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

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.

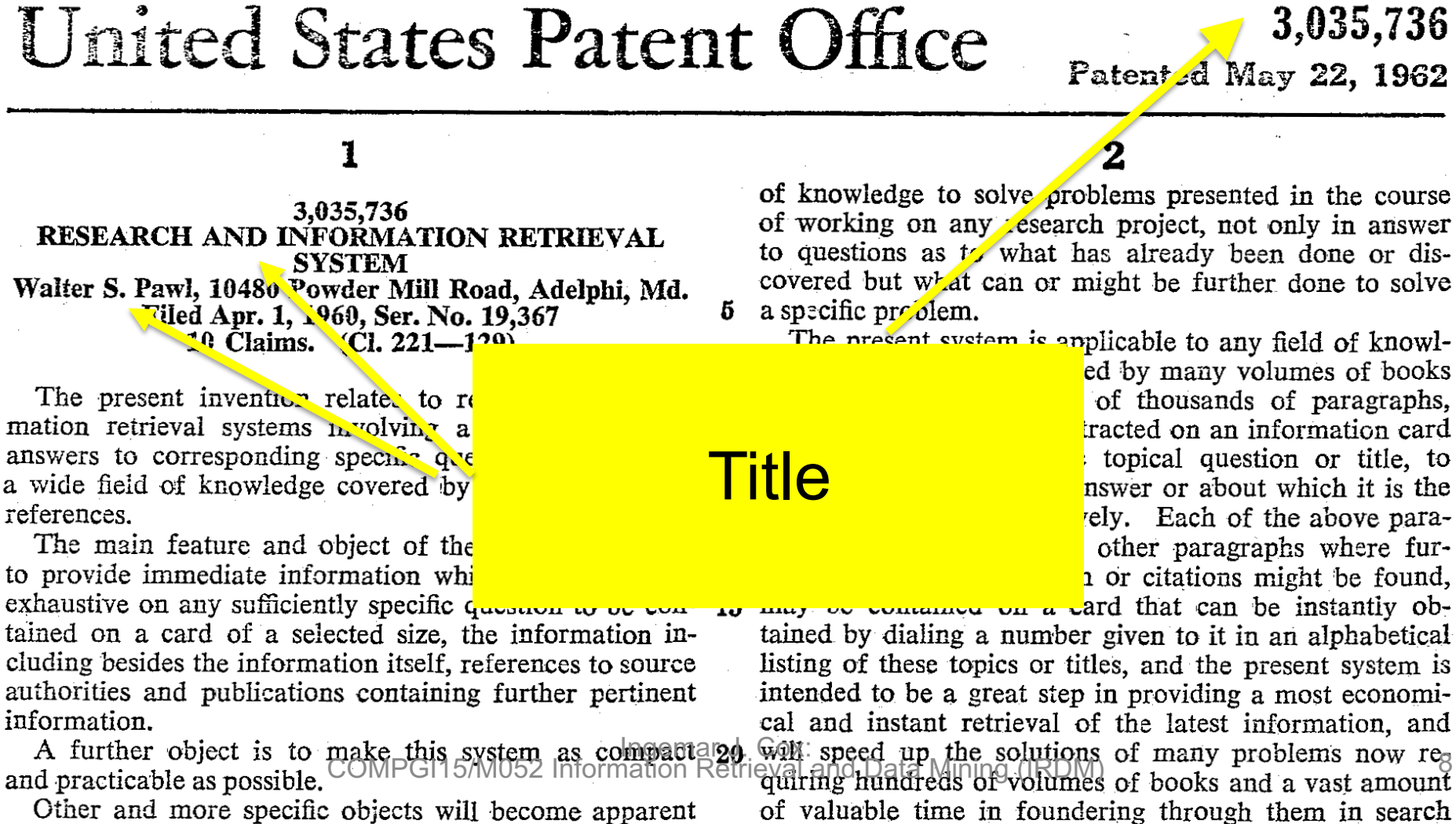
INFORMATION RETRIEVAL: STRUCTURE

Information retrieval: structure

Structure of document

Structure of collection

Information retrieval: Structure



United States Patent Office

3,035,736

Patented May 22, 1962

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—129)

The present invention relates to re
mation retrieval systems involving a
answers to corresponding specific qu
a wide field of knowledge covered by
references.

