CS506/606 - Topics in Information Retrieval

Instructors: Steven Bedrick, Brian Roark,
Emily Prud'hommeaux

Class time: Tu/Th 11:00 a.m. - 12:30 p.m.
September 25 - December 6, 2012

Class location: WCC 403

Office hours: By appointment

Required texts: Readings will come from papers and texts online

http://www.cslu.ogi.edu/people/roark/courses/cse506-TIR/cse506TIRf12.html

Course overview

- Variety of topics in general area of information retrieval (IR).
- Initial lectures: Fundamentals of IR, with particular emphasis on applications of modern NLP techniques and evaluation.
- Remainder of course:
  - Seminar-style review of papers from the IR literature.
  - Visiting speakers.
- Example topics to be covered:
  - practical issues related to web-crawling
  - indexing of raw text and other data, e.g., ASR word lattices
  - query expansion and suggestion methods
  - issues in IR evaluation

Readings

- Most readings will be articles available on the website.
- In addition, some online book chapters will be assigned:

  Christopher Manning, Prabhakar Raghavan P, and Hinrich Schütze H. *Introduction to Information Retrieval.* Cambridge University Press, 2008. `http://www-nlp.stanford.edu/IR-book/`

  Marti Hearst. *Search User Interfaces.* Cambridge University Press, 2009. `http://www.searchuserinterfaces.com`

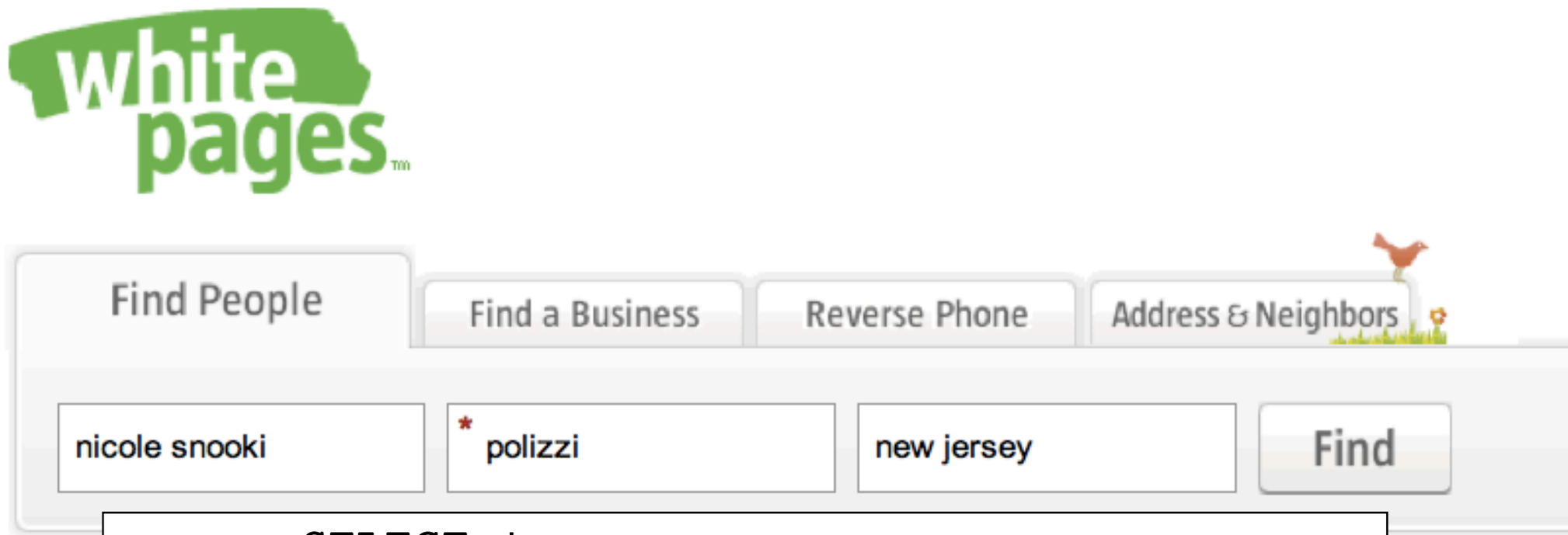
Required coursework

- Active participation in class discussions of research papers (10%)
- Lead discussion of papers in one or more sessions (30%).
- One or more homework assignments involving implementation and evaluation of components of IR systems (20%).
- Term project involving IR (to be approved by one of us, 40%).
  - Cross-language IR
  - Distributed IR
  - Speech IR
  - IR applications

Information retrieval

Manning et al. “Information retrieval (IR) is finding material ... of an **unstructured** nature ... that satisfies an information need from within large collections ...”

