

# **Week 1:**

# **Introduction to Information Retrieval**

Professor Ingemar J. Cox  
Computer Science, UCL

# Acknowledgement

Course slides are based on a combination of material from Jaime Arguello's course on “Information Retrieval” at the University of North Carolina.

[https://ils.unc.edu/courses/2017\\_fall/inls509\\_002/lectures](https://ils.unc.edu/courses/2017_fall/inls509_002/lectures)

and

Pandu Nayak and Prabhakar Raghavan course on  
“Information Retrieval and Web Search

<https://web.stanford.edu/class/cs276/handouts/lecture6-tfidf.ppt>

# WHAT IS INFORMATION RETRIEVAL?

# Information retrieval

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

Gerald Salton, “Automatic Information Organization and Retrieval”, McGraw-Hill, 1968.

# INFORMATION RETRIEVAL: STRUCTURE

# Information retrieval: structure

Structure of document

Structure of collection

# Information retrieval: Structure

## United States Patent Office

1  
3,035,736  
**RESEARCH AND INFORMATION RETRIEVAL SYSTEM**  
Walter S. Pawl, 10480 Powder Mill Road, Adelphi, Md.  
Filed Apr. 1, 1960, Ser. No. 19,367  
10 Claims. (Cl. 221—120)

The present invention relates to information retrieval systems involving answers to corresponding specific questions in a wide field of knowledge covered by references.

The main feature and object of the invention is to provide immediate information which is exhaustive on any sufficiently specific question to be contained on a card of a selected size, the information including besides the information itself, references to source authorities and publications containing further pertinent information.

A further object is to make this system as compact and practicable as possible.

Other and more specific objects will become apparent

3,035,736  
Patented May 22, 1962

2  
of knowledge to solve problems presented in the course of working on any research project, not only in answer to questions as to what has already been done or discovered but what can or might be further done to solve a specific problem.

The present system is applicable to any field of knowledge covered by many volumes of books or thousands of paragraphs, extracted on an information card from a topical question or title, to answer or about which it is relevly. Each of the above paragraphs may be contained in other paragraphs where further or citations might be found,

10 may be contained on a card that can be instantly obtained by dialing a number given to it in an alphabetical listing of these topics or titles, and the present system is intended to be a great step in providing a most economical and instant retrieval of the latest information, and will speed up the solutions of many problems now requiring hundreds of volumes of books and a vast amount of valuable time in fountering through them in search

# Information Retrieval: Structure

AGRIS

Source

Information Systems Division, National Agricultural Library ([click here for contact information](#))  
The National Agricultural Library is one of four national libraries of the United States, with locations in Beltsville, Maryland and Washington, D.C. It houses one of the world's largest and most accessible agr [...]

HOME PAGE: <http://www.nal.usda.gov/>

automatic derivation of information retrieval encodings from machine-readable texts [1959]

Luhn, H. P. (Hans Peter)  
1896-1964  
International Business Machines Corporation [Corporate Author]  
Advanced Systems Development Division. [Corporate Author]

AGRIS SEARCH Find resources... Register Sign in

Other subjects

Navigation text

AGRIS

Food and Agriculture Organization of the United Nations

English Español Français العربية 中文 Русский

AGRIS

Find resources...

automatic derivation of information retrieval encodings from machine-readable texts

Register Sign in

Access the full text: NOT AVAILABLE

Save as:

AGRIS\_AP RIS EndNote(XML)

Lookup the document at:

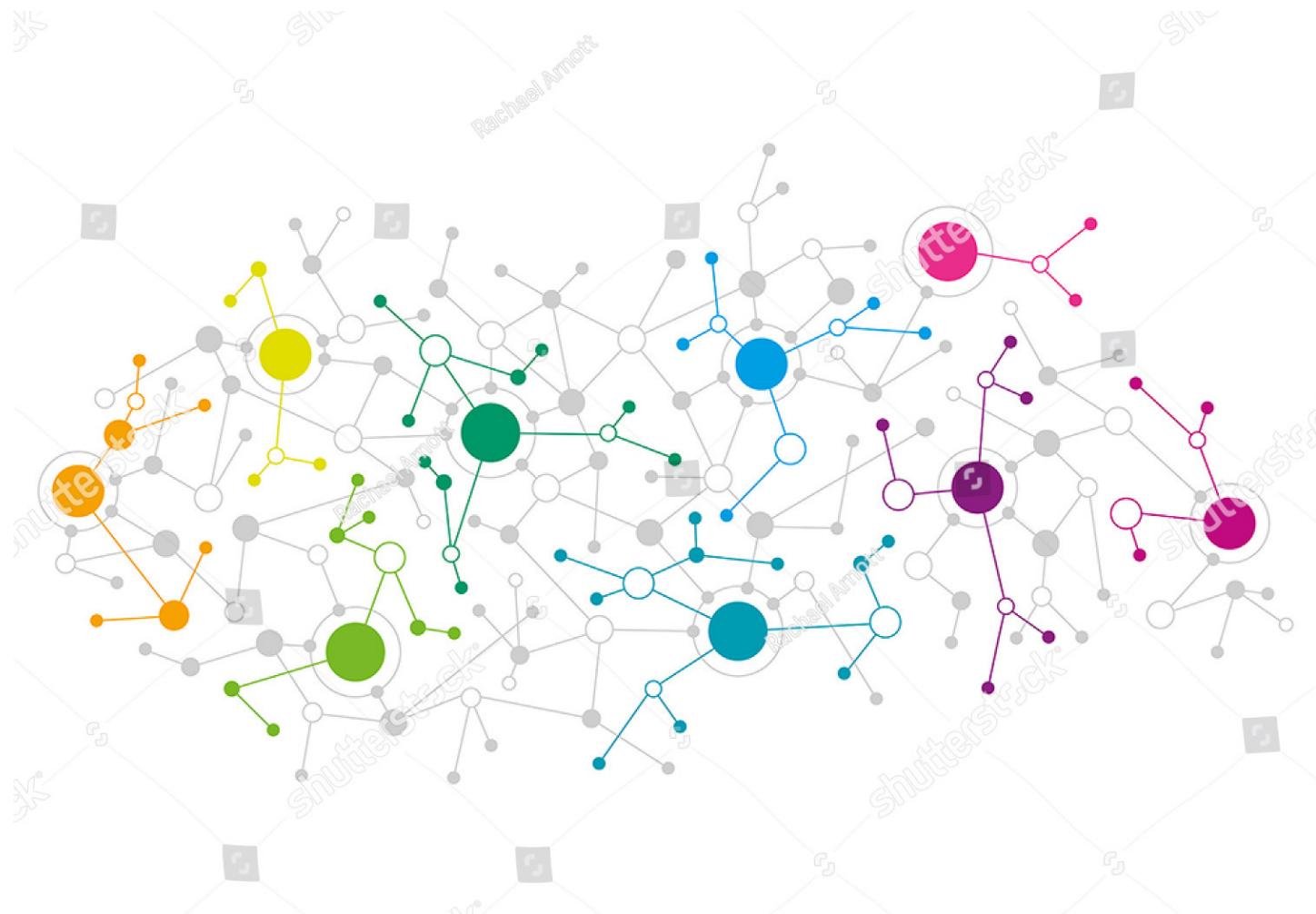
Google scholar

Translate with Google translator

Select Language ▾

This translation tool is powered by Google. FAO is not responsible for the accuracy of translations.

