Evaluation Overview

Jaime Arguello
INLS 509: Information Retrieval
jarguell@email.unc.edu

Information Retrieval

- How would you evaluate a search system?
- How would you determine that system A is better than B?

Information Retrieval Evaluation

- Evaluation is a fundamental issue of information retrieval
  - an area of IR research in its own right
- Evaluation methods:
  - batch evaluation
  - user-study evaluation
  - online evaluation
- Each method has advantages and disadvantages

Batch Evaluation

overview

- Collect a set of queries (to test average performance)
- Construct a more complete description of the information being sought for each query

Batch Evaluation

overview: query + description (example)

- QUERY: pet therapy
- DESCRIPTION: Relevant documents must include details of how pet- and 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.

Batch Evaluation

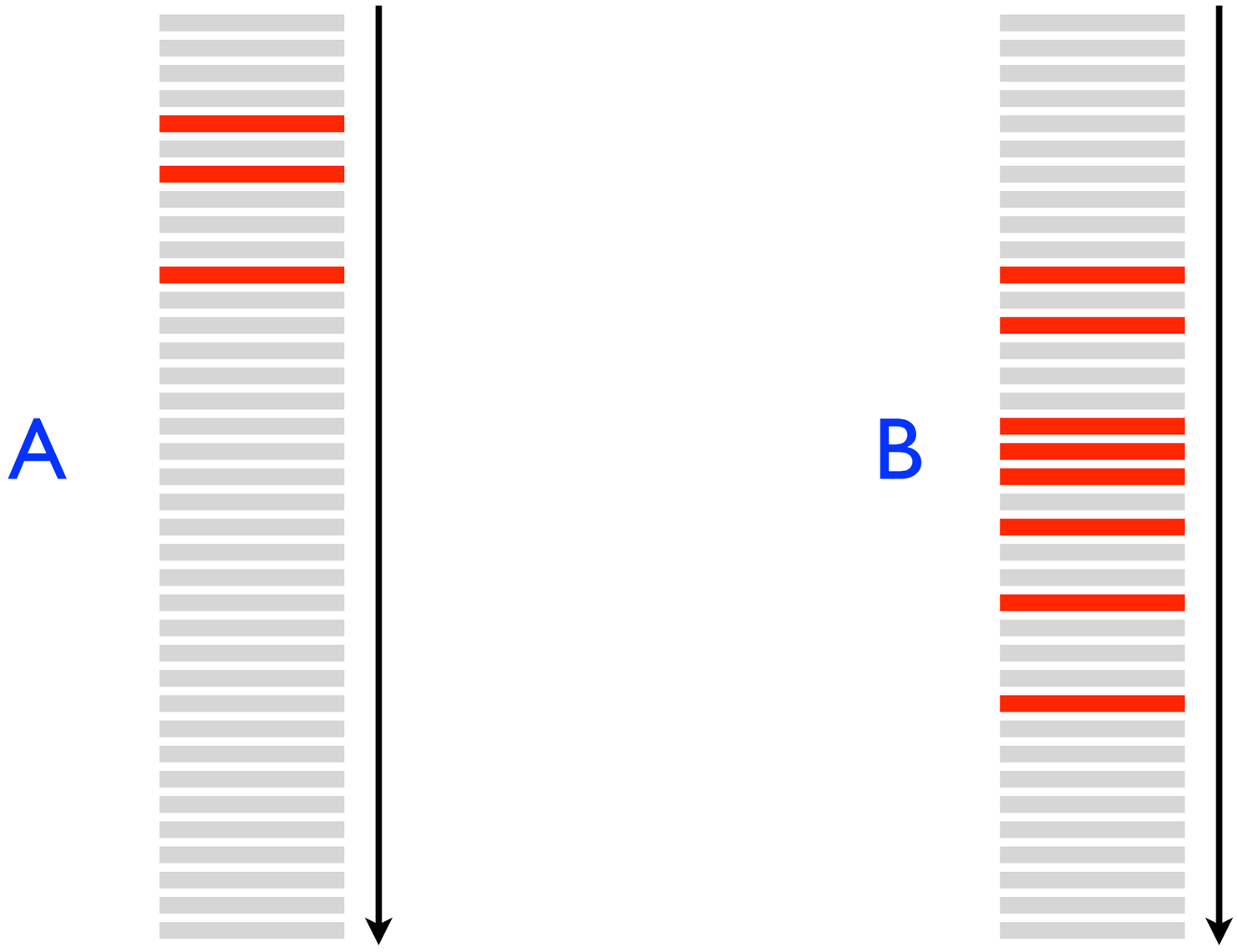
overview

- Using these descriptions, have human assessors determine which documents are relevant for each query
- Evaluate systems based on their ability to retrieve the relevant documents for these queries
  - evaluation metric: a measure of the quality of a particular ranking of results with known relevant/non-relevant documents
  - Evaluation metrics assume that the goal for the system is to displays the most relevant results in ways that users are most likely to engage with them.

Batch Evaluation

overview: metrics

- Which ranking is better?