The main feature and object of the
to provide immediate information whi
exhaustive on any sufficiently specific question to be con
tained on a card of a selected size, the information in-
cluding 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

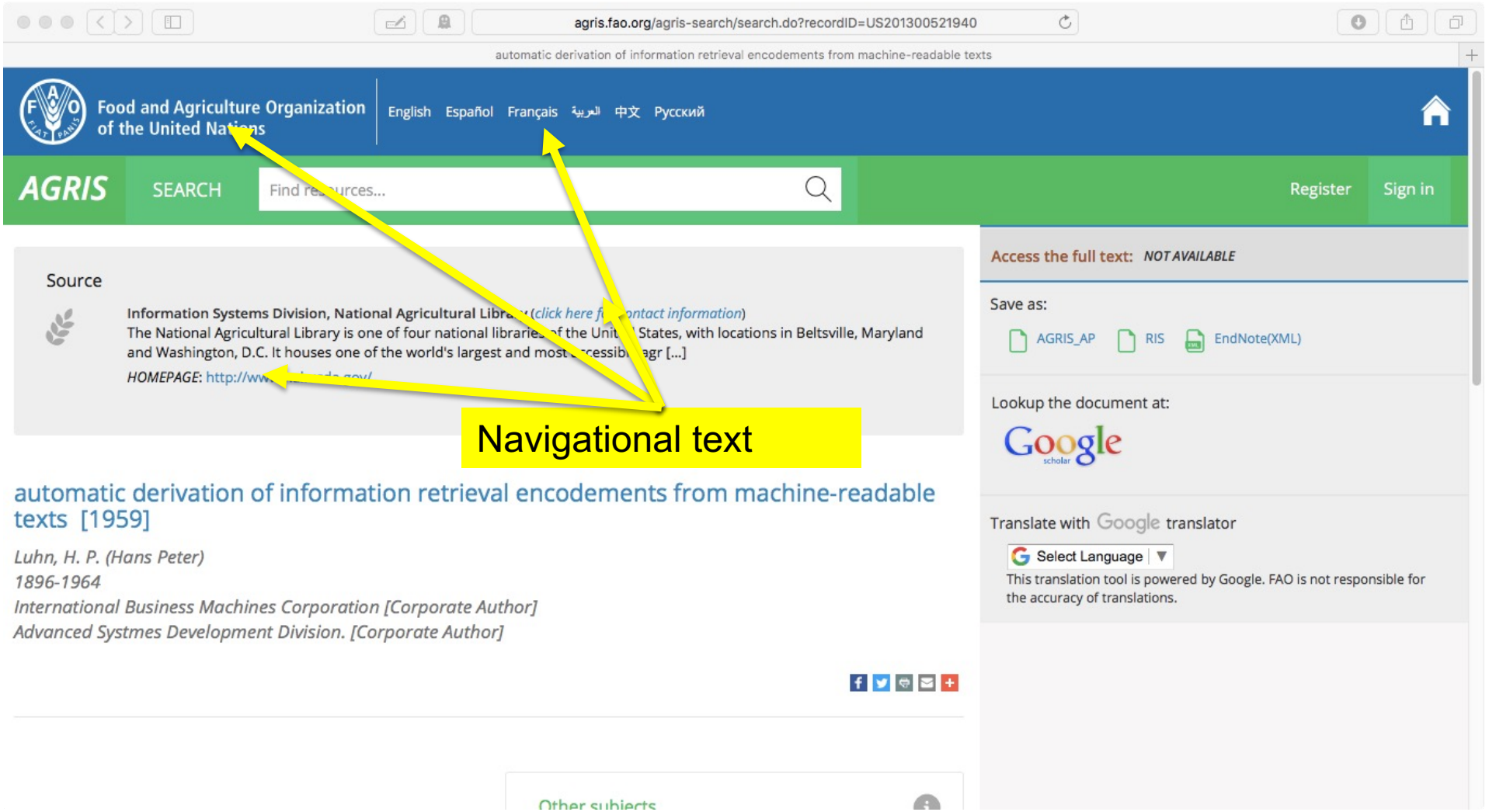
2

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 dis-
covered but what can or might be further done to solve
a specific problem.

