Chapter IR:VIII

VIII. Evaluation

- ❑ Laboratory Experiments
- ❑ Performance Measures
- ❑ Training and Testing
- ❑ Logging

Logging

Query Logs

- ❑ Used for both tuning and evaluating search engines
  - – also for various techniques such as query suggestion
- ❑ Many more queries than for test collections
  - – But less precise
- ❑ Problem: Privacy (especially when shared)
- ❑ Typical contents
  - – User identifier or user session identifier
    - • Login, toolbar, cookie, . . .
  - – Query terms; stored exactly as user entered
  - – Ordered list of URLs of results, their ranks on the result list, and whether they were clicked on
  - – Timestamp(s); records the time of user events such as query submission and result clicks

Logging

Query Logs

- ❑ Clicks are not relevance judgments
  - Although they are highly correlated
  - Biased by a number of factors: rank on result list, snippet, general popularity
- ❑ Other indicators
  - Dwell time: time spent on a clicked result
  - Search exit action: result page, print page, timeout, enter other URL, . . .
- ❑ Can use clickthrough data to predict preferences between pairs of documents
  - Appropriate for tasks with multiple levels of relevance, focused on user relevance
  - Various strategies used to generate preferences

Logging

Example Click Policy

- Skip Above and Skip Next
  - Click data

    d_1

    d_2

    d_3 (clicked)

    d_4

  - Generated preferences

    $d_3 > d_2$

    $d_3 > d_1$

    $d_3 > d_4$

Logging

Query Logs

- Click data can be aggregated to remove noise
- Click distribution information
  - Can be used to identify clicks that have a higher frequency than would be expected
  - High correlation with relevance

- Click deviation $CD(d, p)$ for a result d in position p:

$$CD(d, p) = O(d, p) - E(p)$$

  - $O(d, p)$: observed click frequency for a document in a rank position p over all instances of a given query
  - $E(p)$: expected click frequency at rank p averaged across all queries
  - Use to filter clicks for preference-generation policies