

# INLS 509: Introduction to Information Retrieval

Jaime Arguello  
[jarguell@email.unc.edu](mailto:jarguell@email.unc.edu)

January 8, 2014

# Introductions

- Hello, my name is \_\_\_\_\_.
- However, I'd rather be called \_\_\_\_\_. (optional)
- I'm in the \_\_\_\_\_ program.
- I'm taking this course because I want to \_\_\_\_\_.



# What is Information Retrieval?

- Information retrieval (IR) is the science and practice of developing and evaluating systems that match information seekers with the information they seek.

# What is Information Retrieval?

- This course mainly focuses on search engines
- Given a **query** and a **corpus**, find **relevant** items

**query:** a user's expression of their information need

**corpus:** a repository of retrievable items

**relevance:** satisfaction of the user's information need

# What is Information Retrieval?

- Gerard Salton, 1968:

Information retrieval is a field concerned with the structure, analysis, organization, storage, and retrieval of information.

# Information Retrieval structure

 [Log in / create account](#)



**WIKIPEDIA**  
The Free Encyclopedia

[Main page](#)  
[Contents](#)  
[Featured content](#)  
[Current events](#)  
[Random article](#)  
[Donate to Wikipedia](#)

▼ [Interaction](#)  
[Help](#)  
[About Wikipedia](#)  
[Community portal](#)  
[Recent changes](#)  
[Contact Wikipedia](#)

► [Toolbox](#)  
► [Print/export](#)

▼ [Languages](#)  
[Deutsch](#)  
[Español](#)  
[Bahasa Indonesia](#)

[Article](#) [Discussion](#)

[Read](#) [Edit](#) [View history](#)



## Gerard Salton

From Wikipedia, the free encyclopedia

**Gerard Salton** (8 March 1927 in [Nuremberg](#) - 28 August 1995), also known as Gerry Salton, was a Professor of [Computer Science](#) at [Cornell University](#). Salton was perhaps the leading computer scientist working in the field of [information retrieval](#) during his time. His group at Cornell developed the [SMART Information Retrieval System](#), which he initiated when he was at Harvard.

Salton was born Gerhard Anton Sahlmann on March 8, 1927 in [Nuremberg, Germany](#). He received a Bachelor's (1950) and Master's (1952) degree in mathematics from [Brooklyn College](#), and a Ph.D. from [Harvard](#) in [Applied Mathematics](#) in 1958, the last of [Howard Aiken](#)'s doctoral students, and taught there until 1965, when he joined [Cornell University](#) and co-founded its department of Computer Science.

Salton was perhaps most well known for developing the now widely used [Vector Space Model](#) for Information Retrieval<sup>[1]</sup>. In this model, both documents and queries are represented as vectors of term counts, and the similarity between a document and a query is given by the cosine between the term vector and the document vector. In this paper, he also introduced [TF-IDF](#), or term-frequency-inverse-document frequency, a model in which the score of a term in the a document is the ratio of the number of terms in that document divided by the frequency of the number of documents in which that term occurs. (The concept of inverse document frequency, a measure of specificity, had been introduced in 1972 by [Karen Sparck-Jones](#)<sup>[2]</sup>.) Later in life, he became interested in automatic text summarization and analysis<sup>[3]</sup>, as well as automatic hypertext generation<sup>[4]</sup>. He published over 150 research articles and 5 books during his life.

Salton was editor-in-chief of the [Communications of the ACM](#) and the [Journal of the ACM](#), and chaired [SIGIR](#). He was an associate editor of the [ACM Transactions on Information Systems](#). He was an [ACM Fellow](#) (elected 1995), received an Award of Merit from the [American Society for Information Science](#) (1989), and was the first recipient of the [SIGIR Award](#) for outstanding contributions to study of information retrieval (1983) -- now called the [Gerard Salton Award](#).

## References

[\[edit\]](#)

- <sup>↑</sup> [G. Salton , A. Wong , C. S. Yang, A vector space model for automatic indexing](#) , *Communications of the ACM*, v.18 n.11, p.613-620, Nov. 1975
- <sup>↑</sup> [Spärck Jones, Karen \(1972\), "A statistical interpretation of term specificity and its application in retrieval" !\[\]\(e47d2433872e9ca95cd4b9b9a68aa590\_img.jpg\), \*Journal of Documentation\* \*\*28\*\* \(1\): 11–21, doi:10.1108/eb026526](#) 

# Information Retrieval

## document structure

 [Log in / create account](#)



WIKIPEDIA  
The Free Encyclopedia

Article

Discussion

Read

Edit

View history



## Gerard Salton

From Wikipedia, the free encyclopedia

[Main page](#)

[Contents](#)

[Featured content](#)

[Current events](#)

[Random article](#)

[Donate to Wikipedia](#)

▼ [Interaction](#)

[Help](#)

[About Wikipedia](#)

[Community portal](#)

[Recent changes](#)

[Contact Wikipedia](#)

► [Toolbox](#)

► [Print/export](#)

▼ [Languages](#)

[Deutsch](#)

[Español](#)

[Bahasa Indonesia](#)

**Gerard Salton** (8 March 1927 in [Nuremberg](#) - 28 August 1995), also known as Gerry Salton, was a Professor of [Computer Science](#) at [Cornell University](#). Salton was perhaps the leading computer scientist working in the field of [information retrieval](#) during his time. His group at Cornell developed the [SMART Information Retrieval System](#), which he initiated when he was at Harvard.

Salton was born Gerhard Anton Sahlmann on March 8, 1927 in [Nuremberg, Germany](#). He received a Bachelor's (1950) and Master's (1952) degree in mathematics from [Brooklyn College](#), and a Ph.D. from [Harvard](#) in [Applied Mathematics](#) in 1958, the last of [Howard Aiken](#)'s doctoral students, and taught there until 1965, when he joined [Cornell University](#) and co-founded its department of Computer Science.

Salton was perhaps most well known for developing the now widely used [Vector Space Model](#) for Information Retrieval<sup>[1]</sup>. In this model, both documents and queries are represented as vectors of term counts, and the similarity between a document and a query is given by the cosine between the term vector and the document vector. In this paper, he also introduced [TF-IDF](#), or term-frequency-inverse-document frequency, a model in which the score of a term in the a document is the ratio of the number of terms in that document divided by the frequency of the number of documents in which that term occurs. (The concept of inverse document frequency, a measure of specificity, had been introduced in 1972 by [Karen Sparck-Jones](#)<sup>[2]</sup>.) Later in life, he became interested in automatic text summarization and analysis<sup>[3]</sup>, as well as automatic hypertext generation<sup>[4]</sup>. He published over 150 research articles and 5 books during his life.

Salton was editor-in-chief of the [Communications of the ACM](#) and the [Journal of the ACM](#), and chaired [SIGIR](#). He was an associate editor of the [ACM Transactions on Information Systems](#). He was an [ACM Fellow](#) (elected 1995), received an Award of Merit from the [American Society for Information Science](#) (1989), and was the first recipient of the [SIGIR Award](#) for outstanding contributions to study of information retrieval (1983) -- now called the [Gerard Salton Award](#).

## References

[\[edit\]](#)

- <sup>↑</sup> G. Salton , A. Wong , C. S. Yang, A vector space model for automatic indexing , Communications of the ACM, v.18 n.11, p.613-620, Nov. 1975
- <sup>↑</sup> Spärck Jones, Karen (1972), "A statistical interpretation of term specificity and its application in retrieval" , *Journal of Documentation* **28** (1): 11–21, doi:10.1108/eb026526 

# Information Retrieval

## document structure

 Log in / create account



WIKIPEDIA  
The Free Encyclopedia

Article

Discussion

Read

Edit

View history



## Gerard Salton

From Wikipedia, the free encyclopedia

**Gerard Salton** (8 March 1927 in [Nuremberg](#) - 28 August 1995), also known as [Gerry Salton](#), was a Professor of [Information Retrieval](#) working in the field of [Information Retrieval System](#), which he [described](#) as a [Batch Retrieval System](#). He received a Bachelor's (1950) in [Applied Mathematics](#) in [Cornell University](#) and a [Ph.D.](#) in [Mathematics](#) from [Cornell University](#) and [Cornell University](#) for [Information Retrieval](#)<sup>[1]</sup>. He also published a paper on the similarity between a [term](#) and a [document](#). In this paper, he also introduced [TF-IDF](#), or [term-frequency-inverse-document-frequency](#), a model in which the score of a term in the a document is the ratio of the number of terms in that document divided by the frequency of the number of documents in which that term occurs. (The concept of inverse document frequency, a measure of specificity, had been introduced in 1972 by [Karen Sparck-Jones](#)<sup>[2]</sup>.) Later in life, he became interested in automatic text summarization and analysis<sup>[3]</sup>, as well as automatic hypertext generation<sup>[4]</sup>. He published over 150 research articles and 5 books during his life. Salton was editor-in-chief of the [Communications of the ACM](#) and the [Journal of the ACM](#), and chaired [SIGIR](#). He was an associate editor of the [ACM Transactions on Information Systems](#). He was an [ACM Fellow](#) (elected 1995), received an [Award of Merit](#) from the [American Society for Information Science](#) (1989), and was the first recipient of the [SIGIR Award](#) for outstanding contributions to study of information retrieval (1983) -- now called the [Gerard Salton Award](#).

## References

[\[edit\]](#)

- <sup>↑</sup> G. Salton , A. Wong , C. S. Yang, A vector space model for automatic indexing [\[1\]](#), *Communications of the ACM*, v.18 n.11, p.613-620, Nov. 1975
- <sup>↑</sup> Spärck Jones, Karen (1972), "A statistical interpretation of term specificity and its application in retrieval" [\[2\]](#), *Journal of Documentation* 28 (1): 11–21, doi:10.1108/eb026526 [\[3\]](#)

However, the main content of the page is in the form of natural language text, which has little structure that a computer can understand

[Contact Wikipedia](#)

► [Toolbox](#)

► [Print/export](#)

▼ [Languages](#)

[Deutsch](#)

[Español](#)

[Bahasa Indonesia](#)

# Information Retrieval

## document structure

 [Log in / create account](#)



WIKIPEDIA  
The Free Encyclopedia

Article Discussion

Read Edit View history



## Gerard Salton

From Wikipedia, the free encyclopedia

**Gerard Salton** (8 March 1927 in [Nuremberg](#) - 28 August 1995), also known as [Gerry Salton](#), was a Professor of [Information Science](#) working in the field of [Information Retrieval System](#), which he [co-developed](#). He received a Bachelor's (1950) in [Applied Mathematics](#) in [Cornell University](#) and

However, the main content of the page is in the form of natural language text,

W

As it turns out, it's not necessary for a computer to understand natural language text for it to determine that this document is likely to be relevant to a particular query (e.g., "Gerard Salton")

[Contact Wikipedia](#)

► [Toolbox](#)

► [Print/export](#)

▼ [Languages](#)

[Deutsch](#)

[Español](#)

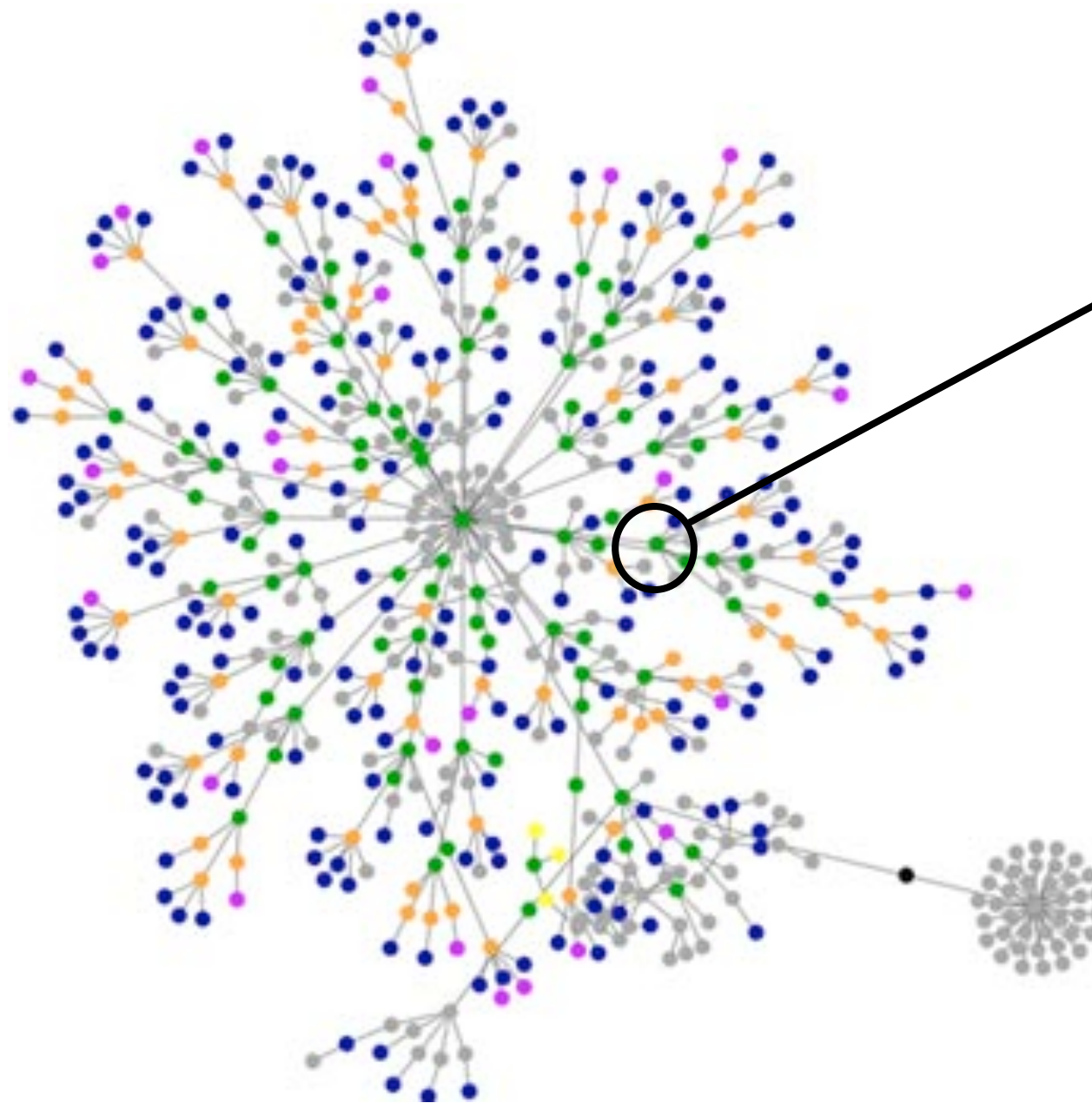
[Bahasa Indonesia](#)

## References

[\[edit\]](#)

- <sup>↑</sup> [G. Salton , A. Wong , C. S. Yang, A vector space model for automatic indexing](#) , *Communications of the ACM*, v.18 n.11, p.613-620, Nov. 1975
- <sup>↑</sup> [Spärck Jones, Karen \(1972\), "A statistical interpretation of term specificity and its application in retrieval"](#) , *Journal of Documentation* **28** (1): 11–21, doi:10.1108/eb026526 

# Information Retrieval collection structure



# Information Retrieval

## analysis: classification and information extraction



The screenshot shows the Wikipedia article for Gerard Salton. The page layout includes a sidebar on the left with navigation links like 'Main page', 'Contents', and 'Interaction'. The main content area has tabs for 'Article' and 'Discussion', and buttons for 'Read', 'Edit', and 'View history'. The article text describes Salton's life and his contributions to information retrieval, specifically mentioning the Vector Space Model and TF-IDF. A 'References' section at the bottom lists two sources.