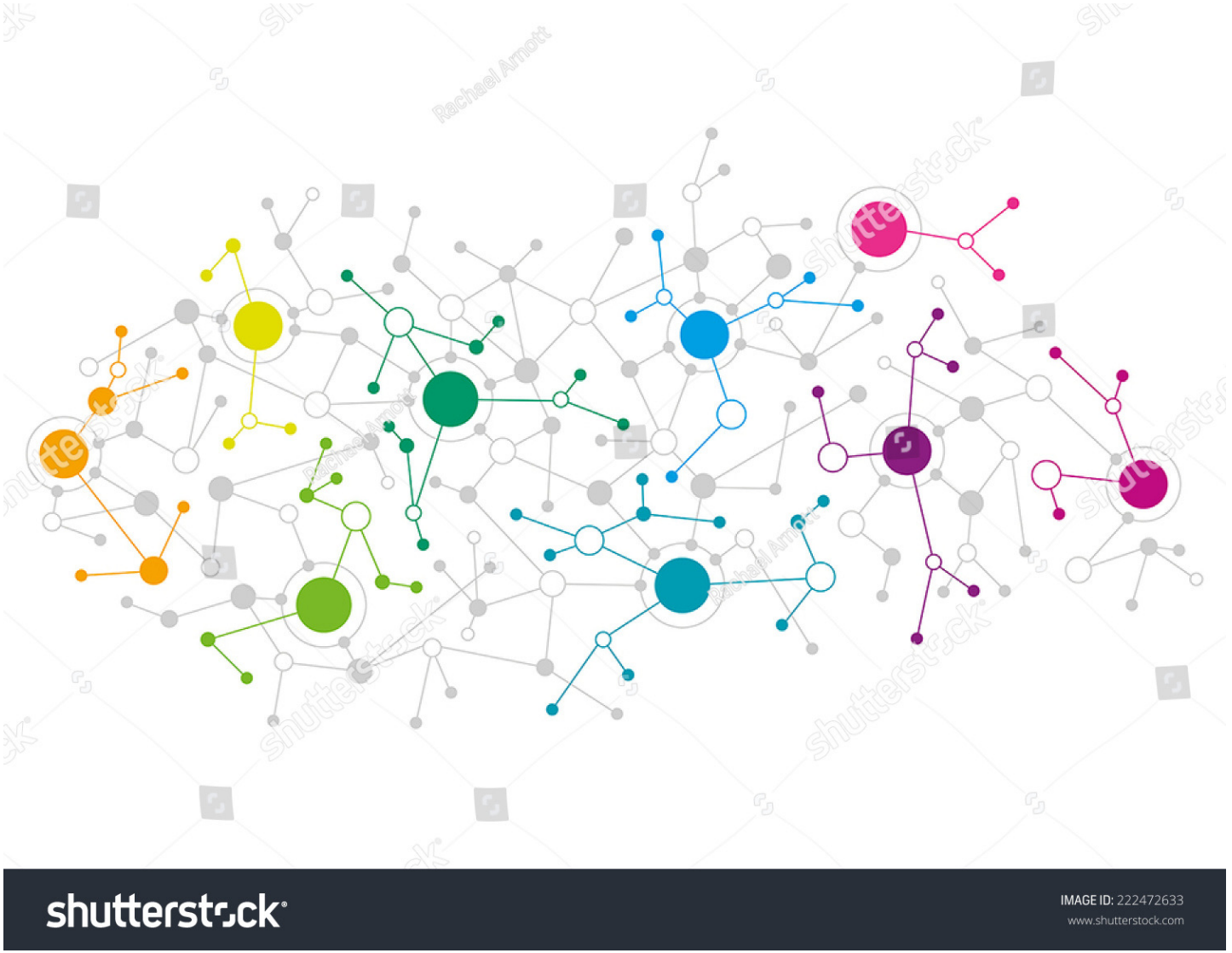
The present system is applicable to any field of knowl-
edge covered by many volumes of books
of thousands of paragraphs,
tracted on an information card
topical question or title, to
nswer or about which it is the
vely. Each of the above para-
other paragraphs where fur-
n or citations might be found,
ard that can be instantly ob-
tained 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 economi-
cal and instant retrieval of the latest information, and
will speed up the solutions of many problems now re-
quiring hundreds of volumes of books and a vast amount
of valuable time in foundering through them in search

Title

Information Retrieval: Structure



Information retrieval: Structure



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

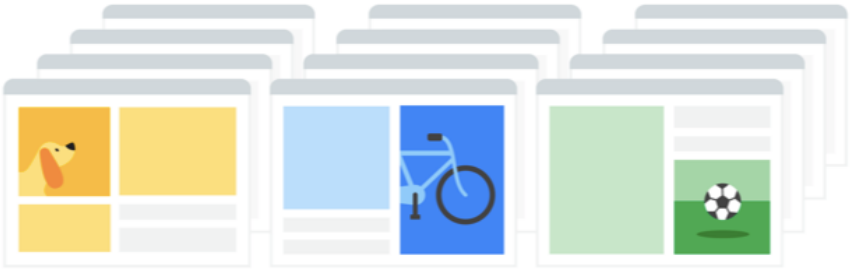
www.google.com/search/howsearchworks/crawling-indexing/

How Google Search Works | Crawling & Indexing

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.



