Document Priors

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

October 7, 2013

Outline

Introduction to language modeling

Language modeling for information retrieval

Query-likelihood Retrieval Model

Smoothing

Priors

Linear Interpolation

Review

$$score(Q, D) = \prod_{i=1}^{n} (\lambda P(q_i|D) + (1 - \lambda) P(q_i|\mathrm{C}))$$

- $P(q_i|D)$ = probability given to query term q_i by the document language model
- $P(q_i|C)$ = probability given to query term q_i by the collection language model

Linearly Interpolated Smoothing
Review

- Doc 1: haikus are easy
- Doc 2: but sometimes they don't make sense
- Doc 3: refrigerator
- Query: haikus make sense

$$score(Q, D) = \prod_{i=1}^{n} (\lambda P(q_i|D) + (1 - \lambda)P(q_i|\mathrm{C}))$$

(source: threadless t-shirt)

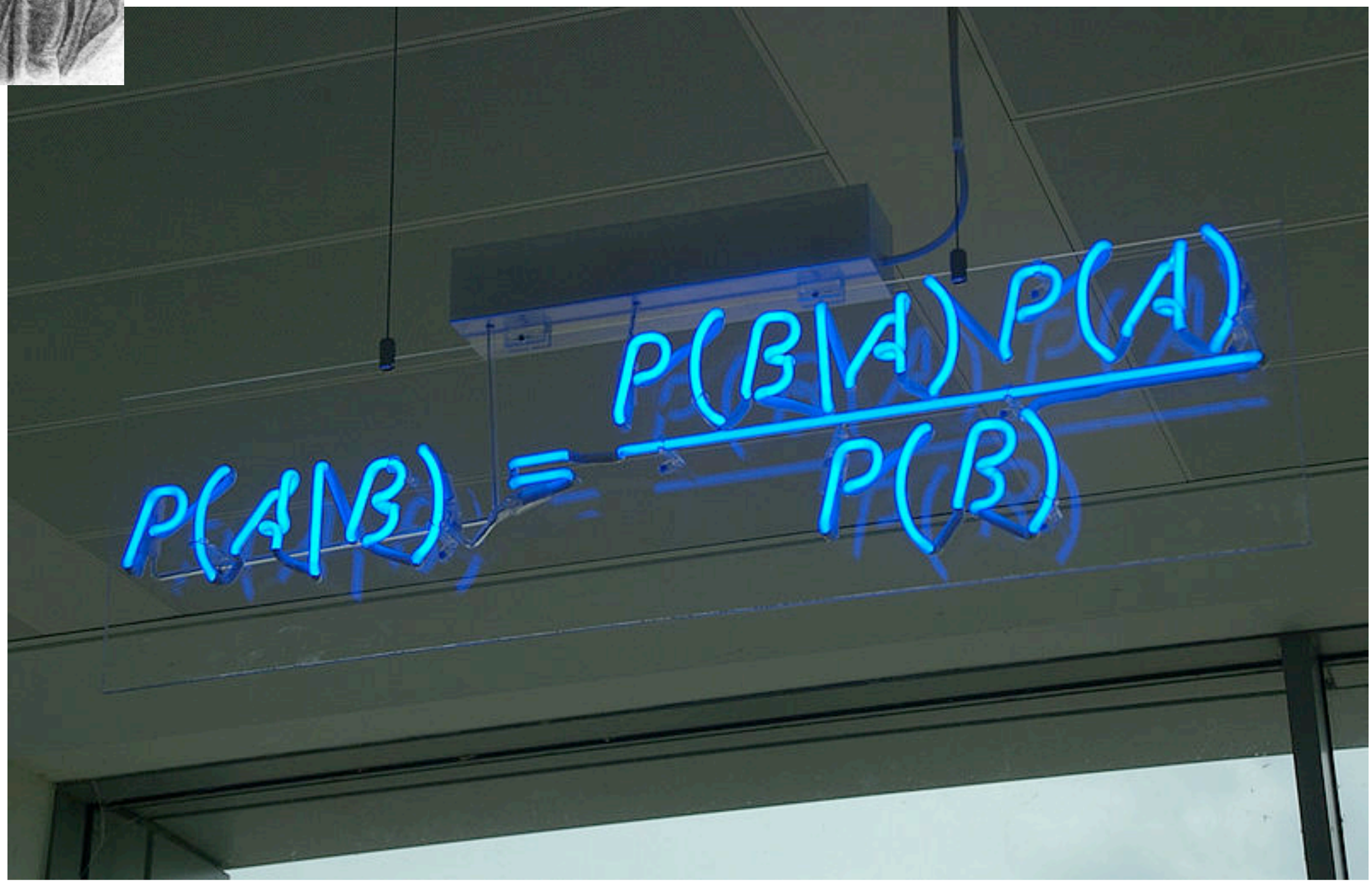
Let's Take A Step Back

- The query likelihood model has a more theoretic motivation than I've portrayed so far

Bayes' Law

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

Bayes' Law



(source: wikipedia)

Bayes' Law Derivation

$$P(A,B) = P(A|B) \times P(B)$$

$$P(A,B) = P(B|A) \times P(A)$$

$$P(A,B) = P(A,B)$$

$$P(A|B) \times P(B) = P(B|A) \times P(A)$$

$$\boxed{P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}}$$

Bayes' Law Applied to Ranking

$$P(D|Q) = \frac{P(Q|D) \times P(D)}{P(Q)}$$

Bayes' Law Applied to Ranking

$$P(D|Q) = \frac{P(Q|D) \times P(D)}{P(Q)}$$

If we're scoring and ranking documents based on this formula, which number doesn't matter?

Query-likelihood Retrieval Model

- Dividing every document score by the same number doesn't change the ranking of documents ...
- So, we can ignore the denominator P(Q)

$$P(D|Q) = \frac{P(Q|D) \times P(D)}{P(Q)}$$

$$P(D|Q) \propto P(Q|D) \times P(D)$$

query-likelihood score
(you already know this)

document prior
(new concept)

Document Prior

$$P(D|Q) \propto P(Q|D) \times P(D)$$

query-likelihood score
(you already know this)

document prior
(new concept)

- The document prior, P(D), is the probability that the document is relevant to any query
- It is a document-specific probability
- It is a query-independent probability

Document Prior

$$P(D|Q) \propto P(Q|D) \times P(D)$$

query-likelihood score
(you already know this)

document prior
(this is a new concept)

- Unknowingly, so far we've assumed that P(D) is the same for all documents
- Under this assumption, the ranking is based only on the query-likelihood given the document language model
- Now, we will assume that P(D) is not uniform
- That is, some documents are more likely to be relevant independent of the query

Document Prior

$$P(D|Q) \propto P(Q|D) \times P(D)$$

- What is it?
- Anything that affects the likelihood that a document is relevant to <u>any</u> query
  - document popularity
  - document authority
  - amount of content (e.g., length)
  - topical cohesion
  - really, you decide ...

Document Prior

$$P(D|Q) \propto P(Q|D) \times P(D)$$

- But, it is a probability, so in a collection of M documents...

$$\sum_{i=1}^{M} P(D_i) = ?$$


THE BEATLES
1

Document Prior

$$P(D|Q) \propto P(Q|D) \times P(D)$$

- Not that difficult...

$$P(D_j) = \frac{score(D_j)}{\sum_{i=1}^{M} score(D_i)}$$

Document Prior

$$P(D|Q) \propto P(Q|D) \times P(D)$$

- What is it?
- Anything that affects the likelihood that a document is relevant to <u>any</u> query
  - document popularity
  - document authority
  - amount of content (e.g., length)
  - topical cohesion
  - really, you decide ...

Document Popularity

- Given user-interaction data, we can determine the popularity of a document based on clicks
- Click-rate:

$$\frac{\text{\# of clicks on the document}}{\text{\# of clicks on any document}}$$

Document Popularity

most clicked urls - aol query-log (2006)

rank	URL	P(URL)	rank	URL	P(URL)
1	http://www.google.com	0.0204	11	http://www.geocities.com	0.0022
2	http://www.myspace.com	0.0093	12	http://www.hotmail.com	0.0022
3	http://mail.yahoo.com	0.0090	13	http://www.ask.com	0.0021
4	http://en.wikipedia.org	0.0066	14	http://www.bizrate.com	0.0017
5	http://www.amazon.com	0.0056	15	http://www.tripadvisor.com	0.0017
6	http://www.mapquest.com	0.0054	16	http://www.msn.com	0.0017
7	http://www.imdb.com	0.0053	17	http://profile.myspace.com	0.0016
8	http://www.ebay.com	0.0044	18	http://www.craigslist.org	0.0015
9	http://www.yahoo.com	0.0030	19	http://disney.go.com	0.0015
10	http://www.bankofamerica.com	0.0027	20	http://cgi.ebay.com	0.0015