**Gerard Salton**

From Wikipedia, the free encyclopedia

**Gerard Salton** (8 March 1927 in [Nuremberg](#) - 28 August 1995), also known as Gerry Salton, was a Professor of Computer Science at [Cornell University](#). Salton was perhaps the leading computer scientist working in the field of [information retrieval](#) during his time. His group at Cornell developed the [SMART Information Retrieval System](#), which he initiated when he was at Harvard.

Salton was born Gerhard Anton Sahlmann on March 8, 1927 in [Nuremberg, Germany](#). He received a Bachelor's (1950) and Master's (1952) degree in mathematics from [Brooklyn College](#), and a Ph.D. from [Harvard](#) in [Applied Mathematics](#) in 1958, the last of [Howard Aiken](#)'s doctoral students, and taught there until 1965, when he joined [Cornell University](#) and co-founded its department of Computer Science.

Salton was perhaps most well known for developing the now widely used [Vector Space Model](#) for Information Retrieval<sup>[1]</sup>. In this model, both documents and queries are represented as vectors of term counts, and the similarity between a document and a query is given by the cosine between the term vector and the document vector. In this paper, he also introduced [TF-IDF](#), or term-frequency-inverse-document frequency, a model in which the score of a term in the a document is the ratio of the number of terms in that document divided by the frequency of the number of documents in which that term occurs. (The concept of inverse document frequency, a measure of specificity, had been introduced in 1972 by [Karen Sparck-Jones](#)<sup>[2]</sup>.) Later in life, he became interested in automatic text summarization and analysis<sup>[3]</sup>, as well as automatic hypertext generation<sup>[4]</sup>. He published over 150 research articles and 5 books during his life.

Salton was editor-in-chief of the [Communications of the ACM](#) and the [Journal of the ACM](#), and chaired [SIGIR](#). He was an associate editor of the [ACM Transactions on Information Systems](#). He was an [ACM Fellow](#) (elected 1995), received an Award of Merit from the [American Society for Information Science](#) (1989), and was the first recipient of the [SIGIR Award](#) for outstanding contributions to study of information retrieval (1983) -- now called the [Gerard Salton Award](#).

**References** [[edit](#)]

- <sup>↑</sup> G. Salton , A. Wong , C. S. Yang, A vector space model for automatic indexing [g](#), *Communications of the ACM*, v.18 n.11, p.613-620, Nov. 1975
- <sup>↑</sup> Spärck Jones, Karen (1972), "A statistical interpretation of term specificity and its application in retrieval" [j](#), *Journal of Documentation* **28** (1): 11–21, doi:10.1108/eb026526 [g](#)

Categories: 1927 births | 1995 deaths | American computer scientists | Computer pioneers | Harvard University alumni | Harvard University faculty | Cornell University faculty | Fellows of the Association for Computing Machinery | Guggenheim Fellows

# Information Retrieval

## organization: cataloguing

 open directory project

In partnership with  
**Aol Search.**

[about dmoz](#) | [dmoz blog](#) | [suggest URL](#) | [help](#) | [link](#) | [editor login](#)

Search [advanced](#)

**Arts**  
[Movies](#), [Television](#), [Music](#)...

**Business**  
[Jobs](#), [Real Estate](#), [Investing](#)...

**Computers**  
[Internet](#), [Software](#), [Hardware](#)...

**Games**  
[Video Games](#), [RPGs](#), [Gambling](#)...

**Health**  
[Fitness](#), [Medicine](#), [Alternative](#)...

**Home**  
[Family](#), [Consumers](#), [Cooking](#)...

**Kids and Teens**  
[Arts](#), [School Time](#), [Teen Life](#)...

**News**  
[Media](#), [Newspapers](#), [Weather](#)...

**Recreation**  
[Travel](#), [Food](#), [Outdoors](#), [Humor](#)...

**Reference**  
[Maps](#), [Education](#), [Libraries](#)...

**Regional**  
[US](#), [Canada](#), [UK](#), [Europe](#)...

**Science**  
[Biology](#), [Psychology](#), [Physics](#)...

**Shopping**  
[Clothing](#), [Food](#), [Gifts](#)...

**Society**  
[People](#), [Religion](#), [Issues](#)...

**Sports**  
[Baseball](#), [Soccer](#), [Basketball](#)...

**World**  
[Català](#), [Dansk](#), [Deutsch](#), [Español](#), [Français](#), [Italiano](#), [日本語](#), [Nederlands](#), [Polski](#), [Русский](#), [Svenska](#)...

[Become an Editor](#) Help build the largest human-edited directory of the web



Copyright © 2011 Netscape

4,916,463 sites - 91,672 editors - over 1,007,856 categories

<http://www.dmoz.org>

# Information Retrieval organization: cataloguing

 open directory project

In partnership with  
**Aol Search.**

[about dmoz](#) | [dmoz blog](#) | [suggest URL](#) | [help](#) | [link](#) | [editor login](#)

Search [advanced](#)

[Arts](#)  
[Movies](#), [Television](#), [Music](#)...

[Games](#)  
[Video Games](#), [RPGs](#), [Gambling](#)...

[Kids and Teens](#)  
[Arts](#), [School Time](#), [Teen Life](#)...

[Reference](#)  
[Maps](#), [Education](#), [Libraries](#)...

[Shopping](#)  
[Clothing](#), [Food](#), [Gifts](#)...

[World](#)  
[Català](#), [Dansk](#), [Deutsch](#), [Español](#), [Français](#), [Italiano](#), [日本語](#), [Nederlands](#), [Polski](#)

[Business](#)  
[Jobs](#), [Real Estate](#), [Investing](#)...

[Health](#)  
[Fitness](#), [Medicine](#), [Alternative](#)...

[News](#)  
[Media](#), [Newspapers](#), [Weather](#)...

[Regional](#)  
[US](#), [Canada](#), [UK](#), [Europe](#)...

[Society](#)  
[People](#), [Religion](#), [Issues](#)...

[Computers](#)  
[Internet](#), [Software](#), [Hardware](#)...

[Home](#)  
[Family](#)...

[Recreation](#)  
[Travel](#), ...

[Science](#)  
[Biology](#)...

[Sports](#)  
[Baseball](#)...

**Top: [Computers: Software: Information Retrieval](#) (96)**

- [Classification@](#) (16)
- [Data Clustering@](#) (166)
- [Fulltext](#) (32)
- [GILS](#) (1)
- [Internet Search Engines@](#) (314)
- [Ranking](#) (35)
- [References](#) (1)
- [Text Clustering@](#) (11)
- [Visual Information](#) (6)
- [Web Clustering](#) (6)

See also:

- [Computers: Software: File Management: Search](#) (37)
- [Computers: Software: Internet: Servers: Search](#) (41)
- [Reference: Knowledge Management: Knowledge Retrieval](#) (40)
- [Reference: Libraries: Library and Information Science: Software](#) (93)

[Become an Editor](#) Help build the largest human-edited directory of the web

Copyright © 2011 Netscape

4,916,463 sites - 91,672 editors - over 1,007,856 categories

<http://www.dmoz.org>

# Information Retrieval

## analysis and organization: reading-level

blues clues

About 12,800 results (0.18 seconds)

Reading level
Advanced

Results by reading level for **blues clues** - [View results for all reading levels](#)

|              |     |  |
|--------------|-----|--|
| Basic        | 66% |  |
| Intermediate | 30% |  |
| Advanced     | 4%  |  |

► [Blue - Wikipedia, the free encyclopedia](#) 🔍

[en.wikipedia.org/wiki/Blue](http://en.wikipedia.org/wiki/Blue) - Cached

**Blue** is a colour, the perception of which is evoked by light having a spectrum dominated by energy with a wavelength of roughly 440–490 nm. ...

[Images for blues clues](#) - Report images



[Researching Blue's Clues: Viewing behavior and impact](#) 🔍

[www.cmch.tv/research/fullrecord.asp?id=1773](http://www.cmch.tv/research/fullrecord.asp?id=1773) - Cached

Year: 2000. Article Title: Researching **Blue's Clues**: Viewing behavior and impact. Journal: Media Psychology. Volume: 2. Edition: Issue: 2. Pages: 179-194 ...

[Do children learn how to watch television? The impact of extensive ...](#) 🔍

[www.cmch.tv/research/fullrecord.asp?id=1932](http://www.cmch.tv/research/fullrecord.asp?id=1932) - Cached

The impact of extensive experience with **Blues Clues** on preschool children's ...

cmchresearch

**Design: Trials comparing TV comprehension, behavior of 2 groups: experienced **Blue's Clues** viewers and inexperienced viewers. Study 1: Subjects viewed unseen ...**

# Information Retrieval

## organization: recommendations



Real people. Real reviews.®

Search for (e.g. taco, cheap dinner, Max's)

cosmic cantina

Near (Address, City, State or Zip)

Chapel Hill, NC

Search

WelcomeAbout MeWrite a ReviewFind ReviewsInvite FriendsMessagingTalkEventsMember Search

### Cosmic Cantina

★★★★☆ 41 reviews [Rating Details](#)

Category: Mexican [\[Edit\]](#)

128 E Franklin St  
Chapel Hill, NC 27514

(919) 960-3955

Price Range: \$

Accepts Credit Cards: Yes

Parking: Street

Attire: Casual

Good for Groups: No

Good for Kids: Yes

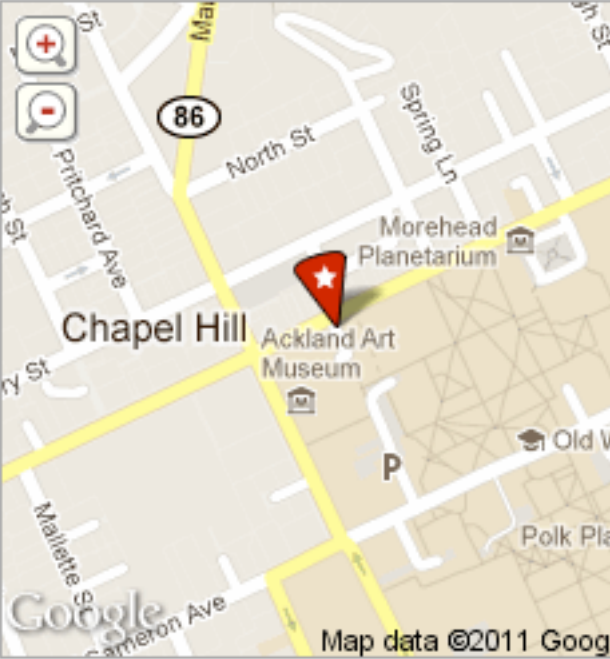
Takes Reservations: No

Delivery: No

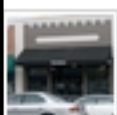
Take-out: Yes

Waiter Service: No





#### People Who Viewed This Also Viewed...

-  **The Cosmic Cantina**  
★★★★☆ 84 reviews  
Durham, NC
-  **Carrburritos**  
★★★★☆ 93 reviews  
Carrboro, NC
-  **Joe's Joint**  
★★★★☆ 8 reviews  
Chapel Hill, NC
-  **Bandido's Mexican Cafe**  
★★★★☆ 20 reviews  
Chapel Hill, NC
-  **Pepper's Pizza**  
★★★★☆ 74 reviews  
Chapel Hill, NC

<http://www.yelp.com/biz/cosmic-cantina-chapel-hill>  
(not actual page)

# Information Retrieval storage



Jaime Arguello  
Assistant Professor  
School of Information & Library Science  
100 Manning Hall  
University of North Carolina at Chapel Hill  
Chapel Hill, NC 27599-3360  
email: [jaime@cs.cmu.edu](mailto:jaime@cs.cmu.edu)  
phone: (630) 781-3898

Office hours: By appointment

### Research Interests

I conduct research in Information Retrieval (IR). Within IR, my main focus has been on aggregated search—the task of providing users with integrated access to multiple systems within a single search interface. In particular, my focus is on aggregated search within the context of Web search. In addition to retrieving Web pages, commercial search engines also function as a single point of access to specialized services known as verticals (e.g., image search, video search, news search, local search, driving directions, weather forecasts, etc.). My methods address the task of deciding which verticals to show (if any) and where to show them. In addition to IR, I am also interested in machine learning, text data-mining, and human-computer interaction.

### Publications

**2011**

Jaime Arguello, Fernando Diaz, and Jamie Callan. Learning to Aggregate Vertical Results into Web Search Results. 20th ACM Conference on Information and Knowledge Management (CIKM'11), To Appear.

Jaime Arguello, Fernando Diaz, Jamie Callan, and Ben Carterette. [A Methodology for Evaluating Aggregated Search Results](#). In Proceedings of the 33rd European Conference on Information Retrieval. (ECIR'11), 2011. (Best Student Paper Award)

**2010**

Jaime Arguello, Fernando Diaz, and Jean-François Paiement. [Vertical Selection in the Presence of Unlabeled Verticals](#). In Proceedings of the 33rd Annual International ACM SIGIR Conference on Research & Development on Information Retrieval (SIGIR'10), 2010.



The New York Times  
Friday, August 5, 2011 Last Update: 11:34 AM ET

Get a Full Times Experience. BECOME A DIGITAL SUBSCRIBER

U.S. Posts Stronger Job Gains Amid Fear  
Economy Added 117,000 Jobs in July; Rate Falls to 9.1%  
By MOTOKO RICH 9:58 AM ET  
The job growth figure, reported by the Labor Department, was the strongest since April, but may still not be high enough to relieve fears of a renewed recession.  
Economic: Government Jobs Down | Seasonal Adjustments  
Post a Comment | Read (148)

Wall Street Rally Quickly Fizzles  
By GRAHAM BOWLEY, MATTHEW SALTmarsh and BETTINA WASSNER 15 minutes ago  
Shares wavered, taking back most of their opening gains after the employment report on Friday.  
Post a Comment | Read (80)

THE LEDE  
Live Blog: The Markets  
By CHRISTINE HAUSER 48 minutes ago  
The E.U.'s top finance official said insufficient information might have been a factor in the recent market turmoil.

RELATED COVERAGE  
Nervous Investors Chase Low-Risk Assets  
Double Dip Recession May Be Happening  
High & Low Finance: Learning to Live With Debt

MARKETS » At 11:44 AM ET  
S&P 500 1,184.58 -15.49 -1.29%  
Dow 11,284.08 -99.60 -0.87%  
Nasdaq 2,508.32 -48.07 -1.88%

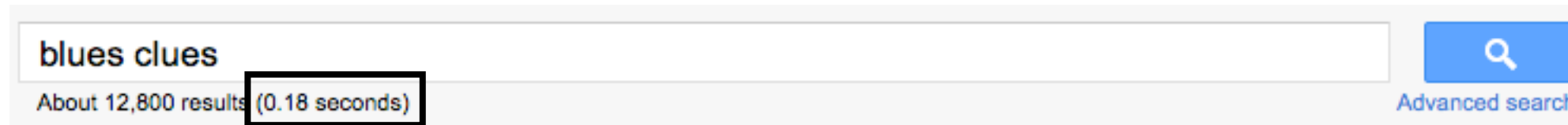
GET QUOTES My Portfolios »  
Stock, ETFs, Funds Go

UPGRADE YOUR TIMES EXPERIENCE.  
Become a Digital Subscriber.  
CLICK FOR DETAILS

- How might a web search engine view these pages differently in terms of storage?

