Evaluating search engines
CE-324: Modern Information Retrieval
Sharif University of Technology

M. Soleymani
Fall 2014

Most slides have been adapted from: Profs. Manning, Nayak & Raghavan (CS-276, Stanford)
Why do we need system evaluation?

- How do we know which of the already introduced techniques are effective in which applications?
  - Should we use stop lists? Should we stem? Should we use inverse document frequency weighting?

- We need evaluation to demonstrate the superior performance of novel techniques on representative document collections.
User happiness is elusive to measure

- The key utility measure is user happiness.
  - How satisfied is each user with the obtained results?
  - The most common proxy to measure human satisfaction is *relevance* of search results to the posed information

- How do you measure relevance?

- Relevance measurement requires 3 elements:
  1. A benchmark doc collection
  2. A benchmark suite of information needs
  3. A usually binary assessment of either *Relevant* or *Nonrelevant* for each information needs and each document
    - Some work on more-than-binary, but not the standard
Evaluating an IR system

- **Note**: The information need is translated into a **query**.

- User happiness can only be measured by **relevance** to an information need, not by relevance to queries.

- Evaluate whether doc addresses information need
  - not whether it has these words

Sec. 8.1
Standard relevance benchmarks

- TREC: NIST has run a large IR test bed for many years
- Reuters and other benchmark doc collections
- “Retrieval tasks” specified
  - sometimes as queries
- Human experts mark, for each query and for each doc, Relevant or Nonrelevant
  - or at least for subset of docs that some systems (participating in the competitions) returned for that query
Unranked retrieval evaluation: Precision and Recall

- **Precision**: $P(\text{relevant}|\text{retrieved})$
  - fraction of retrieved docs that are relevant
- **Recall**: $P(\text{retrieved}|\text{relevant})$
  - fraction of relevant docs that are retrieved

<table>
<thead>
<tr>
<th></th>
<th>Relevant</th>
<th>Nonrelevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retrieved</td>
<td>tp</td>
<td>fp</td>
</tr>
<tr>
<td>Not Retrieved</td>
<td>fn</td>
<td>tn</td>
</tr>
</tbody>
</table>

Precision $P = \frac{tp}{tp + fp}$

Recall $R = \frac{tp}{tp + fn}$
Accuracy measure for evaluation?

- **Accuracy**: fraction of classifications that are correct
  - evaluation measure in machine learning classification works

- The **accuracy** of an engine:
  - \( \frac{(tp + tn)}{(tp + fp + fn + tn)} \)

- Given a query, an engine classifies each doc as “Relevant” or “Nonrelevant”

- Why is this not a very useful evaluation measure in IR?
Why not just use accuracy?

- How to build a 99.9999% accurate search engine on a low budget….
  - The snoogle search engine below always returns 0 results (“No matching results found”), regardless of the query
  - Since many more non-relevant docs than relevant ones

![snoogle.com](snoogle.com)

Search for: 

0 matching results found.

- People *want to find something* and have a certain tolerance for junk.
Precision/Recall

- Retrieving all docs for all queries!
  - High recall but low precision

- Recall is a non-decreasing function of the number of docs retrieved

- In a good system, precision decreases as either the number of docs retrieved or recall increases
  - This is not a theorem, but a result with strong empirical confirmation
A combined measure: $F$

- Combined measure: **F measure**
  - allows us to trade off precision against recall
  - weighted harmonic mean of $P$ and $R$

$$\beta^2 = \frac{1 - \alpha}{\alpha}$$

$$F = \frac{1}{\alpha \frac{1}{P} + (1 - \alpha) \frac{1}{R}} = \frac{(\beta^2 + 1)PR}{\beta^2P + R}$$

- What value range of weights recall higher than precision?
A combined measure: $F$

- People usually use balanced $F$ ($\beta = 1$ or $\alpha = \frac{1}{2}$)

$$F = F_{\beta=1}$$

$$F = \frac{2PR}{P + R}$$

- harmonic mean of $P$ and $R$: $\frac{1}{F} = \frac{1}{2} \left( \frac{1}{P} + \frac{1}{R} \right)$
Why harmonic mean

- Why don’t we use a different mean of P and R as a measure?
  - e.g., the arithmetic mean

- The simple (arithmetic) mean is 50% for “return-everything” search engine, which is too high.

- Desideratum: Punish really bad performance on either precision or recall.
  - Taking the minimum achieves this.
  - But minimum is not smooth and hard to weight.
  - F (harmonic mean) is a kind of smooth minimum.
$F_1$ and other averages

Harmonic mean is a conservative average
We can view the harmonic mean as a kind of soft minimum
Evaluating ranked results

- Precision, recall and F are measures for (unranked) sets.
  - We can easily turn set measures into measures of ranked lists.