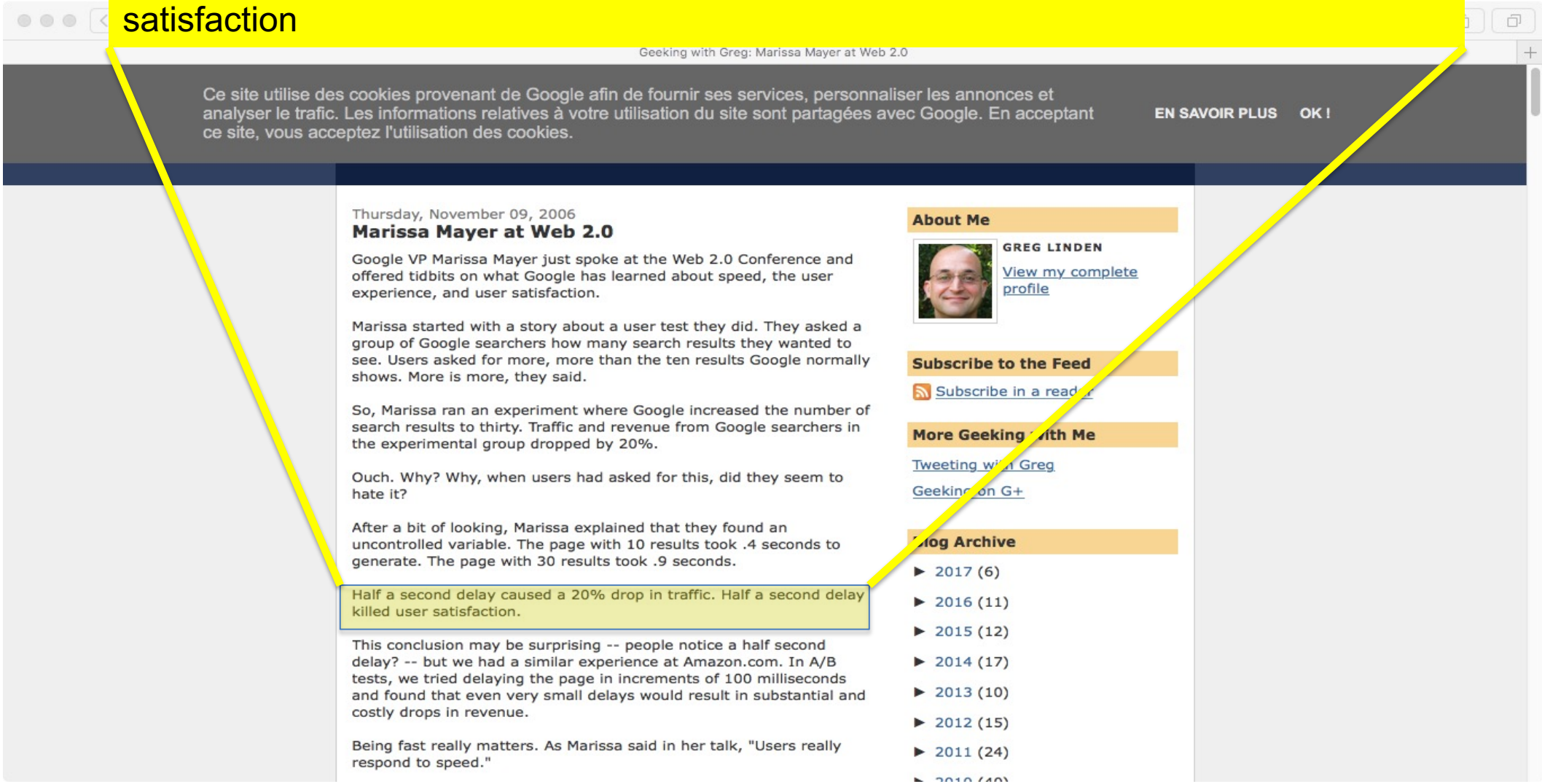
INFORMATION RETRIEVAL: RETRIEVAL

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

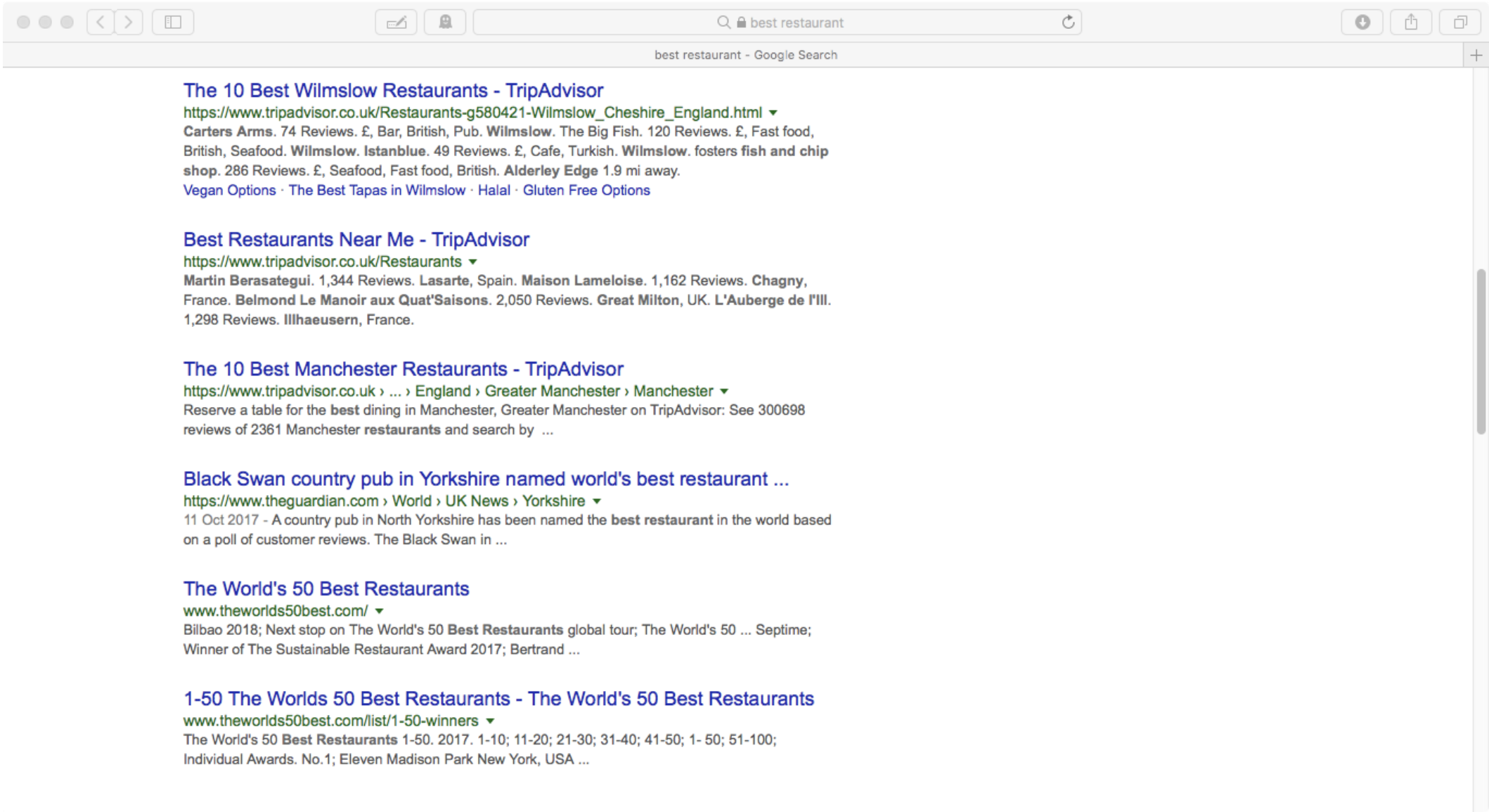


Information retrieval: Retrieval

Effectiveness

Information retrieval: Retrieval

Effectiveness



Information retrieval: Retrieval

Effectiveness

Precision

Recall

Precision

The proportion of retrieved documents that are relevant

Precision

$$\frac{\text{number of retrieved docs that are relevant}}{\text{number of retrieved docs}}$$

