Chapter 8

Evaluation and Result Summaries
Measures for an IR System

- How fast does it index
  - Number of documents/bytes per hour

- How fast does it search
  - Latency as a function of index size / queries per second

- What is the cost per query?
  - Given certain requirements, e.g., a 20-billion-page index
Measures for an IR System (2)

- All of the preceding criteria are *measurable*: we can quantify speed / size / money

- However, the key measure for a search engine is *user happiness*

- What is user happiness and how do we measure it?

  - Factors include:
    - Speed of response
    - Size of index
    - Uncluttered UI
    - Most important: *relevance*

- Note that none of these is sufficient: blindingly fast, but useless answers won’t make a user happy.
Measuring User Happiness

Most common definition of user happiness: relevance of returned documents

How do we measure the quality of what is returned by an IR system?

In IR there are two measures:

- Precision $P$: the fraction of retrieved documents that are relevant
- Recall $R$: the fraction of relevant documents that are retrieved
Precision and Recall

Let us give a more formal definition

Let $A$ be the set of retrieved documents, $D$ be the set of relevant documents and $D_A$ the set of relevant documents retrieved, then

$$P = \frac{|D_A|}{|A|} \quad \text{and} \quad R = \frac{|D_A|}{|D|}$$

relevant documents retrieved ($D_A$)

relevant documents ($D$) retrieved documents ($A$)
Alternative Definition

<table>
<thead>
<tr>
<th>Relevant</th>
<th>Nonrelevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retrieved</td>
<td>true positives (TP)</td>
</tr>
<tr>
<td>Not retrieved</td>
<td>false negatives (FN)</td>
</tr>
</tbody>
</table>

\[ P = \frac{TP}{(TP + FP)} \] and \[ R = \frac{TP}{(TP + FN)} \]

relevant documents retrieved (\(D_A\))
Accuracy

- Why do we use complex measures like precision and recall?
- Why not something simple like accuracy?
- Accuracy is the fraction of decisions (relevant/nonrelevant) that are correct.
- In terms of the contingency table above, accuracy = \((TP + TN)/(TP + FP + FN + TN)\).
- There's a problem with that . . .
Accuracy (2)

Simple trick to maximize accuracy in IR: always say no and return nothing.

You then get 99.99% accuracy on most queries (there is a huge number of true negatives you get right).

Searchers on the web (and in IR in general) want to find something and have a certain tolerance for junk.
Precision/Recall Tradeoff

- You can increase recall by returning more docs
- Recall is a non-decreasing function of the number of docs retrieved
  - A system that returns all docs has 100% recall!
- The converse is (usually) also true: You can increase precision by returning fewer docs
  - A system that only returns documents that have a very high score (usually) has a high precision
Depending on the application one or the other may be more important:

- Typical web surfers would like every result on the first page to be relevant (high precision)
  are not interested in looking at every document that might be relevant (there might be millions)
- Various professional searchers such as paralegals and intelligence analysts
  are usually very concerned with trying to get as high recall as possible
  will tolerate fairly low precision results in order to get it
Precision-recall Curve

- Precision/recall are measures for *unranked sets*.
- We can easily turn set measures into measures of *ranked lists*.
- Just compute the set measure for each “prefix”: the top 1, top 2, top 3, top 4, ... results.
- Doing this for precision and recall gives you a *precision-recall curve*. 

Information Retrieval and Organization – p. 245/320
Example

- Assume following documents are relevant for query $q$: 
  \{d_3, d_5, d_9, d_{25}, d_{39}, d_{44}, d_{56}, d_{71}, d_{89}, d_{123}\}

- IR system gives back this ranked list:

<table>
<thead>
<tr>
<th>Ranking</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. $d_{123}$ ←</td>
<td>10%</td>
<td>100%</td>
</tr>
<tr>
<td>2. $d_{84}$</td>
<td>10%</td>
<td>50%</td>
</tr>
<tr>
<td>3. $d_{56}$ ←</td>
<td>20%</td>
<td>67%</td>
</tr>
<tr>
<td>4. $d_6$</td>
<td>20%</td>
<td>50%</td>
</tr>
<tr>
<td>5. $d_8$</td>
<td>20%</td>
<td>40%</td>
</tr>
<tr>
<td>6. $d_9$ ←</td>
<td>30%</td>
<td>50%</td>
</tr>
<tr>
<td>7. $d_{511}$</td>
<td>30%</td>
<td>43%</td>
</tr>
<tr>
<td>8. $d_{129}$</td>
<td>30%</td>
<td>38%</td>
</tr>
<tr>
<td>9. $d_{187}$</td>
<td>30%</td>
<td>33%</td>
</tr>
<tr>
<td>10. $d_{25}$ ←</td>
<td>40%</td>
<td>40%</td>
</tr>
<tr>
<td>11. $d_{38}$</td>
<td>40%</td>
<td>36%</td>
</tr>
<tr>
<td>12. $d_{48}$</td>
<td>40%</td>
<td>33%</td>
</tr>
<tr>
<td>13. $d_{250}$</td>
<td>40%</td>
<td>31%</td>
</tr>
<tr>
<td>14. $d_{113}$</td>
<td>40%</td>
<td>29%</td>
</tr>
<tr>
<td>15. $d_3$ ←</td>
<td>50%</td>
<td>33%</td>
</tr>
</tbody>
</table>
As the recall is increasing monotonically, we can plot the precision in relation to the recall:
Interpolation

- Examining the entire precision-recall curve can be very informative, but often we only want an overview.
- The traditional way of doing this is the 11-point interpolated average precision.
- For each test query, the interpolated precision is measured at the 11 recall levels of 0.0, 0.1, 0.2, ..., 1.0.
- We then plot the average precision at each level.
- Where do we get the queries from with which to test the system?
  - We’ll talk about this in just a minute...
Queries vs. Information Needs

- We still haven’t defined when a document is relevant.
- Who decides when a document is relevant and relevant to what?
- “Relevance to a query” is very problematic.
- A user starts out with an information need, not a query.
Let’s look at an example:

Information need $i$: You are looking for information on whether drinking red wine is more effective at reducing your risk of heart attacks than white wine.

This is an information need, not a query.

(Possible) query $q$: wine AND red AND white AND heart AND attack

Consider document $d'$: He then launched into the heart of his speech and attacked the wine industry lobby for downplaying the role of red and white wine in drunk driving.

$d'$ is relevant to the query $q$ . . .

$d'$ is not relevant to the information need $i$. 

Queries vs. Information Needs (2)
Queries vs. Information Needs (3)

- User happiness can only be measured by relevance to an information need, not by relevance to queries.
- We’ve been a bit sloppy with our terminology:
  - We talk about query/document relevance judgments even though we mean
  - information-need/document relevance judgments.
**Benchmarks**

- What we need is a *benchmark*

- A benchmark for IR systems consists of
  - A collection of documents
    - Documents must be representative of the documents we expect to see in reality.
  - A collection of information needs
    - ... which we will often incorrectly refer to as queries
    - Information needs must be representative of the information needs we expect to see in reality.
Benchmarks (2)

And last but not least:

- Human relevance assessments
  - We need to hire/pay “judges” or assessors to do this.
  - Expensive, time-consuming
  - Judges must be representative of the users we expect to see in reality.
Standard Relevance Benchmark: Cranfield

- Pioneering: first testbed allowing precise quantitative measures of information retrieval effectiveness
- Late 1950s, UK
- 1398 abstracts of aerodynamics journal articles, a set of 225 queries, exhaustive relevance judgments of all query-document-pairs
- Too small, too untypical for serious IR evaluation today
Standard Relevance Benchmark: TREC

- TREC = Text Retrieval Conference (TREC)
- Organized by the U.S. National Institute of Standards and Technology (NIST)
- TREC is actually a set of several different relevance benchmarks.
- Best known: TREC Ad Hoc, used for first 8 TREC evaluations between 1992 and 1999
- 1.89 million documents, mainly newswire articles, 450 information needs
Example of a TREC Collection