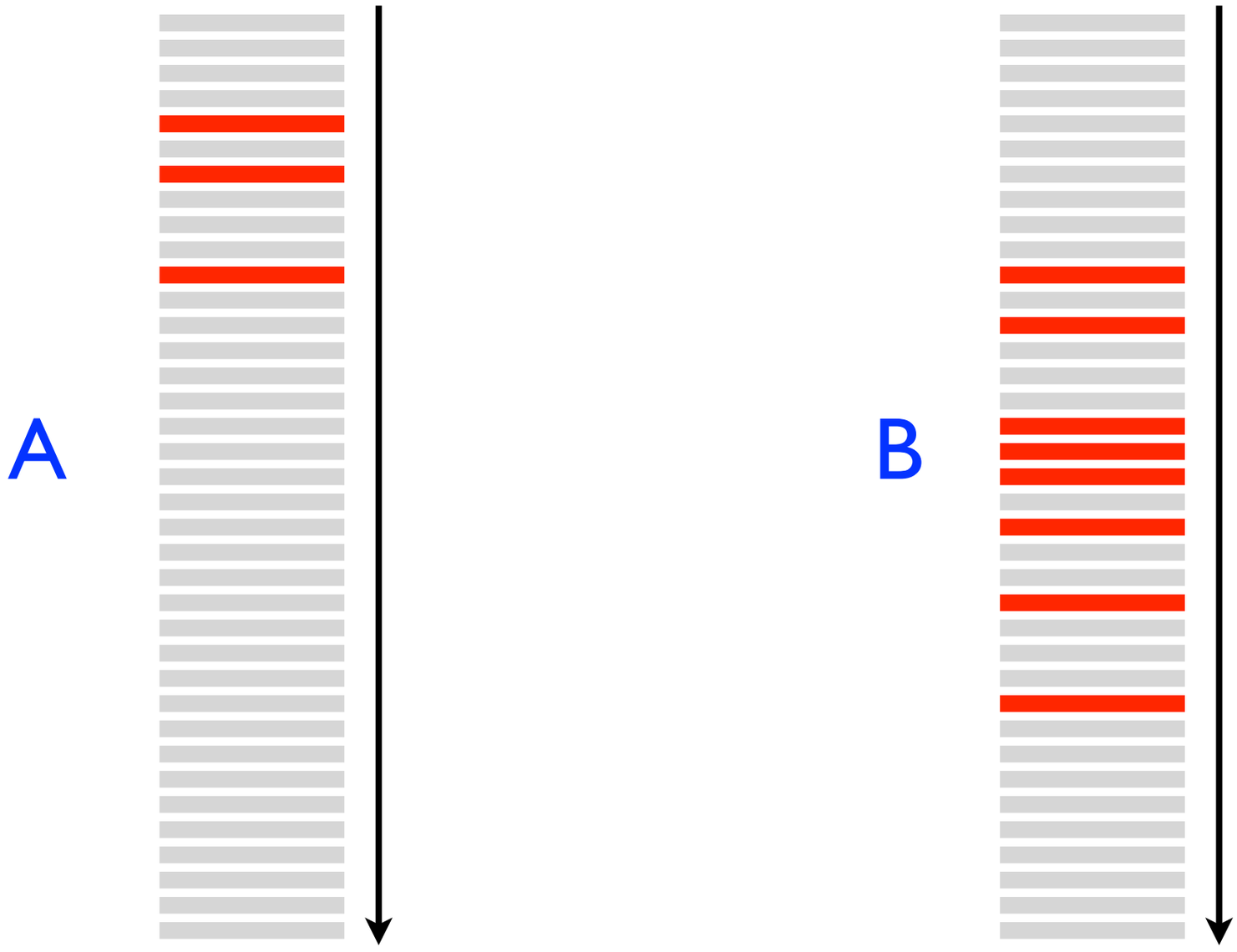


- The **red** documents are relevant

Batch Evaluation

overview: metrics

- Which ranking is better?