# Information Retrieval

## retrieval



- **Efficiency:** retrieving results in this lifetime (or, better yet, in 0.18 seconds)
- **Effectiveness:** retrieving results that satisfy the user's information need (more on this later)
- We will focus more on effectiveness
- However, we will also discuss in some detail how search engines retrieve results as fast as they do

# Many Types of Search Engines

bing™



# Many Types of Search Engines



bing



eHarmony

PANDORA

Google



match.com

mapquest m<sup>q</sup>



YAHOO! ANSWERS



LinkedIn

flickr



Westlaw

The New York Times



YouTube  
Broadcast Yourself



# The Search Task

- Given a **query** and a **corpus**, find **relevant** items
  - query**: user's expression of their information need
  - corpus**: a repository of retrievable items
  - relevance**: satisfaction of the user's information need

# Search Engines

## web search

query

results

facebook and productivity

Study: Facebook use cuts productivity at work - Computerworld  
www.computerworld.com > Internet > Web 2.0 and Web Apps - Cached  
Jul 22, 2009 - A Nucleus Research study found that Facebook work in the workplace is cutting employee productivity.

Pulling the Plug on Facebook, Productivity/Time Management Article ...  
www.inc.com > Leadership and Managing > Human Resources - Cached  
Pulling the Plug on Facebook, Productivity/Time Management Article - All that friending and superpoking wastes a lot of time at the office -- and could be ...

Twitter and Facebook: The New Tools of Productivity or Distraction ...  
www.briansolis.com/.../twitter-and-facebook-the-new-tools-of-prod... - Cached  
Mar 26, 2010 - RT Twitter and Facebook: Tools of Productivity or Distraction .... RT @PRSAcolo: Twitter & Facebook: New tools of productivity or ...

Twitter, Facebook Can Improve Work Productivity | PCWorld Business ...  
www.pcworld.com/.../twitter\_facebook\_can\_improve\_work\_produc... - Cached  
Apr 2, 2009 - Reach Older Users on Facebook and Twitter - The Web's Best Productivity Sites. According to a study by the Australian University, ...

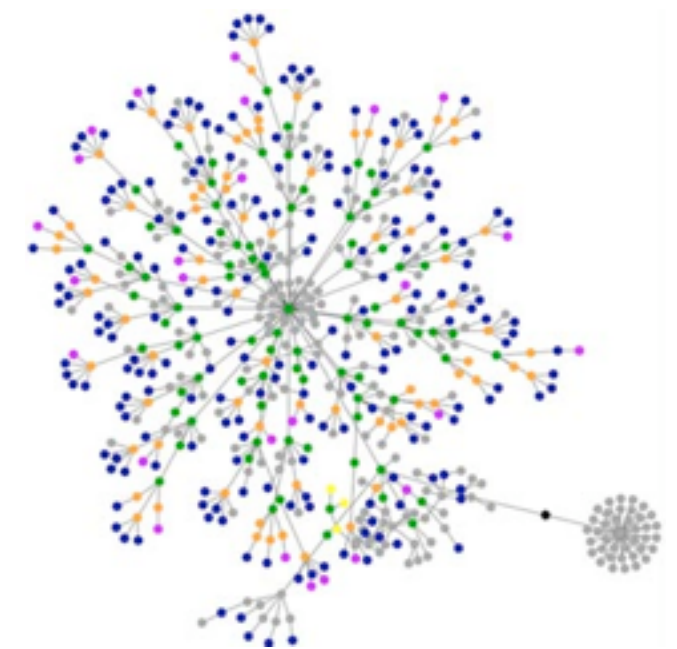
Is Facebook Killing Your Employees' Productivity? | WebProNews  
www.webpronews.com/is-facebook-killing-your-employees-produc... - Cached  
Jul 21, 2009 - On the heels of a study indicating that social media can significantly impact a brand's bottom line positively, another one has come out ...

Productivity Strategies | Facebook  
www.facebook.com/beproductive - Cached  
Productivity Strategies - To learn more about the Productive Today "Content Collaborative" faculty, click the "Info" tab or this direct link: | Facebook.

Butt Out IT! Facebook "Productivity Loss" Is No Concern of Yours  
blogs.gartner.com/.../butt-out-it-facebook-productivity-loss-is-no-co... - Cached  
Facebook "Productivity Loss" Is No Concern of Yours. by Brian Prentice | November 23, 2008 | 10 Comments. Like my colleague Anthony Bradley, I also speak to ...

Productivity Levels Plummet After Yale Student Makes Facebook Look ...  
www.betabeat.com/.../yale-student-makes-facebook-look-like-excel... - Cached  
5 days ago - Productivity Levels Plummet After Yale Student Makes Facebook Look Like Excel. By Rebecca Panovka 7/28 6:11pm ...

corpus



web pages

# Search Engines

## digital library search

query

results

- 1 [Effective teaching practices using free Google services: conference tutorial](#)  
[Paul Gestwicki](#), [Brian McNely](#)  
October 2010 **Journal of Computing Sciences in Colleges** , Volume 26 Issue 1  
**Publisher:** Consortium for Computing Sciences in Colleges  
Full text available:  [Pdf](#) (22.76 KB)  
**Bibliometrics:** Downloads (6 Weeks): 2, Downloads (12 Months): 48, Downloads (Overall): 48,  
  
In this 90-minute tutorial, we will share our experiences using free Web services from Google teaching effectiveness. Participants will engage with these services as part of the tutorial. We have used and studied these technologies, ...

- 2 [Model-Based Engineering of Software: Three Productivity Perspectives](#)  
[Shawn A. Bohner](#), [Sriram Mohan](#)  
October 2009 **SEW '09: Proceedings of the 2009 33rd Annual IEEE Software Engineering Workshop**  
**Publisher:** IEEE Computer Society  
Full text available:  [Publisher Site](#)  
**Bibliometrics:** Downloads (6 Weeks): n/a, Downloads (12 Months): n/a, Downloads (Overall): n/a, Citation Count: 0  
  
Evolving software products is a tricky business, especially when the domain is complex and changing rapidly. Like other fields of engineering, software engineering productivity advances have come about largely through abstraction reuse, process, and ...  
**Keywords:** Agent-Based Software Systems, Model-Driven Architecture, Model-Driven Development, Model-Based Software Development, Model-Based Software Engineering

- 3 [Absolute Beginner's Guide to Computer Basics, 5th edition](#)  
[Michael Miller](#)  
September 2009 **Absolute Beginner's Guide to Computer Basics, 5th edition**  
**Publisher:** Que Publishing Company  
**Bibliometrics:** Downloads (6 Weeks): n/a, Downloads (12 Months): n/a, Downloads (Overall): n/a, Citation Count: 0

Everything casual users need to know to get the most out of their new Windows 7 PCs, software, and the Internet  
The best-selling beginner's guide, now completely updated for Windows 7 and today's most popular Internet tools -

corpus



scientific  
publications

# Search Engines

## news search

query

results

### [Tropical Storm Emily Heads Toward Haiti, Dominican Republic](#)

Voice of America - 23 minutes ago

August 03, 2011 Tropical Storm **Emily** Heads Toward Haiti, Dominican Republic VOA News Tropical storm warnings and watches are posted for parts of the ...

[Video: Tropical Storm Emily on Path Toward Haiti](#)  The Associated Press

Local: [Emily could reach hurricane strength](#) Sarasota Herald-Tribune (blog)

Blog: [Mubarak Trial Begins; Tropical Storm Emily Threatens East Coast](#) NPR (blog)

[The Guardian](#) - [Fox News](#)

[all 1804 news articles »](#)

### [Emily Heads For Hispaniola](#)

WAVY-TV (blog) - Jeremy Wheeler - 2 hours ago

**Emily** has had little change in strength since yesterday. She has winds of 50mph. The pressure is down a little to 1003 mb (millibars of pressure). ...

[Tropical Storm Emily remains weak, hits Haiti tonight](#) Examiner.com

August 2, 2011 Midday Tropics Update: TS **Emily** Now Moving More ... Wakulla.com

[Emily the Effervescent](#) Caribbean Hurricane Network

[all 4 news articles »](#)

### ['Snow White' Writer to Pen Universal's 'Emily the Strange' \(Exclusive\)](#)

Hollywood Reporter - Borys Kit - 13 hours ago

Melisa Wallack, who wrote the script that became Relativity's high-profile Snow White project, has been drafted by Universal to pen **Emily** the Strange. ...

[Early Edition: Emily the Strange to Darken Big Screen; More News](#) Moviefone (blog)

[Writer Melisa Wallack Will Follow SNOW WHITE With EMILY THE STRANGE](#) Collider.com

[Details on the upcoming 'Emily the Strange' movie](#) Examiner.com

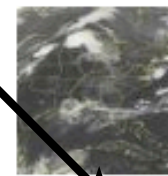
[ComingSoon.net](#) - [411mania.com](#)

[all 7 news articles »](#)

### [High heat in Midwest and South](#)

Reuters - Tim Sharp - Kevin Murphy - 22 hours ago

By Wendell Marsh WASHINGTON (Reuters) - Record-breaking heat continued to broil central and southern states on Tuesday as Tropical Storm **Emily** threatened to ...



CSA.ca



Collider.com



Reuters

corpus

news articles

# Search Engines

## local business search

query

results

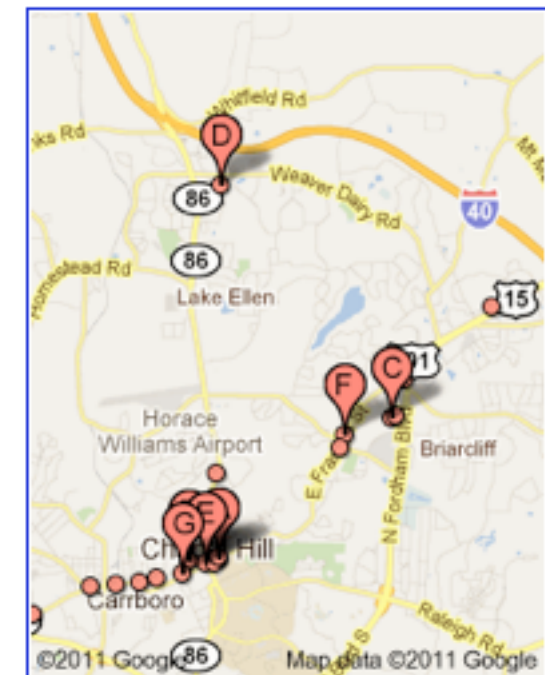
### Places for **mexican food** near Chapel Hill, NC

- A** [Bandido's Mexican Cafe & Cantina](#)  - ★★★★★ 14 reviews - [Place page](#)  
[www.bandidoscafe.com](http://www.bandidoscafe.com) - 159 1/2 East Franklin Street, Chapel Hill - (919) 967-5048
- B** [Las Potrillos Mexican Restaurant](#)  - ★★★★★ 9 reviews - [Place page](#)  
[www.lospotrillos.net](http://www.lospotrillos.net) - 220 West Rosemary Street, Chapel Hill - (919) 932-4301
- C** [monterrey mexican restaurant](#)  - ★★★★★ 17 reviews - [Place page](#)  
[monterreychapelhill.com](http://monterreychapelhill.com) - 237 South Elliot Road, Chapel Hill - (919) 969-8750
- D** [Margaret's Cantina](#)  - ★★★★★ 19 reviews - [Place page](#)  
[www.margaretscantina.com](http://www.margaretscantina.com) - 1129 Weaver Dairy Road, Chapel Hill - (919) 942-4745
- E** [Qdoba Mexican Grill](#)  - ★★★★★ 19 reviews - [Place page](#)  
[www.qdoba.com](http://www.qdoba.com) - 100 West Franklin Street, Chapel Hill - (919) 929-8998
- F** [Cinco de Mayo](#)  - ★★★★★ 11 reviews - [Place page](#)  
[www.cincomayorestaurants.net](http://www.cincomayorestaurants.net) - 1502 East Franklin Street, Chapel Hill - (919) 929-6566
- G** [Chipotle Mexican Grill](#)  - ★★★★★ 15 reviews - [Place page](#)  
[www.chipotle.com](http://www.chipotle.com) - 301 W. Franklin St., Chapel Hill - (919) 942-2091

corpus

Google maps

curated/synthesized  
business listings

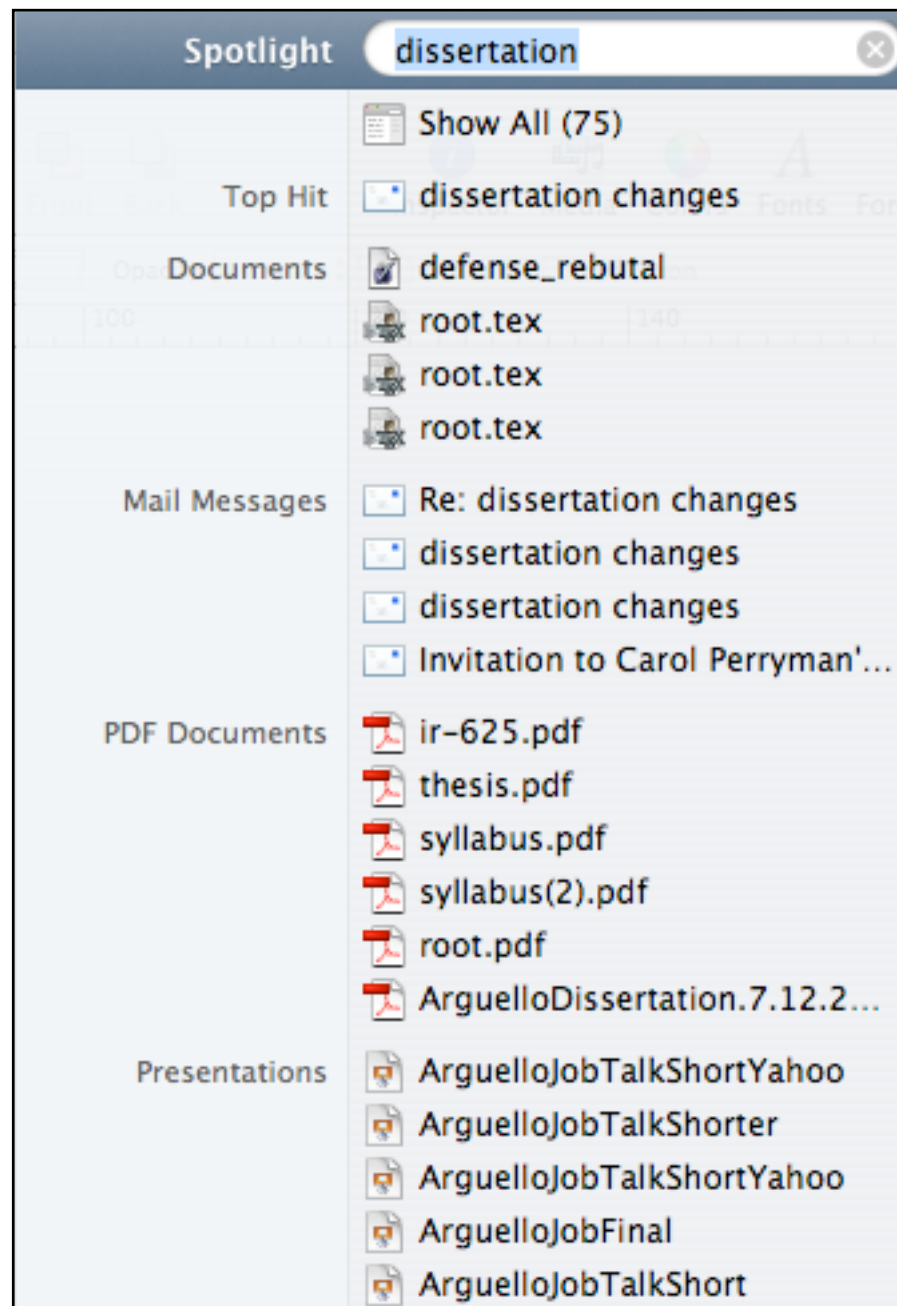


# Search Engines

## desktop search

query

results



corpus



files in my laptop

# Search Engines

## micro-blog search

query

results

**neenjames** Neen James  
Productivity tip: Follow ppl on **Twitter** that inspire, challenge and inform you - delete the clutter!  
4 minutes ago

**mr\_Ostentatious** Jason Pitts  
Took a day off from **twitter** to increase my productivity and ended up having a productive day!  
1 hour ago

**adamwiebe** Adam Wiebe  
Social media at work is here. Be wary of what is and is not productive. <http://lnkd.in/DW3z8J>  
3 hours ago

**ViggosDaddy** Gert van der Linde  
A brief look: To tweet, or not to tweet? - How does **Twitter** affect our productivity, influence and how informe... <http://tinyurl.com/3wbz3rn>  
6 hours ago

**IncorrectMystic** Raghavender | raGz  
**#productivity** day - So going be off **twitter** and other social networks till work is over :) bye tweeples for a while  
6 hours ago

corpus

twitter



tweets

# Search Engines

## people/profile search

query

results

michael jordan



**Michael Jordan**

Page

12,856,455 people like this.



