Evaluation in IR

Introduction to Information Retrieval
CS 221
Donald J. Patterson

Content adapted from Hinrich Schütze
http://www.informationretrieval.org
Exercise

- If my system returns A, C, D, E to query q....

<table>
<thead>
<tr>
<th>Document</th>
<th>Relevant(q)</th>
<th>Not Relevant(q)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td></td>
<td>√</td>
</tr>
<tr>
<td>C</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>√</td>
<td></td>
</tr>
</tbody>
</table>

- What do I want Accuracy to be?

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(\frac{1}{2})</td>
<td>(\frac{2}{3})</td>
<td>(\frac{1}{2})</td>
</tr>
</tbody>
</table>
Exercise

If my system returns A,C,D,E to query q....

<table>
<thead>
<tr>
<th>Document</th>
<th>Relevant(q)</th>
<th>Not Relevant(q)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>B</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>C</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>D</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>E</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>F</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

What do I want Accuracy to be?

<table>
<thead>
<tr>
<th></th>
<th>Relevant</th>
<th>Not Relevant</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>

\[
\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}
\]
Evaluation in IR

Unranked retrieval - Accuracy
Evaluation in IR

Unranked retrieval - Accuracy

- Welcome to my search engine
Unranked retrieval - Accuracy

- Welcome to my search engine
- I guarantee a 99.9999% accuracy.
Unranked retrieval - Accuracy

- Welcome to my search engine
  - I guarantee a 99.9999% accuracy.
- Bring on the venture capital
Unranked retrieval - Accuracy

• Welcome to my search engine
• I guarantee a 99.9999% accuracy.
• Bring on the venture capital
Evaluation in IR

Unranked retrieval - Accuracy

• Welcome to my search engine
  • I guarantee a 99.9999% accuracy.
  • Bring on the venture capital

Search for: 

Beta Wednesday, March 10, 2010
Evaluation in IR

Unranked retrieval - Accuracy

• Welcome to my search engine
• I guarantee a 99.9999% accuracy.
• Bring on the venture capital

Search for: 

0 matching results found
Unranked retrieval - Accuracy

\[
\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}
\]

\[
\text{Accuracy} = \frac{0+ \uparrow}{0 + 0 + \epsilon + \uparrow}
\]
Unranked retrieval - Accuracy

- Most people want to find something and can tolerate some junk

\[
\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}
\]

\[
\text{Accuracy} = \frac{0 + \uparrow}{0 + 0 + \epsilon + \uparrow}
\]
Unranked retrieval - ROC curve

Receiver Operating Characteristic (ROC) curve

Really good precision and recall
Best you can do
Likely impossible in practice

Precision
\[ \frac{TP}{TP + FP} \]

Recall
\[ \frac{TP}{TP + FN} \]

100%

0%

More junk

Return everything
Very high recall
No value for the user

Less junk

A system can deliver results at any point on its ROC curve by trading off more results that include more junk versus less results that include less junk. Picking the right point requires deciding how much the application/user can tolerate junk.

PitterPattersonFinder
No results
Very high precision
Very high accuracy for the web corpus
No value for the user
Unranked retrieval - ROC curve

Example Histogram of Documents versus relevance score

Receiver Operating Characteristic (ROC) curve

- Precision
- Recall

More junk: 0%, 90%, 80%, 70%, 60%, 50%, 40%, 30%, 20%, 10%

Less junk: 0%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 100%

Relevant
Irrelevant
Evaluation in IR

Unranked retrieval - ROC curve

Example Histogram of Documents versus relevance score

Receiver Operating Characteristic (ROC) curve

- Relevant
- Irrelevant

Precision

Recall

More junk

Less junk

Wednesday, March 10, 2010
Evaluation in IR

Unranked retrieval - ROC curve

Example Histogram of Documents versus relevance score

Receiver Operating Characteristic (ROC) curve

Precision

Recall

0% 100%

0% 100%

Less junk

More junk

Relevant

Irrelevant

0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100

0 5 10 15 20 25 30 35 40 45 50 60 70 80 90

Wednesday, March 10, 2010
Ranked Retrieval

- Precision and Recall are **set-based measures**
- They are computed independent of order
- But, web search return things in lists
- Lists have order.
- A better metric of user happiness/relevance is warranted
Ranked Retrieval

- Let’s use our existing metrics and extend them to ranked retrieval
- In one system we can get many samples
- We can get the top $X$ results:
  - $X = 10, 20, 30, 40, \text{ etc...}$
  - Each one of those sets has a precision and recall value
  - Each of those sets corresponds to a point on the ROC curve.
Ranked Retrieval