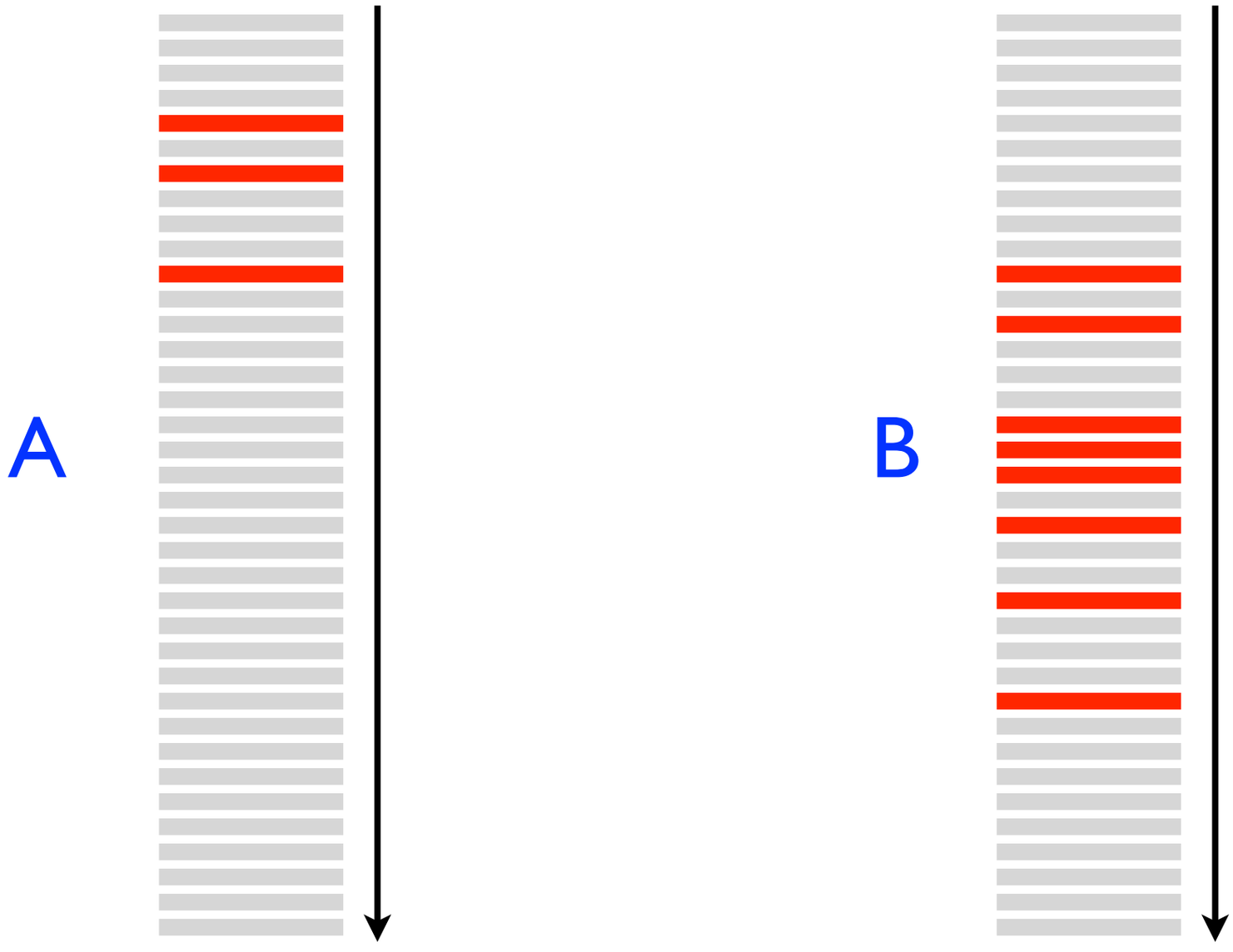


- rank of the first relevant document (lower value is better)

Batch Evaluation

overview: metrics

- Which ranking is better?