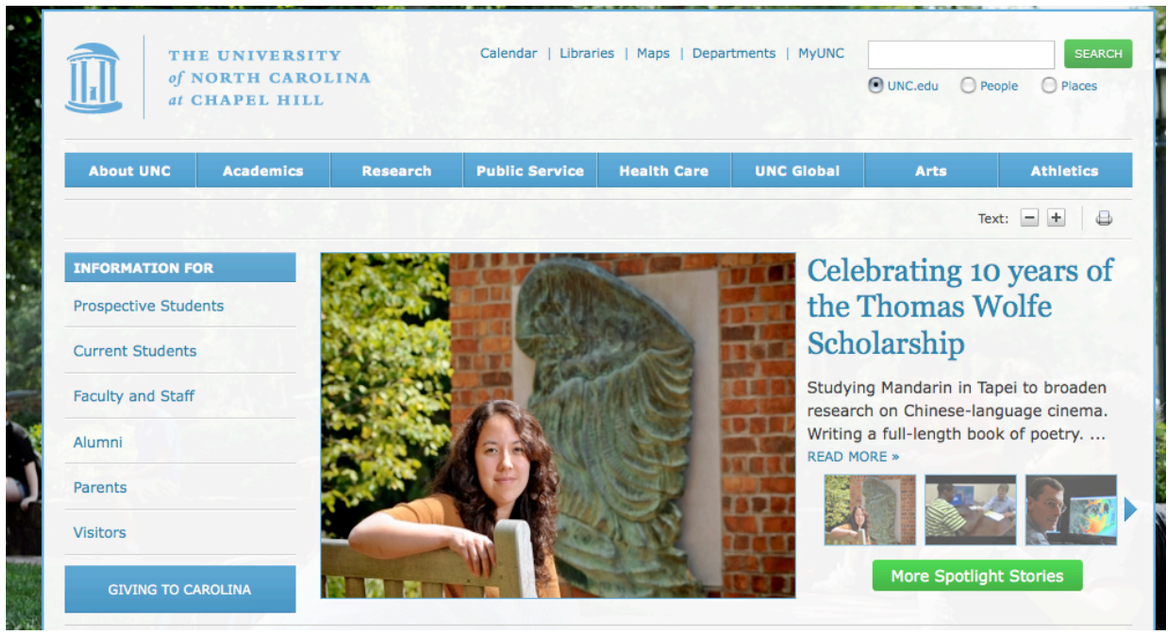
Document Popularity

least clicked urls - aol query-log (2006)

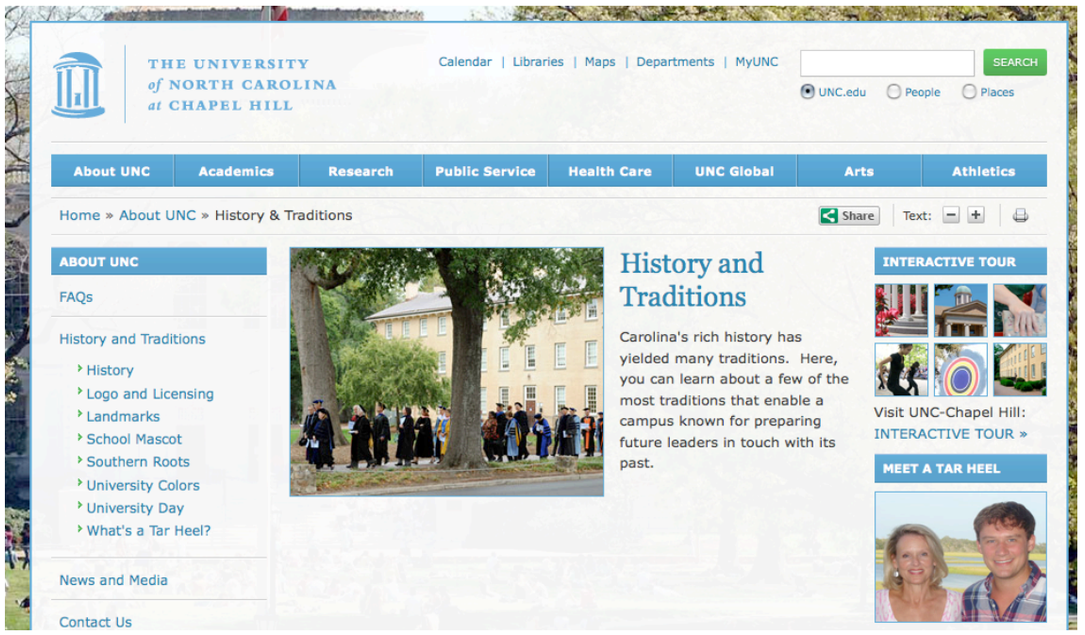
rank	URL	P(URL)	rank	URL	P(URL)
1501087	http://www.live4soccer.com	0.0000	1501097	http://www.toymod.com	0.0000
1501088	http://www.smalltowngallery.com	0.0000	1501098	http://www.aaabarcodes.com	0.0000
1501089	http://1239.8wmc5l.info	0.0000	1501099	http://www.stubaidirect.com	0.0000
1501090	http://silverjews.lyrics-online.net	0.0000	1501100	http://rtbknox.no-ip.biz	0.0000
1501091	http://www2.glenbrook.k12.il.us	0.0000	1501101	http://www.panontheweb.com	0.0000
1501092	http://www.palmerschools.org	0.0000	1501102	http://4395.bsxnf57.info	0.0000
1501093	http://www.rainbowridgefarmequestriancenter.com	0.0000	1501103	http://www.calco.com	0.0000
1501094	http://mncable.net	0.0000	1501104	http://www.sharpe.freshair.org	0.0000
1501095	http://www.modem-software.com	0.0000	1501105	http://www.opium.co.za	0.0000
1501096	http://www.clevelandrugby.com	0.0000	1501106	http://grediagnostic.ets.org	0.0000

Document Popularity

http://www.unc.edu



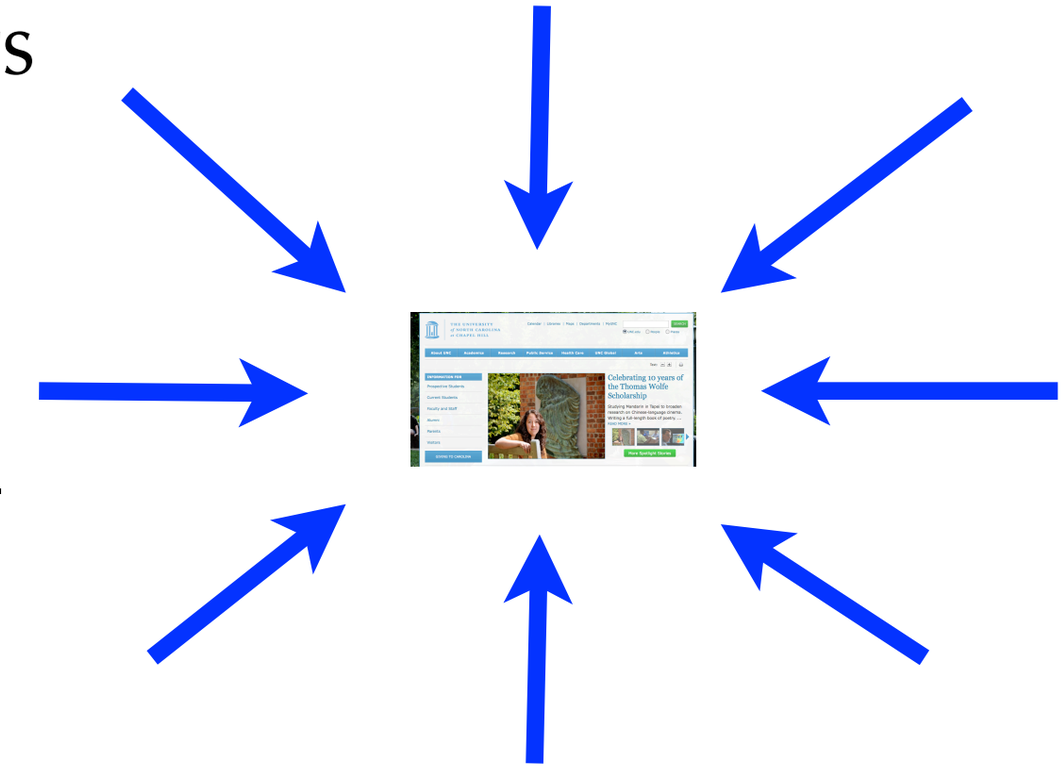
http://www.unc.edu/about/history-traditions



- URL depth
  - website entry-pages tend to be more popular than those that are deep within the domain
- Count the number of "/" in the URL