**Michael Jordan**

Carnegie Mellon



**Michael Jordan**

1 mutual friend



**Michael Jordan**

Page

215,268 people like this.



**Michael jordan**

Page

225,371 people like this.



**MICHAEL JORDAN**

Page

190,013 people like this.



**Michael Jordan**

Page

58,003 people like this.

corpus



profiles

# Information Retrieval Tasks and Applications

digital library search

web search

enterprise search

news search

local business search

image search

video search

(micro-)blog search

community Q&A search

desktop search

question-answering

federated search

social search

expert search

product search

patent search

recommender systems

opinion mining

# The Search Task

- Given a **query** and a **corpus**, find **relevant** items
  - query**: user's expression of their information need
  - corpus**: a repository of retrievable items
  - relevance**: satisfaction of the user's information need

# The Search Task

## in this course

- Given a query and a corpus, find relevant items

**query:** user's expression of their information need

- ▶ a textual description of what the user wants

**corpus:** a repository of retrievable items

- ▶ a collection of textual documents

**relevance:** satisfaction of the user's information need

- ▶ the document contains information the user wants

# Why is IR fascinating?

- Information retrieval is an **uncertain** process
  - ▶ users don't know what they want
  - ▶ users don't know how to convey what they want
  - ▶ computers can't elicit information like a librarian
  - ▶ computers can't understand natural language text
  - ▶ the search engine can only guess what is relevant
  - ▶ the search engine can only guess if a user is satisfied
  - ▶ over time, we can only guess how users adjust their short- and long-term behavior for the better

# Queries and Relevance

August 7, 2006 9:59 AM PDT

## AOL apologizes for release of user search data

By Dawn Kawamoto and Elinor Mills

Staff Writers, CNET News

Last modified: August 7, 2006 2:30 PM PDT

### Related Stories

[Should Google be forced to hand over data?](#)

March 14, 2006

[Judge to help feds against Google](#)

March 14, 2006

[Google, feds face off over search records](#)

March 14, 2006

AOL apologized on Monday for releasing search log data on subscribers that had been intended for use with the company's newly launched research site.

The randomly selected data, which focused on 658,000 subscribers and posted 10 days ago, was among the tools intended for use on the recently launched AOL Research site. But the Internet giant has since removed the search logs from public view.

"This was a screw-up, and we're angry and upset about it. It was an innocent enough attempt to reach out to the academic community with new research tools, but it was obviously not appropriately vetted, and if it had been, it would have been stopped in an instant," AOL, a unit of Time Warner, said in a statement. "Although there was no personally identifiable data linked to these accounts, we're absolutely not defending this. It was a mistake, and we apologize. We've launched an internal investigation into what happened, and

we are taking steps to ensure that this type of thing never happens again."

Although AOL had used identification numbers rather than names or user IDs when listing the search logs, that did not quell concerns of privacy advocates, who said that anyone among the 658,000 could easily be identified based on the searches each individual conducted.

"We think it's a major privacy concern, and we're glad to see AOL is taking it seriously," said Ari Schwartz, deputy director of the [Center for Democracy and Technology](#). "Companies that deal in search results have to understand that they carry very sensitive information, even if it doesn't have what we would traditionally consider to be personally identifiable information involved."

# Queries and Relevance

- ▶ 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
- ▶ marie selby botanical gardens
- ▶ 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
- ▶ riverbank run

(AOL query-log)

# Queries and Relevance

- A query is an impoverished description of the user's information need
- Highly ambiguous to anyone other than the user

# Queries and Relevance

the input to the system

- Query 435: curbing population growth

what is in the user's head

- **Description:** What measures have been taken worldwide and what countries have been effective in curbing population growth? A relevant document must describe an actual case in which population measures have been taken and their results are known. Reduction measures must have been actively pursued. Passive events such as decease, which involuntarily reduce population, are not relevant.

(from TREC 2005 HARD Track)

# Queries and Relevance

- Query 435: curbing population growth
- Description:

[illegible]

(from TREC 2005 HARD Track)

# Queries and Relevance

- Query 435: curbing population growth
- Can we imagine a relevant document without all these query terms?

# Queries and Relevance

- Query 435: curbing population growth
- The same concept can be expressed in different ways

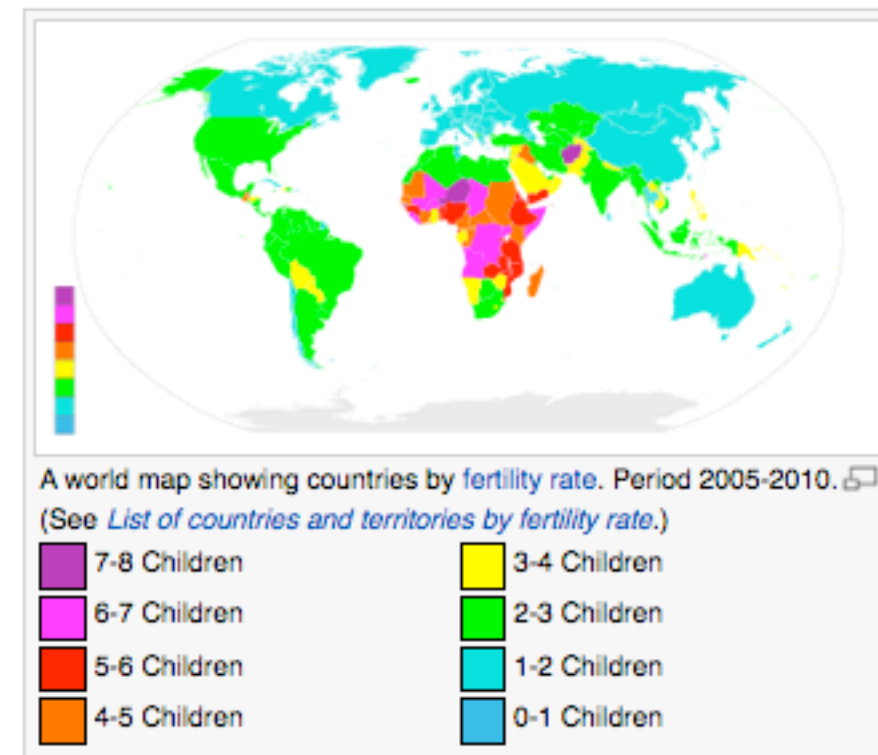
## Human population control

From Wikipedia, the free encyclopedia

**Human population control** is the practice of artificially altering the rate of growth of a human population.

Historically, human population control has been implemented by limiting the population's [birth rate](#), usually by government mandate, and has been undertaken as a response to factors including high or increasing levels of [poverty](#), [environmental concerns](#), [religious reasons](#), and [overpopulation](#). While population control can involve measures that improve people's lives by giving them greater control of their reproduction, some programs have exposed them to exploitation.<sup>[1]</sup>

Worldwide, the population control movement was active throughout the 1960s and 1970s, driving many [reproductive health](#) and [family planning](#) programs. In the 1980s, tension grew between population control advocates and women's health activists who advanced women's [reproductive rights](#) as part of a [human rights](#)-based approach.<sup>[2]</sup> Growing opposition to the narrow population control focus led to a significant change in population control policies in the early 1990s.<sup>[3]</sup>



# Queries and Relevance

- Query 435: curbing population growth
- Can we imagine a non-relevant document with all these query terms?

# Queries and Relevance

- Query 435: curbing population growth
- The query concept can have different “senses”

## 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.



# Queries and Relevance

- This is why IR is difficult (and fascinating!)
- Croft, Metzler, & Strohman:

Understanding how people compare text and designing computer algorithms to accurately perform this comparison is at the core of information retrieval.
- IR does not seek a deep “understanding” of the document text
- It uses statistical properties of the text to predict whether a document is relevant to a query
  - ▶ easier and often times sufficient

# Predicting Relevance

- What types of evidence can we use to predict that a document is relevant to a query?
  - ▶ **query-document evidence:** a property of the query-document pair (e.g., a measure of similarity)
  - ▶ **document evidence:** a property of the document (same for all queries)

# Query-Document Evidence

- Query: bathing a cat

## How to Bathe a Cat

Edited by Kimberly and 107 others

Article

Edit

Discuss

History

Tweet

22

Like

50

Even though they mostly keep themselves clean, most cats need a bath every now and then. Everyone knows that cats hate getting wet. Cats generally keep themselves clean, and therefore should not be bathed any more often than is absolutely necessary. But they sometimes get especially dirty, get bombarded by fleas or ticks, or get into substances that are toxic or otherwise harmful. On such occasions, it is a good idea to bathe your cat. Here's how to keep your feline fresh and lovely as painlessly as possible!



Ads by Google

[Pet Allergy Info](#) Information On The Symptoms And Triggers For Indoor Allergies.  
ZYRTEC.com

### Steps

Edit

- 1 **Decide, whether your cat really needs a bath.** There may be other ways to clean your cat, such as [brushing the cat](#), combing or even rubbing it down with a cloth.
- 2 **Wear appropriate clothing.** Not only is it important to wash your cat but it is important to be safe from any possible scratches, minor or major. A jumper or a long-sleeved shirt must be worn so that your cat has no bare skin to scratch. If you don't have one, you could wear long-sleeved gloves. It is also a good idea to wear clothing that isn't new.
- 3 **Get at least two people involved in washing your cat, especially if your cat is rather strong and can kick and wriggle its way out of your hands.** One person should hold all four legs and hold the cat's jaw so it can't open its mouth to bite you, but be sure you don't hold it really tight so it can't breathe. Hold your cat firmly so it cannot wriggle out from your grip.
- ⋮
- 16 **Reward your cat.** Give him/her their favorite canned food or catnip or treats, and he/she will come to realize that there is a good side to being bathed.



# Query-Document Evidence

- Query: **bathing** a **cat**
- The important query terms occur frequently
- Both terms occur
- Terms occur close together
- Terms occur in the title
- Terms occur in the URL

[www.wikihow.com/bathe-your-cat](http://www.wikihow.com/bathe-your-cat)

- Any other ideas?

**How to Bathe a Cat**  
Edited by Kimberly and 107 others

Article Edit Discuss History

Tweet 22 Like 50

Even though they mostly keep themselves clean, most **cats** need a **bath** every now and then. Everyone knows that **cats** hate getting wet. **Cats** generally keep themselves clean, and therefore should not be **bathed** any more often than is absolutely necessary. But they sometimes get especially dirty, get bombarded by fleas or ticks, or get into substances that are toxic or otherwise harmful. On such occasions, it is a good idea to **bathe** your **cat**. Here's how to keep your feline fresh and lovely as painlessly as possible!



Ads by Google  
[Pet Allergy Info](#) Information On The Symptoms And Triggers For Indoor Allergies. ZYRTEC.com

**Steps** Edit

- 1 **Decide, whether your **cat** really needs a **bath**.** There may be other ways to clean your **cat**, such as **brushing the cat**, combing or even rubbing it down with a cloth.
- 2 **Wear appropriate clothing.** Not only is it important to wash your **cat** but it is important to be safe from any possible scratches, minor or major. A jumper or a long-sleeved shirt must be worn so that your **cat** has no bare skin to scratch. If you don't have one, you could wear long-sleeved gloves. It is also a good idea to wear clothing that isn't new.
- 3 **Get at least two people involved in washing your **cat**, especially if your **cat** is rather strong and can kick and wriggle its way out of your hands.** One person should hold all four legs and hold the **cat's** jaw so it can't open its mouth to bite you, but be sure you don't hold it really tight so it can't breathe. Hold your cat firmly so it cannot wriggle out from your grip.
- ⋮
- 16 **Reward your **cat**.** Give him/her their favorite canned food or catnip or treats, and he/she will come to realize that there is a good side to being **bathed**.



# Query-Document Evidence

- Terms occur in hyperlinks pointing to the page

## How to Raise Angora Cats

Edited by Elizabeth Knudsen and 3 others

- 3 **Groom the cat regularly.** The angora cats usually have medium length fur, and need to be groomed at least 3-4 times at week. This does not include bathing, you should only **bathe the cat** if it gets specially dirty, or for a show.

## How to Bathe a Cat

Edited by Kimberly and 107 others

Article

Edit

Discuss

History

Tweet 22

Like 50

Even though they mostly keep themselves clean, most **cats** need a **bath** every now and then. Everyone knows that **cats** hate getting wet. **Cats** generally keep themselves clean, and therefore should not be **bathed** any more often than is absolutely necessary. But they sometimes get especially dirty, get bombarded by fleas or ticks, or get into substances that are toxic or otherwise harmful. On such occasions, it is a good idea to **bathe** your **cat**. Here's how to keep your feline fresh and lovely as painlessly as possible!



Ads by Google

**Pet Allergy Info** Information On The Symptoms And Triggers For Indoor Allergies. ZYRTEC.com

## Steps

Edit

- 1 **Decide, whether your **cat** really needs a **bath**.** There may be other ways to clean your **cat**, such as **brushing the cat**, combing or even rubbing it down with a cloth.
- 2 **Wear appropriate clothing.** Not only is it important to wash your **cat** but it is important to be safe from any possible scratches, minor or major. A jumper or a long-sleeved shirt must be worn so that your **cat** has no bare skin to scratch. If you don't have one, you could wear long-sleeved gloves. It is also a good idea to wear clothing that isn't new.
- 3 **Get at least two people involved in washing your **cat**, especially if your **cat** is rather strong and can kick and wriggle its way out of your hands.** One person should hold all four legs and hold the **cat's** jaw so it can't open its mouth to bite you, but be sure you don't hold it really tight so it can't breathe. Hold your cat firmly so it cannot wriggle out from your grip.
- ⋮
- 16 **Reward your **cat**.** Give him/her their favorite canned food or catnip or treats, and he/she will come to realize that there is a good side to being **bathed**.



# Query-Document Evidence

- Does not contain “.com”
- [verb] [article] [noun]
- Not one of the most popular queries
- Does not contain the term “news”

## How to Bathe a Cat

Edited by Kimberly and 107 others

Article

Edit

Discuss

History

Tweet

22

Like

50

Even though they mostly keep themselves clean, most cats need a bath every now and then. Everyone knows that cats hate getting wet. Cats generally keep themselves clean, and therefore should not be bathed any more often than is absolutely necessary. But they sometimes get especially dirty, get bombarded by fleas or ticks, or get into substances that are toxic or otherwise harmful. On such occasions, it is a good idea to bathe your cat. Here's how to keep your feline fresh and lovely as painlessly as possible!



