Information Retrieval

INFO 4300 / CS 4300

- Evaluation
  - ➡ – Evaluation corpus and logging
  - – Metrics
  - – Training, testing

Evaluation

- Evaluation is key to building ***effective*** and ***efficient*** search engines
  - measurement usually carried out in controlled laboratory experiments
  - ***online*** testing can also be done
- Effectiveness, efficiency and ***cost*** are related
  - e.g., if we want a particular level of effectiveness and efficiency, this will determine the cost of the system configuration
  - efficiency and cost targets may impact effectiveness

Evaluation Corpus

- ***Test collections*** consist of documents, queries, and relevance judgments, e.g.,
  - CACM: Titles and abstracts from the Communications of the ACM from 1958-1979. Queries and relevance judgments generated by computer scientists.
  - AP: Associated Press newswire documents from 1988-1990 (from TREC disks 1-3). Queries are the title fields from TREC topics 51-150. Topics and relevance judgments generated by government information analysts.
  - GOV2: Web pages crawled from websites in the .gov domain during early 2004. Queries are the title fields from TREC topics 701-850. Topics and relevance judgments generated by government analysts.

Test Collections

Collection	Number of documents	Size	Average number of words/doc.
CACM	3,204	2.2 Mb	64
AP	242,918	0.7 Gb	474
GOV2	25,205,179	426 Gb	1073

Collection	Number of queries	Average number of words/query	Average number of relevant docs/query
CACM	64	13.0	16
AP	100	4.3	220
GOV2	150	3.1	180

TREC Topic Example

```
<top>
<num> Number: 794

<title> pet therapy

<desc> Description:
How are pets or animals used in therapy for humans and what are the
benefits?

<narr> Narrative:
Relevant documents must include details of how pet- or animal-assisted
therapy is or has been used.  Relevant details include information
about pet therapy programs, descriptions of the circumstances in which
pet therapy is used, the benefits of this type of therapy, the degree
of success of this therapy, and any laws or regulations governing it.

</top>
```

Relevance Judgments

- Obtaining relevance judgments is an expensive, time-consuming process
  - who does it?
  - what are the instructions?
  - what is the level of agreement?
- TREC judgments
  - depend on task being evaluated
  - generally binary (e.g. CACM)
    - GOV2: not relevant, relevant, highly relevant
  - agreement good because of “narrative”

Pooling

- Exhaustive judgments for all documents in a collection is not practical
- Pooling technique is used in TREC
  - top *k results (for TREC, k varied between 50 and* 200) from the rankings obtained by different search engines (or retrieval algorithms) are merged into a pool
  - duplicates are removed
  - documents are presented in some random order to the relevance judges
- Produces a large number of relevance judgments for each query, although still incomplete

Query Logs

- Used for both tuning and evaluating search engines
  - also for various techniques such as query suggestion
- Typical contents
  - User identifier or user session identifier
  - Query terms - stored exactly as user entered
  - 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, clicks

Query Logs

- Clicks are not relevance judgments
  - although they are correlated
  - biased by a number of factors such as rank on result list
- Can use clickthough data to predict ***preferences*** between pairs of documents
  - appropriate for tasks with multiple levels of relevance, focused on user relevance
  - various "policies" used to generate preferences

Example Click Policy

- *Skip Above and Skip Next*
  - click data

$$\begin{array}{l} d_1 \\ d_2 \\ d_3 \text{ (clicked)} \\ d_4 \end{array}$$

  - generated preferences

$$\begin{array}{l} d_3 > d_2 \\ d_3 > d_1 \\ d_3 > d_4 \end{array}$$

Query Logs

- Click data can also 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
  - e.g., using ***click deviation*** to filter clicks for preference-generation policies