# Information retrieval: Structure



shutterstock®

IMAGE ID: 222472633  
www.shutterstock.com

# INFORMATION RETRIEVAL: ANALYSIS

# Information retrieval: Analysis

Classification

Topic

Reviews – positive/negative

Reading level

# INFORMATION RETRIEVAL: ORGANISATION

# Information retrieval: Organisation

Cataloguing/tagging

Topic

Location

Likes

# INFORMATION RETRIEVAL: STORAGE

# Information retrieval: Storage

Inverted index

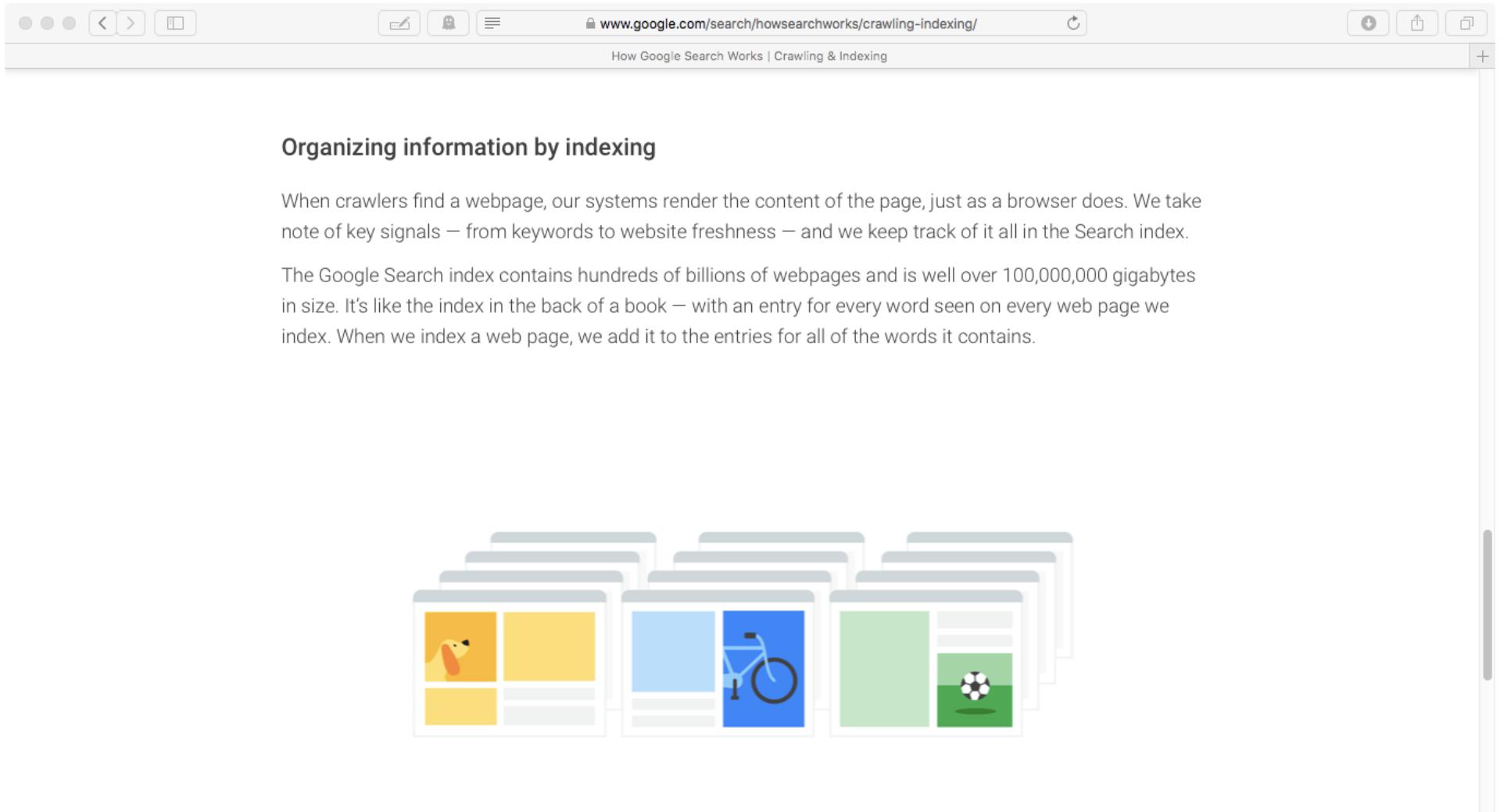
Distributed storage

Peer-to-peer

Privacy, security, censorship

...

# Information retrieval: Storage

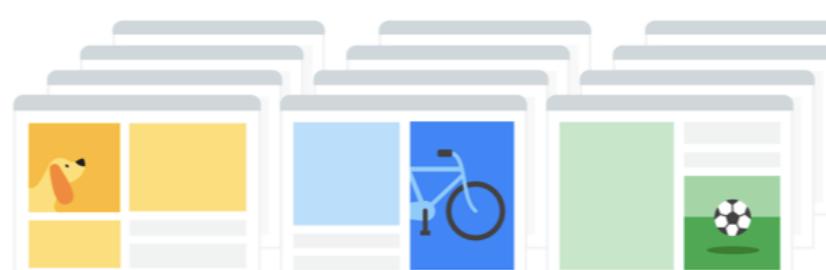


The screenshot shows a web browser window with the URL [www.google.com/search/howsearchworks/crawling-indexing/](https://www.google.com/search/howsearchworks/crawling-indexing/) in the address bar. The page title is "How Google Search Works | Crawling & Indexing". The main content on the page is titled "Organizing information by indexing". It contains two paragraphs of text. Below the text is a diagram of three filing cabinets, each with multiple drawers, containing icons of a dog, a bicycle, and a soccer ball.

**Organizing information by indexing**

When crawlers find a webpage, our systems render the content of the page, just as a browser does. We take note of key signals — from keywords to website freshness — and we keep track of it all in the Search index.

The Google Search index contains hundreds of billions of webpages and is well over 100,000,000 gigabytes in size. It's like the index in the back of a book — with an entry for every word seen on every web page we index. When we index a web page, we add it to the entries for all of the words it contains.



# INFORMATION RETRIEVAL: RETRIEVAL

# Information retrieval: Retrieval

- Efficiency
  - Computational cost
  - Response time
  - Storage
- Effectiveness
  - User satisfaction

# Information retrieval: Retrieval

Half a second delay caused a 20% drop in traffic. Half a second delay killed user satisfaction

Geeking with Greg: Marissa Mayer at Web 2.0

Ce site utilise des cookies provenant de Google afin de fournir ses services, personnaliser les annonces et analyser le trafic. Les informations relatives à votre utilisation du site sont partagées avec Google. En acceptant ce site, vous acceptez l'utilisation des cookies.

EN SAVOIR PLUS OK I

Thursday, November 09, 2006

## Marissa Mayer at Web 2.0

Google VP Marissa Mayer just spoke at the Web 2.0 Conference and offered tidbits on what Google has learned about speed, the user experience, and user satisfaction.

Marissa started with a story about a user test they did. They asked a group of Google searchers how many search results they wanted to see. Users asked for more, more than the ten results Google normally shows. More is more, they said.

So, Marissa ran an experiment where Google increased the number of search results to thirty. Traffic and revenue from Google searchers in the experimental group dropped by 20%.

Ouch. Why? Why, when users had asked for this, did they seem to hate it?

After a bit of looking, Marissa explained that they found an uncontrolled variable. The page with 10 results took .4 seconds to generate. The page with 30 results took .9 seconds.