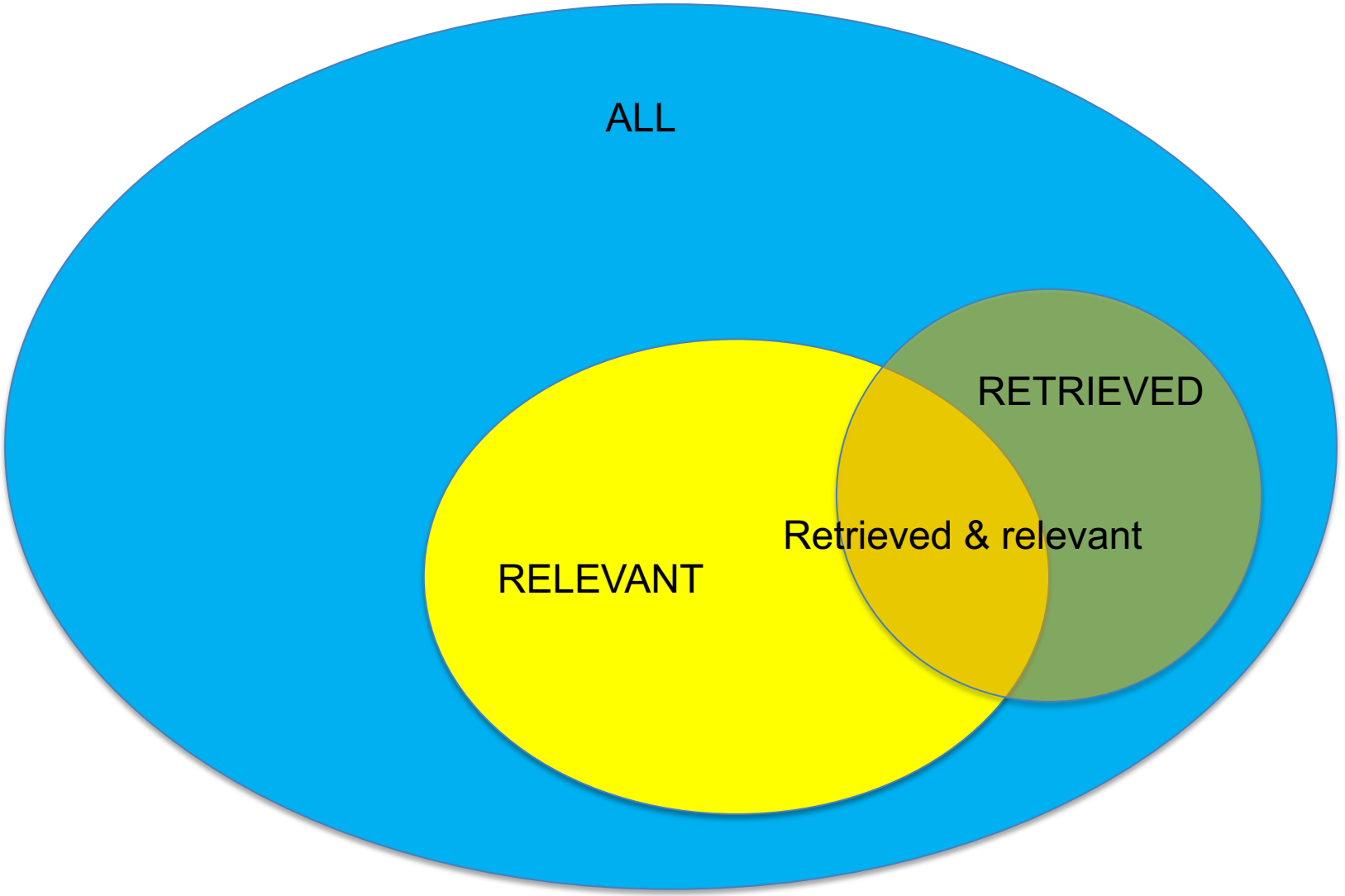
Recall

The proportion of relevant documents actually retrieved

Recall

$$\frac{\text{number of retrieved docs that are relevant}}{\text{total number of relevant docs}}$$

Precision and Recall