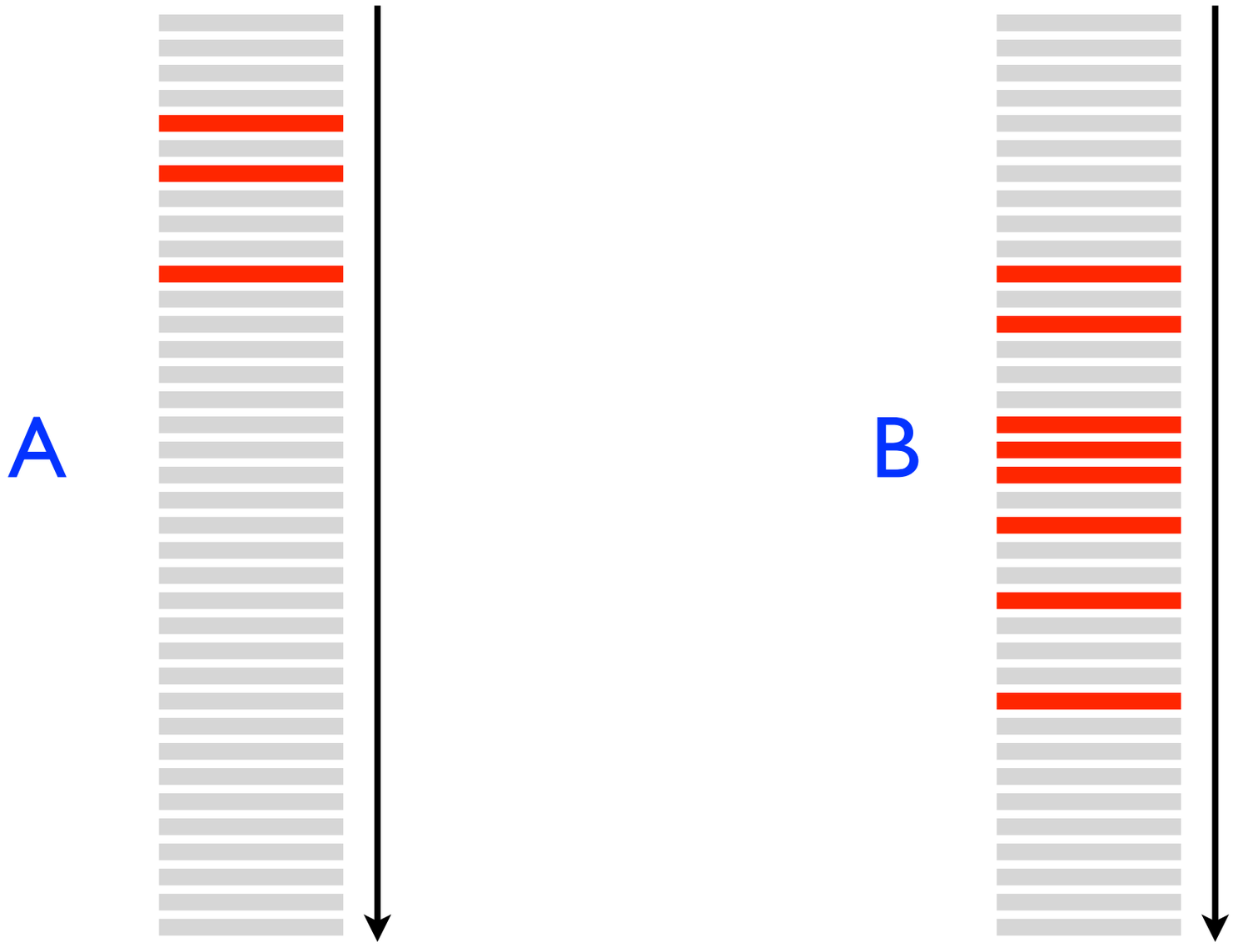


- precision at rank 10 (higher value is better)

Batch Evaluation

overview: metrics

- Which ranking is better?

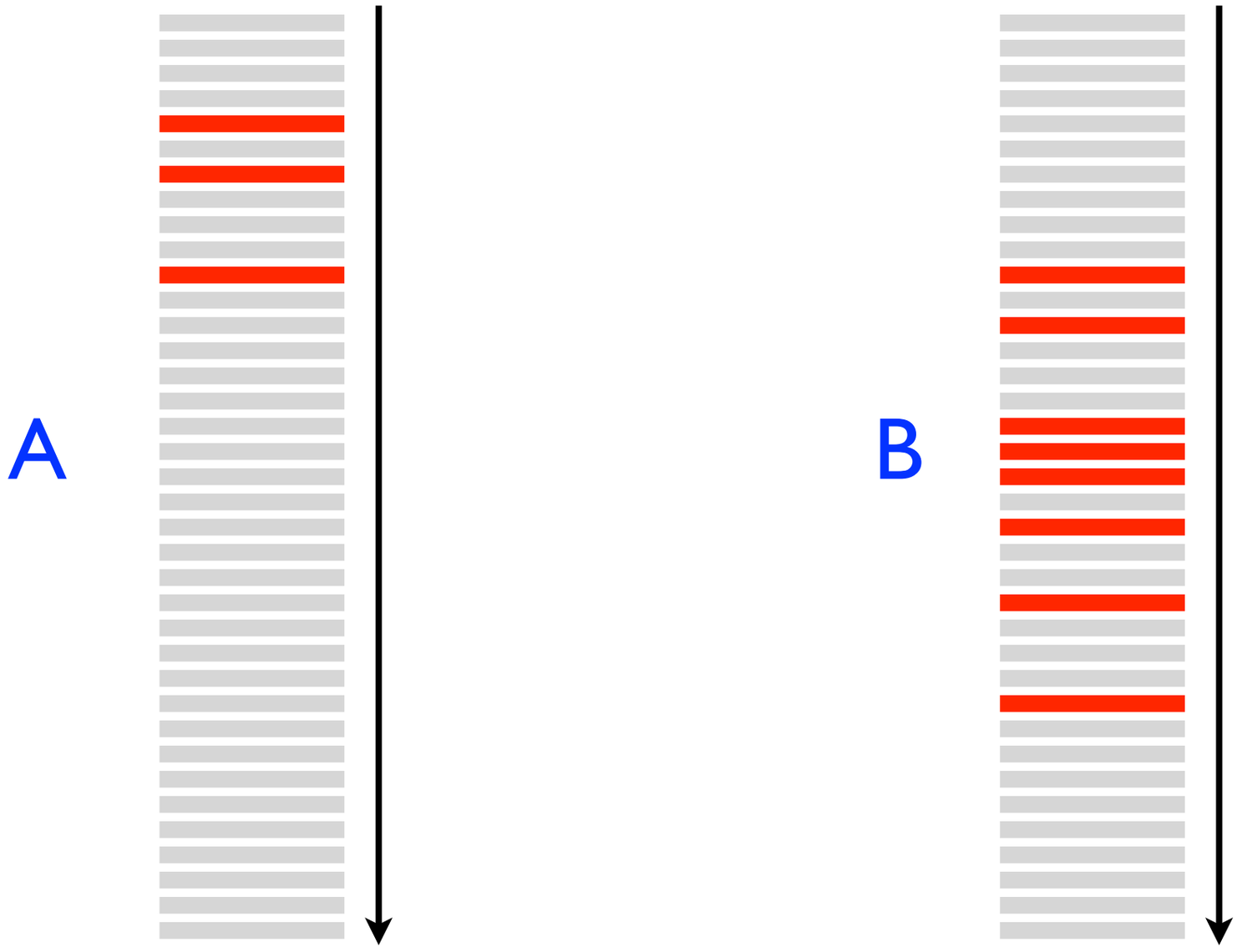


- precision at rank 1 (higher value is better)

Batch Evaluation

overview: metrics

- Which ranking is better?