Half a second delay caused a 20% drop in traffic. Half a second delay killed user satisfaction.

This conclusion may be surprising -- people notice a half second delay? -- but we had a similar experience at Amazon.com. In A/B tests, we tried delaying the page in increments of 100 milliseconds and found that even very small delays would result in substantial and costly drops in revenue.

Being fast really matters. As Marissa said in her talk, "Users really respond to speed."

**About Me**  
GREG LINDEN  
[View my complete profile](#)

**Subscribe to the Feed**  
[Subscribe in a reader](#)

**More Geekin' with Me**  
[Tweeting with Greg](#)  
[Geekin' on G+](#)

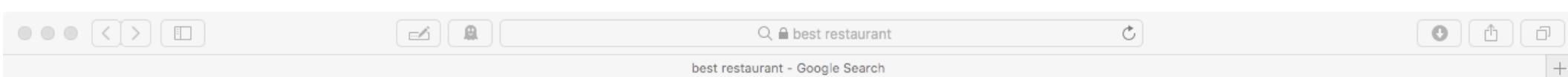
**Blog Archive**

- 2017 (6)
- 2016 (11)
- 2015 (12)
- 2014 (17)
- 2013 (10)
- 2012 (15)
- 2011 (24)
- 2010 (60)

# Information retrieval: Retrieval

## Effectiveness

# Information retrieval: Retrieval Effectiveness



## The 10 Best Wilmslow Restaurants - TripAdvisor

[https://www.tripadvisor.co.uk/Restaurants-g580421-Wilmslow\\_Cheshire\\_England.html](https://www.tripadvisor.co.uk/Restaurants-g580421-Wilmslow_Cheshire_England.html) ▾  
Carters Arms. 74 Reviews. £, Bar, British, Pub. Wilmslow. The Big Fish. 120 Reviews. £, Fast food, British, Seafood. Wilmslow. Istanblue. 49 Reviews. £, Cafe, Turkish. Wilmslow. fosters fish and chip shop. 286 Reviews. £, Seafood, Fast food, British. Alderley Edge 1.9 mi away.  
Vegan Options · The Best Tapas in Wilmslow · Halal · Gluten Free Options

## Best Restaurants Near Me - TripAdvisor

<https://www.tripadvisor.co.uk/Restaurants> ▾  
Martin Berasategui. 1,344 Reviews. Lasarte, Spain. Maison Lameloise. 1,162 Reviews. Chagny, France. Belmond Le Manoir aux Quat'Saisons. 2,050 Reviews. Great Milton, UK. L'Auberge de l'Ill. 1,298 Reviews. Illhaeusern, France.

## The 10 Best Manchester Restaurants - TripAdvisor

[https://www.tripadvisor.co.uk/.../England/Greater\\_Manchester/Manchester](https://www.tripadvisor.co.uk/.../England/Greater_Manchester/Manchester) ▾  
Reserve a table for the best dining in Manchester, Greater Manchester on TripAdvisor: See 300698 reviews of 2361 Manchester restaurants and search by ...

## Black Swan country pub in Yorkshire named world's best restaurant ...

[https://www.theguardian.com/World/UK\\_News/Yorkshire](https://www.theguardian.com/World/UK_News/Yorkshire) ▾  
11 Oct 2017 - A country pub in North Yorkshire has been named the best restaurant in the world based on a poll of customer reviews. The Black Swan in ...

## The World's 50 Best Restaurants

[www.theworlds50best.com/](http://www.theworlds50best.com/) ▾  
Bilbao 2018; Next stop on The World's 50 Best Restaurants global tour; The World's 50 ... Septime; Winner of The Sustainable Restaurant Award 2017; Bertrand ...

## 1-50 The Worlds 50 Best Restaurants - The World's 50 Best Restaurants

[www.theworlds50best.com/list/1-50-winners](http://www.theworlds50best.com/list/1-50-winners) ▾  
The World's 50 Best Restaurants 1-50. 2017. 1-10; 11-20; 21-30; 31-40; 41-50; 1- 50; 51-100;  
Individual Awards. No.1; Eleven Madison Park New York, USA ...

# Information retrieval: Retrieval

Effectiveness

Precision

Recall

# Precision

The proportion of retrieved documents that are relevant

# Precision

number of retrieved docs that are relevant  
\_\_\_\_\_  
number of retrieved docs