A note on structured information

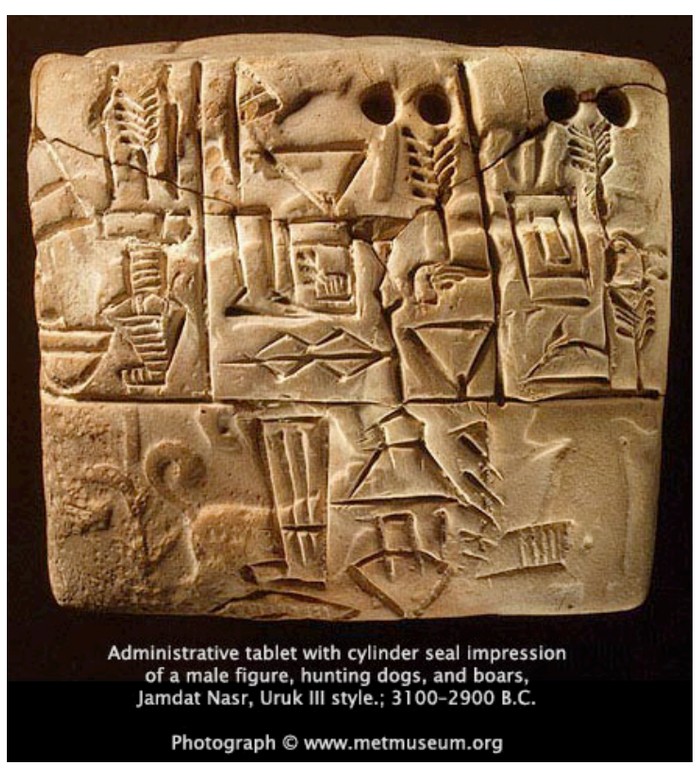


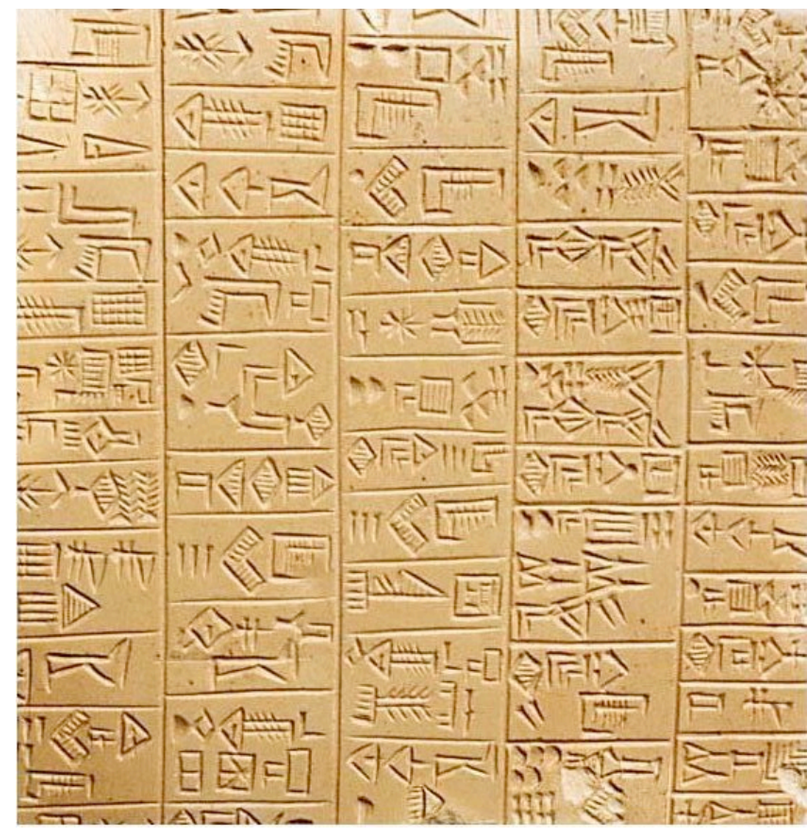

```
SELECT *
FROM whitepages
WHERE lastname='polizzi'
AND firstname='nicole snooki'
AND state='new jersey'
```

In this class, we'll be talking about unstructured information rather than structured information (e.g., information stored in a database).