Ads by Google

[Pet Allergy Info](#) Information On The Symptoms And Triggers For Indoor Allergies.  
ZYRTEC.com

### Steps

Edit

- 1 **Decide, whether your cat really needs a bath.** There may be other ways to clean your cat, such as [brushing the cat](#), combing or even rubbing it down with a cloth.
- 2 **Wear appropriate clothing.** Not only is it important to wash your cat but it is important to be safe from any possible scratches, minor or major. A jumper or a long-sleeved shirt must be worn so that your cat has no bare skin to scratch. If you don't have one, you could wear long-sleeved gloves. It is also a good idea to wear clothing that isn't new.
- 3 **Get at least two people involved in washing your cat, especially if your cat is rather strong and can kick and wriggle its way out of your hands.** One person should hold all four legs and hold the cat's jaw so it can't open its mouth to bite you, but be sure you don't hold it really tight so it can't breathe. Hold your cat firmly so it cannot wriggle out from your grip.
- ⋮
- 16 **Reward your cat.** Give him/her their favorite canned food or catnip or treats, and he/she will come to realize that there is a good side to being bathed.



# Query-Document Evidence

- We can also use previous user interactions, e.g.:
- The query is similar to other queries associated with clicks on this document
- The document is similar to other documents associated with clicks for this query

## How to Bathe a Cat

Edited by Kimberly and 107 others

Article

Edit

Discuss

History

Tweet

22

Like

50

Even though they mostly keep themselves clean, most cats need a bath every now and then. Everyone knows that cats hate getting wet. Cats generally keep themselves clean, and therefore should not be bathed any more often than is absolutely necessary. But they sometimes get especially dirty, get bombarded by fleas or ticks, or get into substances that are toxic or otherwise harmful. On such occasions, it is a good idea to bathe your cat. Here's how to keep your feline fresh and lovely as painlessly as possible!



Ads by Google

[Pet Allergy Info](#) Information On The Symptoms And Triggers For Indoor Allergies.  
ZYRTEC.com

### Steps

Edit

- 1 **Decide, whether your cat really needs a bath.** There may be other ways to clean your cat, such as [brushing the cat](#), combing or even rubbing it down with a cloth.
- 2 **Wear appropriate clothing.** Not only is it important to wash your cat but it is important to be safe from any possible scratches, minor or major. A jumper or a long-sleeved shirt must be worn so that your cat has no bare skin to scratch. If you don't have one, you could wear long-sleeved gloves. It is also a good idea to wear clothing that isn't new.
- 3 **Get at least two people involved in washing your cat, especially if your cat is rather strong and can kick and wriggle its way out of your hands.** One person should hold all four legs and hold the cat's jaw so it can't open its mouth to bite you, but be sure you don't hold it really tight so it can't breathe. Hold your cat firmly so it cannot wriggle out from your grip.
- ⋮
- 16 **Reward your cat.** Give him/her their favorite canned food or catnip or treats, and he/she will come to realize that there is a good side to being bathed.



# Document Evidence

- Lots of in-links (endorsements)
- Non-spam properties:
  - ▶ grammatical sentences
  - ▶ no profanity
- Has good formatting
- Anything other ideas?

## How to Bathe a Cat

Edited by Kimberly and 107 others

Article

Edit

Discuss

History

Tweet

22

Like

50

Even though they mostly keep themselves clean, most cats need a bath every now and then. Everyone knows that cats hate getting wet. Cats generally keep themselves clean, and therefore should not be bathed any more often than is absolutely necessary. But they sometimes get especially dirty, get bombarded by fleas or ticks, or get into substances that are toxic or otherwise harmful. On such occasions, it is a good idea to bathe your cat. Here's how to keep your feline fresh and lovely as painlessly as possible!



Ads by Google

[Pet Allergy Info](#) Information On The Symptoms And Triggers For Indoor Allergies. ZYRTEC.com

### Steps

Edit

- 1 **Decide, whether your cat really needs a bath.** There may be other ways to clean your cat, such as [brushing the cat](#), combing or even rubbing it down with a cloth.
- 2 **Wear appropriate clothing.** Not only is it important to wash your cat but it is important to be safe from any possible scratches, minor or major. A jumper or a long-sleeved shirt must be worn so that your cat has no bare skin to scratch. If you don't have one, you could wear long-sleeved gloves. It is also a good idea to wear clothing that isn't new.
- 3 **Get at least two people involved in washing your cat, especially if your cat is rather strong and can kick and wriggle its way out of your hands.** One person should hold all four legs and hold the cat's jaw so it can't open its mouth to bite you, but be sure you don't hold it really tight so it can't breathe. Hold your cat firmly so it cannot wriggle out from your grip.
- ⋮
- 16 **Reward your cat.** Give him/her their favorite canned food or catnip or treats, and he/she will come to realize that there is a good side to being bathed.



# Document Evidence

- Author attributes
- Peer-reviewed by many
- Reading-level appropriate for user community
- Has pictures
- Recently modified (fresh)
- Normal length
- From domain with other high-quality documents

## How to Bathe a Cat

Edited by Kimberly and 107 others

Article

Edit

Discuss

History

Tweet

22

Like

50

Even though they mostly keep themselves clean, most cats need a bath every now and then. Everyone knows that cats hate getting wet. Cats generally keep themselves clean, and therefore should not be bathed any more often than is absolutely necessary. But they sometimes get especially dirty, get bombarded by fleas or ticks, or get into substances that are toxic or otherwise harmful. On such occasions, it is a good idea to bathe your cat. Here's how to keep your feline fresh and lovely as painlessly as possible!



Ads by Google

[Pet Allergy Info](#) Information On The Symptoms And Triggers For Indoor Allergies.  
ZYRTEC.com

### Steps

Edit

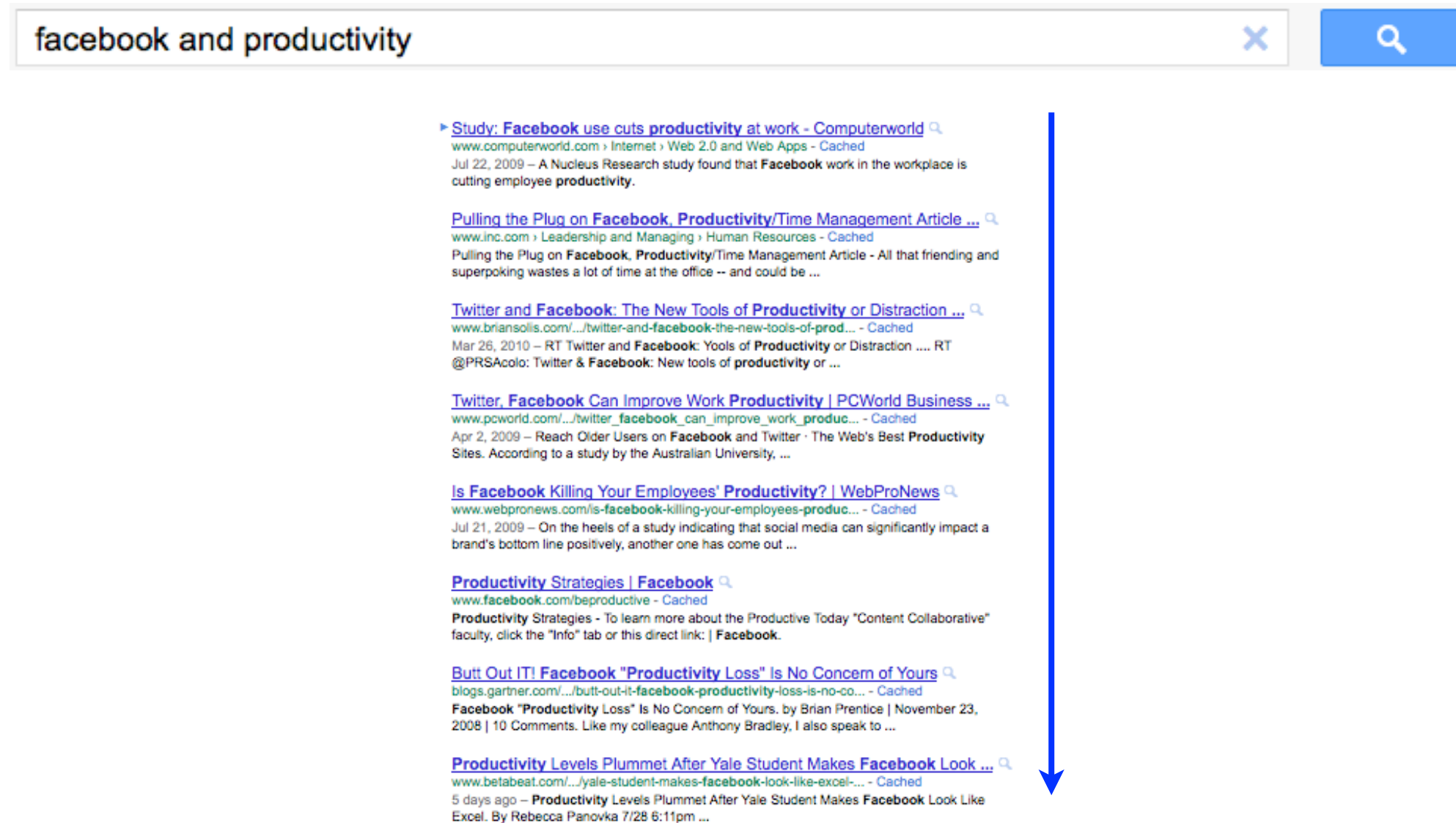
- 1 **Decide, whether your cat really needs a bath.** There may be other ways to clean your cat, such as [brushing the cat](#), combing or even rubbing it down with a cloth.
- 2 **Wear appropriate clothing.** Not only is it important to wash your cat but it is important to be safe from any possible scratches, minor or major. A jumper or a long-sleeved shirt must be worn so that your cat has no bare skin to scratch. If you don't have one, you could wear long-sleeved gloves. It is also a good idea to wear clothing that isn't new.
- 3 **Get at least two people involved in washing your cat, especially if your cat is rather strong and can kick and wriggle its way out of your hands.** One person should hold all four legs and hold the cat's jaw so it can't open its mouth to bite you, but be sure you don't hold it really tight so it can't breathe. Hold your cat firmly so it cannot wriggle out from your grip.
- ⋮
- 16 **Reward your cat.** Give him/her their favorite canned food or catnip or treats, and he/she will come to realize that there is a good side to being bathed.



# Predicting Relevance

- IR does not require a deep “understanding” of information
- We can get by using shallow sources of evidence, which can be generated from the query-document pair or just the document itself.

# The Search Task



- **Output:** a ranking of items in descending order of predicted relevance (simplifies the task)
- **Assumption:** the user scans the results from top to bottom and stops when he/she is satisfied or gives up

# Evaluating a Ranking

facebook and productivity

Study: Facebook use cuts productivity at work - Computerworld  
www.computerworld.com > Internet > Web 2.0 and Web Apps - Cached  
Jul 22, 2009 - A Nucleus Research study found that Facebook work in the workplace is cutting employee productivity.

Pulling the Plug on Facebook, Productivity/Time Management Article ...  
www.inc.com > Leadership and Managing > Human Resources - Cached  
Pulling the Plug on Facebook, Productivity/Time Management Article - All that friending and superpoking wastes a lot of time at the office -- and could be ...

Twitter and Facebook: The New Tools of Productivity or Distraction ...  
www.briansolis.com/.../twitter-and-facebook-the-new-tools-of-prod... - Cached  
Mar 26, 2010 - RT Twitter and Facebook: Yools of Productivity or Distraction .... RT @PRSAcolo: Twitter & Facebook: New tools of productivity or ...

Twitter, Facebook Can Improve Work Productivity | PCWorld Business ...  
www.pcworld.com/.../twitter\_facebook\_can\_improve\_work\_produc... - Cached  
Apr 2, 2009 - Reach Older Users on Facebook and Twitter · The Web's Best Productivity Sites. According to a study by the Australian University, ...

Is Facebook Killing Your Employees' Productivity? | WebProNews  
www.webpronews.com/is-facebook-killing-your-employees-produc... - Cached  
Jul 21, 2009 - On the heels of a study indicating that social media can significantly impact a brand's bottom line positively, another one has come out ...

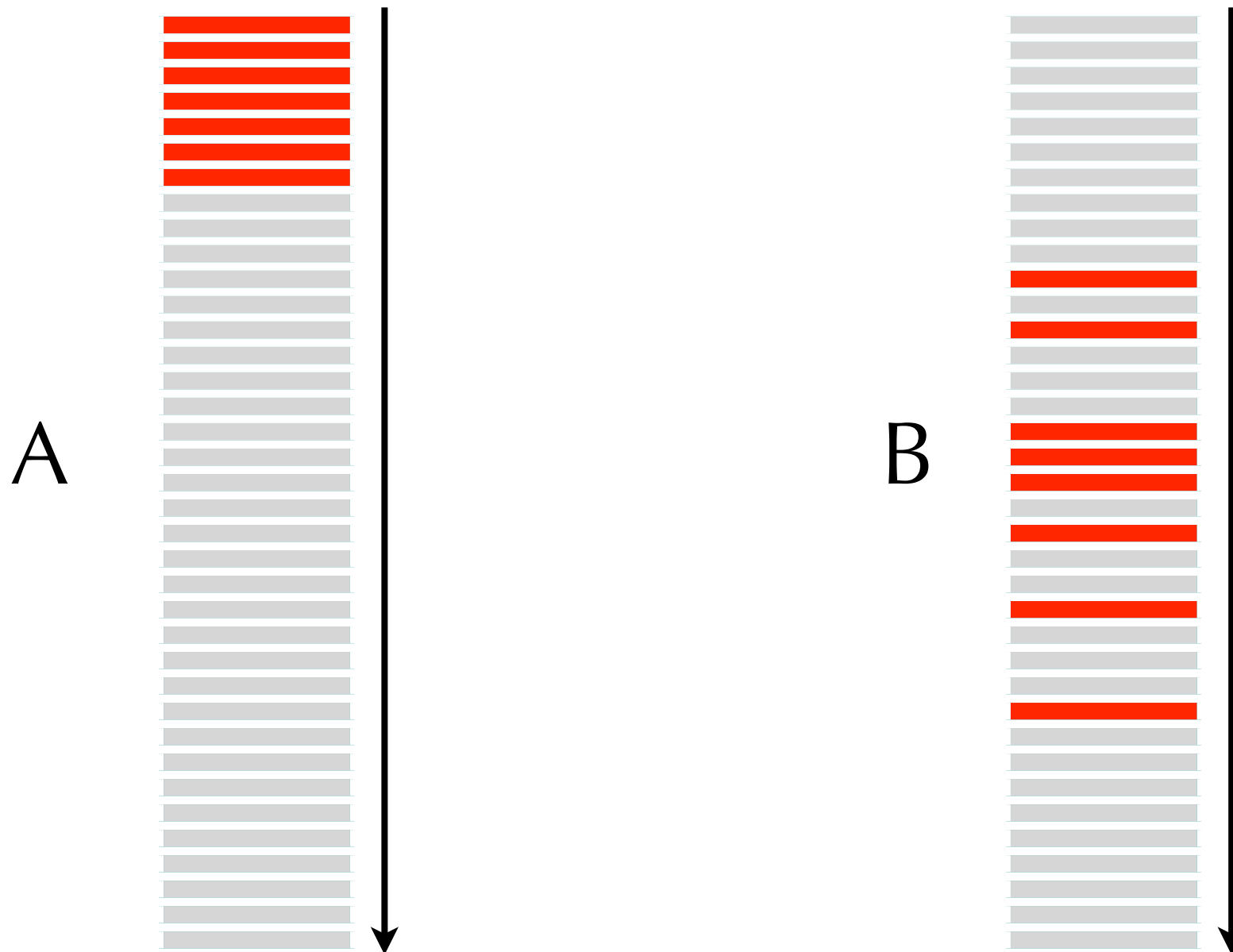
Productivity Strategies | Facebook  
www.facebook.com/beproductive - Cached  
Productivity Strategies - To learn more about the Productive Today "Content Collaborative" faculty, click the "Info" tab or this direct link: | Facebook.

