.

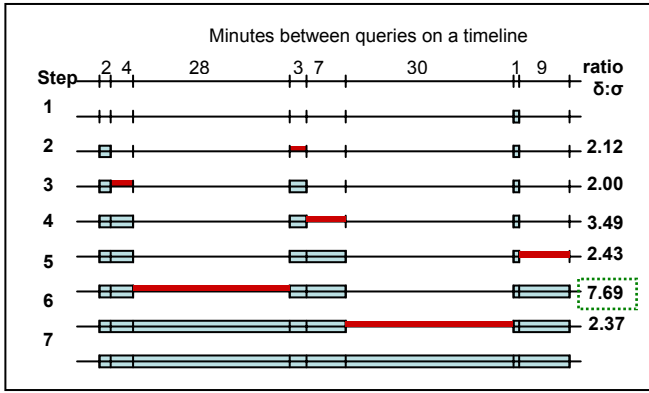


Figure 1: Illustration of HAC clustering steps in first loop. Subsequently longer intervals are included into sessions and the max ratio is found of deviation from prior mean over prior standard deviation.

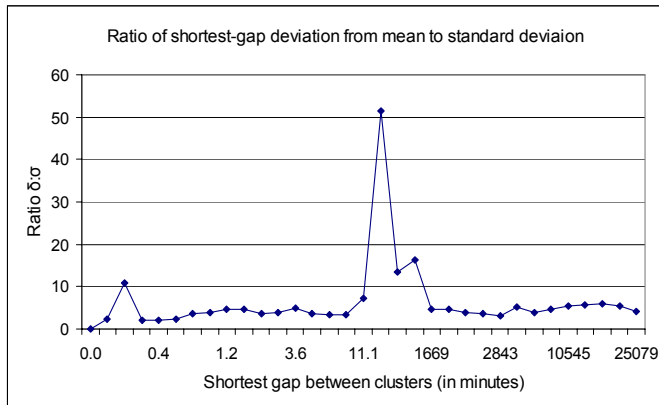


Figure 3: A sudden rise and fall in the ratio of deviation from previous mean to the standard deviation of previously intervals yields an informative discriminatory interval.

```

FOR each user
  FOR each new query
    SET gap interval = (current query time – last query time)
    PUSH gap interval onto list of gap intervals
  END FOR

  SORT gap intervals from shortest to longest
  INIT max $\delta:\sigma$  = 0
  INIT argmax( $\delta:\sigma$ ) = null
  INIT clustered intervals = null set

  FOR each sorted gap interval  $i$ 
    IF number of clustered intervals > 1
      SET  $\mu$  = mean of clustered intervals
      SET  $\sigma$  = standard deviation of clustered intervals
      SET  $\delta_i = (i - \mu)$ 
      IF ( $\delta_i / \sigma$ ) > max $\delta:\sigma$ 
        SET max $\delta:\sigma$  =  $\delta_i / \sigma$ 
        SET argmax( $\delta:\sigma$ ) =  $i$ 
      ENDIF
    END IF
    ADD gap interval  $i$  to set of clustered intervals
  END FOR

  FOR each query  $q$ 
    SET gap interval = (current query time – last query time)
    IF gap interval < argmax( $\delta:\sigma$ )
      ADD  $q$  to current cluster
    ELSE
      INIT new cluster = [ $q$ ]
    END IF
  END FOR
END FOR

```

Figure 2: HAC algorithm for clustering queries with stopping criterion based on standard deviation

Table 1 - HAC vs. fixed time intervals

	HAC	12 min.	15 min.	20 min.
precision	0.973	0.781	0.810	0.846
recall	0.762	0.997	0.996	0.995

Note that here we are describing a post hoc analysis of the user. Naturally, any on-line application that would leverage this information should include an additional buffer of time in the session criterion. For example, consider the timing data for the user in Figure 3. For display purposes we plot the gaps on a disproportionate scale. The extreme spike occurs between a gap of 690 seconds (11.5 minutes) and a gap of several thousand seconds. Triggering a new “session” in a model of this user after 12 minutes of inactivity would seem far too aggressive. But, when the HAC algorithm closed the gap of 11.5 minutes, the standard deviation of clustered intervals for this user was 160 seconds (2.33 minutes). If we believe that the gaps of activity under 11 minutes are representative of within-session intervals, it would be reasonable to include intervals that deviated from that mean as much as two standard deviations (~16 minute gaps) as also being within-session intervals. Based on this user’s statistics, when this user stops searching, he/she does not typically start again for at least two hours. If this user were to show a 15 minute gap in activity, it would still not be informative given this user’s history, and it would be safer to assume that the user was still engaged in the previous “session” of activity. If we had relied on a fixed interval, we might have arbitrarily split sessions or merged sessions together.