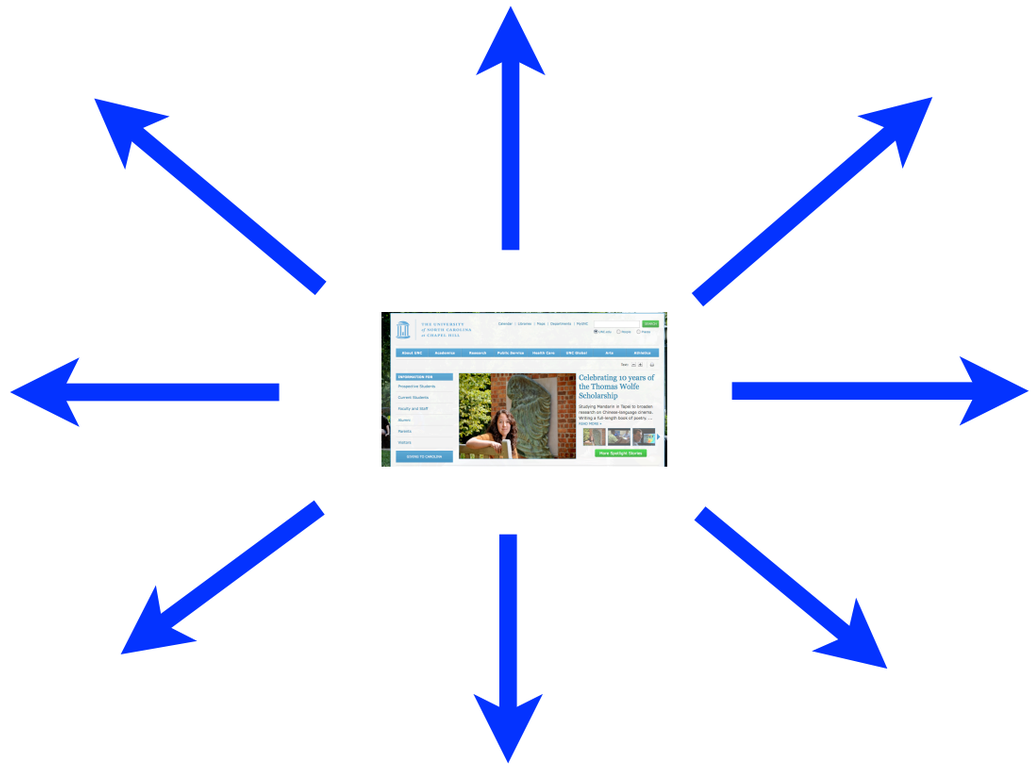
Document Authority

- Number of "endorsements"
  - scientific search: number of citations in other papers
  - web search: number of incoming hyperlinks
  - blog search: number user-generated comments
  - twitter search: number of followers
  - review search: number of times someone found the review useful



Document Authority

- "HUB" score
  - scientific search: number citations of other papers
  - web search: number of outgoing hyperlinks
  - blog search: number of links to other bloggers
  - twitter search: number of people followed by author
  - review search: number of reviews written by the reviewer

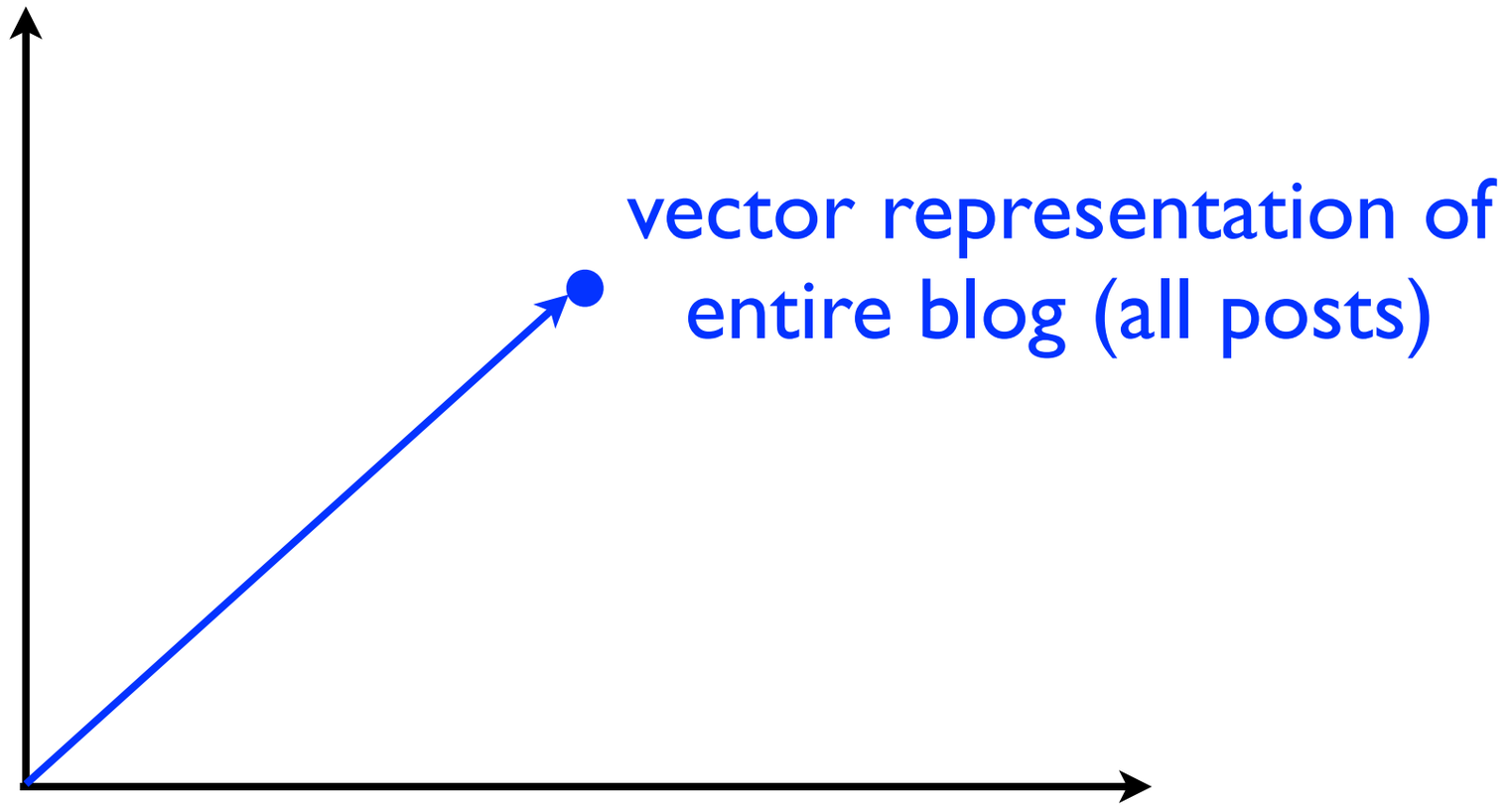


Topical Focus

- Example: blog retrieval
- Objective: favor blogs that focus on a coherent, recurring topic
- How might we do this? (HINT: vector space model)