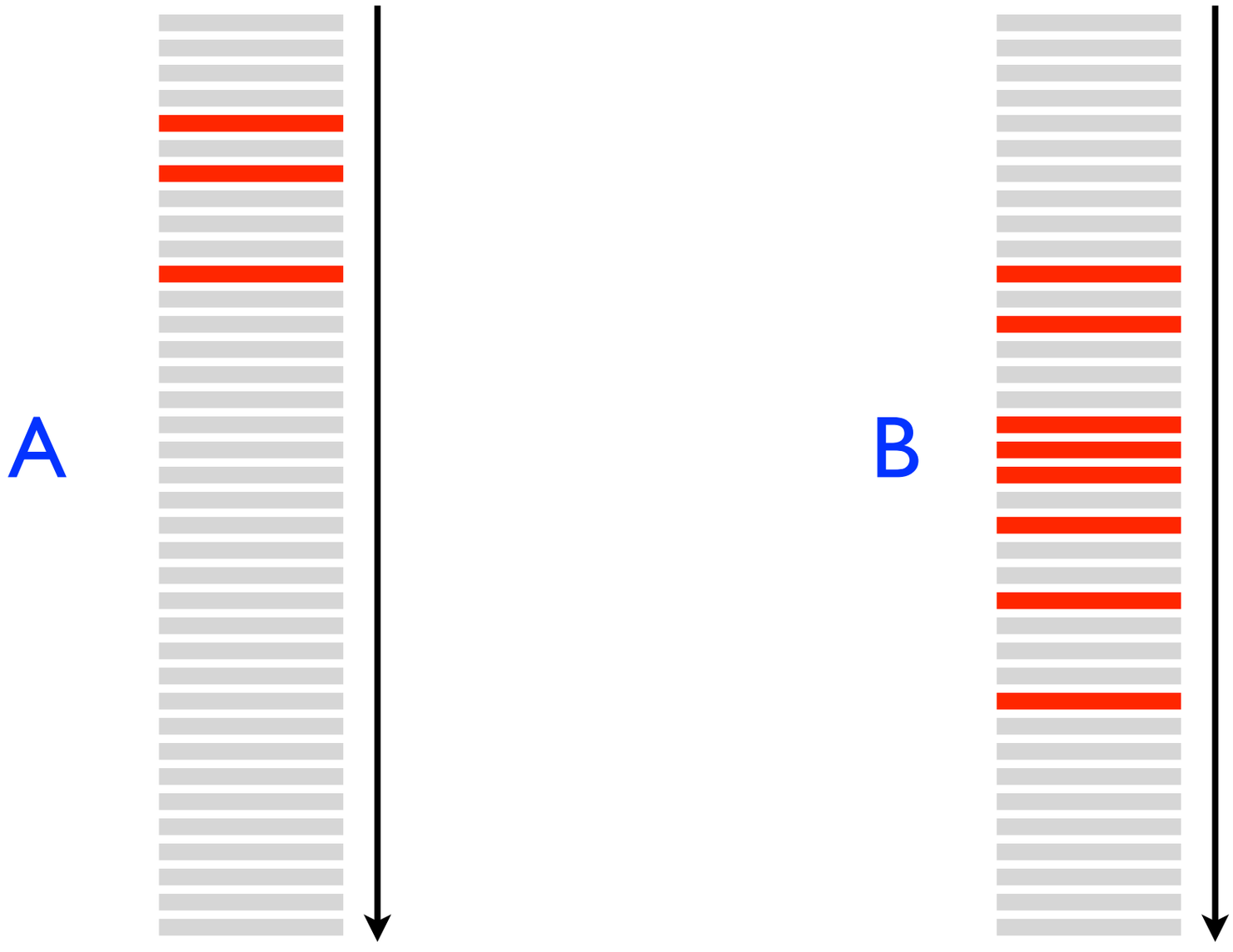


- recall at rank 10 (higher value is better)

Batch Evaluation

overview: metrics

- Which ranking is better?



- recall at rank 30 (higher value is better)

Batch Evaluation

overview: trade-offs

- Advantages:
  - inexpensive (once the test collection is constructed)
  - the experimental condition is fixed; same queries, and same relevance judgements
  - evaluations are reproducible; keeps us “honest”
  - by experimenting on the same set of queries and judgements, we can: (1) understand the strengths and weaknesses of a system and (2) measure small improvements to a baseline system

Batch Evaluation
overview: trade-offs

- Disadvantages:
  - high initial cost. human assessors (the ones who judge documents relevant/non-relevant) are expensive
  - human assessors are not the users; judgements are made “out of context”
  - assumes that relevance is the same, independent of the user and the user’s context

Batch Evaluation

overview: trade-offs

- Many factors affect whether a document satisfies a particular user's information need
- Topicality, novelty, freshness, authority, formatting, reading level, assumed level of expertise
- Topical relevance: the document is on the same topic as the query
- User relevance: everything else
- Which kind of relevance does batch-evaluation address?