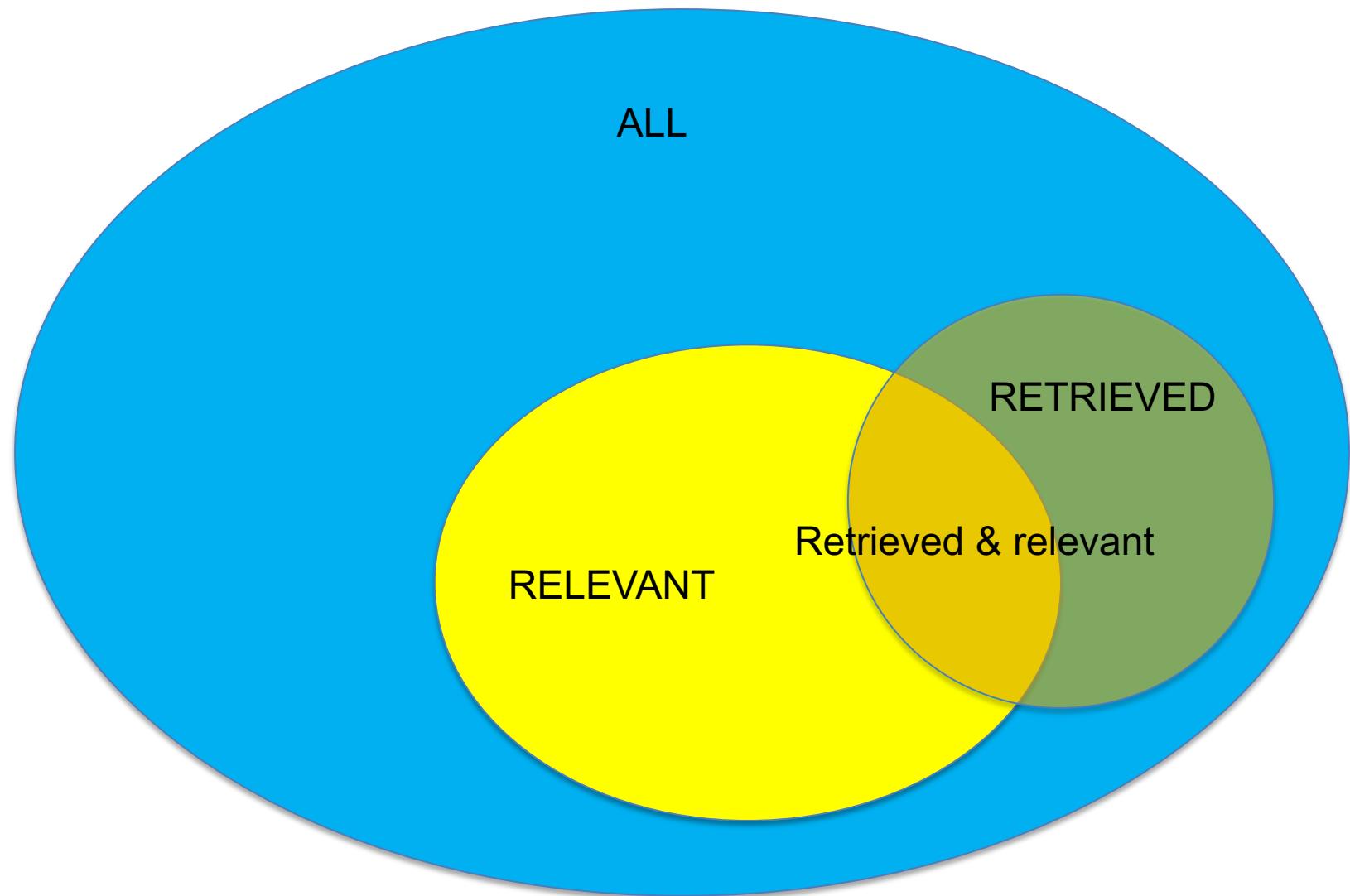
# Recall

The proportion of relevant documents actually retrieved

# Recall

number of retrieved docs that are relevant  
\_\_\_\_\_  
total number of relevant docs

# Precision and Recall



# Information retrieval

Crawling

Indexing

Boolean search

Ranking

# Types of search engines

- Web
- Enterprise
- Desktop
- Specialized
  - Patent search
  - Expert search
  - Music
  - Video
  - Dating

