

# Identification of User Sessions Using Hierarchical Agglomerative Clustering



G. Craig Murray, Jimmy Lin and Abdur Chowdhury

gcraigm@umd.edu • jimmylin@umd.edu • abdur@ir.iit.edu

COLLEGE of INFORMATION STUDIES  
UNIVERSITY OF MARYLAND  
<http://www.clis.umd.edu>

## Research Objectives

- Find a user-centered method for determining breaks between sessions of search activity.
- Find session breaks without reference to the content of search queries.
- Explore alternatives to assigning session boundaries based on simple fixed time thresholds.

## Theoretical Framework

This work is part of a larger program of research. Within our developing framework, search behavior can be modeled at three separate levels of interaction:

*Physical* – what users do and when, physical activities that facilitate search, user sessions, etc.

*Topical* – topics of interest, clarity of ideas, importance and persistence of need, etc.

*Semantic* – expressions of need, query term relationships, search results, document representation, etc.

## Experimental Method

Assign Session Boundaries:

Make two passes on a large data set of users' search histories (8.3M queries) using *only the timestamp data*.

On first pass:

Apply Hierarchical Agglomerative Clustering [HAC]. Analyze changes in the variance of clustered time intervals along the way.

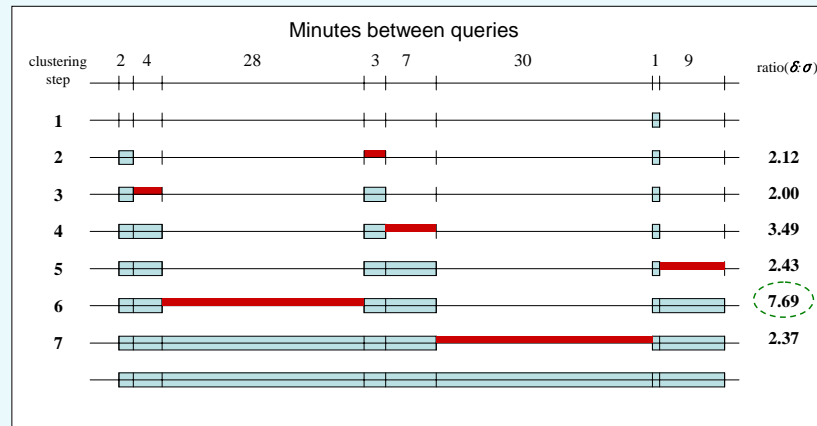
On second pass:

Assign session boundaries based on the changes of variance found in each individual user's history.

Evaluate Results:

Compare boundaries assigned by HAC to those assigned by human judges.

## Clustering Example\*



## Algorithm pseudo-code

```

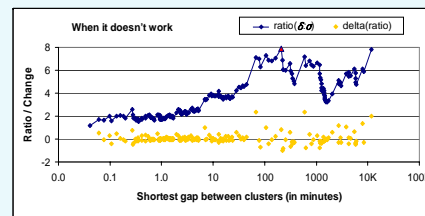
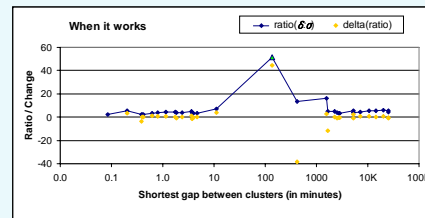
FOR each user
  FOR each new query
    SET gap interval = (current query time - last query time)
    PUSH gap interval onto list of gap intervals
  END FOR

  SORT gap intervals from shortest to longest
  INIT maxσ = 0
  INIT argmax(δ, σ) = null
  INIT clustered intervals = null set

  FOR each sorted gap interval i
    IF number of clustered intervals > 1
      SET μ = mean of clustered intervals
      SET σ = standard deviation of clustered intervals
      SET δ = (i - μ)
      IF (δ / σ) > maxσ
        SET maxσ = δ / σ
        SET argmax(δ, σ) = i
      END IF
    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(δ, σ)
      ADD q to current cluster
    ELSE
      INIT new cluster = {q}
    END IF
  END FOR
END FOR
    
```

## Variance Examples\*



\* For clarity of illustration and for reasons of institutional policy, the details in our examples are altered from the original data.

## Evaluation

102 user histories randomly selected

- ♦ limited to minimum of 20 queries in 3 months
- ♦ yielded 1593 query pairs to be judged

Two human judges marked queries that appeared to be the start of a new session.

HAC algorithm marked queries that appeared to be the start of a new session.

## Preliminary Results

The HAC algorithm marked 854 session breaks

- 831 were "correct" compared to human judges
- 23 were considered to be false alarms
- 263 session breaks were "missed"

	HAC	12 min	15 min	20 min
Precision	<b>0.973</b>	0.781	0.810	0.846
Recall	0.762	<b>0.997</b>	0.996	0.995
F-measure	0.868	0.889	0.903	<b>0.921</b>

The HAC approach achieved higher accuracy than fixed time thresholds at a cost in overall recall.

However, 60% (158) of the "missed" boundaries were from only 3% of the user histories.

## Implications

- Patterns of search activity can be identified without reference to the contents of a query.
- Search behavior can be independently modeled at the physical interaction layer of our theoretical framework.
- Evidence from the physical interaction layer can be combined with topic models and semantic models.