<table>
<thead>
<tr>
<th>Disk</th>
<th>Contents</th>
<th>Size</th>
<th>Number</th>
<th>Words/Doc. (median)</th>
<th>Words/Doc. (mean)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>WSJ, 1987-1989</td>
<td>267</td>
<td>98,732</td>
<td>245</td>
<td>434.0</td>
</tr>
<tr>
<td></td>
<td>AP, 1989</td>
<td>254</td>
<td>84,678</td>
<td>446</td>
<td>473.9</td>
</tr>
<tr>
<td></td>
<td>ZIFF</td>
<td>242</td>
<td>75,180</td>
<td>200</td>
<td>473.0</td>
</tr>
<tr>
<td></td>
<td>FR, 1989</td>
<td>260</td>
<td>25,960</td>
<td>391</td>
<td>1315.9</td>
</tr>
<tr>
<td></td>
<td>DOE</td>
<td>184</td>
<td>226,087</td>
<td>111</td>
<td>120.4</td>
</tr>
<tr>
<td>2</td>
<td>WSJ, 1990-1992</td>
<td>242</td>
<td>74,520</td>
<td>301</td>
<td>508.4</td>
</tr>
<tr>
<td></td>
<td>AP, 1988</td>
<td>237</td>
<td>79,919</td>
<td>438</td>
<td>468.7</td>
</tr>
<tr>
<td></td>
<td>ZIFF</td>
<td>175</td>
<td>56,920</td>
<td>182</td>
<td>451.9</td>
</tr>
<tr>
<td></td>
<td>FR, 1988</td>
<td>209</td>
<td>19,860</td>
<td>396</td>
<td>1378.1</td>
</tr>
<tr>
<td>3</td>
<td>SJMN, 1991</td>
<td>287</td>
<td>90,257</td>
<td>379</td>
<td>453.0</td>
</tr>
<tr>
<td></td>
<td>AP, 1990</td>
<td>237</td>
<td>78,321</td>
<td>451</td>
<td>478.4</td>
</tr>
<tr>
<td></td>
<td>ZIFF</td>
<td>345</td>
<td>161,021</td>
<td>122</td>
<td>295.4</td>
</tr>
<tr>
<td></td>
<td>PAT, 1993</td>
<td>243</td>
<td>6,711</td>
<td>4,445</td>
<td>5391.0</td>
</tr>
<tr>
<td></td>
<td>FR, 1994</td>
<td>395</td>
<td>55,630</td>
<td>588</td>
<td>644.7</td>
</tr>
<tr>
<td></td>
<td>CR, 1993</td>
<td>235</td>
<td>27,922</td>
<td>288</td>
<td>1373.5</td>
</tr>
<tr>
<td>5</td>
<td>FBIS</td>
<td>470</td>
<td>130,471</td>
<td>322</td>
<td>543.6</td>
</tr>
<tr>
<td></td>
<td>LAT</td>
<td>475</td>
<td>131,896</td>
<td>351</td>
<td>526.5</td>
</tr>
<tr>
<td>6</td>
<td>FBIS</td>
<td>490</td>
<td>120,653</td>
<td>348</td>
<td>581.3</td>
</tr>
</tbody>
</table>

WSJ=Wall Street Journal
AP=Associated Press
ZIFF=Computer Selects, Ziff-Davis
FR=Federal Register
DOE=US DOE Publications
SJMN=San Jose Mercury News
PAT=US Patents
FT=Financial Times
CR=Congressional Record
FBIS=Foreign Broadcast Information Service
LAT=LA Times
Example Document

<doc>
  <docno>WSJ880406-0090</docno>
  <hl>AT&T Unveils Services to Upgrade Phone Networks Under Global Plan</hl>
  <author>Janet Guyon (WSJ Staff)</author>
  <dateline>New York</dateline>
  <text>American Telephone & Telegraph Co. introduced the first of a new generation of phone services with broad ...</text>
</doc>

- Documents contain SGML-Markup tags
- Important fields like document number (<docno>) and text (<text>) can be found in all documents
Example Information Need

<Top>
<num> Number: 168
<title> Topic: Financing AMTRAK
<desc> Description:
A document will address the role of the Federal Government in
financing the operation of the National Railroad Transportation Cor-
poration (AMTRAK).