- Each of those sets corresponds to a point on the ROC curve.
Ranked Retrieval

- One option is to average the precision scores at discrete points on the ROC curve
- But which points?
- We want to evaluate the system, not the corpus
- So it can’t be based on number of documents returned
Evaluation in IR

Ranked Retrieval - **11 point precision**

- Evaluate based on precision at defined recall points
- Average the precision at 11 points
- This can be compared across corpora
  - because it isn’t based on corpus size or number of results returned
Ranked Retrieval - Mean Average Precision

- Why just 11 points?
- Why not average over all points?
- This is roughly equivalent to measuring the area under the curve.
Ranked Retrieval - Precision at k

- Users don’t care about results past a page or two.
- So area under the curve is too naive.
- Let’s evaluate precision with k results instead.
- Highly dependent on number of relevant documents.
- If k is 20 and relevant docs is 8.
- Best score is $8/(8+12) = 0.4$. 

Recall

Precision

0%

100%

100%

0%

Less junk

More junk

Top K results could fall anywhere
Ranked Retrieval - **Precision at R**

- We know the number of relevant documents, $r$, so rather than looking at $k$ results let’s look at the top $r$ results.
- If $r$ is 20
  - best score is $20/(20) = 1.0$
  - best score is always 1.0
Ranked Retrieval - **Precision at R**

- It turns out that Precision at R is the break-even point
- When Precision and Recall are equal
- Do we care about this point for any rational reason?
Critiques of relevance

• Is the relevance of one document independent of another?
• Is a gold standard possible?
  • Is a gold standard static?
  • Uniform?
  • Binary?
• Perhaps relevance as a ranking is better.
• Relevance versus marginal relevance
  • what does another document add?
Refining a deployed system

- Once you have a system, with metrics, how do you consider changing the system to improve the metrics?
- A common approach is A/B testing.
  - This is done by Google for clients and Amazon for itself and probably many others.

- The idea:
  - Treat a small number of your users as experiments.
  - Have them use the different system.
  - Evaluate metrics on experimental group.
Evaluation in IR

- Gold standard approach

Diagram:
- Corpus Queries
- Gold Standard
- Standard Search Engine
- Collect Metrics
- Statistical Comparison of Metrics
- Experimental Validation or Reject of Alternative
- Proposed Alternative Search Engine
- Collect Metrics
Online A/B approach

- Requires
  - users
  - an infrastructure to support testing
  - metrics that don’t require a gold standard

![Diagram showing the process of evaluating search engines using an online A/B approach.](chart)

User Base

- majority of traffic
- experimental traffic

Group A
- Standard Search Engine
- Collect Metrics

Group B
- Proposed Alternative Search Engine
- Collect Metrics

Statistical Comparison of Metrics

Experimental Validation or Reject of Alternative
Introducing Kindle: Amazon’s Revolutionary Wireless Reading Device

Amazon is excited to introduce Kindle—a wireless, portable reading device with instant access to more than 100,000 books, blogs, newspapers, and magazines. Whether you’re in bed or on the train, Kindle lets you think of a book and get it in less than a minute.

Learn more

Get Yourself a Little Something

Save $50 on Select Toshiba Laptops

For a limited time save $50 on select Toshiba laptops offered by Amazon.com, with power Intel or AMD processors, large hard drives, and 13- or 15.4-inch display sizes. Hurry—sales end March 17.

Add to Your Collection

Captain Underpants And...

In this worthy sequel to The Adventures of Captain Underpants, Pilkey maintains the... Read more

See more items in the Captain Underpants series
Evaluation in IR

Snippets

• Little bits of text that summarize the page

[Informatics - Wikipedia, the free encyclopedia]
Informatics includes the science of information, the practice of information ... Informatics studies the structure, behavior, and interactions of natural and ...
en.wikipedia.org/wiki/Informatics - 35k - Cached

• They function as an implicit tool for users to rank the results on their own (among those visible)

• The user does the final ranking

• Users are still biased by presented order though.
Snippets

- The goal of snippet generation is
  - present the most informative bit of a document in light of the query
  - present something which is self-contained
    - i.e., a clause or a sentence
  - present something short enough to fit in output
  - be fast, accurate (where are the snippets stored?)

- Challenges
  - Multiple occurrences of keyword in document
  - Poor English (or other language) grammar
Evaluation in IR

Snippets

• Snippets can be static
  • A snippet for a web page is precompiled and always the same.

• Snippets can be dynamic
  • Depends on the query
  • “informatics”
  • “informatics definition”
Snippets

- Snippets may contain
  - A few sentences from the web page
  - Meta data about the page
    - Author, Date, Title
  - Output of a text-summarization algorithm
    - Advanced technology that attempts to write snippets
  - Images from the document