- Evaluation of ranked results:
  - Taking various numbers of top returned docs (recall levels)
    - Sets of retrieved docs are given by the top k retrieved docs.
      - Just compute the set measure for each “prefix”: the top 1, top 2, top 3, top 4, and etc results
  - Doing this for precision and recall gives you a precision-recall curve
A precision-recall curve
Interpolated precision

- Interpolation: Take maximum of all future points
- Rationale for interpolation: The user is willing to look at more stuff if both precision and recall get better.
  - If locally precision increases with increasing recall, then you should get to count that…

![Graphs](image-url)
An interpolated precision-recall curve

\[ p_{interp}(r) = \max_{r' \geq r} p(r') \]
Averaging over queries

- Precision-recall graph for one query
  - It isn’t a very sensible thing to look at

- Average performance over a whole bunch of queries.

- But there’s a technical issue:
  - Precision-recall: only place some points on the graph
  - How do you determine a value (interpolate) between the points?
Evaluation

- Graphs are good, but people want summary measures!
  - 11-point interpolated average precision
  - Precision at fixed retrieval level
  - MAP
  - R-precision
11-point interpolated average precision

- The standard measure in the early TREC competitions

- Precision at 11 levels of recall varying from 0 to 1
  - by tenths of the docs using interpolation and average them

- Evaluates performance at all recall levels (0, 0.1, 0.2, ..., 1)
Typical (good) 11 point precisions

- SabIR/Cornell 8A1
  - 11 pt precision from TREC 8 (1999)
Precision-at-k

- **Precision-at-k**: Precision of top $k$ results

- Perhaps appropriate for most of web searches
  - people want good matches on the first one or two results pages

- Does not need any estimate of the size of relevant set
  - But: averages badly and has an arbitrary parameter of $k$
Mean Average Precision (MAP)

- Mean Average Precision (MAP)
  - Average precision is obtained for the top \( k \) docs, each time a relevant doc is retrieved

- Avoids interpolation and use of fixed recall levels

- MAP for query collection is arithmetic average
  - Macro-averaging: each query counts equally
Mean Average Precision (MAP)

- \( Q \): set of information needs
- Set of relevant docs to \( q_j \in Q \): \( d_{j1}, d_{j2}, \ldots, d_{jm_j} \)
- \( R_{jk} \): set of ranked retrieval results from the top until reaching \( d_{jk} \)

\[
MAP(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{i=1}^{m_j} \text{Precision}(R_{kj})
\]
R-precision

- *Rel*: A known (though perhaps incomplete) set of relevant docs

- Calculate precision of the top $|Rel|$ docs returned
  - $r$ relevant among the top $|Rel|$ results \Rightarrow for this set
    \[ P = R = \frac{r}{|Rel|} \]

- Perfect system could score 1.0.
Variance

“The variance in performance of the same system across queries”

is much greater than

“the variance of different systems on the same query.”

- There are easy information needs and hard ones!
Creating Test Collections
for IR Evaluation
TREC

- TREC Ad Hoc task from first 8 TRECgs is standard IR task
  - 50 detailed information needs for each year
  - Human evaluation of pooled results returned

- A TREC query (TREC 5): Example

  `<top>
  <num> Number: 225
  <desc> Description:
  What is the main function of the Federal Emergency Management Agency (FEMA) and the funding level provided to meet emergencies? Also, what resources are available to FEMA such as people, equipment, facilities?
  </top>`
Other standard relevance benchmarks

- **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)**
  - European languages and cross-language information retrieval.
From doc collections to test collections

- Test queries (information needs)
  - Must be germane to docs available
  - Best designed by domain experts
  - Random query terms generally not a good idea

- Relevance assessments
  - Human judges, time-consuming
  - Pooling
  - Are human panels perfect?
Kappa measure for inter-judge (dis)agreement

- **Kappa measure**
  - Agreement measure among judges
  - Designed for categorical judgments
  - Corrects for chance agreement

\[
Kappa = \frac{P(A) - P(E)}{1 - P(E)}
\]

- \( P(A) \): proportion of time judges agree
- \( P(E) \): what agreement would be by chance

\( Kappa = 0 \) for chance agreement, \( 1 \) for total agreement.
### Kappa measure: example

\[ P(A) \text{?} \quad P(E) ? \]