Information retrieval

Crawling

Indexing

Boolean search

Ranking

Types of search engines

- Web

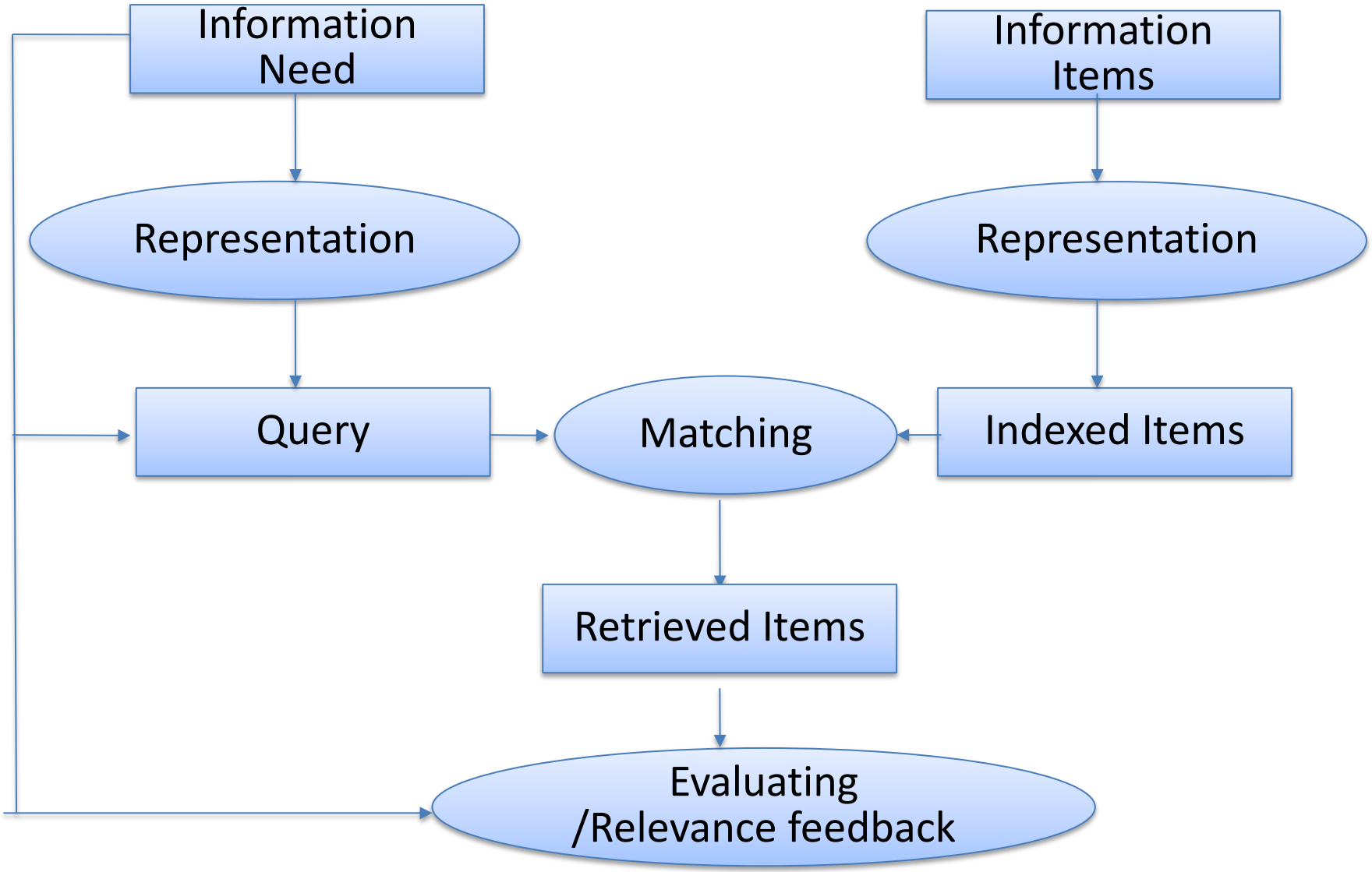
- Enterprise
- Desktop
- Specialized
  - Patent search
  - Expert search
  - Music
  - Video
  - Dating