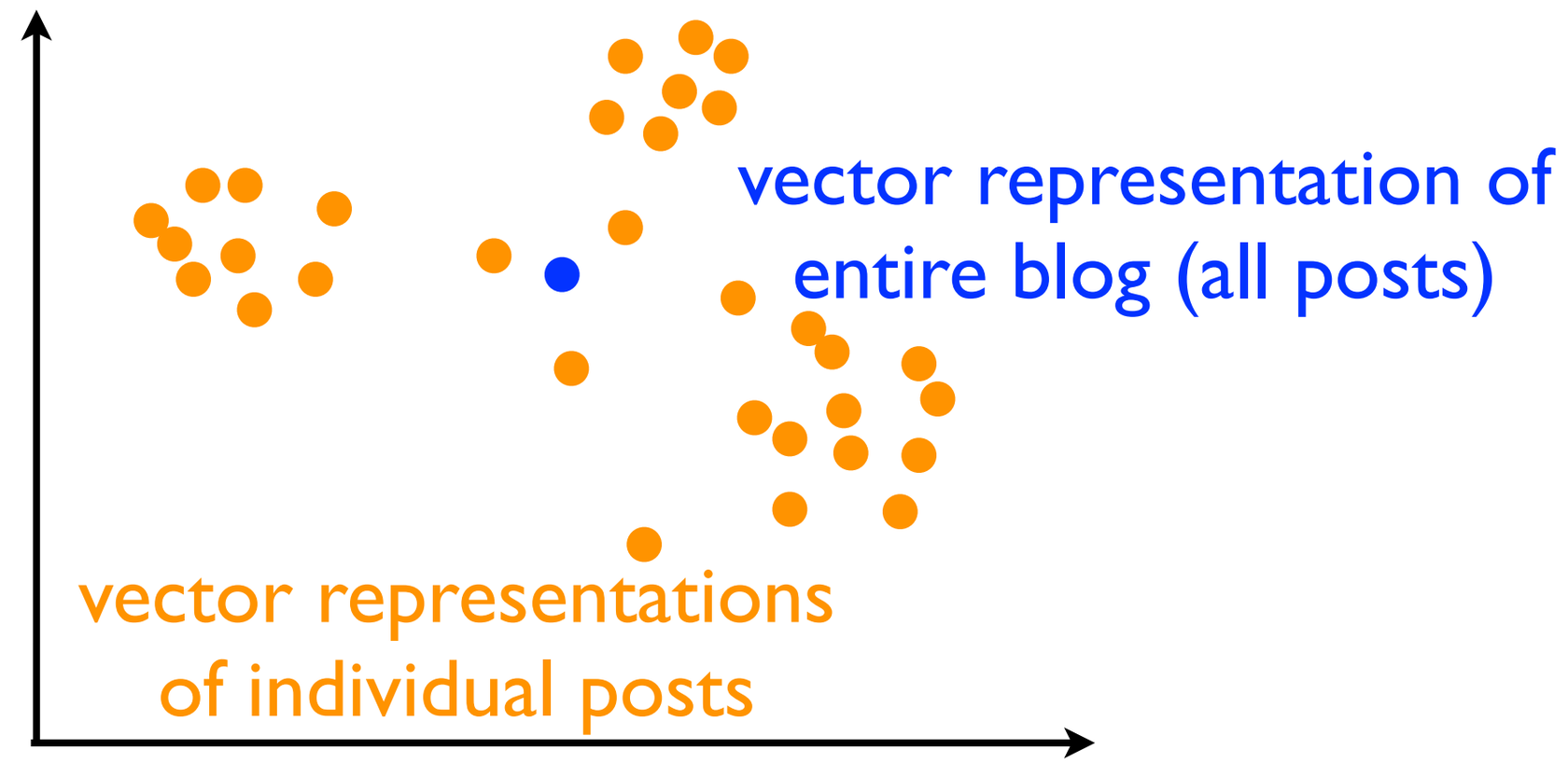
Topical Focus

- Example: blog retrieval
- Objective: favor blogs that focus on a coherent, recurring topic
- How might we do this? (HINT: vector space model)



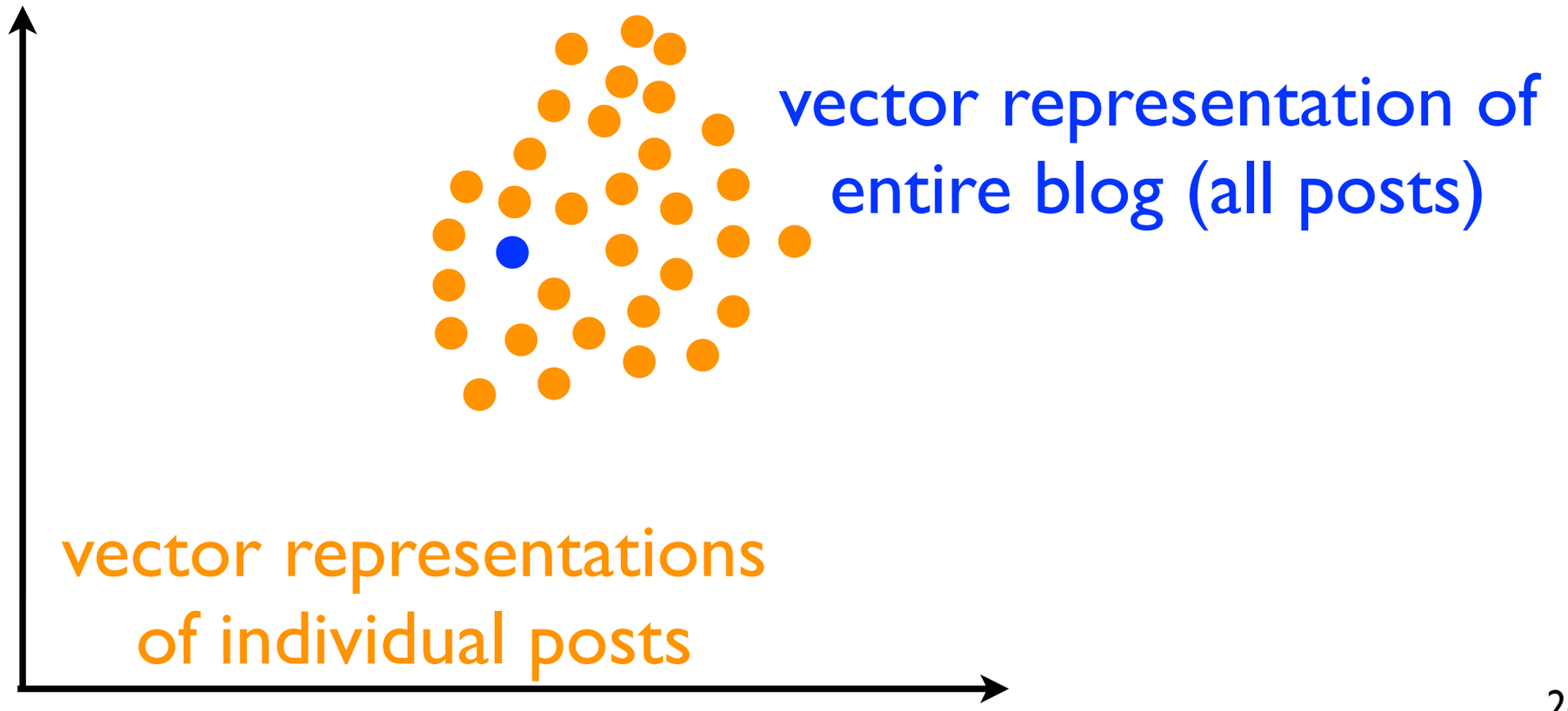
Topical Focus

- How might we do this? (HINT: vector space model)
- Compute <u>average</u> cosine similarity between the posts and the entire blog



Topical Focus

- How might we do this? (HINT: vector space model)
- Compute <u>average</u> cosine similarity between the posts and the entire blog



Document Prior

$$P(D|Q) \propto P(Q|D) \times P(D)$$

- What is it?
- Anything you want.
  - document popularity
  - document authority
  - amount of content (e.g., length)
  - topical focus
  - really, you decide

What document priors would you use?























Remember Smoothing?

- YOU: Are there mountain lions around here?
- YOUR FRIEND: Nope.
- YOU: How can you be so sure?
- YOUR FRIEND: Because I've been hiking here five times before and have never seen one.
- MOUNTAIN LION: You should have learned about smoothing by taking INLS 509. Yum!

Remember Smoothing?

- When estimating probabilities, we tend to ...
  - Over-estimate the probability of observed outcomes
  - Under-estimate the probability of unobserved outcomes
- The goal of smoothing is to ...
  - Decrease the probability of observed outcomes
  - Increase the probability of unobserved outcomes
- Smoothing P(D) is very important!

Example: Click-Rate

$$\frac{\text{\# of clicks on the document}}{\text{\# of clicks on any document}}$$

$$P(D|Q) \propto P(Q|D) \times P(D)$$

- Do we really want to always give documents that have never been clicked a score of zero?
- How could we smooth this probability?

Example: Click-Rate

$$\frac{\text{\# of clicks on the document}}{\text{\# of clicks on any document}}$$

$$P(D|Q) \propto P(Q|D) \times P(D)$$

- Do we really want to always give documents that have never been clicked a score of zero?
- Add-one smoothing!

$$\frac{\text{(\# of clicks on the document)} + 1}{\text{(\# of clicks on any document)} + \text{(\# of documents)}}$$

Outline

Introduction to language modeling

Language modeling for information retrieval

Query-likelihood Retrieval Model

Smoothing

Priors

Evaluation Overview

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

October 9, 2013

(Some slides courtesy of Fernando Diaz, Yahoo! Research)

Information Retrieval

- Given a query and a corpus, find relevant documents

  Query: user's textual description of their information need

  Corpus: a repository of textual documents

  Relevance: satisfaction of the user's information need

Information Retrieval

- In some disciplines, you can prove that solution A is better than B without experimentation.
- Is information retrieval one of these disciplines?
- If not, why not?

Uncertainty in Information Retrieval

Uncertainty in Information Retrieval

- The user may not know what they want
- Even if they do, they may not know how to describe it

Uncertainty in Information Retrieval

- soft surroundings
- trains interlocking dog sheets
- belly dancing music
- christian dior large bag
- best western airport sea tac
- www.bajawedding.com
- marie selby botanical gardens
- big chill down coats
- www.magichat.co.uk
- broadstone raquet club
- seadoo utopia
- seasons white plains condo
- priority club.com
- aircat tools
- epicurus evil
- instructions
- hinds county city of jackson
- last searches on aol a to z

(AOL query-log)