Batch Evaluation

overview: trade-offs

- Many factors affect whether a document satisfies a particular user's information need
- Topicality, novelty, freshness, authority, formatting, reading level, assumed level of expertise
- Topical relevance: the document is on the same topic as the query
- User relevance: everything else
- Which kind of relevance does batch-evaluation address?
- Mostly topical relevance

User-Study Evaluation
overview

- Give participants access to one or more systems
- Ask them to complete several (potentially different) tasks
- Learn about system performance by:
  - observing what they do
  - asking about their actions and thought processes
  - measuring success (task completion, time, etc.)
  - measuring perceived success (questionnaire data)

User-Study Evaluation
overview: trade-offs

- Advantages:
  - very detailed data about users' reaction to systems
  - search is done to accomplish a higher-level task
  - tasks can be manipulated and studied
  - we can learn about users' needs, behaviors, and challenges
  - can help us design and evaluate tools to better support searchers (perhaps in specific contexts)

User-Study Evaluation
overview: trade-offs

- Disadvantages:
  - user studies are expensive (pay users/subjects, scientist's time, data coding)
  - difficult to generalize from small studies to broad populations
  - the laboratory setting is not the user's normal environment
  - need to re-run experiment every time a new system is considered

User-Study Evaluation
class discussion

- **Why are behavioral studies necessary?**
- Why can't we simply ask people:
  - what are your needs?
  - how do you behave?
  - what challenges do you face?
  - do you prefer system A or B?
  - help me help you!

On-line Evaluation
overview

- Given a search service with an existing user population (e.g., Google, Yahoo!, Bing) ...
- Have x% of query traffic use system A and y% of query-traffic use system B
- Compare systems using logged data (implicit feedback)
  - clicks: surrogates for perceived relevance (good)
  - skips: surrogates for perceived non-relevance (bad)

Implicit Feedback

implicit feedback in information retrieval

Advanced search

[PDF] **Implicit Feedback** for Interactive **Information Retrieval**
research.microsoft.com/en-us/um/people/ryenw/papers/thesis.pdf
File Format: PDF/Adobe Acrobat
by RW White - 2004 - Cited by 30 - Related articles
Implicit Feedback for Interactive. **Information Retrieval**. Ryen William White. Department of Computing Science. Faculty of Computing Science, Mathematics and ...

[PDF] Context-Sensitive **Information Retrieval** Using **Implicit Feedback**
citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.61.987...
File Format: PDF/Adobe Acrobat - Quick View
by X Shen - 2005 - Cited by 217 - Related articles
ploit **implicit feedback** information, including previous queries and clickthrough information, to improve retrieval accuracy in an in- teractive **information retrieval** ...

Context-Sensitive **Information Retrieval** Using **Implicit Feedback**
citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.61.987
by X Shen - 2005 - Cited by 217 - Related articles
CiteSeerX - Document Details (Isaac Councill, Lee Giles): A major limitation ...

Show more results from psu.edu

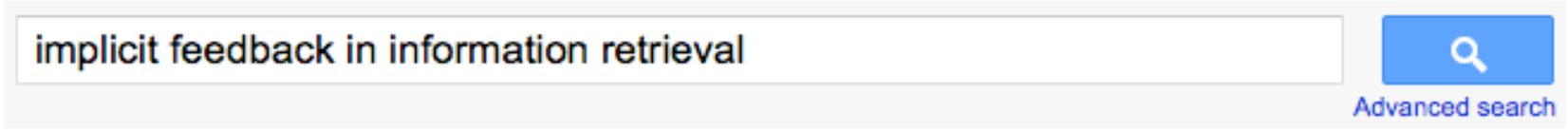
Relevance **feedback** - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Relevance_**feedback**
Relevance feedback is a feature of some **information retrieval** systems. The idea ... 1 Explicit feedback; 2 **Implicit feedback**; 3 Blind feedback; 4 Using relevance ...

A Search Engine that Learn from **Implicit Feedback**
striver.joachims.org/
OSMOT - Learning **Retrieval** Functions from **Implicit Feedback**. ... Such observable behavior gives weak and noisy feedback **information** about which links the ...

click!

what does this tell us?

Implicit Feedback



Implicit Feedback - Under the Reading Lamp
bcao.wikidot.com/**implicit-feedback**
Xuehua Shen, Bin Tan, and ChengXiang Zhai, "Context-sensitive **information retrieval** using **implicit feedback**," in Proceedings of the 28th annual ...

[PDF] **Implicit Feedback** for Interactive **Information Retrieval**
research.microsoft.com/en-us/um/people/ryenw/papers/thesis.pdf
File Format: PDF/Adobe Acrobat
by RW White - 2004 - Cited by 30 - Related articles
Implicit Feedback for Interactive. **Information Retrieval**. Ryen William White. Department of Computing Science. Faculty of Computing Science, Mathematics and ...

[PDF] Context-Sensitive **Information Retrieval** Using **Implicit Feedback**
citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.61.987...
File Format: PDF/Adobe Acrobat - Quick View
by X Shen - 2005 - Cited by 217 - Related articles
ploit **implicit feedback** information, including previous queries and clickthrough information, to improve retrieval accuracy in an in- teractive **information retrieval** ...

Context-Sensitive **Information Retrieval** Using **Implicit Feedback**
citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.61.987
by X Shen - 2005 - Cited by 217 - Related articles
CiteSeerX - Document Details (Isaac Councill, Lee Giles): A major limitation ...