<narr> Narrative: A relevant document must provide information on
the government’s responsibility to make AMTRAK an economically
viable entity. It could also discuss the privatization of AMTRAK as
an alternative to continuing government subsidies. Documents com-
paring government subsidies given to air and bus transportation with
those provided to AMTRAK would also be relevant.
</top>

- TREC defines information needs (topics) in natural
  language
- These have to be translated into a query and then
  processed
Relevance in TREC

- No exhaustive relevance judgments:
  - That would be too expensive
  - NIST assessors’ relevance judgments are available only for the documents that were among the top-$K$
  - This means the top-$K$ of systems entered in the TREC evaluation for which the information need was developed
Standard Relevance Benchmarks: Others

- GOV2
  - Another TREC/NIST collection
  - 25 million web pages
  - Largest collection that is easily available
  - But still 3 orders of magnitude smaller than what Google/Yahoo/MSN index

- NTCIR
  - East Asian language and cross-language information retrieval

- Cross Language Evaluation Forum (CLEF)
  - This evaluation series has concentrated on European languages and cross-language information retrieval.
Evaluation of Large IR Systems

How do you measure recall on the web?

Search engines often use precision at top-$K$, e.g., $K = 10$ …

… or measures that reward you more for getting rank 1 right than for getting rank 10 right.

Search engines also use non-relevance-based measures.

Example 1: clickthrough on first result

Not very reliable if you look at a single clickthrough

… but pretty reliable in the aggregate.

Example 2: Ongoing studies of user behavior in the lab
A/B Testing

- Purpose: Test a single innovation
- Have most users use old system (prerequisite: You have a large search engine up and running)
- Divert a small proportion of traffic (e.g., 1%) to the new system that includes the innovation
- Evaluate with an “automatic” measure like clickthrough on first result
- Now we can directly see if the innovation does improve user happiness.
- Probably the evaluation methodology that large search engines trust most
- Variant: Give users the option to switch to new algorithm/interface
Result Summaries

- How do we present results to the user?
  - Most often: as a list – aka “10 blue links”
  - How should each document in the list be described?
  - This description is crucial.
  - User can identify good hits (= relevant hits) based on description.
  - No need to “click” on all documents sequentially
Doc Description in Result List

Most commonly: doc title, url, some metadata …

… and a summary

How do we “compute” the summary?
Summaries

- Two basic kinds: (i) static (ii) dynamic

- A static summary of a document is always the same, regardless of the query that hit the document.

- Dynamic summaries are query-dependent. They attempt to explain why the document was retrieved for the query at hand.
Static Summaries

- Simplest form of summary takes e.g. the first two sentences or 50 words of a document
- May also extract information from a particular zone of the document or from metadata, e.g. title and author
- Typically extracted and cached at indexing time, so that it can be retrieved and presented quickly
- There are more sophisticated approaches using natural language processing (NLP)
  - Many of these are still subject to research and not within the scope of this lecture
Dynamic Summaries

- Dynamic summaries display one or more “windows” on the document.

- Usually windows contain query terms, and so are often referred to as keyword-in-context or KWIC snippets.
  - If the query is found as a phrase, occurrences of the phrase in the document will be shown as the summary.
  - If not, windows within the document that contain multiple query terms will be selected.
  - These windows may just stretch some number of words to the left and right of the query terms.
  - NLP can also be employed usefully: users prefer snippets that read well because they contain complete phrases.
Dynamic Summaries (2)

- Dynamic summaries are liked by users: scan them to decide if you want to click (e.g. Google provides them).
- However, not easy to implement as they cannot be precomputed.
- Reconstructing the context with only a positional index is also difficult and time-consuming.
- However, generating snippets must be fast since many snippets are typically generated for each query.
- Caching the whole documents is not feasible, it is common to cache a fixed-size prefix.
  - For short documents, the whole document is cached.
  - For longer documents we assume that prefix will contain some summary.
Summary

In this chapter, we looked at:

- how to evaluate the retrieval quality of an IR system:
  - Discussing different measures for doing so
  - Explaining how relevance of documents is determined
- how to present summaries of the document answer set to a user