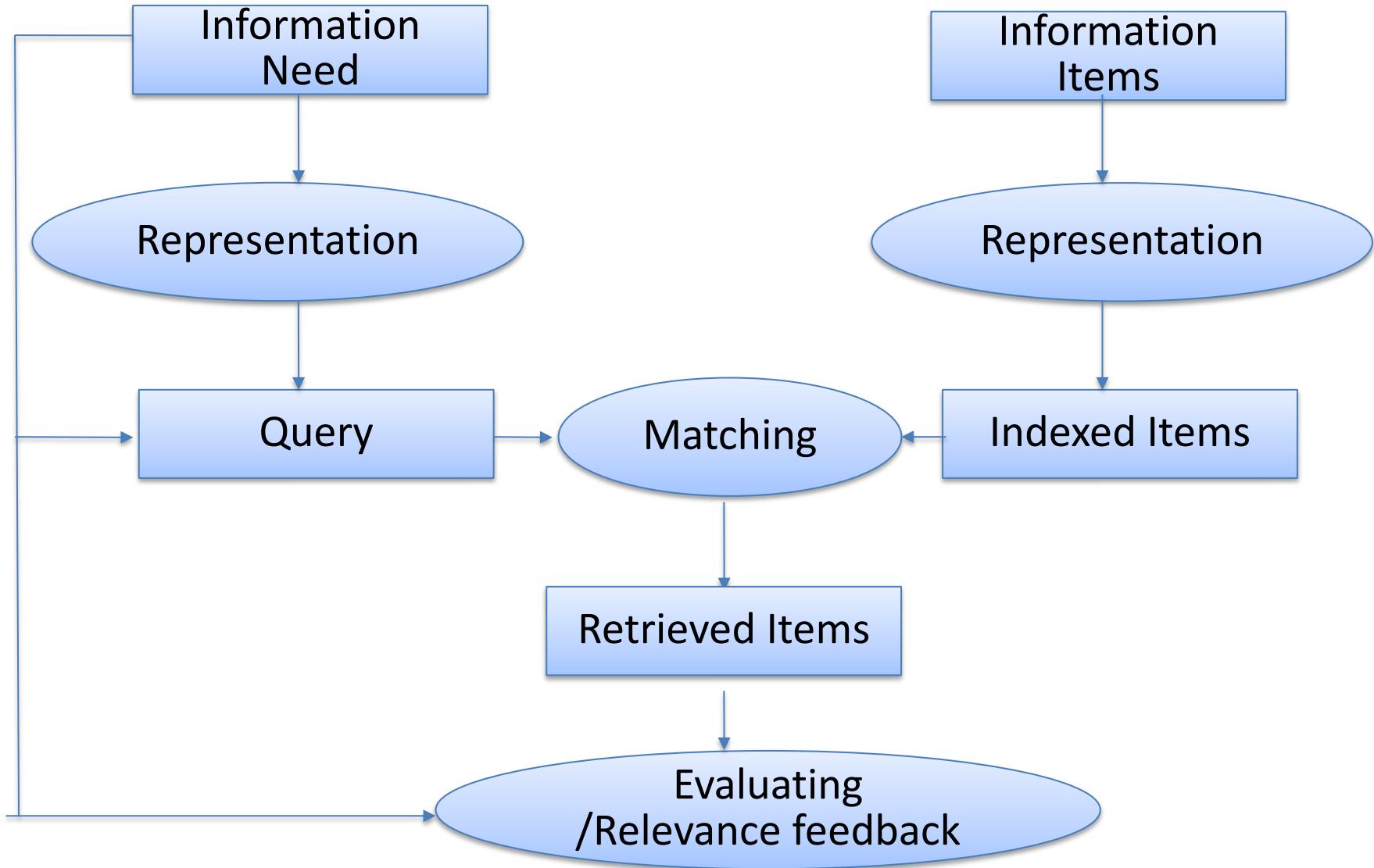


# Types of search engines

- Web
- Enterprise
- Desktop
- Specialized
  - Patent search
  - Expert search
  - Music
  - Video
  - Dating

User satisfaction

# Basic elements



# Information items

- Text
  - Patents
  - Hypertext doc
  - Email
  - ...

# Information items

- Images
- Video
- Music
- ...

# Information need

Only the user knows

# An example of an information need

Who is the writer responsible for the line “Where there’s life, there’s threat.” and is this a play on the Roman “Where there’s life there’s hope”?



Who is the writer responsible for the line “Where there's life, there's threat.” and is t X



All Images News Videos Shopping More Tools

About 2,830 results (0.67 seconds)

No results found for Who is the writer responsible for the line “Where there's life, there's threat.” and is this a play on the Roman “Where there's life there's hope”?

Results for Who is the writer responsible for the line Where theres life, theres threat. and is this a play on the Roman Where theres life theres hope? (without quotes):

[https://issuu.com/docs/new\\_hope\\_5](https://issuu.com/docs/new_hope_5) ::

### Where's There's Hope, There's Life by Teaching English Ideas ...

Aug 3, 2020 — “Christ, Pino, and what the hell's Mauro doing **there**?” ... The Morandi Bridge was the city's **life-line** to and from Milan and the north and ...

<https://www.jstor.org/stable/> ::

### The Scandal of the Arena - jstor

by CA Barton · 1989 · Cited by 328 — tiarius (the hunter of the arena) provoked Cyprian, **writing** in the third century ... In all of **Roman life** there was no more severe commitment that could be.

<https://www.joyofquotes.com/inspirational-quotes-by-...> ::

### Inspirational Quotes to Live By: Listed by Author

The grand essentials to happiness in this **life** are something to do, something to love ... **There** is no greater agony than bearing an untold **story** inside you.

[https://www.preceptaustin.org/gods\\_word\\_of\\_hope](https://www.preceptaustin.org/gods_word_of_hope) ::

### God's Word of Hope | Precept Austin

Jan 22, 2020 — Where **there's life, there's hope**. Peter teaches that the truth is exactly the opposite for where **there** is genuine Biblical hope, **there** is real ...

# Representation of documents

A document is an ordered sequence of words,  
i.e. a time series

# Representation of documents

Order matters.

The cat killed the mouse.

The mouse killed the cat.

So how do we represent documents?