Show more results from psu.edu

Relevance **feedback** - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Relevance_**feedback**
Relevance feedback is a feature of some **information retrieval** systems. The idea ... 1 Explicit feedback; 2 **Implicit feedback**; 3 Blind feedback; 4 Using relevance ...

skip!

click!

what does this tell us?

Implicit Feedback

implicit feedback in information retrieval

Advanced search

Images for **lemurs** - Report images





click!

[PDF] **Implicit Feedback** for Interactive **Information Retrieval**
research.microsoft.com/en-us/um/people/ryenw/papers/thesis.pdf
File Format: PDF/Adobe Acrobat
by RW White - 2004 - Cited by 30 - Related articles
Implicit Feedback for Interactive. **Information Retrieval**. Ryen William White. Department of Computing Science. Faculty of Computing Science, Mathematics and ...

[PDF] Context-Sensitive **Information Retrieval** Using **Implicit Feedback**
citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.61.987...
File Format: PDF/Adobe Acrobat - Quick View
by X Shen - 2005 - Cited by 217 - Related articles
ploit **implicit feedback** information, including previous queries and clickthrough information, to improve retrieval accuracy in an in- teractive **information retrieval** ...

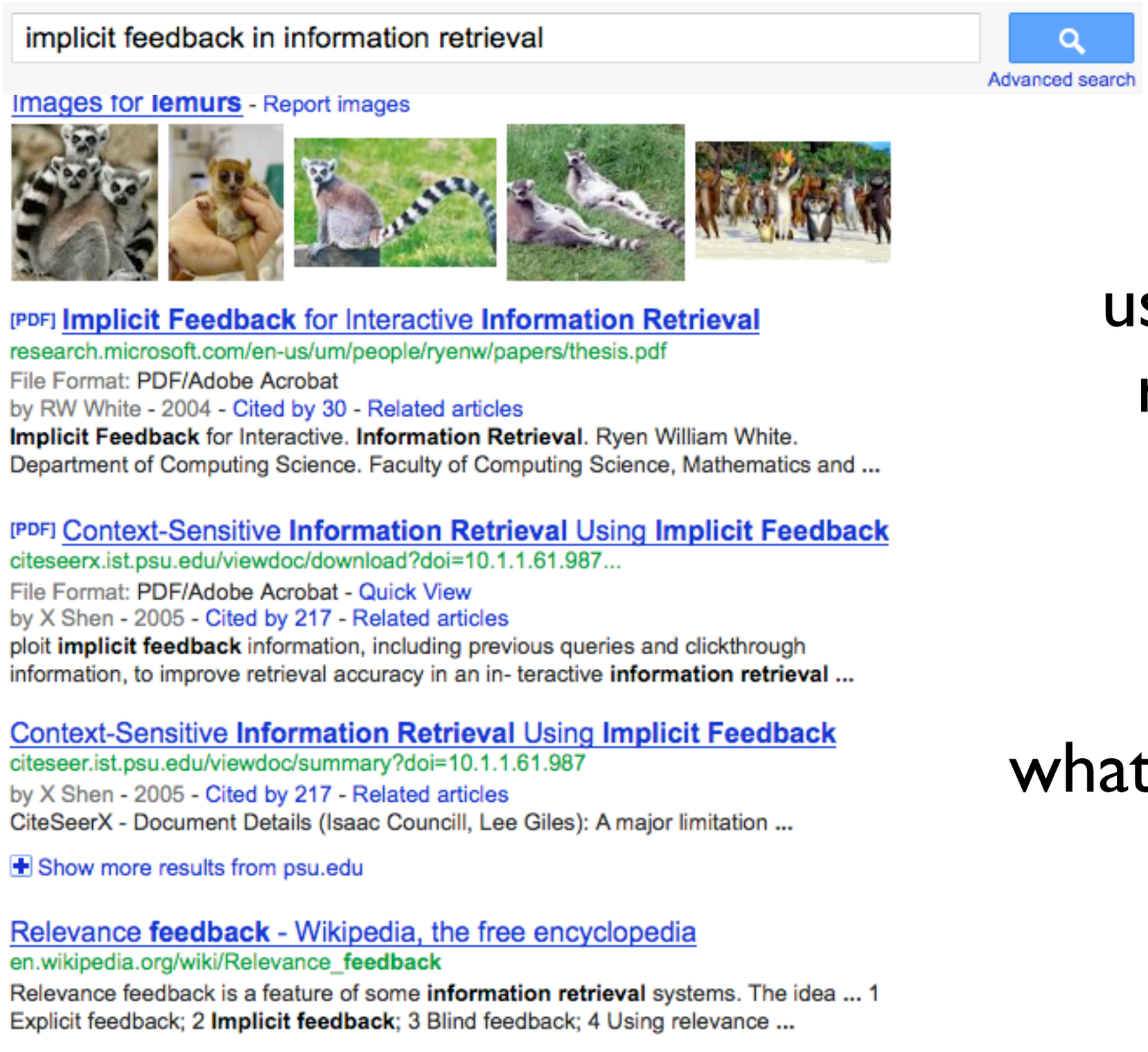
Context-Sensitive **Information Retrieval** Using **Implicit Feedback**
citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.61.987
by X Shen - 2005 - Cited by 217 - Related articles
CiteSeerX - Document Details (Isaac Councill, Lee Giles): A major limitation ...