Uncertainty in Information Retrieval

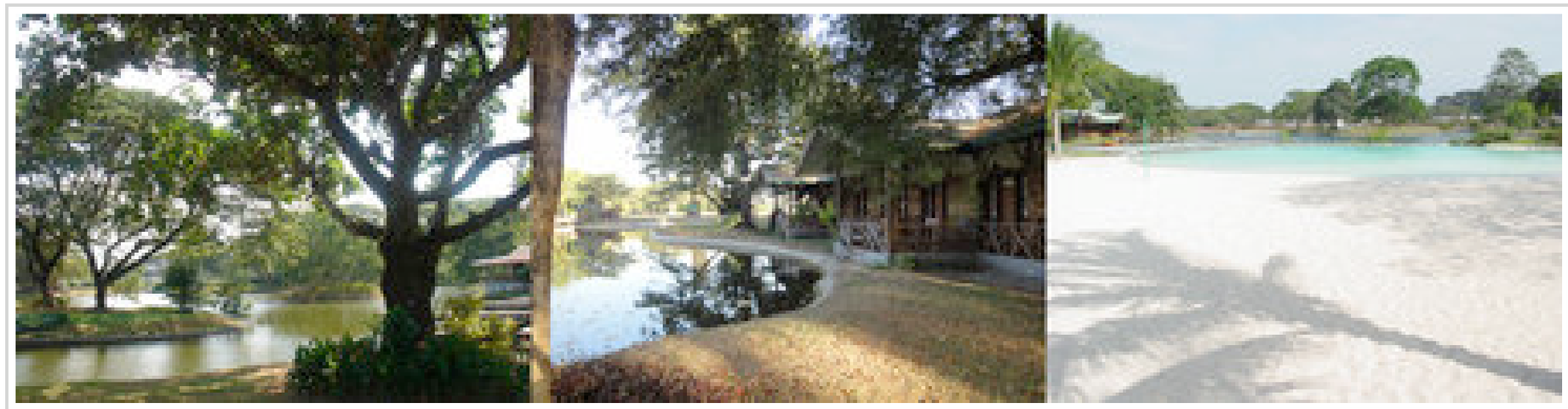
- Even if the user knows how to describe their information need, it will probably still be ambiguous
  - to the system
  - to a human trying to see if system A is better than B

Uncertainty in Information Retrieval

- Query: curbing population growth

Dengue in Philippines taken seriously at one hotel in Clark Philippines

Jul 18, 2010 23:58 EDT



Philippines government department of Health has issued warning of Dengue epidemic in July 2010 which is expected to last until the end of the year.

While most hotels and resorts are slow to take steps to address this health and safety issues, one hotel in Clark Philippines is already well on its way to implementing proactive measures to prevent the spread of Dengue by curbing population growth of mosquito.

Uncertainty in Information Retrieval

- Document relevance is subjective
  - many factors affect whether a document is relevant to a query
  - relevance is difficult to define in a way that generalizes across users, tasks, and environments

Uncertainty in Information Retrieval

- Query: cures for anatidaephobia

Anatidaephobia - The Fear That You are Being Watched by a Duck

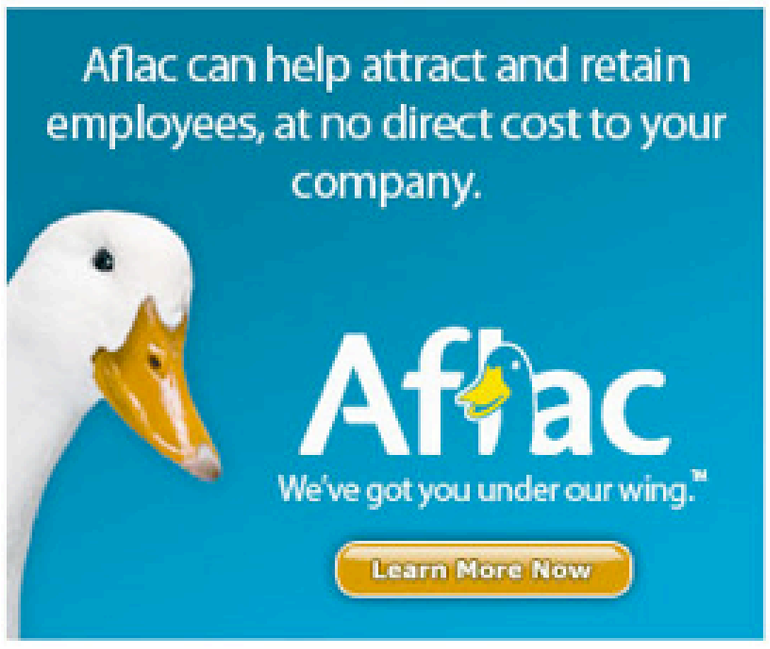
December 08, 2008 by **Tammy Duffey**

Single page Font Size Read comments (44) Share

What Is Anatidaephobia?

Anatidaephobia is defined as a pervasive, irrational fear that one is being watched by a duck. The anatidaephobic individual fears that no matter where they are or what they are doing, a duck watches.

Anatidaephobia is derived from the Greek word "anatidae", meaning ducks, geese or swans and "phobos" meaning fear.



What Causes Anatidaephobia?

As with all phobias, the person coping with Anatidaephobia has experienced a real-life trauma. For the anatidaephobic individual, this trauma most likely occurred during childhood.

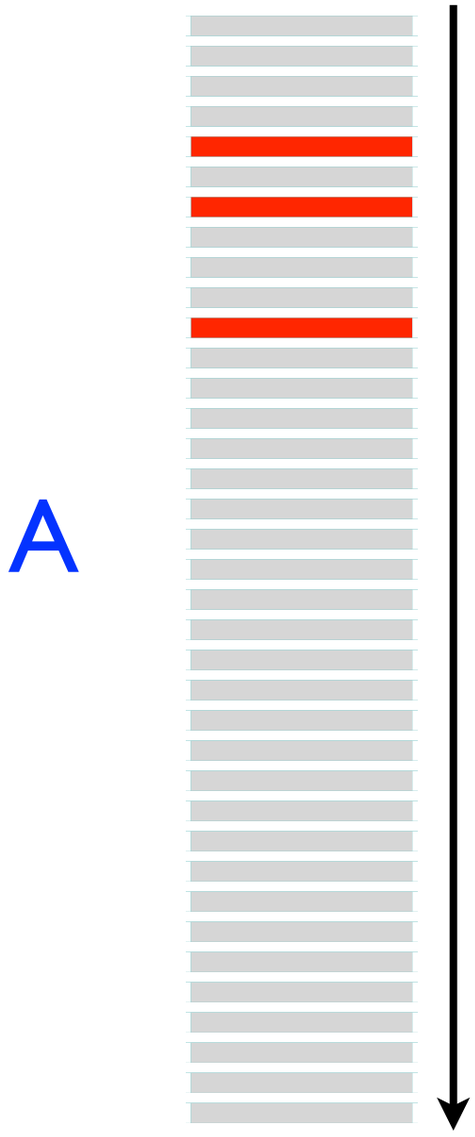
Perhaps the individual was intensely frightened by some species of water fowl. Geese and swans are relatively well known for their aggressive tendencies and perhaps the anatidaephobic person was actually bitten or flapped at. Of course, the Far Side comics did little to minimize the fear of being watched by a duck.

Uncertainty in Information Retrieval

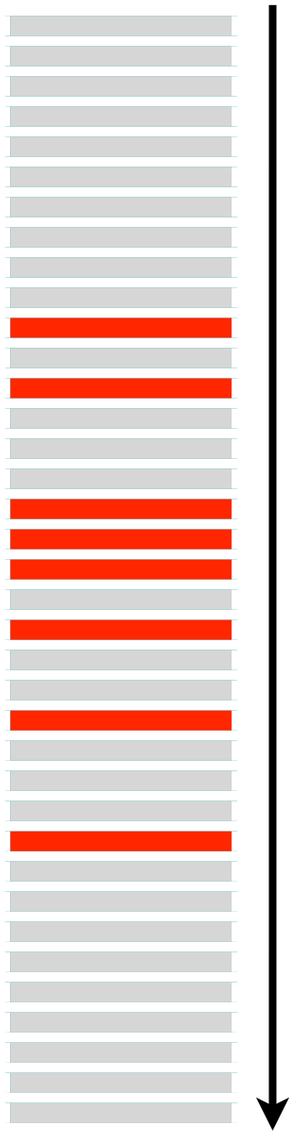
- Even if we know which documents are relevant, the preferred ranking may be difficult to determine without understanding the user population and their objectives

Uncertainty in Information Retrieval

- Which ranking is better?



A



B

Information Retrieval

- In some disciplines, you can prove that solution A is better than B without experimentation.
- Is information retrieval one of these disciplines?

