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.
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 $\frac{8}{8+12} = 0.4$
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
Online A/B approach

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

Get 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—savings end March 17.
Evaluation in IR

Amazon
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

Shop Irish Jewelry at Amazon.com

Shop all Irish jewelry

Claddagh
Clovers & Shamrocks
Celtic

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—savings end March 17.

Save up to 30%

Save on Xyron’s helpful and handy craft products and expand your crafting repertoire.

$5 or $10 Bonus on Southern Living

Subscribe to Southern Living this month, and by April 15...
Evaluation in IR

Google
# Website Optimizer (beta): Experiment List

**Congratulations!** You've signed up for Website Optimizer and are ready to create your experiment. As you move through the steps, your information will be saved.

+ **Create a new experiment**

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Status</th>
<th>Page Visitors</th>
<th>Conv.</th>
<th>Conv. Rate</th>
<th>Finish Time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

No experiments

Click **Create experiment** to get started.

- **My Client Center User Access**

- **Tell us about Website Optimizer**

- **Take a survey**

---

**Common Questions**

- How do I create an experiment?
- Do I need my webmaster's help?

**Helpful Links**

- Website Optimizer Help
- Website Optimizer Demo (Flash - English Only)
- Website Optimizer User Guide (Flash - English Only)
- Quick Start Guide
- Installation Guide
- Testing Guides and Strategies
- Discussion Forum
- Website Optimizer Home Page
- Send Us Feedback

---

©2007 Google - AdWords Home - Editorial Guidelines - Privacy Policy - Contact Us
Evaluation in IR

Snippets

• Little bits of text that summarize the page

• 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.
Evaluation in IR

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
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”
Evaluation in IR

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