Information retrieval of yore

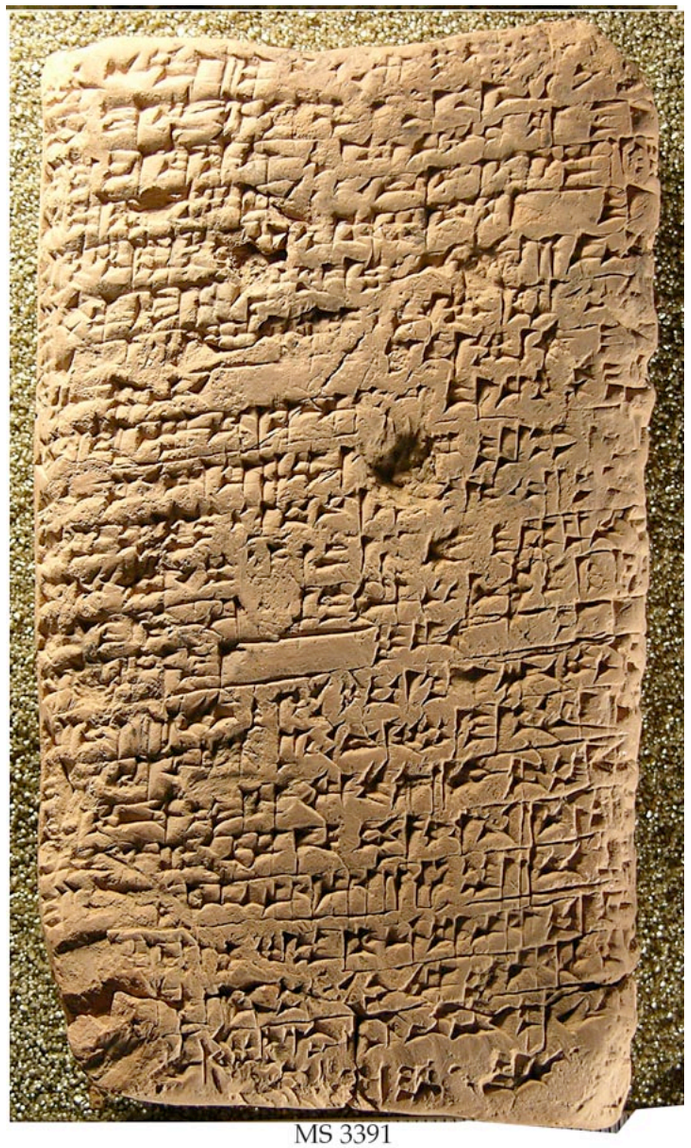
Manning et al. “Information retrieval (IR) is finding material ... of an unstructured nature ... that satisfies an information need from within large collections ...”





Information retrieval of yore

Manning et al. “Information retrieval (IR) is finding material ... of an unstructured nature ... that satisfies an information need from within large collections ...”



MS 3391
Library catalogue. Babylonia, 2000-1600 BC

Information retrieval of yore

Manning et al. “Information retrieval (IR) is finding material ... of an unstructured nature ... that satisfies an information need from within large collections ...”

Information retrieval of yore

Manning et al. “Information retrieval (IR) is finding material ... of an unstructured nature ... that satisfies an information need from within large collections ...”

Information retrieval today

Manning et al. “Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).”

Information retrieval today

Manning et al. “Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).”

PubMed

LexisNexis®

uspto.GOV
The United States Patent and Trademark Office
an agency of the Department of Commerce

Information retrieval today

Has this ever happened to you?