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

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

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

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

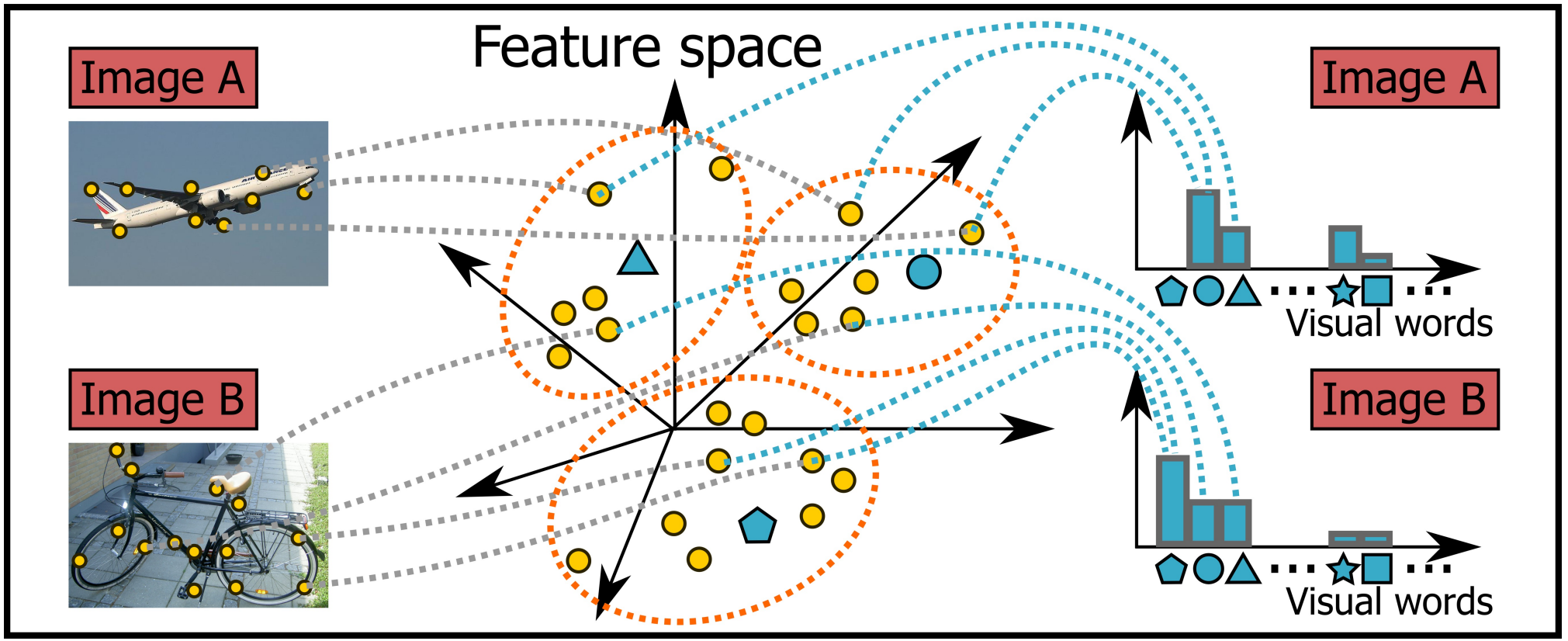
Purely statistical

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

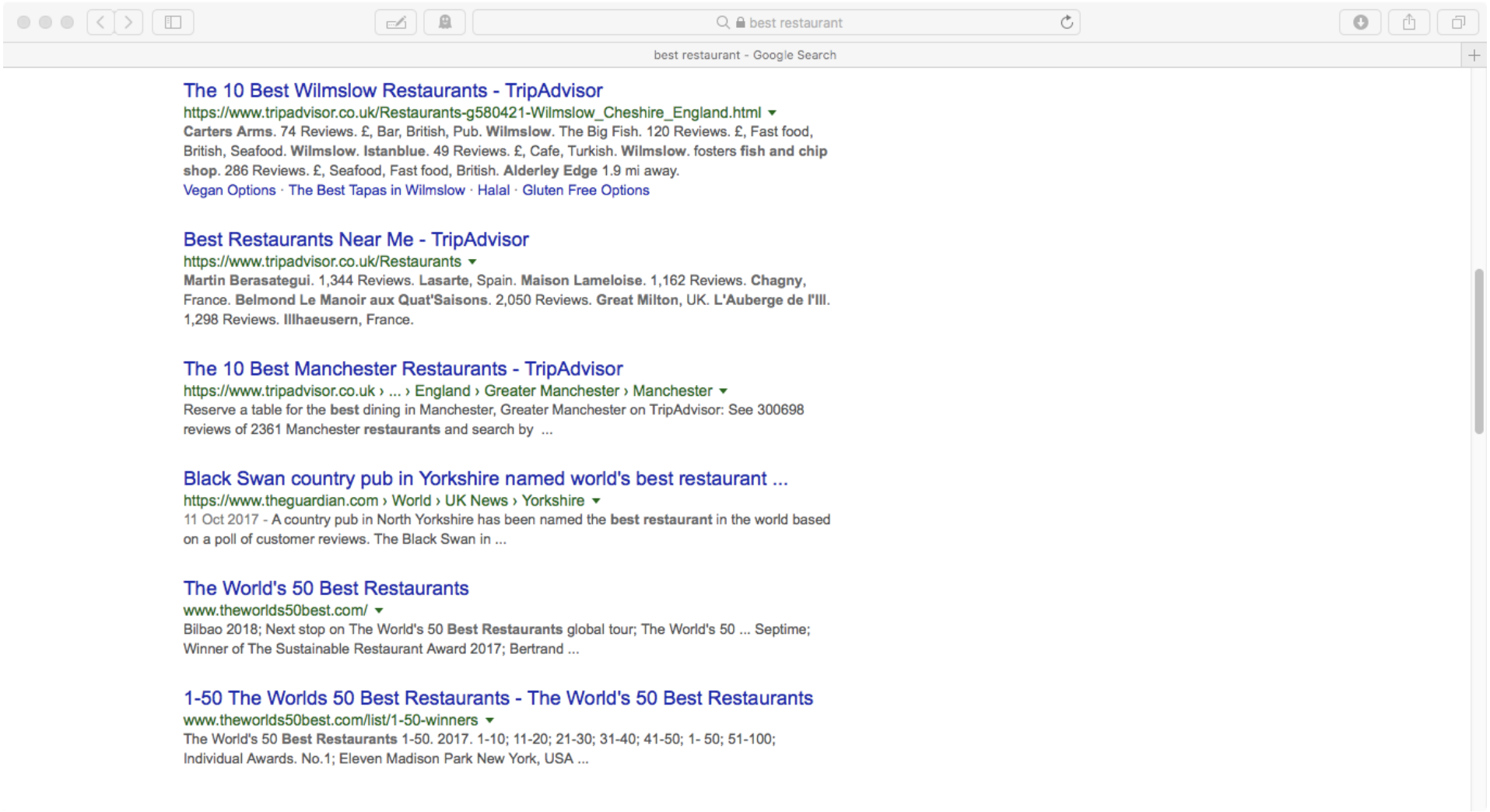


From “Bag of Features”, Yi Gao, Hsiang-Yun Wu, Kazuo Misue,Kazuyo Mizuno, and Shigeo Takahashi, The 7th International Symposium on Visualization & Interaction, 2014

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”



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