Show more results from psu.edu

Relevance **feedback** - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Relevance_**feedback**
Relevance feedback is a feature of some **information retrieval** systems. The idea ... 1 Explicit feedback; 2 **Implicit feedback**; 3 Blind feedback; 4 Using relevance ...

what does this tell us?

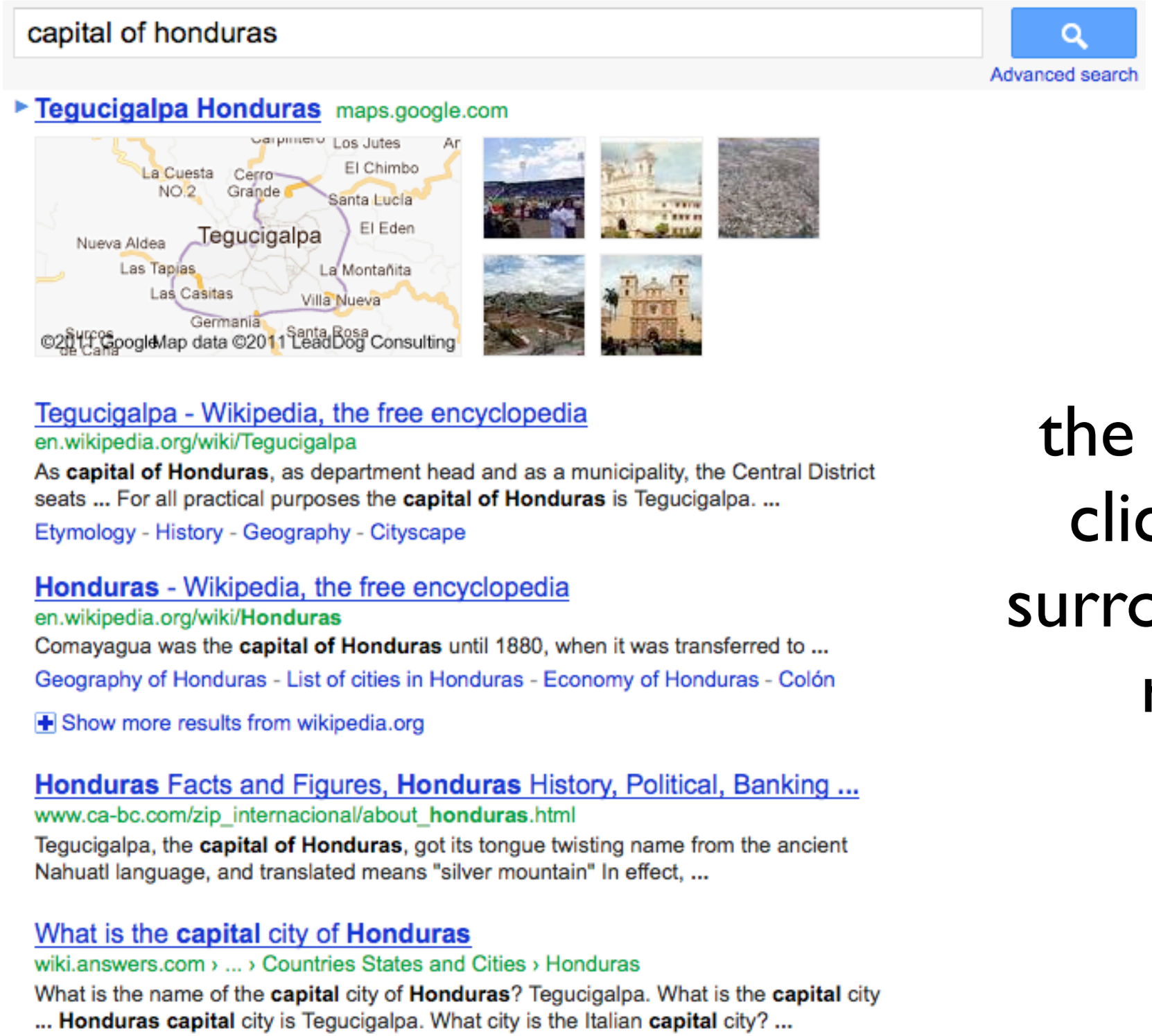
Implicit Feedback



user sees the results and closes the browser

what does this tell us?

Implicit Feedback



the absence of a click is a noisy surrogate for non-relevance

On-Line Evaluation

overview: trade-offs

- Advantages:
  - system usage is naturalistic; users are situated in their natural context and often don't know that a test is being conducted
  - evaluation can include lots of users

On-Line Evaluation

overview: trade-offs

- Disadvantages:
  - requires a service with lots of users (enough of them to potential hurt performance for some)
  - this is often referred to as the “cold-start problem”
  - requires a good understanding on how different implicit feedback signals correlate with positive and negative user experiences
  - experiments are difficult to repeat

Information Retrieval Evaluation

- Evaluation is a fundamental issue of information retrieval
  - an area of IR research in its own right
- Evaluation methods:
  - batch evaluation
  - user-study evaluation
  - online evaluation
- Each method has advantages and disadvantages