Filtering Clicks

- ***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***

Information Retrieval

INFO 4300 / CS 4300

- Evaluation
  - Evaluation corpus and logging
  - Metrics
  - Training, testing

Effectiveness Measures

A is set of relevant documents,
B is set of retrieved documents

	Relevant	Non-Relevant
Retrieved	$A \cap B$	$\overline{A} \cap B$
Not Retrieved	$A \cap \overline{B}$	$\overline{A} \cap \overline{B}$

$$Recall = \frac{|A \cap B|}{|A|}$$

$$Precision = \frac{|A \cap B|}{|B|}$$

Classification Errors

- ***False Positive*** (Type I error)
  - a non-relevant document is retrieved

$$Fallout = \frac{|\overline{A} \cap B|}{|\overline{A}|}$$

- ***False Negative*** (Type II error)
  - a relevant document is not retrieved
  - 1- *Recall*
- *Precision* is used when probability that a positive result is correct is important

F Measure

- *Harmonic mean* of recall and precision

$$F = \frac{1}{\frac{1}{2}(\frac{1}{R}+\frac{1}{P})} = \frac{2RP}{(R+P)}$$

  - harmonic mean emphasizes the importance of small values, whereas the arithmetic mean is affected more by outliers that are unusually large
- More general form

$$F_\beta = (\beta^2 + 1)RP/(R + \beta^2 P)$$

  - β is a parameter that determines relative importance of recall and precision

Focusing on Top Documents

- Users tend to look at only the top part of the ranked result list to find relevant documents
- Some search tasks have only one relevant document
  - e.g., navigational search, question answering
- Recall not appropriate
  - instead need to measure how well the search engine does at retrieving relevant documents at very high ranks

Focusing on Top Documents

- Precision at Rank R
  - R typically 5, 10, 20
  - easy to compute, average, understand
  - not sensitive to rank positions less than R
- Reciprocal Rank
  - reciprocal of the rank at which the first relevant document is retrieved
  - *Mean Reciprocal Rank (MRR)* is the average of the reciprocal ranks over a set of queries
  - very sensitive to rank position

Discounted Cumulative Gain

- Popular measure for evaluating web search and related tasks
- Two assumptions:
  - Highly relevant documents are more useful than marginally relevant document
  - The lower the ranked position of a relevant document, the less useful it is for the user, since it is less likely to be examined

Discounted Cumulative Gain

- Uses ***graded relevance*** as a measure of the usefulness, or ***gain****,* from examining a document
- Gain is accumulated starting at the top of the ranking and may be reduced, or ***discounted***, at lower ranks
- Typical discount is 1/*log (rank)*
  - With base 2, the discount at rank 4 is 1/2, and at rank 8 it is 1/3

Discounted Cumulative Gain

- *DCG* is the total gain accumulated at a particular rank *p*:

$$DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2 i}$$

- Alternative formulation:

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{log(1+i)}$$

  – used by some web search companies
  – emphasis on retrieving highly relevant documents

DCG Example

DCG Example

- 10 ranked documents judged on 0-3 relevance scale:

  3, 2, 3, 0, 0, 1, 2, 2, 3, 0

- discounted gain:

  3, 2/1, 3/1.59, 0, 0, 1/2.59, 2/2.81, 2/3, 3/3.17, 0

  = 3, 2, 1.89, 0, 0, 0.39, 0.71, 0.67, 0.95, 0

- DCG:

  3, 5, 6.89, 6.89, 6.89, 7.28, 7.99, 8.66, 9.61, 9.61

Normalized DCG

- DCG numbers are averaged across a set of queries at specific rank values
  – e.g., DCG at rank 5 is 6.89 and at rank 10 is 9.61
- DCG values are often *normalized* by comparing the DCG at each rank with the DCG value for the *perfect ranking*
  – makes averaging easier for queries with different numbers of relevant documents

NDCG Example

- Perfect ranking:

  3, 3, 3, 2, 2, 2, 1, 0, 0, 0

- ideal DCG values:

  3, 6, 7.89, 8.89, 9.75, 10.52, 10.88, 10.88, 10.88, 10

- NDCG values (divide actual by ideal):

  1, 0.83, 0.87, 0.76, 0.71, 0.69, 0.73, 0.8, 0.88, 0.88

  – NDCG ≤ 1 at any rank position