Butt Out IT! Facebook "Productivity Loss" Is No Concern of Yours  
blogs.gartner.com/.../butt-out-it-facebook-productivity-loss-is-no-co... - Cached  
Facebook "Productivity Loss" Is No Concern of Yours. by Brian Prentice | November 23, 2008 | 10 Comments. Like my colleague Anthony Bradley, I also speak to ...

Productivity Levels Plummet After Yale Student Makes Facebook Look ...  
www.betabeat.com/.../yale-student-makes-facebook-look-like-excel-... - Cached  
5 days ago - Productivity Levels Plummet After Yale Student Makes Facebook Look Like Excel. By Rebecca Panovka 7/28 6:11pm ...

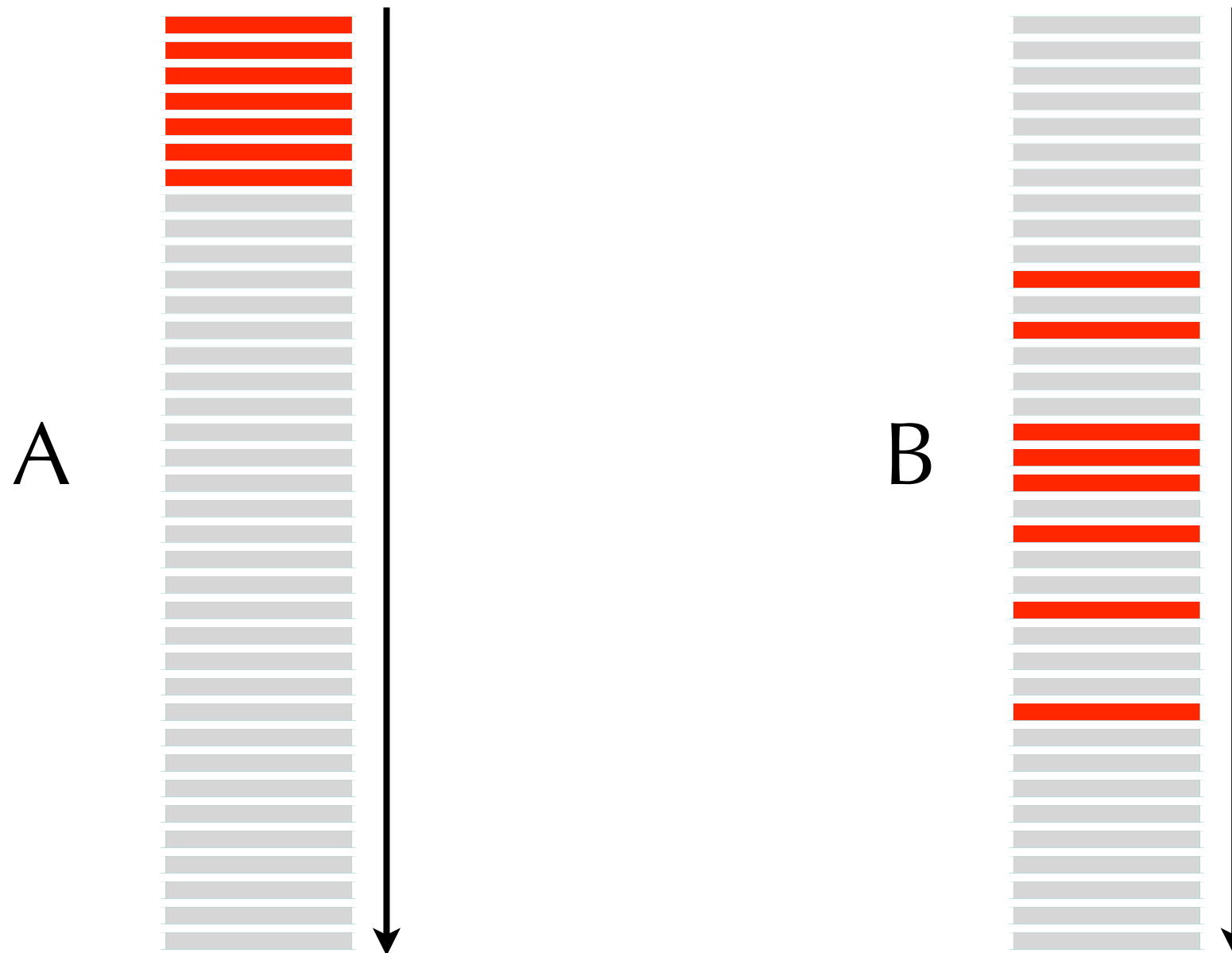
- So, how good is a particular ranking?
- Suppose we know which documents are truly relevant to the query...

# Evaluating a Ranking



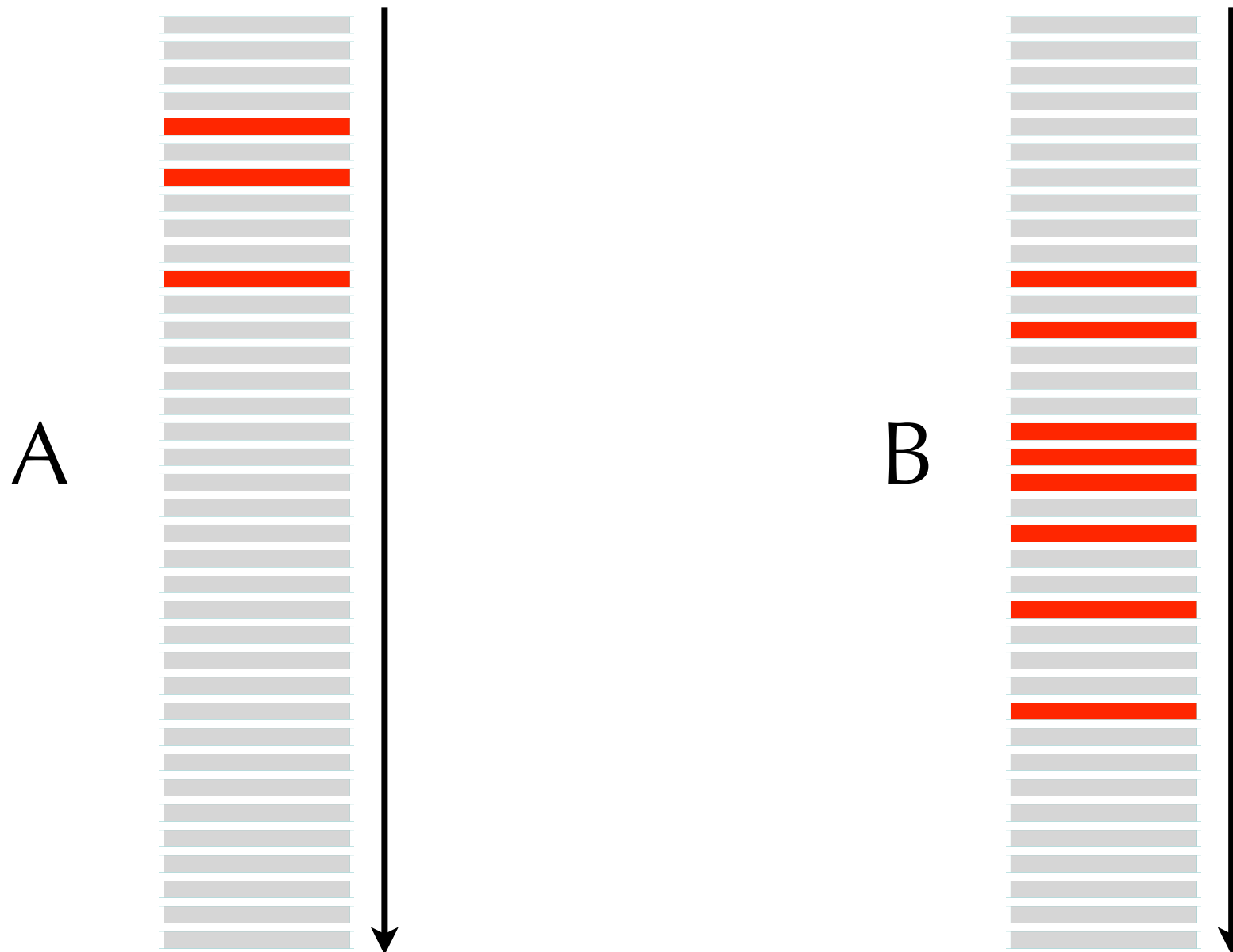
- Which ranking is better?

# Evaluating a Ranking



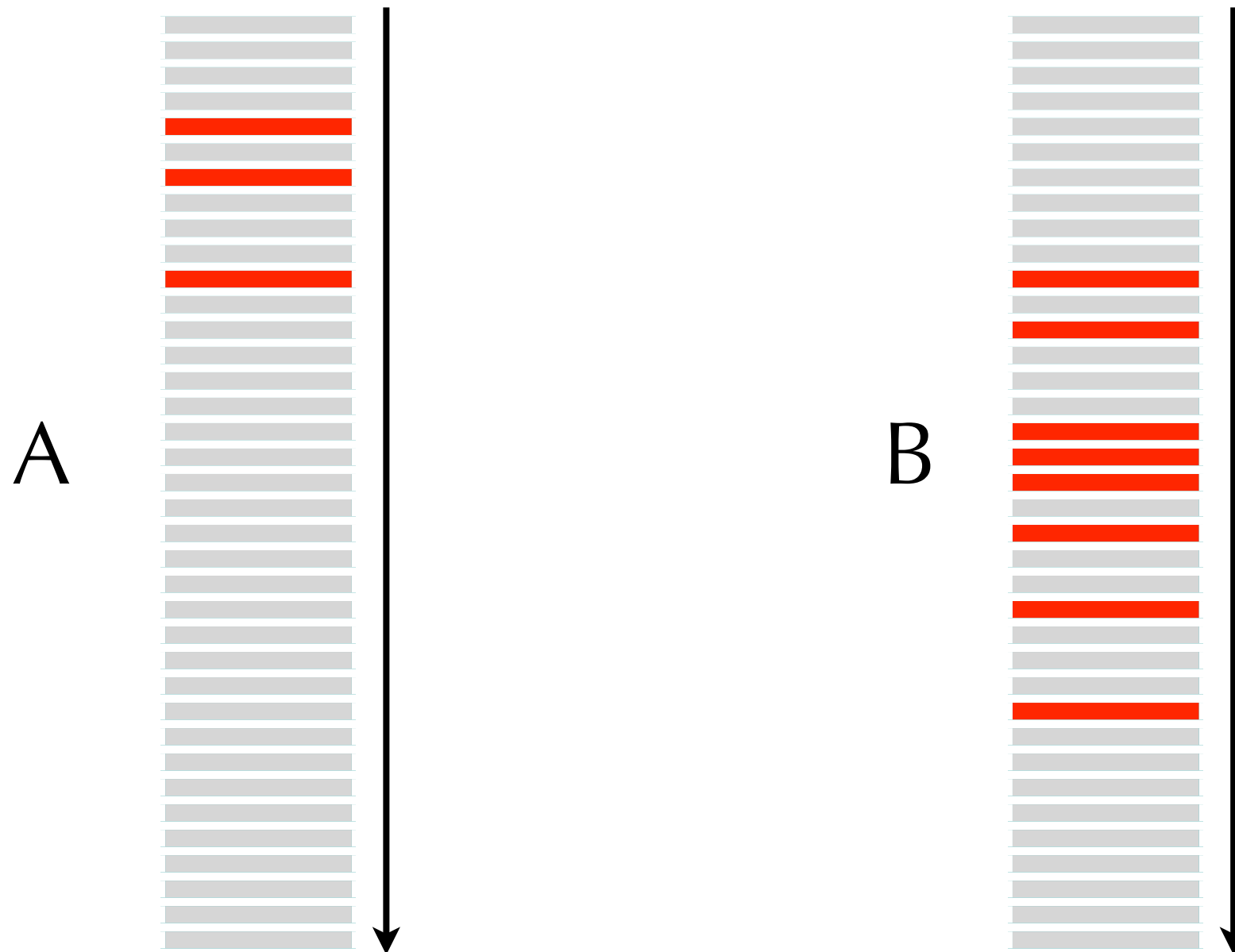
- In general, a ranking with all the relevant documents at the top is best (A is better than B)

# Evaluating a Ranking



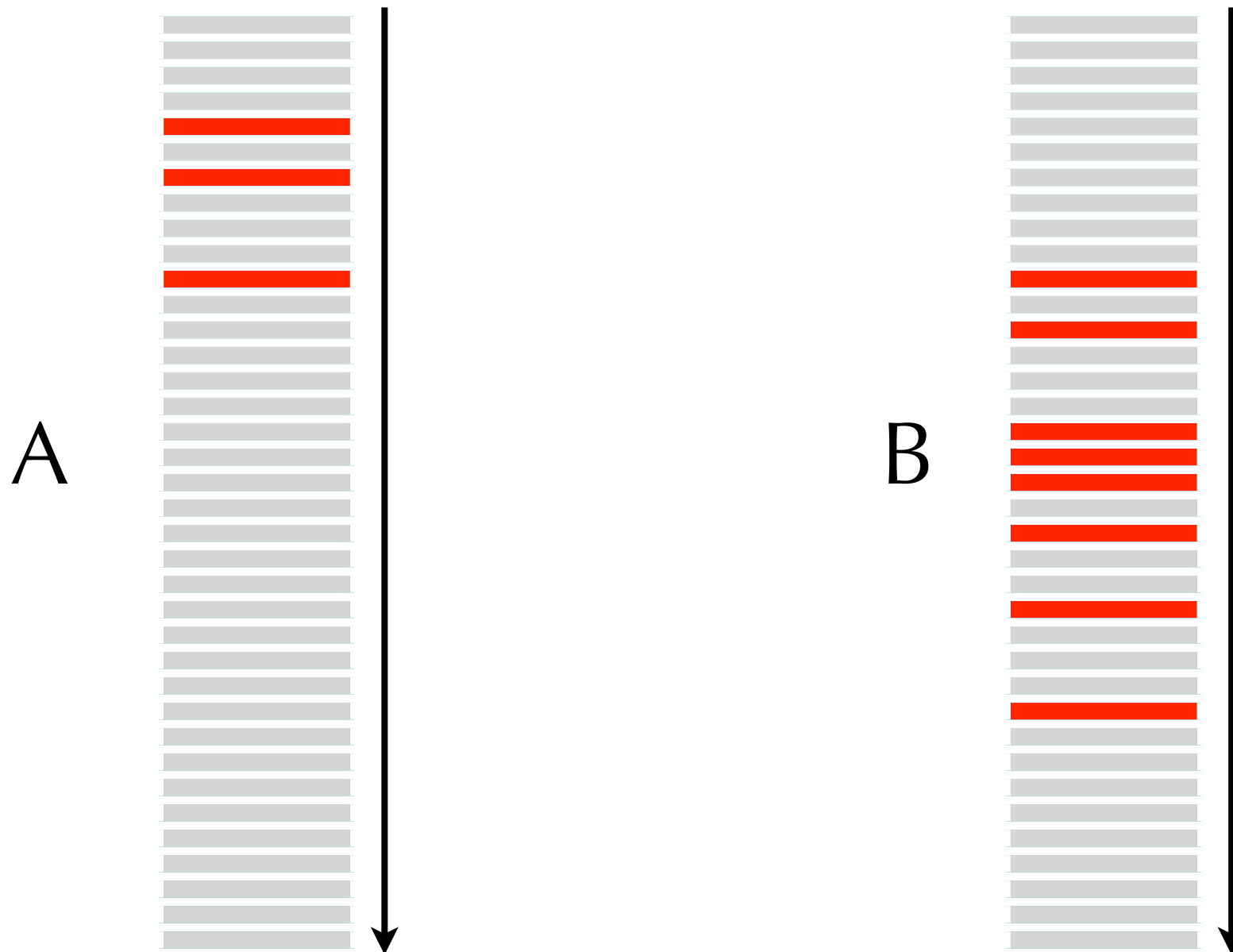
- Which ranking is better?