- No, it is not. There is too much uncertainty
- For now, the only way to show that A is better than B, is through extensive experimentation

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

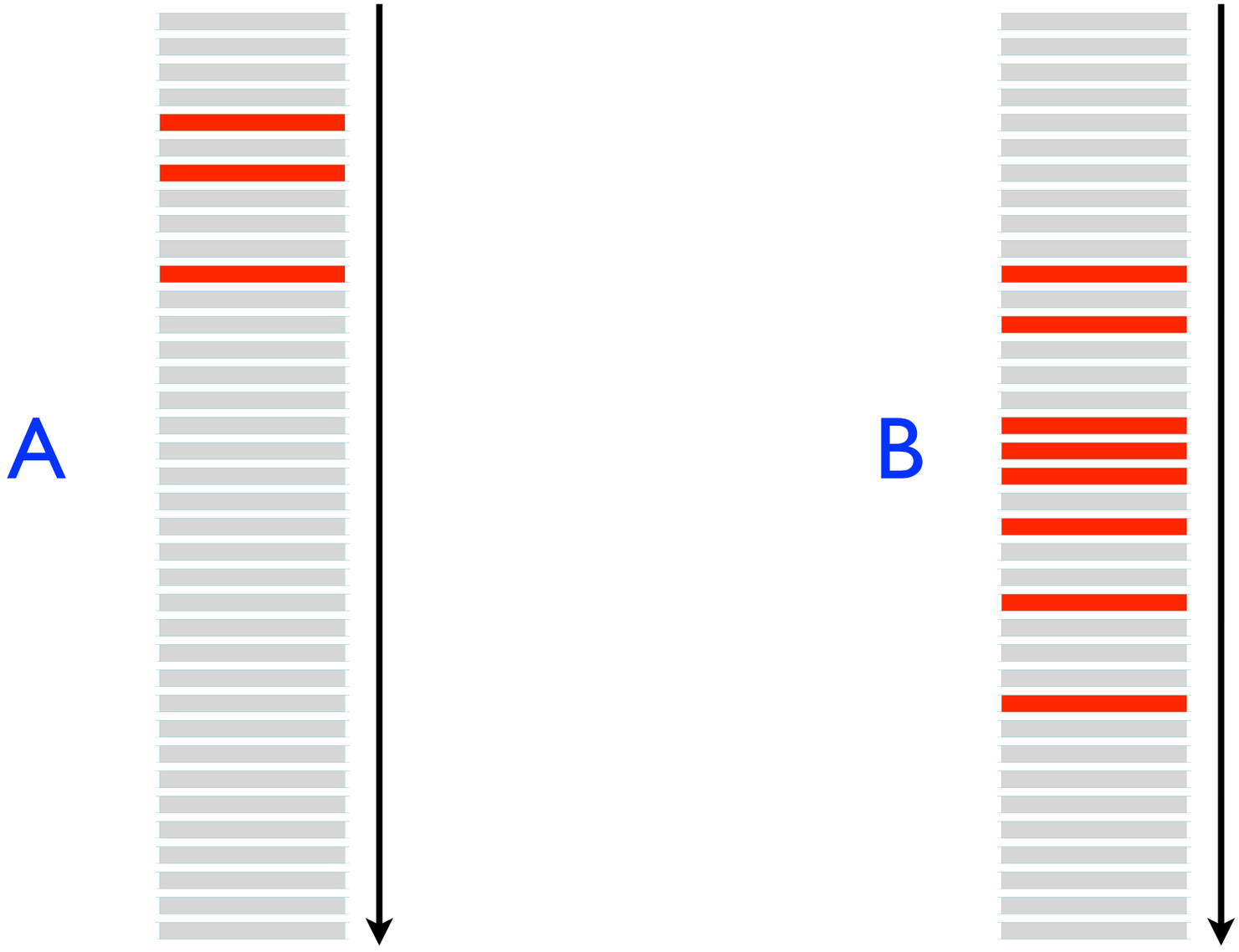
overview

- Using these descriptions, have human judges 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 measurement that quantifies the quality of a particular ranking of results with known relevant/non-relevant documents

Batch Evaluation

overview: metrics

- Which ranking is better?

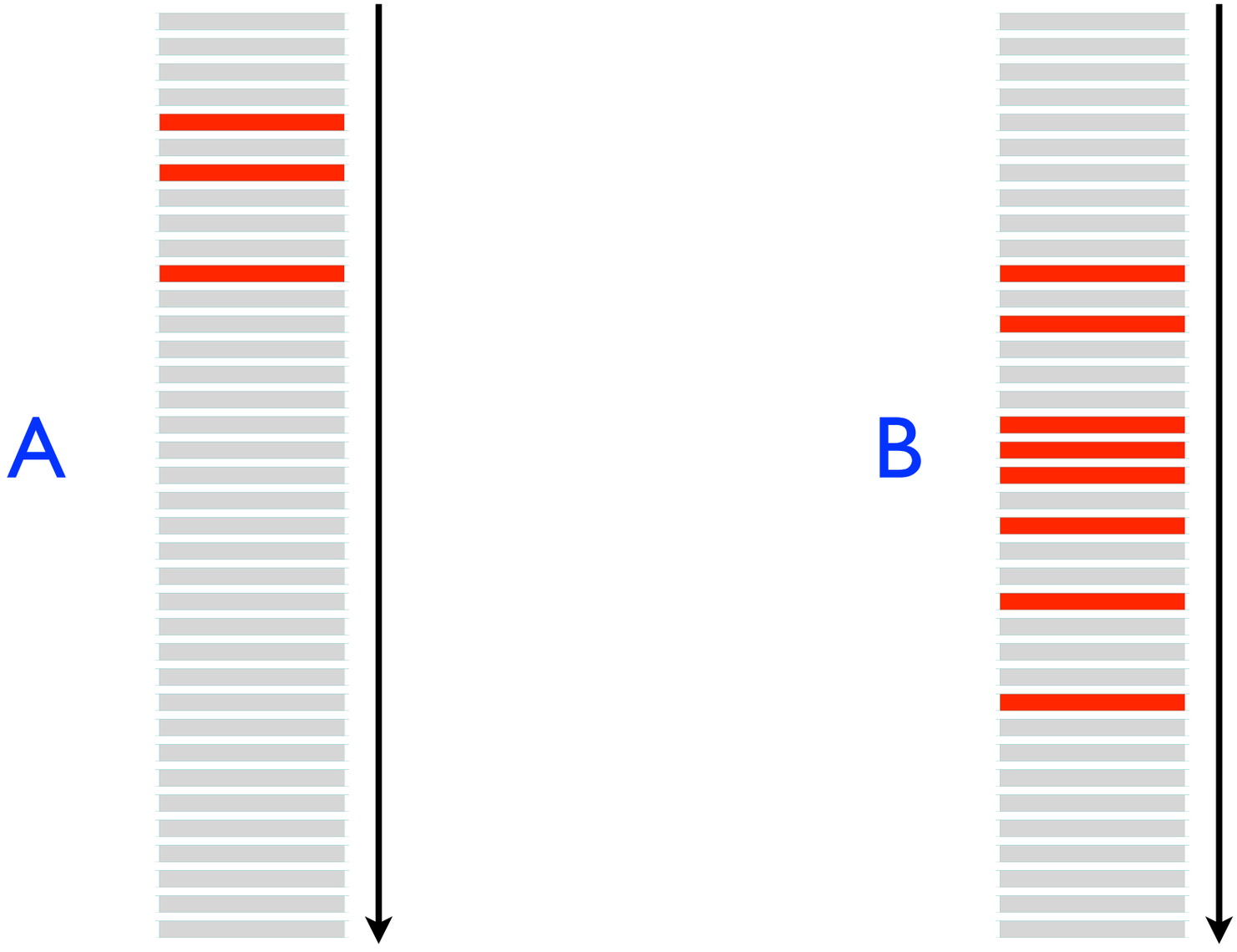


- rank of the first relevant document (lower value is better)

Batch Evaluation

overview: metrics

- Which ranking is better?