Process Name	User	CPU	# Threads	Real Memory
mdworker	Ryan	68.6	3	3.16 MB
uTorrent	Ryan	4.1	7	23.86 MB
SystemUIServer	Ryan	3.7	6	12.45 MB
Activity Monitor	Ryan	1.5	6	12.32 MB
screencapture	Ryan	0.1	1	2.28 MB
Quicksilver	Ryan	0.0	5	14.55 MB
Finder	Ryan	0.0	7	9.81 MB
loginwindow	Ryan	0.0	3	5.86 MB
iStat menus Helper	Ryan	0.0	1	2.55 MB
Dock	Ryan	0.0	2	6.64 MB
launchd	Ryan	0.0	3	528.00 KB
				3.95 MB
				2.94 MB
				2.28 MB
				588.00 KB
				2.06 MB

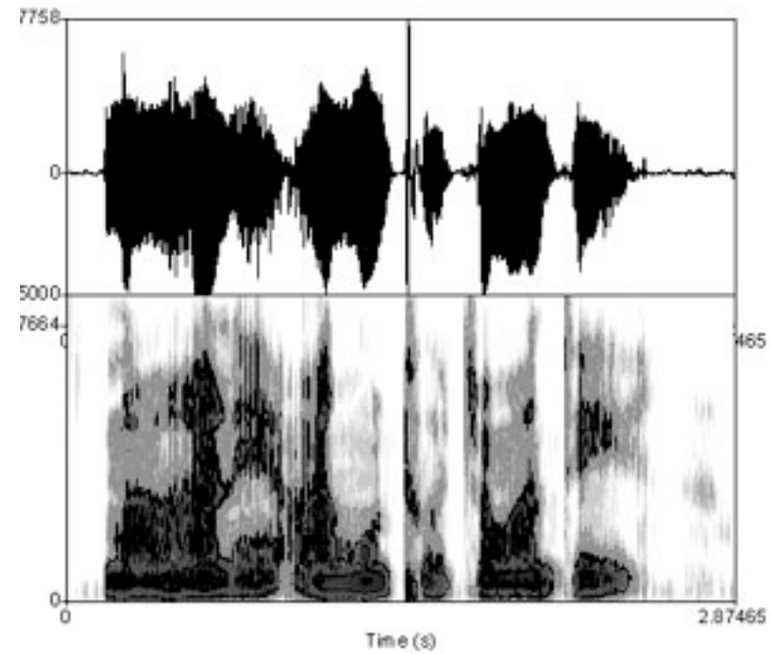
`mdworker` is part of the indexer for Mac OS X, which is indexing your stuff so that it can be easily searched with Spotlight, the desktop search feature on OS X.