# Evaluating a Ranking



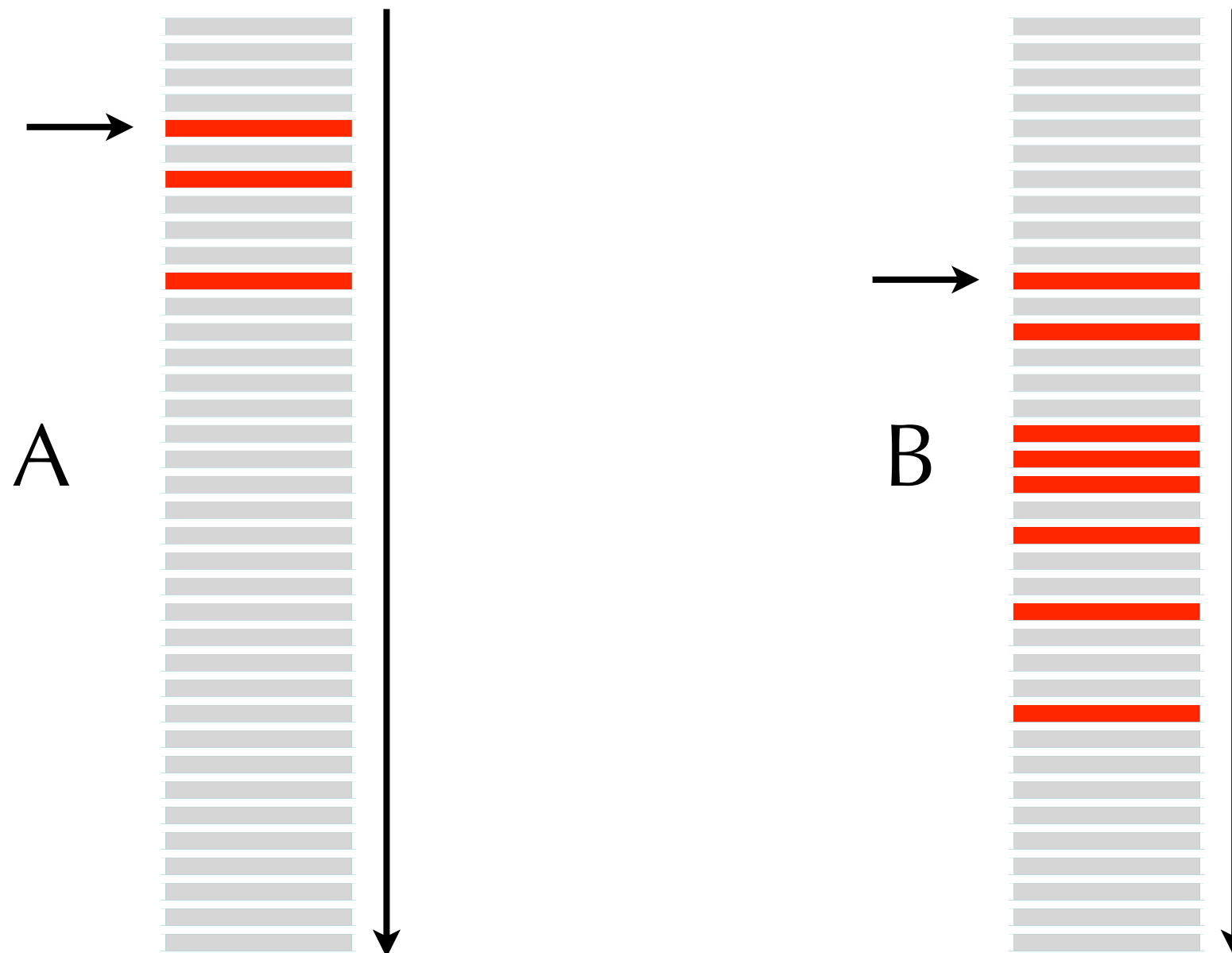
- Oftentimes the (relative) quality of a ranking is unclear and depends on the task

# Evaluating a Ranking



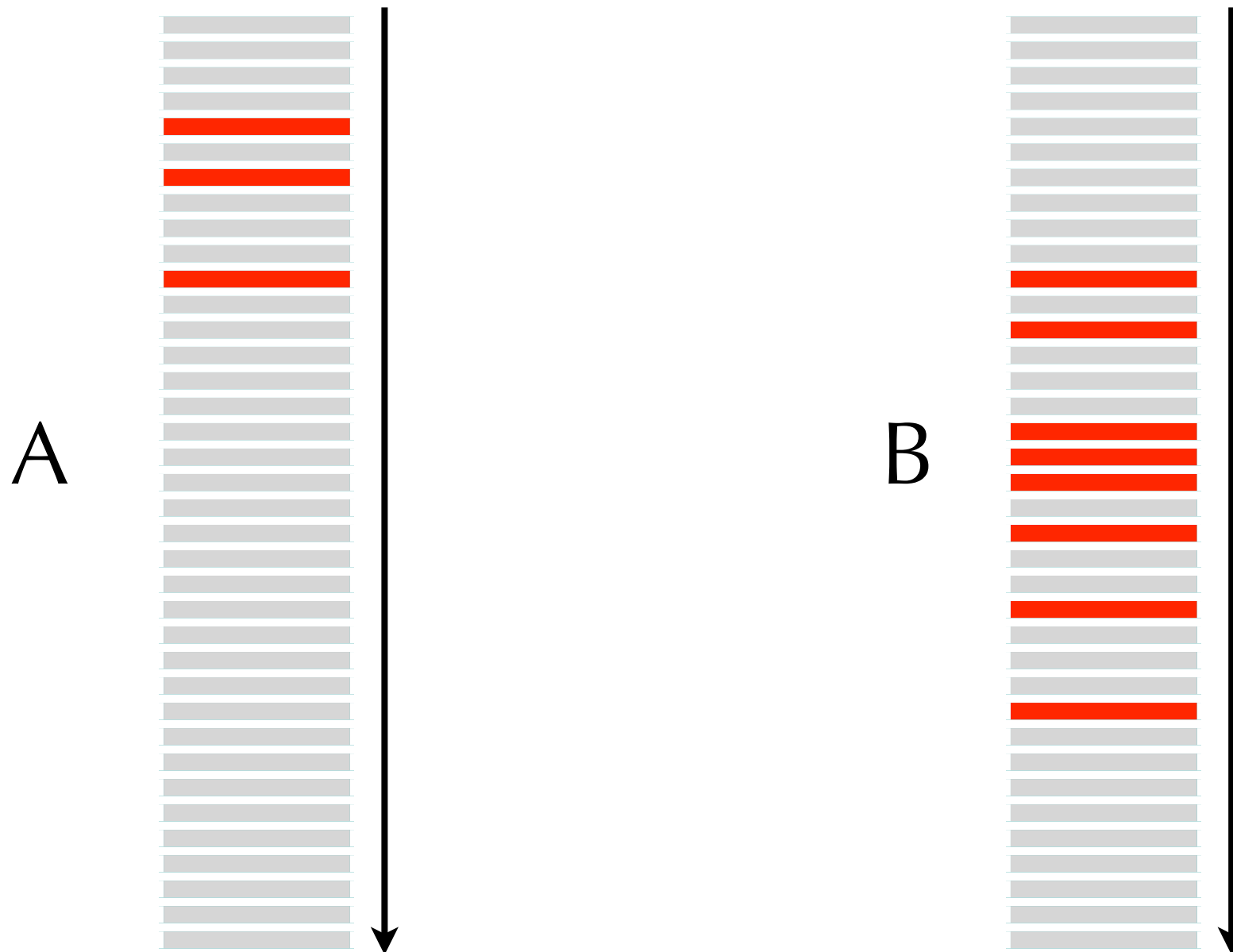
- Web search: ???????

# Evaluating a Ranking



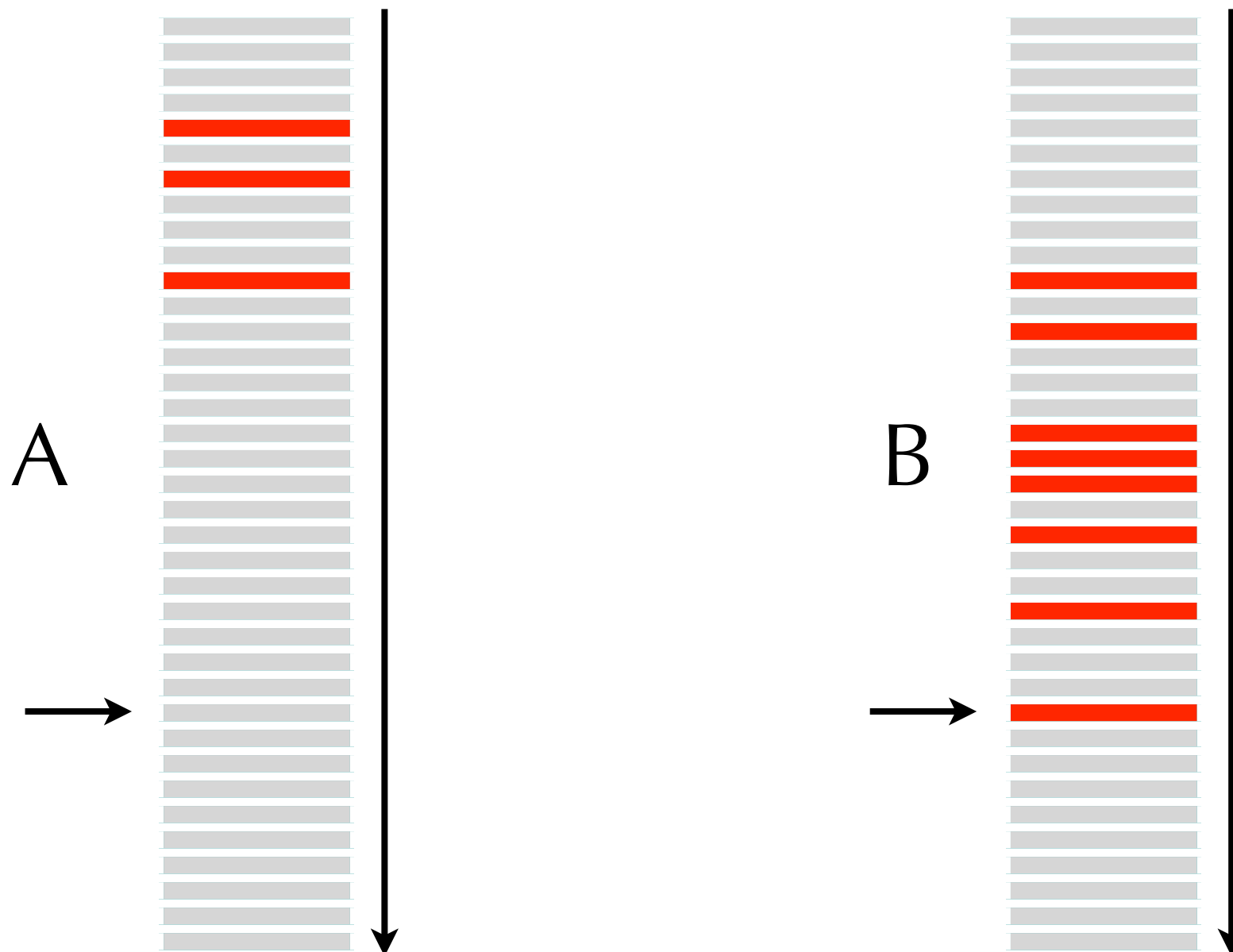
- **Web search:** A is better than B
- Many documents (redundantly) satisfy the user; the higher the first relevant document, the better

# Evaluating a Ranking



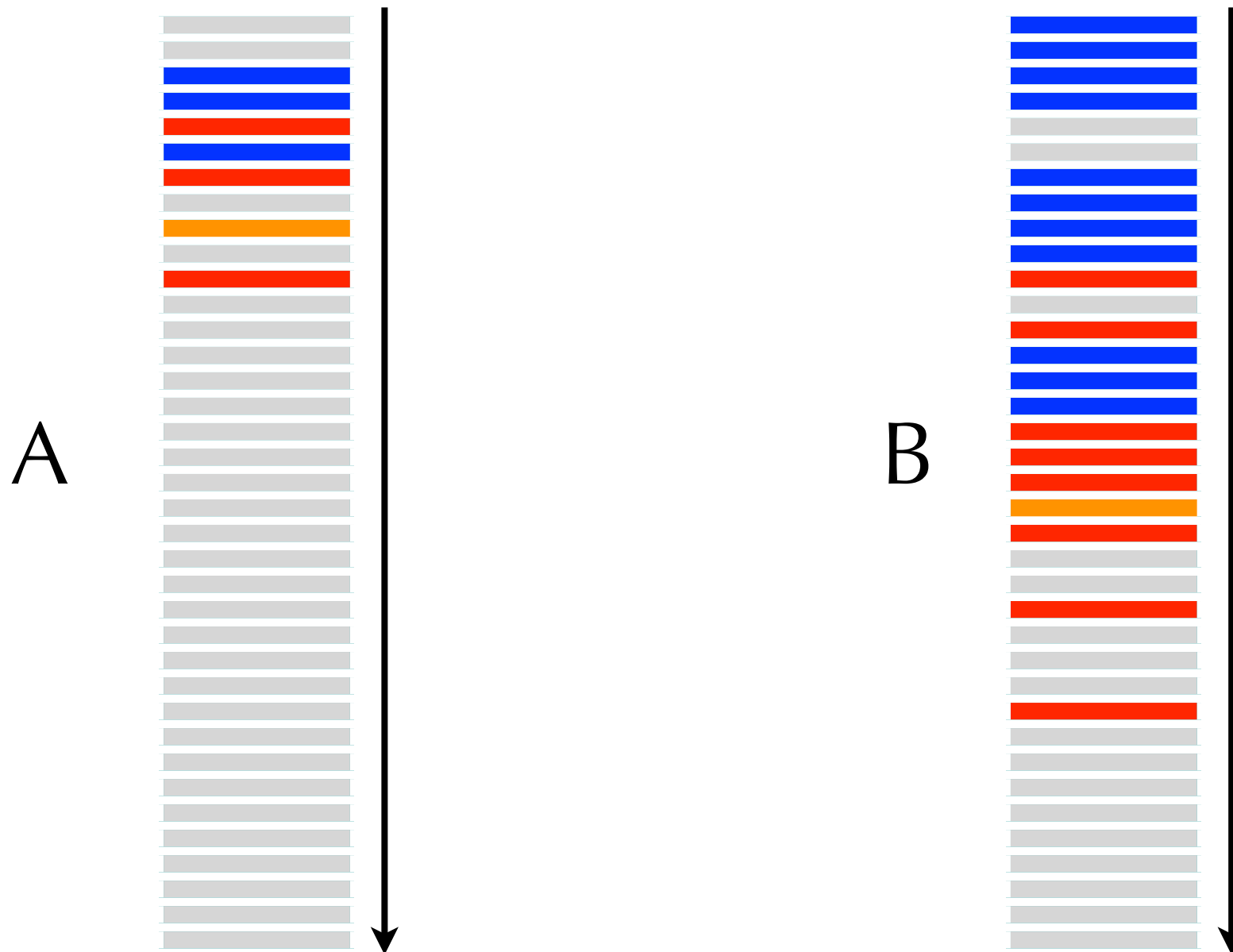
- Patent search: ???????