<table>
<thead>
<tr>
<th>Number of docs</th>
<th>Judge 1</th>
<th>Judge 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>300</td>
<td>Relevant</td>
<td>Relevant</td>
</tr>
<tr>
<td>70</td>
<td>Nonrelevant</td>
<td>Nonrelevant</td>
</tr>
<tr>
<td>20</td>
<td>Relevant</td>
<td>Nonrelevant</td>
</tr>
<tr>
<td>10</td>
<td>Nonrelevant</td>
<td>Relevant</td>
</tr>
</tbody>
</table>
Kappa example

\[ P(A) = \frac{370}{400} = 0.925 \]

\[ P(\text{nonrelevant}) = \frac{10 + 20 + 70 + 70}{800} = 0.2125 \]

\[ P(\text{relevant}) = \frac{10 + 20 + 300 + 300}{800} = 0.7878 \]

\[ P(E) = 0.2125^2 + 0.7878^2 = 0.665 \]

\[ Kappa = \frac{0.925 - 0.665}{1 - 0.665} = 0.776 \]
Kappa

- $Kappa > 0.8$
  - good agreement
- $0.67 < Kappa < 0.8$
  - “tentative conclusions” (Carletta ’96)
- $Kappa < 0.67$
  - A dubious basis for evaluation

Precise cutoffs depends on purpose of study

For >2 judges: average pairwise kappas
Impact of inter-judge agreement

- Impact on absolute performance measure can be significant (0.32 vs 0.39)

- Little impact on ranking of different systems or relative performance

- “Algorithm A is better than algorithm B?”
  - A standard information retrieval experiment will give us a reliable answer to this question.
Difficulties in (Precision/Recall) system evaluation

- Should average over large doc collection/query ensembles
- Need human relevance assessments
  - People aren’t reliable assessors
- Assessments have to be binary
  - Nuanced assessments?
- Heavily skewed by collection/authorship
  - Results may not translate from one domain to another
Critique of pure relevance

- Docs are scored against queries, not against each other

Relevance vs. Marginal Relevance
- A doc can be redundant even if it is highly relevant
- The same information from different sources (Duplicates)
- Marginal relevance is a better measure of utility

- Using facts/entities as evaluation units more directly measures true relevance.
  - But harder to create evaluation set
Can we avoid human judgment?

- No. However, needing human judgment makes experimental work hard
  - Especially on a large scale

- In some very specific settings, can use proxies
  - Example: for approximate vector space retrieval
    - compare cosine distance closeness of the closest docs to those found by an approximate retrieval algorithm
Evaluation at large search engines

- Test collections of queries and hand-ranked results

- Recall is difficult to measure 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.
    - NDCG (Normalized Cumulative Discounted Gain)

- Search engines also use non-relevance-based measures.
  - Click-through on first result
    - Not very reliable if you look at a single click-through … but pretty reliable in the aggregate.
  - Studies of user behavior in the lab

- A/B testing
A/B testing: refining a deployed system

- **Purpose**: Test a single innovation

- **Prerequisite**: You have a large search engine up and running.

- **Method**: Divert a small proportion of traffic (e.g., 1%) to the new system that includes the innovation
  - So most users use old system
A/B testing: refining a deployed system

- Evaluate with an “automatic” measure
  - E.g., click-through on first result

- Does innovation improve user happiness?

- Probably the evaluation methodology that large search engines trust most
A broader perspective for IR system evaluation

- System issues
- User utility
System issues

- How fast does it index?
  - Number of documents (or bytes) per hour
- How fast does it search?
  - Latency as a function of queries per second
- How large is its document collection?
- Expressiveness of query language
  - Ability to express complex information needs
  - Speed on complex queries

All of the preceding criteria are *measurable*: we can quantify speed/size we can also make expressiveness precise
User utility

- **The key measure is user happiness**

- Factors of **user happiness** include:
  - Speed of response
  - Uncluttered User Interface
  - Most important: **relevance**
    - Speed of response and size of index are factors but blindingly fast, useless answers won’t make a user happy

- Quantifying aggregate user happiness based on relevance, speed, and user interface of the system

- User satisfaction can be measured by running user studies
Measuring user happiness

“Issue: who is the user we are trying to make happy?”

- Web search engine: searcher
- Web search engine: advertiser
- Ecommerce: buyer
- Ecommerce: seller
- Enterprise: CEO
Measuring user happiness

“Issue: who is the user we are trying to make happy?”

- **Web search engine: searcher**
  - Success: Searcher finds what she was looking for
  - Measure: rate of return to this search engine

- **Web search engine: advertiser**
  - Success: Searcher clicks on ad.
  - Measure: clickthrough rate

- Ecommerce: buyer
- Ecommerce: seller
- Enterprise: CEO
Measuring user happiness

“Issue: who is the user we are trying to make happy?”