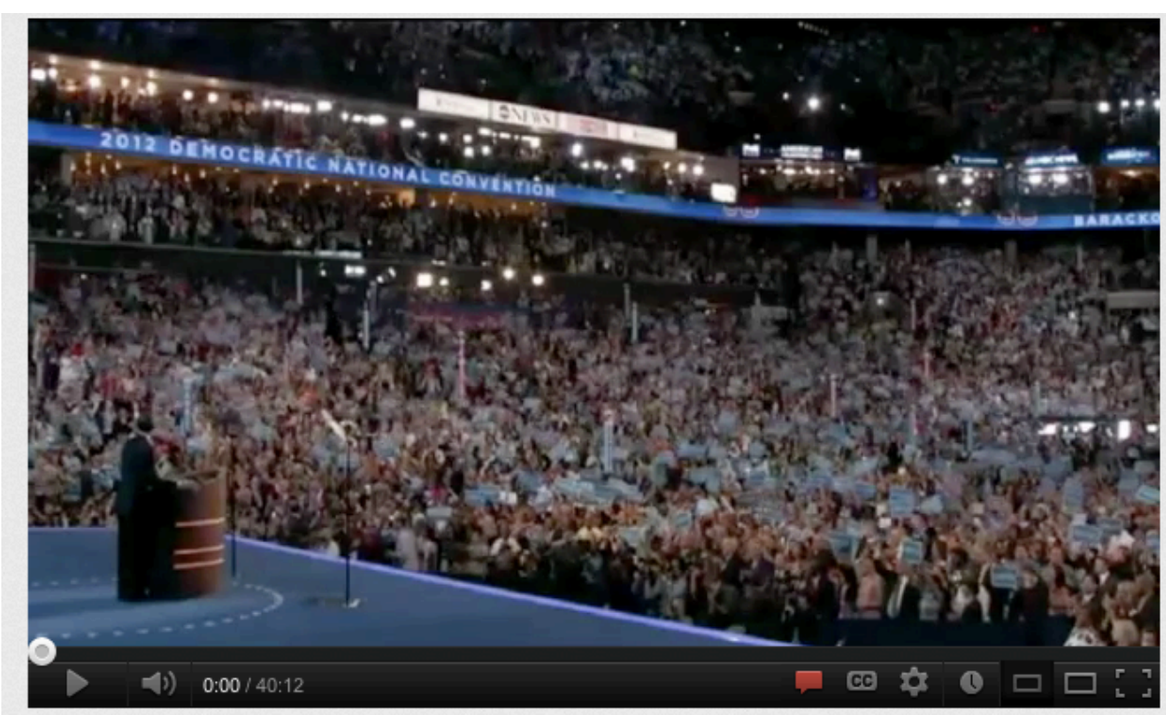
Usage Ne

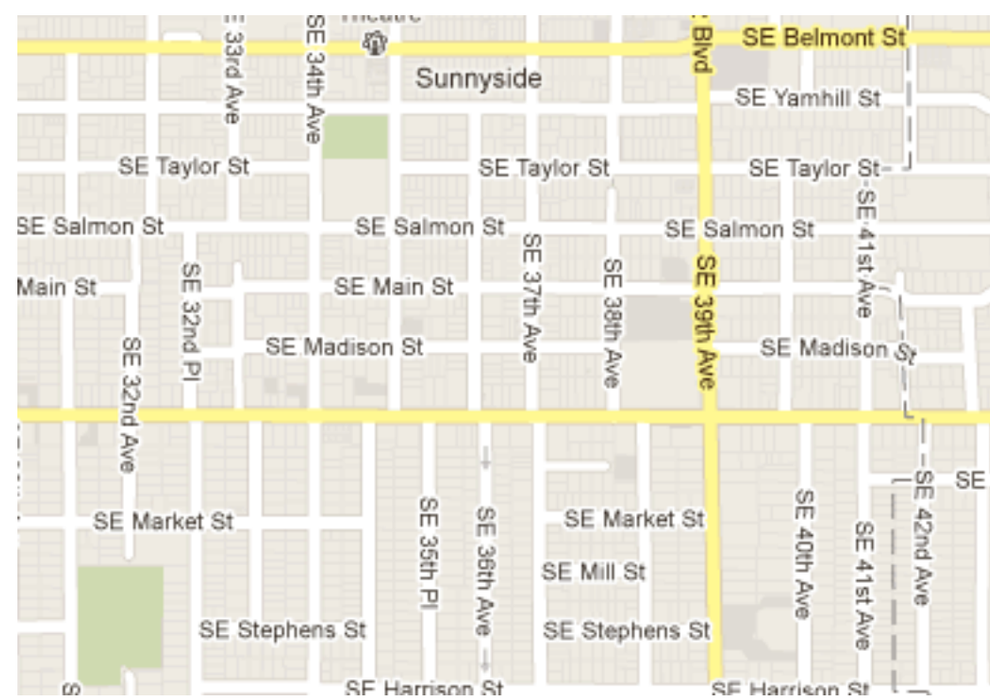
% User: 35.50
ystem: 6.00
% Nice: 0.00
% Idle: 58.50