# Representation of documents

- Bag of words
  - List of unique words in a document

# Bag of words

- “The brown fox jumped over the brown dog and the black dog”

Word	Frequency
and	1
brown	2
dog	2
fox	1
jumped	1
over	1
the	3

# Bag of words

- “The brown dog and the black dog jumped over the brown fox”

Word	Frequency
and	1
brown	2
dog	2
fox	1
jumped	1
over	1
the	3

# Bag of words

Does not capture meaning

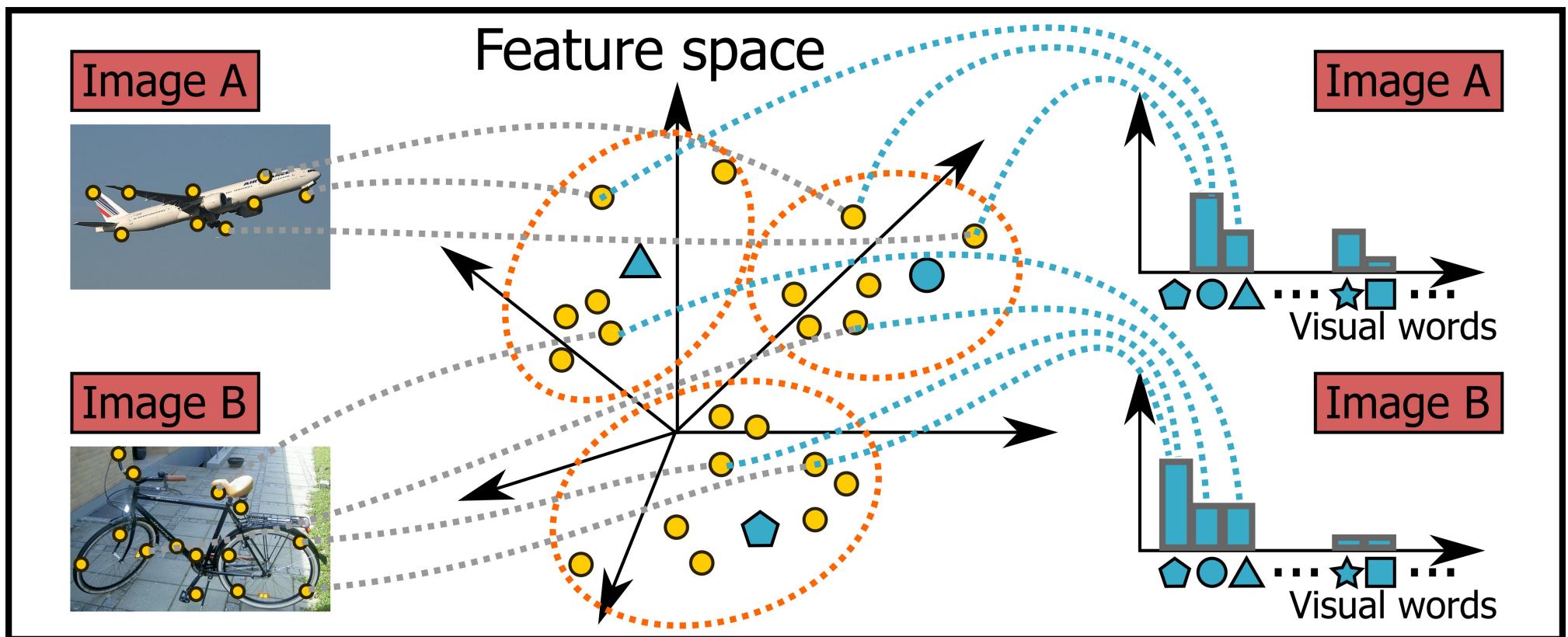
Purely statistical

Yet works surprisingly well

# Bag of features

- Same approach can be used for image and audio search
- Geometric and temporal information is discarded

# Bag of features



From “Bag of Features”, Yi Gao, Hsiang-Yun Wu, Kazuo Misue, Kazuyo Mizuno, and Shigeo Takahashi, The 7th International Symposium on Visualization & Interaction, 2014

# Representation of user need: Query

- Explicit
  - “best restaurant”
- Implicit – context
  - Location (local restaurants)
  - Time (breakfast, lunch, dinner)
  - User profile (John likes French cuisine)

# Query: “best restaurant”

The screenshot shows a web browser window with the search bar containing "best restaurant". The results are as follows:

- The 10 Best Wilmslow Restaurants - TripAdvisor**  
[https://www.tripadvisor.co.uk/Restaurants-g580421-Wilmslow\\_Cheshire\\_England.html](https://www.tripadvisor.co.uk/Restaurants-g580421-Wilmslow_Cheshire_England.html) ▾  
Carters Arms. 74 Reviews. £, Bar, British, Pub. Wilmslow. The Big Fish. 120 Reviews. £, Fast food, British, Seafood. Wilmslow. Istanblue. 49 Reviews. £, Cafe, Turkish. Wilmslow. fosters fish and chip shop. 286 Reviews. £, Seafood, Fast food, British. Alderley Edge 1.9 mi away.  
Vegan Options · The Best Tapas in Wilmslow · Halal · Gluten Free Options
- Best Restaurants Near Me - TripAdvisor**  
<https://www.tripadvisor.co.uk/Restaurants> ▾  
Martin Berasategui. 1,344 Reviews. Lasarte, Spain. Maison Lameloise. 1,162 Reviews. Chagny, France. Belmond Le Manoir aux Quat'Saisons. 2,050 Reviews. Great Milton, UK. L'Auberge de l'Illi. 1,298 Reviews. Illhaeusern, France.
- The 10 Best Manchester Restaurants - TripAdvisor**  
<https://www.tripadvisor.co.uk/.../England/Greater%20Manchester/Manchester> ▾  
Reserve a table for the best dining in Manchester, Greater Manchester on TripAdvisor: See 300698 reviews of 2361 Manchester restaurants and search by ...
- Black Swan country pub in Yorkshire named world's best restaurant ...**  
<https://www.theguardian.com/World/UK/News/Yorkshire> ▾  
11 Oct 2017 - A country pub in North Yorkshire has been named the best restaurant in the world based on a poll of customer reviews. The Black Swan in ...
- The World's 50 Best Restaurants**  
[www.theworlds50best.com/](http://www.theworlds50best.com/) ▾  
Bilbao 2018; Next stop on The World's 50 Best Restaurants global tour; The World's 50 ... Septime; Winner of The Sustainable Restaurant Award 2017; Bertrand ...
- 1-50 The Worlds 50 Best Restaurants - The World's 50 Best Restaurants**  
[www.theworlds50best.com/list/1-50-winners](http://www.theworlds50best.com/list/1-50-winners) ▾  
The World's 50 Best Restaurants 1-50. 2017. 1-10; 11-20; 21-30; 31-40; 41-50; 1-50; 51-100; Individual Awards. No.1; Eleven Madison Park New York, USA ...

# What is Relevance?

- **Relevance** is the “correspondence” between information needs (queries) and information items (documents, webpages, images etc)
- But, the exact meaning of relevance depends on applications:
  - = useful
  - = topical (about)
  - = interesting
  - = ?
- Predicting **relevance** is the central goal of IR

# Relevance

Abstract: This paper reports on a novel technique for literature indexing and searching in a mechanized library system. The notion of **relevance** is taken as a key concept in the theory of information retrieval.

Journal ACM, 1960

*The RAND Corporation, Santa Monica, California*

AND

J. L. KUHNS

*Ramo-Wooldridge, Canoga Park, California*

*Abstract.* This paper reports on a novel technique for literature indexing and searching in a mechanized library system. The notion of *relevance* is taken as the key concept in the theory of information retrieval and a comparative concept of relevance is explicated in terms of the theory of probability. The resulting technique called "Probabilistic Indexing," allows a computing machine, given a request for information, to make a statistical inference and derive a number (called the "relevance number") for each document, which is a measure of the probability that the document will satisfy the given request. The result of a search is an ordered list of those documents which satisfy the request ranked according to their probable relevance.

The paper goes on to show that whereas in a conventional library system the cross-referencing ("see" and "see also") is based solely on the "semantical closeness" between index terms, statistical measures of closeness between index terms can be defined and computed. Thus, given an arbitrary request consisting of one (or many) index term(s), a machine can elaborate on it to increase the probability of selecting relevant documents that would not otherwise have been selected.

Finally, the paper suggests an interpretation of the whole library problem as one where the request is considered as a clue on the basis of which the library system makes a concatenated statistical inference in order to provide as an output an ordered list of those documents which most probably satisfy the information needs of the user.

## 1. *Introduction*

One of the really remarkable characteristics of human beings is their ability to communicate with and operate on information formulated in ordinary language. We somehow are able to determine the meanings of words and sentences so as to make judgments about sameness of meaning, redundancy, inconsistency,

# Retrieval Models

- A retrieval model
  - abstracts away from the real world
  - is a mathematical representation of the essential aspects of a retrieval system
  - aims at computing relevance and retrieving relevant documents
  - thus, either explicitly or implicitly, defines relevance

# Predicting relevance

What information can we use to predict relevance?

# Predicting relevance

Query

Query context

User profile

Content-dependent information

Content-independent information

# Predicting relevance

Query “Mexican food”

Query context

User profile

Content-dependent information

Content-independent information

# Representation of information need

- Textual queries
  - Boolean:
    - “(information AND retrieval) OR (machine AND learning)”
  - free text: “movie matrix review”
- User profiles/preferences
  - set of rated movies
  - music playlist
- Query by example

# Evaluation

- Precision and Recall: two widely used measures for evaluating the quality of retrieval results
  - **Precision:** the number of relevant documents retrieved divided by the total number of documents retrieved
  - **Recall:** the number of relevant documents retrieved divided by the total number of existing relevant documents

# Evaluation

Precision and Recall are just two considerations  
What if the query is ambiguous, e.g. “jaguar”?

# Relevance feedback

Iterate search based on user feedback