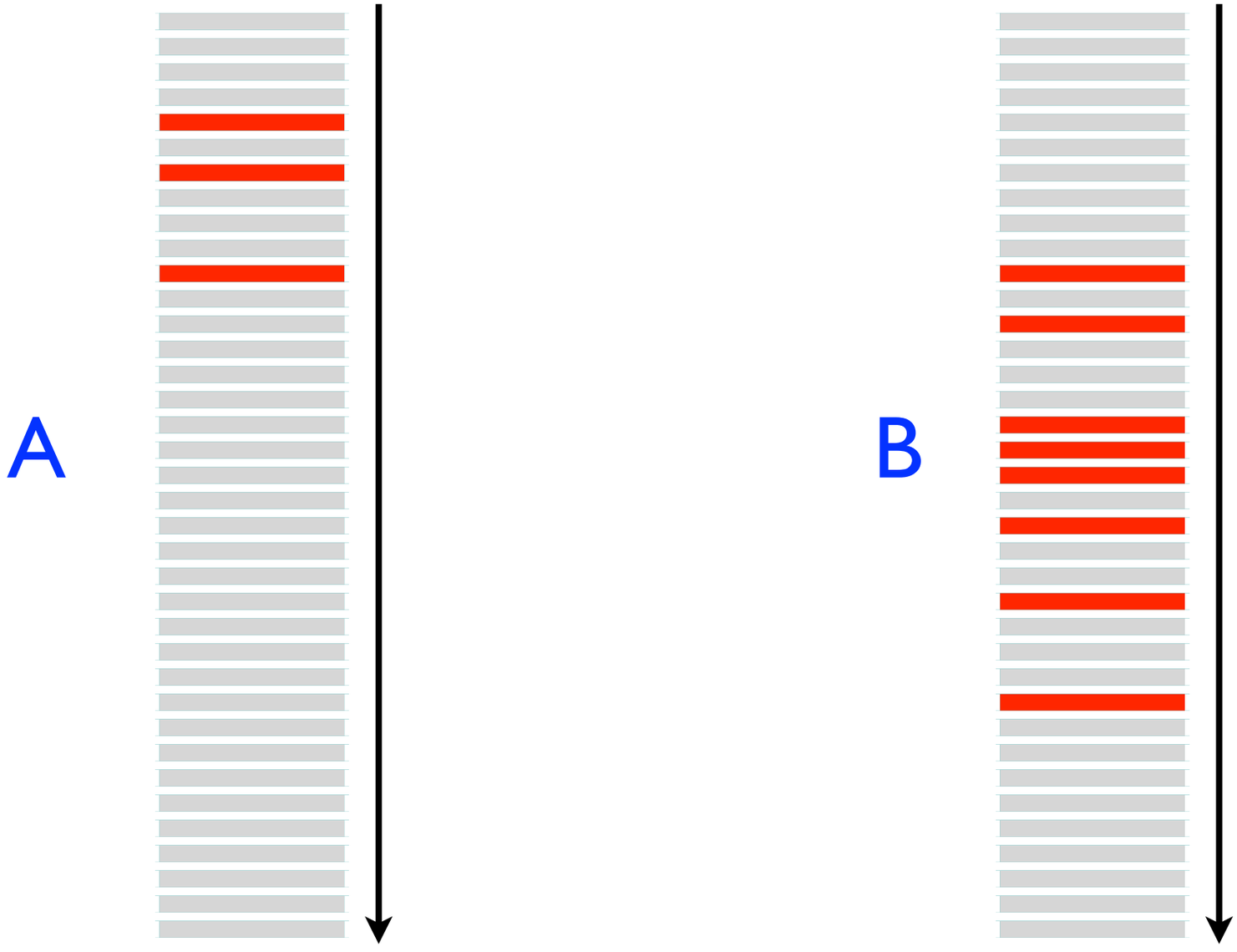


- precision at rank 10 (higher value is better)

Batch Evaluation

overview: metrics

- Which ranking is better?

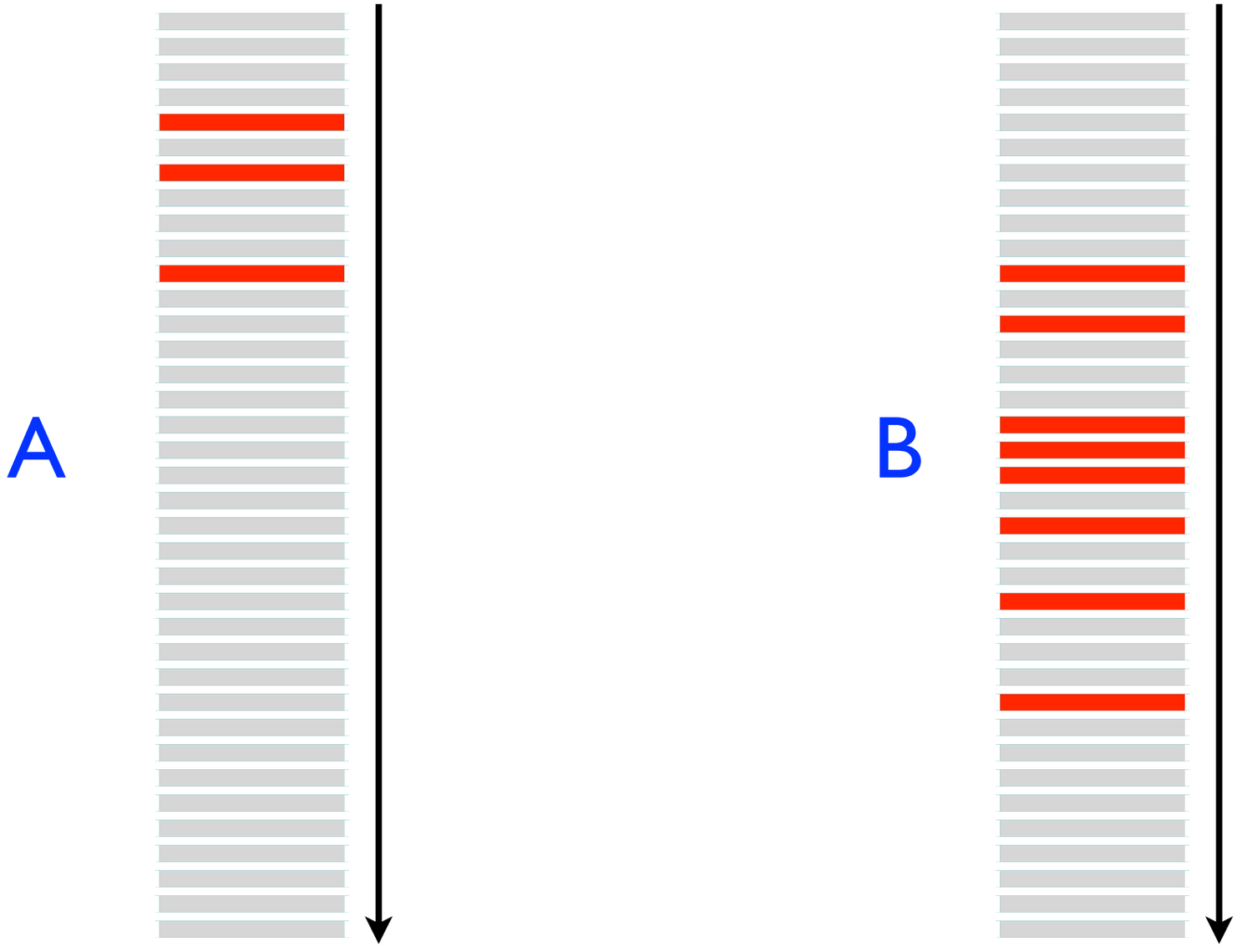


- precision at rank 1 (higher value is better)

Batch Evaluation

overview: metrics

- Which ranking is better?

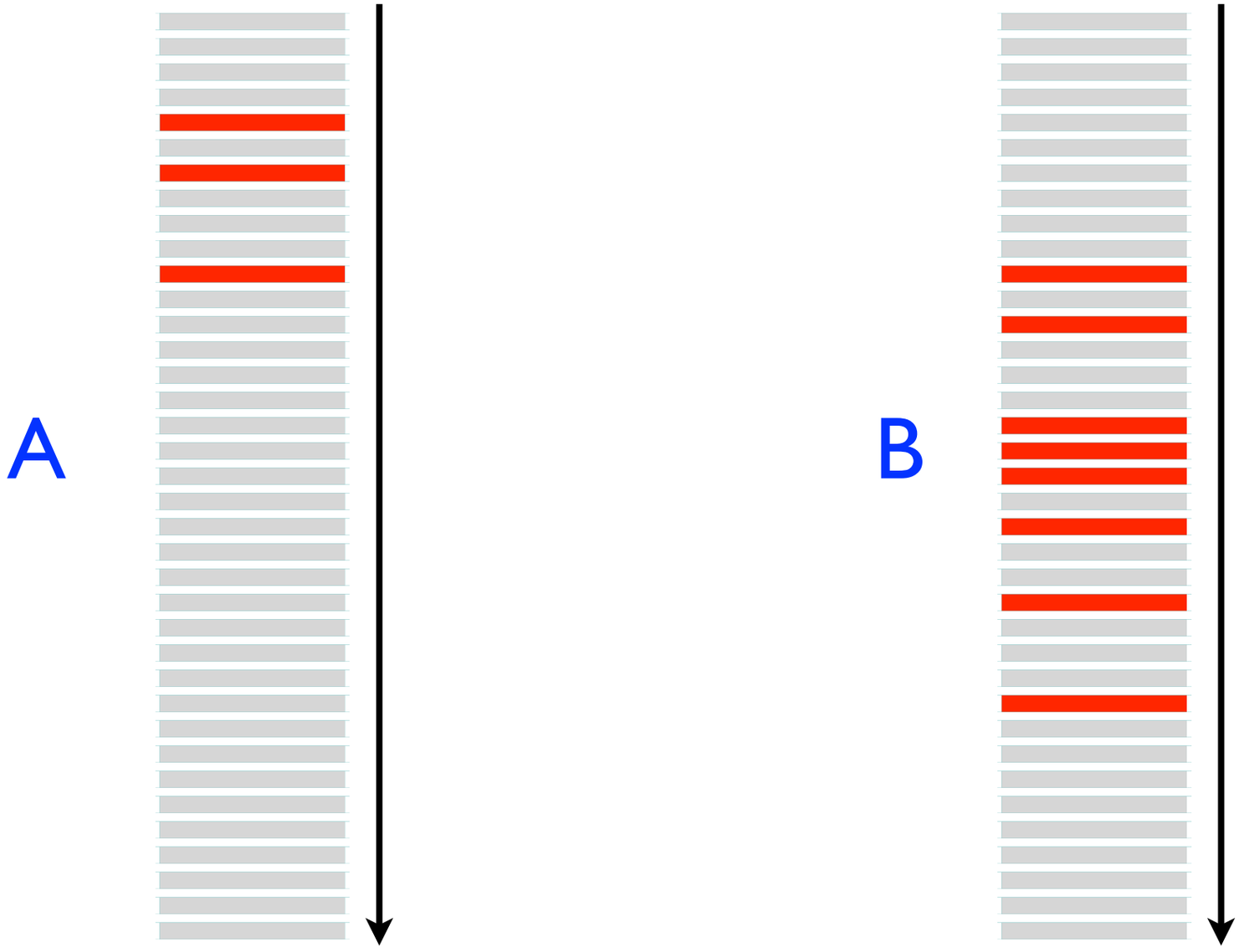


- recall at rank 10 (higher value is better)