Threads: 190
Processes: 43

IR for non-textual media







Beyond textual queries

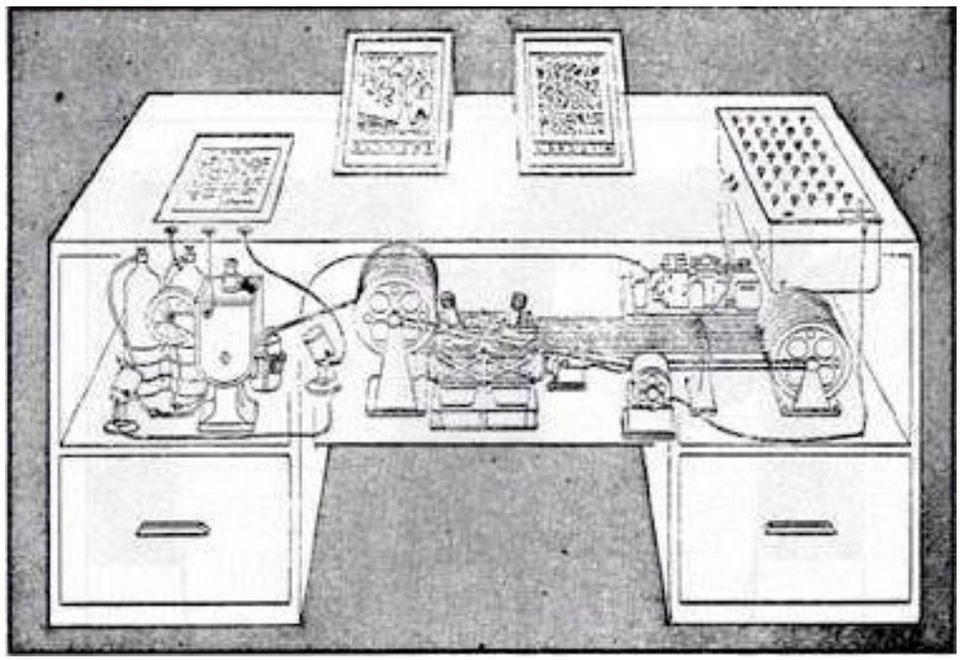
TinEye
Reverse Image Search

SHAZAM®

Highlights from the history of IR

1945

Memex: “a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility"



As we may think
Vannevar Bush
1945

Highlights from the history of IR

Alan Kent: literature searching systems, precision/recall evaluation.

Calvin Mooers: allegedly coins the term *information retrieval*.

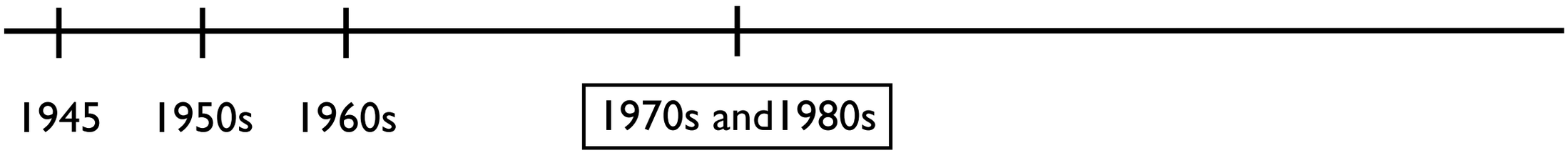
Highlights from the history of IR

Cranfield experiments: computer-based information retrieval evaluation experiments.

MEDLARS: machine readable database at NLM.

SMART: information retrieval system by Gerard Salton.

Highlights from the history of IR



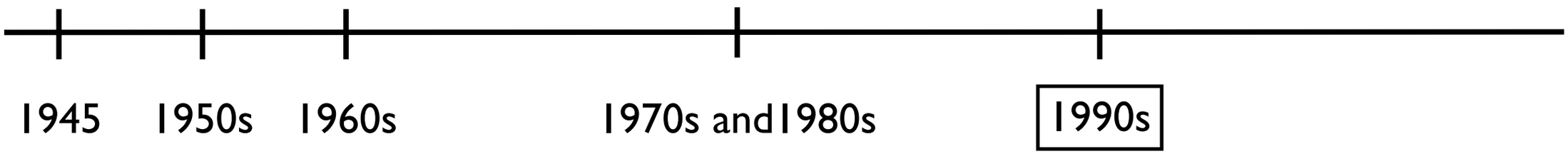
Gerard Salton: indexing, term weighting, vector space model.

SIGIR: special interest research group on IR.

Online IR systems: MEDLINE, AIM-TWX, Dialog.

LexisNexis: database for legal cases and news articles.

Highlights from the history of IR



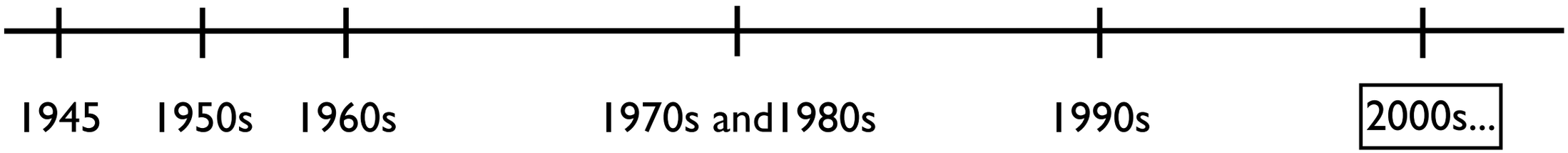
NIST TREC: Text REtrieval Conference

Archie: early search engine for FTP archives.