- Web search engine: searcher
- Web search engine: advertiser
- **Ecommerce: buyer**
  - Success: Buyer buys something
  - Measures: time to purchase, fraction of “conversions” of searchers to buyers
- **Ecommerce: seller**
  - Success: Seller sells something
  - Measure: profit per item sold
- Enterprise: CEO
Measuring user happiness

“Issue: who is the user we are trying to make happy?”

- Web search engine: searcher
- Web search engine: advertiser
- Ecommerce: buyer
- Ecommerce: seller
- **Enterprise: CEO**
  - Success: Employees are more productive (because of effective search)
  - Measure: profit of the company
Result summary or snippet

- Having ranked docs matching a query, we wish to present a results list that is informative to the user
  - Usually, a list of doc titles plus a short summary (snippet)

- **Snippet**: a short summary of the document that is designed so as to allow the user to decide its relevance
News for president rouhani

Academic Freedoms In Iran Should Grow, President Rouhani Says
Huffington Post - 4 days ago
Iransans celebrate the victory of moderate presidential candidate
Hassan Rouhani (portrait) in the presidential elections at Vanak
square in ...

President Rouhani: Iran to Maintain Peaceful Interaction with World
Tasnim News Agency - 4 days ago
Rouhani promises academic freedom at Iranian universities
Asharq Alawsat English - 2 days ago

Hassan Rouhani - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Hassan_Rouhani
Jump to Presidential campaign - Main article: Hassan Rouhani presidential
campaign, 2013. See also: Iranian presidential election, 2013. Our centrifuges ...
Mohammad Bagher Ghalibaf - Itihad - Glasgow Caledonian University - Sorkheh

Hassan Rouhani (HassanRouhani) on Twitter
https://twitter.com/HassanRouhani
The latest from Hassan Rouhani (@HassanRouhani), Iranian President’s Sole
English Account; Persian @Rouhani_ir; media@rouhani.ir. Tehran, Iran.

Images for president rouhani - Report images

President Rouhani: Iran has nothing to hide - The iran project
theiranproject.com/blog/2013/.../president-rouhani-iran-has-nothing-to-hid...
Result summery or snippet

- Title is often automatically extracted from doc metadata.
  - Or field and zone

- What about summaries?
  - This description is crucial.
  - User can identify good/relevant hits based on description.

- Two basic kinds:
  - Static
  - Dynamic
Summaries

- **Static summary** of a doc is always the same, regardless of the query that hit the doc.

- **Dynamic summary** is a *query-dependent* attempt to explain why doc was retrieved for query at hand.
Static summaries

- In typical systems, static summary is a subset of doc.
  - **Simplest heuristic**: e.g., title & the first 50 words of the doc
    - Summary cached at indexing time
  - **More sophisticated**: extract from each doc a set of “key” sentences
    - Simple NLP heuristics to score each sentence and summary is made up of top-scoring sentences.
  - **Most sophisticated**: NLP used to synthesize a summary
    - Seldom used in IR; cf. text summarization work
Dynamic summaries

- Present one or more “windows” within the doc that contain several of the query terms
  - “KWIC” snippets: Keyword in Context
- Requires a high disk space to save docs or at-least their prefixes
  - However, they can greatly improve the usability of IR systems.
Techniques for dynamic summaries

- Find small windows in doc that contain query terms
  - Requires fast window lookup in a doc cache

- Score each window wrt query
  - Use various features such as window width, position, etc.
  - Combine features through a scoring function

- Challenges in evaluation: judging summaries
  - Pairwise comparisons rather than binary relevance assessments
Quicklinks

- Example navigational query: **united airlines**
- user’s need likely satisfied on [www.united.com](http://www.united.com)
- Quicklinks provide navigational cues on that home page
Alternative results presentations?

United Airlines - Airline Tickets, Travel Deals and Flights on ...
https://www.united.com/
Find travel deals and flights on united.com. Book airline tickets and MileagePlus award tickets to more than 350 international and U.S. destinations.

Reservations
Book your flight reservations, hotel, rental car, cruise and vacation ...

Flight Search
Note: MileagePlus Premier members may receive better ...

Baggage Information
Find links to all the baggage information you need for your ...

Flight Check-in
Find more information and help on how to get started with your ...

Flight Status & Information
... schedule. Check flight status, find your gate and access other ...

Sign In
To sign in, please enter your MileagePlus number and ...

United Airlines - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/United_Airlines
United Airlines, Inc. (commonly referred to simply as "United") is an American major airline headquartered in Chicago, Illinois. In the late 1920s, just prior to the ...
Resources for this lecture

- IIR 8
- MIR Chapter 3
- MG 4.5