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 better understand how system A is better than B

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, presence of a duck lurking in the shadows (of an insurance ad.), etc
- Topical relevance: the document is on the same topic as the query
- User relevance: everything else (quack!)
- 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, presence of a duck lurking in the shadows (of an insurance ad.), etc
- Topical relevance: the document is on the same topic as the query
- User relevance: everything else (quack!)
- Which kind of relevance does batch-evaluation address?
- Whether the document contains the sought-after information

User-Study Evaluation
overview

- Provide a small set of users with several retrieval systems
- Ask them to complete several (potentially different) search tasks

- Learn about system performance by:
  - observing what they do
  - asking why they do what they do

User-Study Evaluation
overview: trade-offs

- Advantages:
  - very detailed data about users' reaction to systems
  - in reality, a search is done to accomplish a higher-level task
  - in user studies, this task can be manipulated and studied
  - in other words, the experimental 'starting-point' need not be the query

User-Study Evaluation
overview: trade-offs

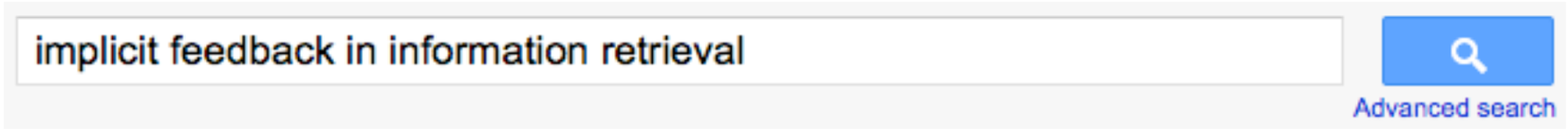
- Disadvantages:
  - user studies are expensive (pay users/subjects, scientist's time, data coding)
  - difficult to generalize from small studies to broad populations
  - the laboratory setting is not the user's normal environment
  - need to re-run experiment every time a new system is considered

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 system effects on logged user interactions (implicit feedback)
  - clicks: surrogates for perceived relevance (good)
  - skips: surrogates for perceived non-relevance (bad)

Implicit Feedback



[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!

can we say that the first result is more relevant than the second?

Implicit Feedback

implicit feedback in information retrieval

Advanced search

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 ...

skip!

[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 ...

click!

[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 ...

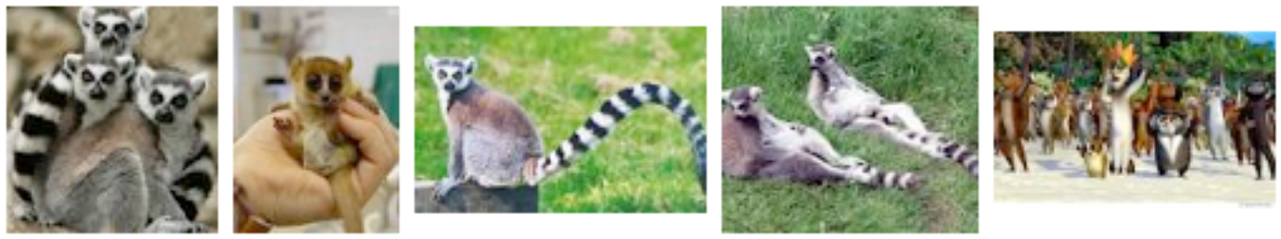
can we say that the second result is more relevant than the first?

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 ...

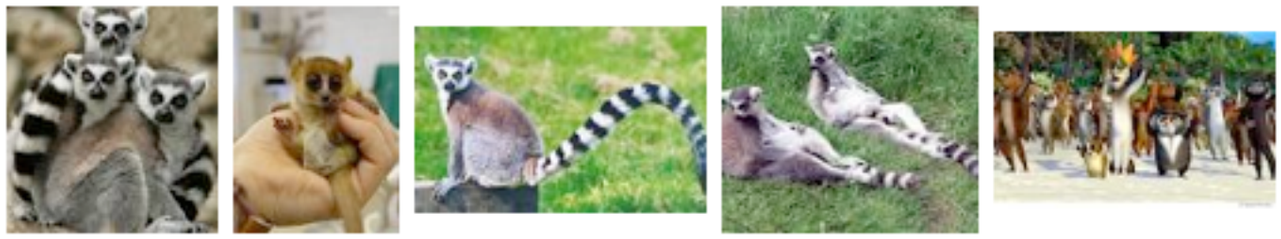
a click is a noisy surrogate for relevance!

Implicit Feedback

capital of honduras

Advanced search

Images for **lemurs** - Report images



[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** ...

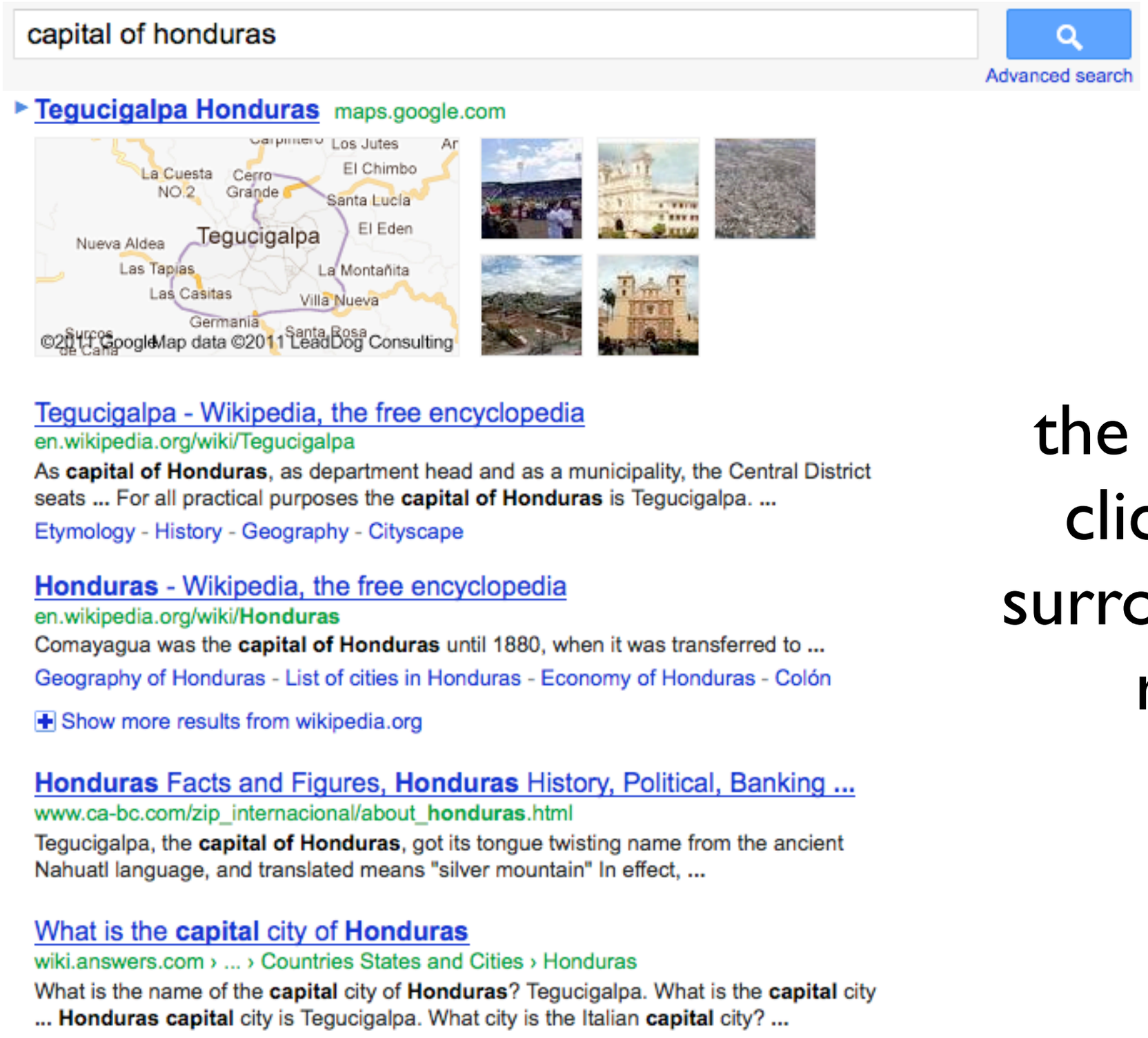
Context-Sensitive **Information Retrieval** Using **Implicit Feedback**
citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.61.987
by X Shen - 2005 - Cited by 217 - Related articles
CiteSeerX - Document Details (Isaac Councill, Lee Giles): A major limitation ...

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 ...

user sees the results and closes the browser

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 predict 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