Altavista, Yahoo, Excite: first WWW search engines.

PageRank: Google's web page ranking algorithm

Highlights from the history of IR



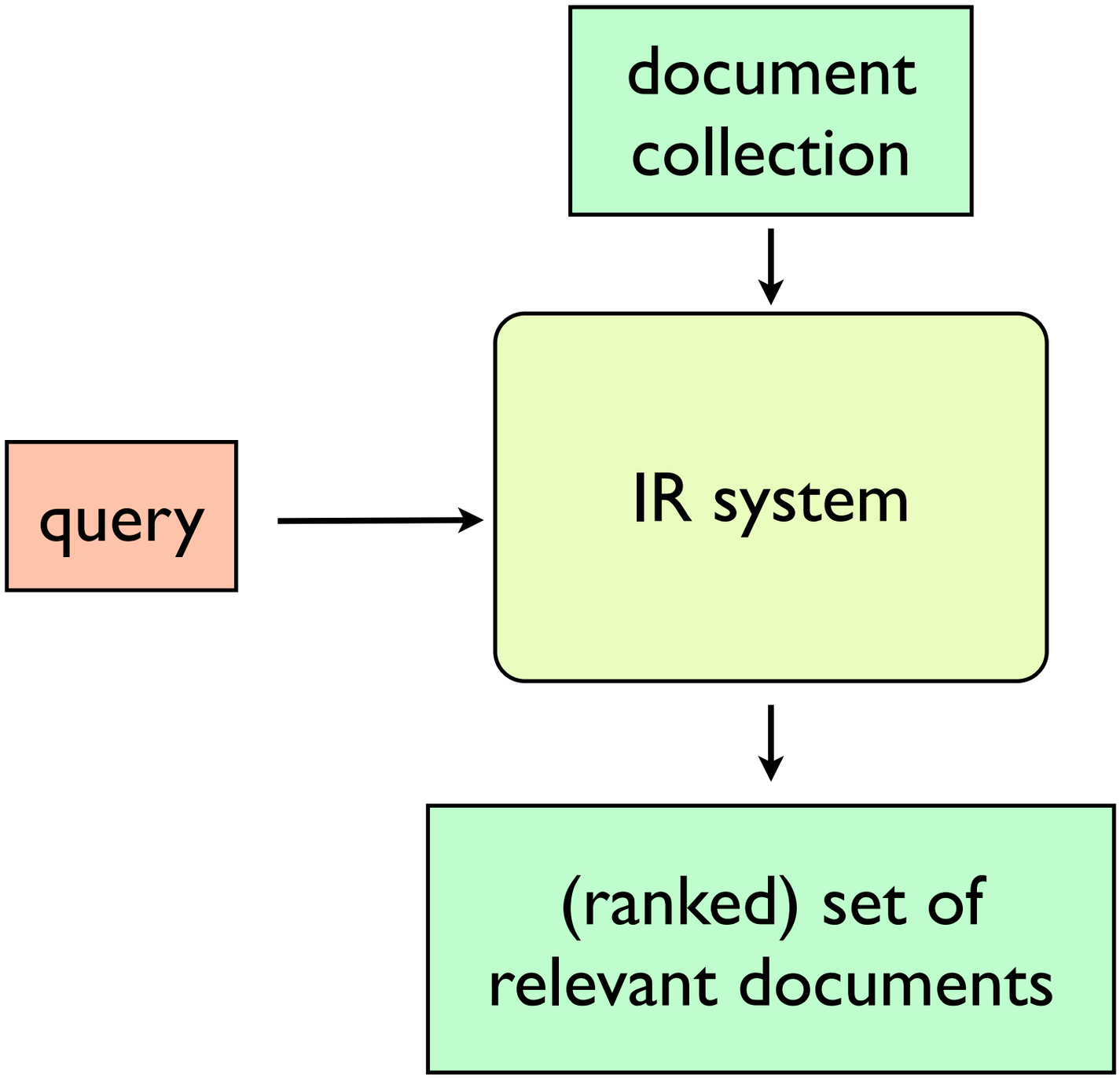
Google: web search and specialized search (scholar, news).