4. Results

To verify the validity of our approach we evaluated the clustering algorithms performance on 102 users’ query histories. We began by randomly selecting 500 users from our data set. From these we excluded any users that had issued fewer than 20 queries in the 3 month time span of the query logs. This resulted in 1593 query pairs to be judged. We implemented the HAC algorithm described above, which generated a total of 855 session clusters for these users. The clustering results of the HAC approach were then compared to human judgments. Each chronological pair of queries for a given user was considered to be a potential break point

between sessions. Human participants viewed each user history and noted where gaps between pairs of queries should have been identified as breaks between sessions.

Out of 854 session breaks identified by the HAC approach, 831 were “correct” compared to human judgments and only 23 were considered to be false alarms. There were a total of 263 “missed” session breaks. On further analysis we found that 60% (158) of these were accounted for by less than 3% of the users. We calculated IR metrics of precision and recall for the total set of session breaks. In this case, precision is the number of correctly identified session breaks over the total number of session breaks marked by the HAC algorithm (831/854). Recall is the number of correctly identified session breaks over the total number of actual session breaks as identified by human judges (831/1094). Our total precision was 0.973 and our total recall was 0.760. Our average precision per user was 0.988 and average recall was 0.884. Although our sample set was small, by these numbers we achieved 99% accuracy in session identification for this set of users. Compare this to the performance of a fixed time interval (see Table 1). Although our approach has lower recall, it is far more accurate in its predictions. The probability of a false alarm when the algorithm claims that a session break has occurred is less than 0.015 for our sample set. We take this to be highly indicative of the validity of such user centered approaches.

5. Conclusion

We have shown that an algorithm based on hierarchical agglomerative clustering that models the burstiness of user activities can accurately identify session boundaries. Many users in our data imbed one search activity within the other or search for two different things simultaneously. We also find users who issue a query, log off for 1 or 2 days, and later return to issue the exact same or similar query. There is good reason to separate query content from query frequency. We find that the HAC approach identifies boundaries and thresholds that are more user-centered. Moreover, it does this at a layer of activity that is independent from semantics or topicality. This is a promising result for our three layered model of search behavior. Possible applications include combining this activity model with content analysis of queries at the semantic layer to find search within search, or with a topical layer to improve user modeling of persistent needs. By leveraging users’ search histories we were able to identify session boundaries using a different criterion for each user. It is no longer necessary to set a singular threshold of time for all users and break activities into sessions based on that threshold, nor is the privacy of users violated as the queries themselves are not stored or analyzed at the activity layer.

6. Reference List

Arlitt, M. (2000). Characterizing Web user sessions. *ACM SIGMETRICS Performance Evaluation Review*, 28(2), 50-63.

Beeferman, D., & Berger, A. (2000). Agglomerative clustering of a search engine query log. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 407-416). Boston, Massachusetts, August 2000. ACM Press: New York.

He, D., & Göker, A. (2000). Detecting session boundaries from web user logs. In *Proceedings of the BCS/IRSG 22nd Annual Colloquium on Information Retrieval Research* (pp. 57-66). Cambridge, UK, April 2000.

He, D. Q., Göker, A., & Harper, D. J. (2002). Combining evidence for automatic Web session identification. *Information Processing & Management*, 38(5), 727-742.

Jansen, B. J., & Spink, A. (2003). An analysis of Web documents retrieved and viewed. In *Proceedings of the 4th International Conference on Internet Computing* (pp. 65-69). Las Vegas, Nevada, June 2003.

Lau, T., & Horvitz, E. (1999). Patterns of search: Analyzing and modeling Web query refinement. In *Proceedings of the Seventh International Conference on User Modeling* (pp. 119-128). Banff, Canada, June 1999. Springer: New York.

Spink, A., Jansen, B. J., & Ozmutlu, H. C. (2000). Use of query reformulation and relevance feedback by Excite users. *Internet Research: Electronic Networking Applications and Policy*, 10(4), 317-328.

Wang, W., & Zaiane, O. R. (2002). Clustering Web sessions by sequence alignment. In *Proceedings of the 13th International Workshop on Database Expert Systems Applications* (pp. 394-398). Aix-en-Provence, France, September 2-6, 2002. IEEE: Los Alamitos, CA.

Willett, P. (1988). Recent trends in hierarchic document clustering: a critical review. *Information Processing & Management*, 24(5), 577-597.