# Evaluating a Ranking



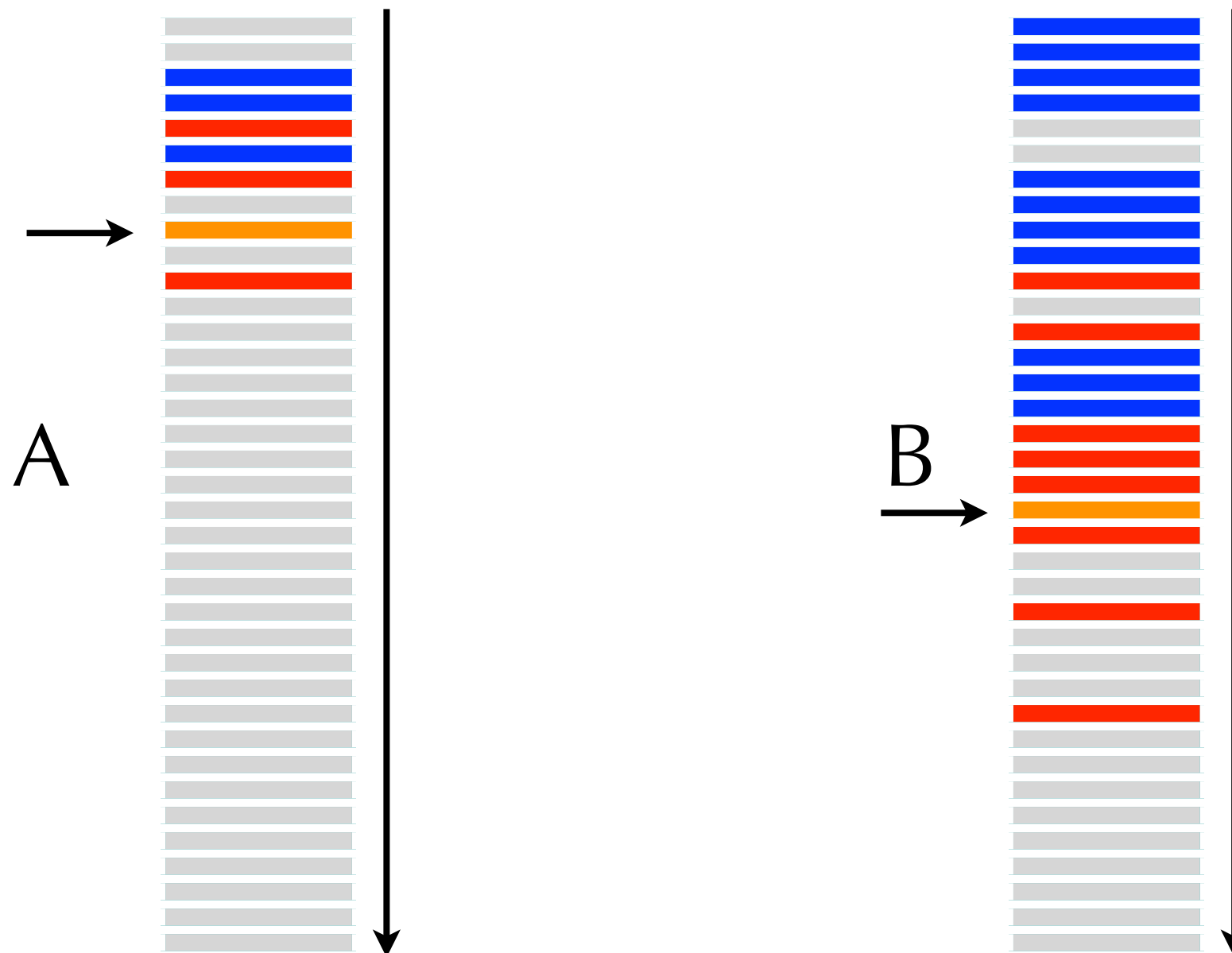
- Patent search: B is better than A
- User wants to see everything in the corpus that is related to the query (high cost in missing something)

# Evaluating a Ranking



- Exploratory search: ???????

# Evaluating a Ranking



- **Exploratory search:** A is better than B
- Satisfying the information need requires information found in different documents

# Evaluating a Ranking

## evaluation metrics

- Given a ranking with known relevant/non-relevant documents, an **evaluation metric** outputs a quality score
- Many, many metrics
- Different metrics make different assumptions
- Choosing the “right one” requires understanding the task
- Often, we use several (sanity check)

# Summary

- The goal of information retrieval is to match information-seekers with the information they seek.
- IR involves analysis, organization, storage, and retrieval
- There are many types of search engines
- There is uncertainty at every step of the search process
- Simple heuristics don't work, so IR systems make predictions about relevance!
- IR systems use “superficial” evidence to make predictions
- Users expect different things, depending on the task
- Evaluation requires understanding the user community.
- My goal is convince you that IR is a fascinating science

# Course Overview

Jaime Arguello  
[jarguell@email.unc.edu](mailto:jarguell@email.unc.edu)

January 8, 2014

# Course Objectives

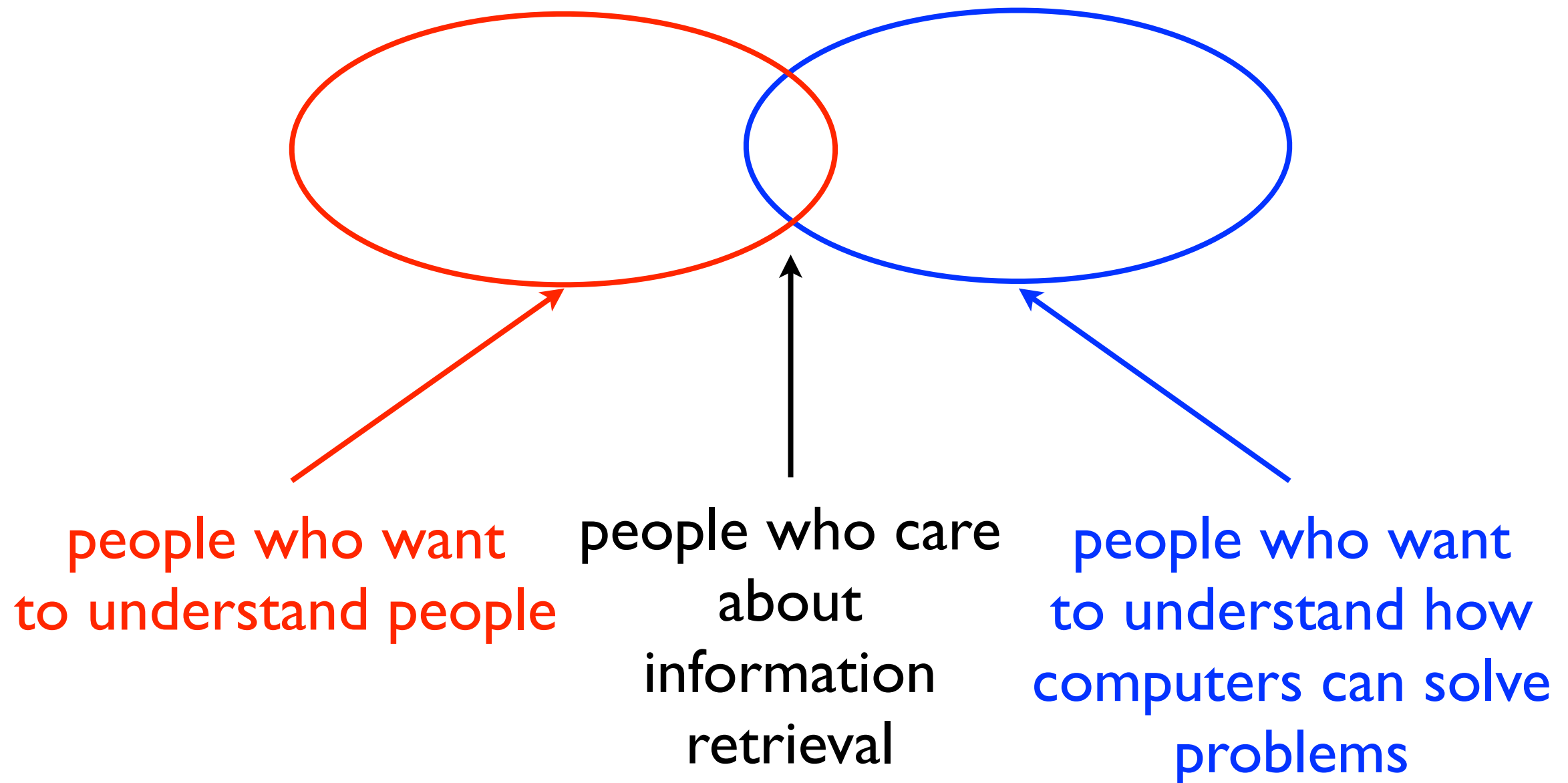
- How do search engines work?
  - ▶ effectiveness and efficiency
- How do users behave with them?
  - ▶ how do users determine usefulness of information?
  - ▶ how can a search engine mimic this process?
- Why do search engines fail?
  - ▶ the user? the corpus? the system? something else?
- How can they be evaluated (off-line)?
- How can they be monitored and tuned (on-line)?

# Why are these important questions?

- Most of the world's information is in natural language text
  - ▶ the world wide web
  - ▶ scientific publications
  - ▶ books
  - ▶ social media interactions
- The amount of this information is growing quickly; human capacity is not (evolution doesn't move that fast)
- We need smarter tools
- IR provides tools for analyzing and organizing content to facilitate search, discovery, and learning

# Course Structure

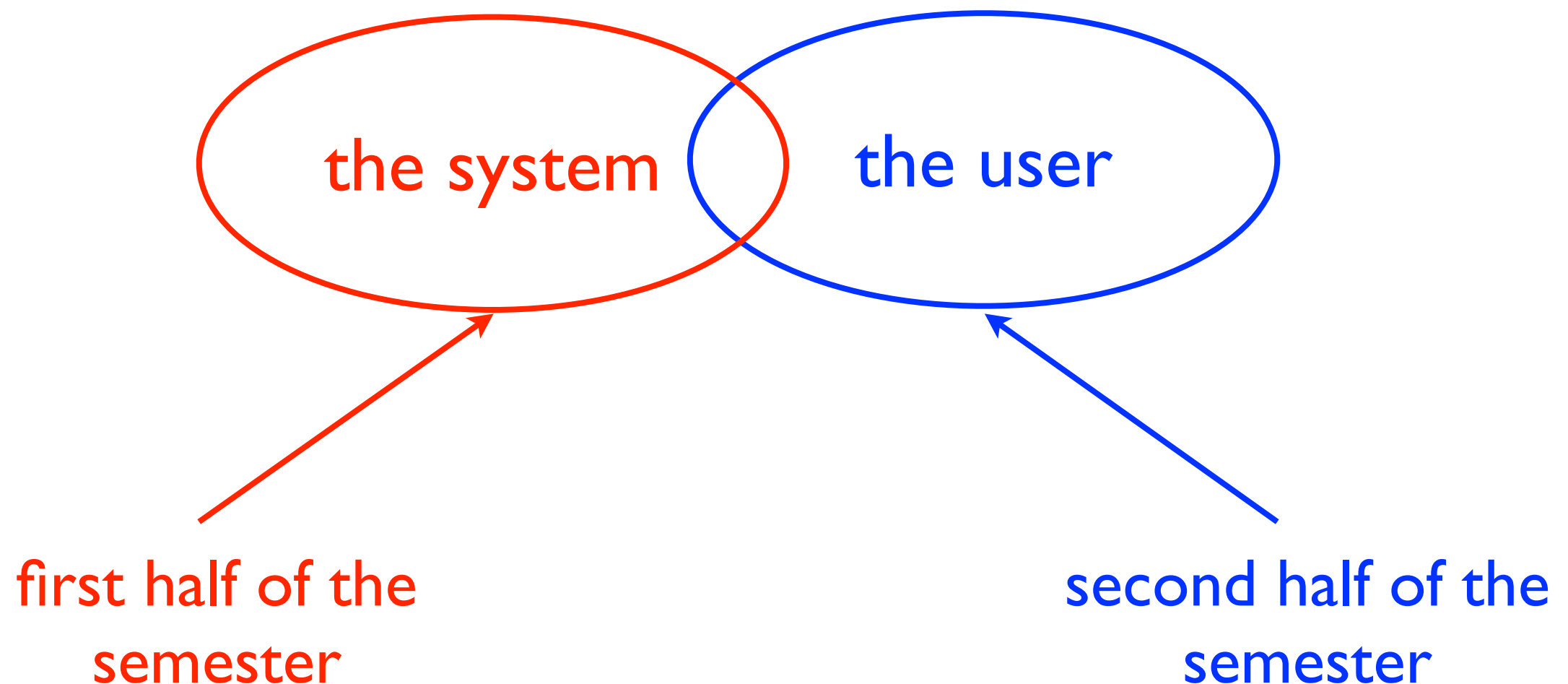
- Information retrieval is an interdisciplinary problem



- We need to understand both ends of the spectrum

# Course Structure

- IR: computer-based solutions to a human problem



- Understanding IR systems requires math!

# Road Map

- Introduction to ad-hoc retrieval
  - ▶ controlled vocabularies
  - ▶ full-text indexing
- Boolean retrieval
- Indexing and query processing
- Statistical Properties of Text
- Document Representation
- Retrieval Models
  - ▶ vector space model
  - ▶ language modeling
  - ▶ others (depending on how quickly we progress)

# Road Map

- Evaluation
  - ▶ test-collection construction
  - ▶ evaluation metrics
  - ▶ experimentation
  - ▶ user studies
  - ▶ search-log analysis
- Studies of search behavior
- Federated Search
- Clustering
- Text Classification

# Grading

- 30% homework
  - ▶ 10% each
- 15% midterm
- 15% final exam
- 30% literature review
  - ▶ 5% proposal
  - ▶ 10% presentation
  - ▶ 15% paper
- 10% (and chocolates) participation

# Grading for Graduate Students

- H: 95-100%
- P: 80-94%
- L: 60-79%
- F: 0-59%

# Grading for Undergraduate Students

- A+: 97-100%
- A: 94-96%
- A-: 90-93%
- B+: 87-89%
- B: 84-86%
- B-: 80-83%
- C+: 77-79%
- C: 74-76%
- C-: 70-73%
- D+: 67-69%
- D: 64-66%
- D-: 60-63%
- F:  $\leq 59\%$

# Homework vs. Midterm vs. Final



- The homework will be challenging. It should be, you have more time.

# Literature Review

- See description on the syllabus
- Form groups of 2 or 3
- Choose an IR task (next slide)
- Write a short proposal (mostly for feedback)
- Review the literature
  - ▶ not just the different solutions to the problem
  - ▶ the best solutions to the problem!
- Write a paper (~30 pages double-spaced)
- Make a presentation
  - ▶ 10 minute presentation + 5 minutes Q&A

# Literature Review

## example tasks

- Personalized information retrieval
- Session-based information retrieval
- Clustering of search results
- Book search
- Multimedia search (over items not inherently associated with text)
- Social-media data for forecasting and event-detection
- Query-log analysis for forecasting and event-detection
- Faceted search
- Federated search

# Literature Review

## tips

- Be thorough
- Be scientific
  - ▶ don't focus on the writing of the papers you review
  - ▶ focus on the science (the method and the evaluation)
- Be constructive
- Contribute new insight and structure
  - ▶ your literature review shouldn't read like a "list"
  - ▶ connect dots that haven't been connected
- Say what you think!

# Course Tips

- Work hard
- Do the assigned readings
- Do other readings
- Be patient and have reasonable expectations
  - ▶ you're not supposed to understand everything we cover in class during class
- Seek help sooner rather than later
  - ▶ office hours: manning 305, T, Th 9:30-10:30am
  - ▶ questions via email
- Keep laptop usage to a minimum (live in the present)

Questions?