Multimedia IR: images, video, audio.

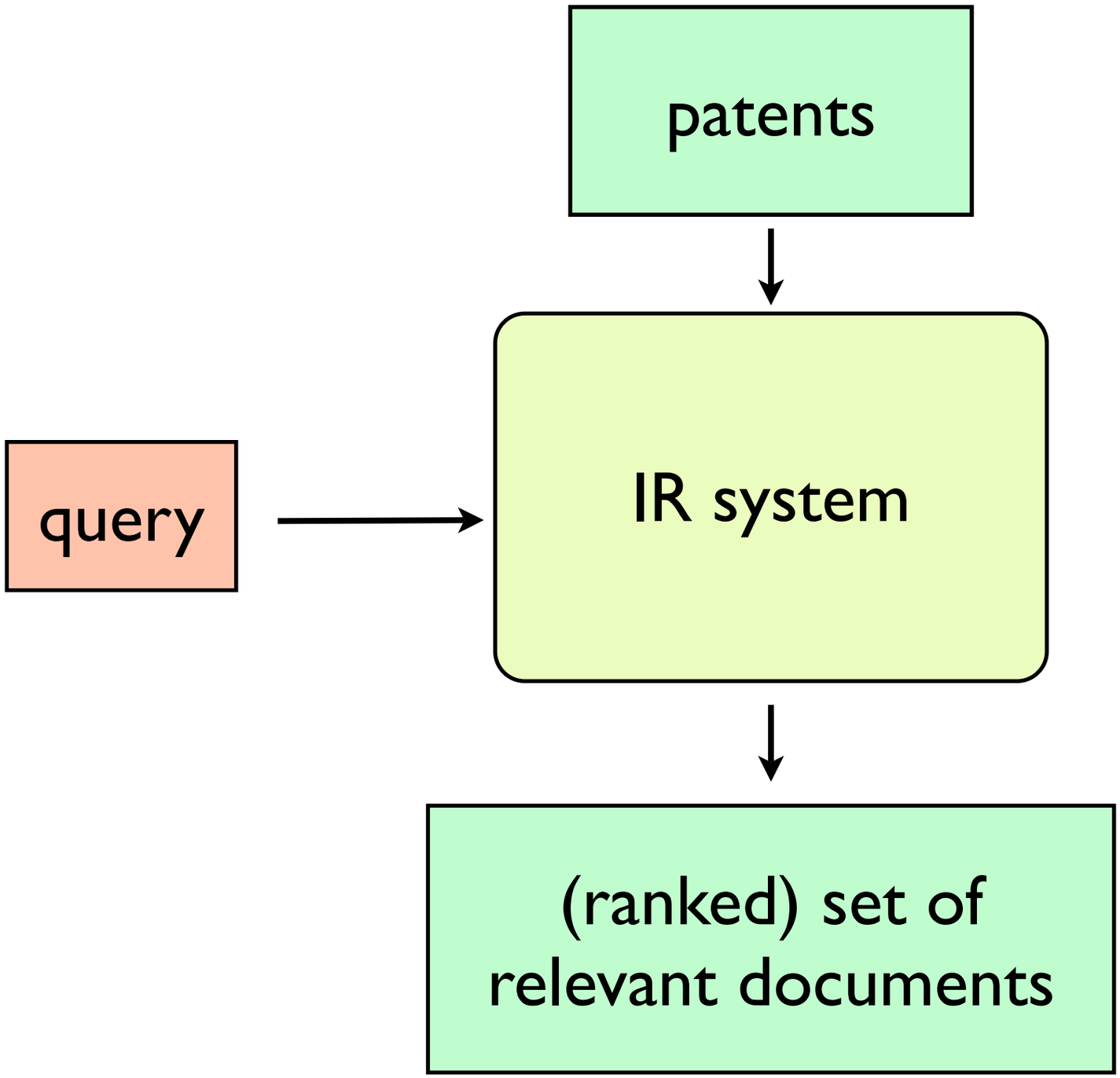
Multilingual IR: CLEF intiative.

Recommendation systems: Amazon, Netflix challenge.

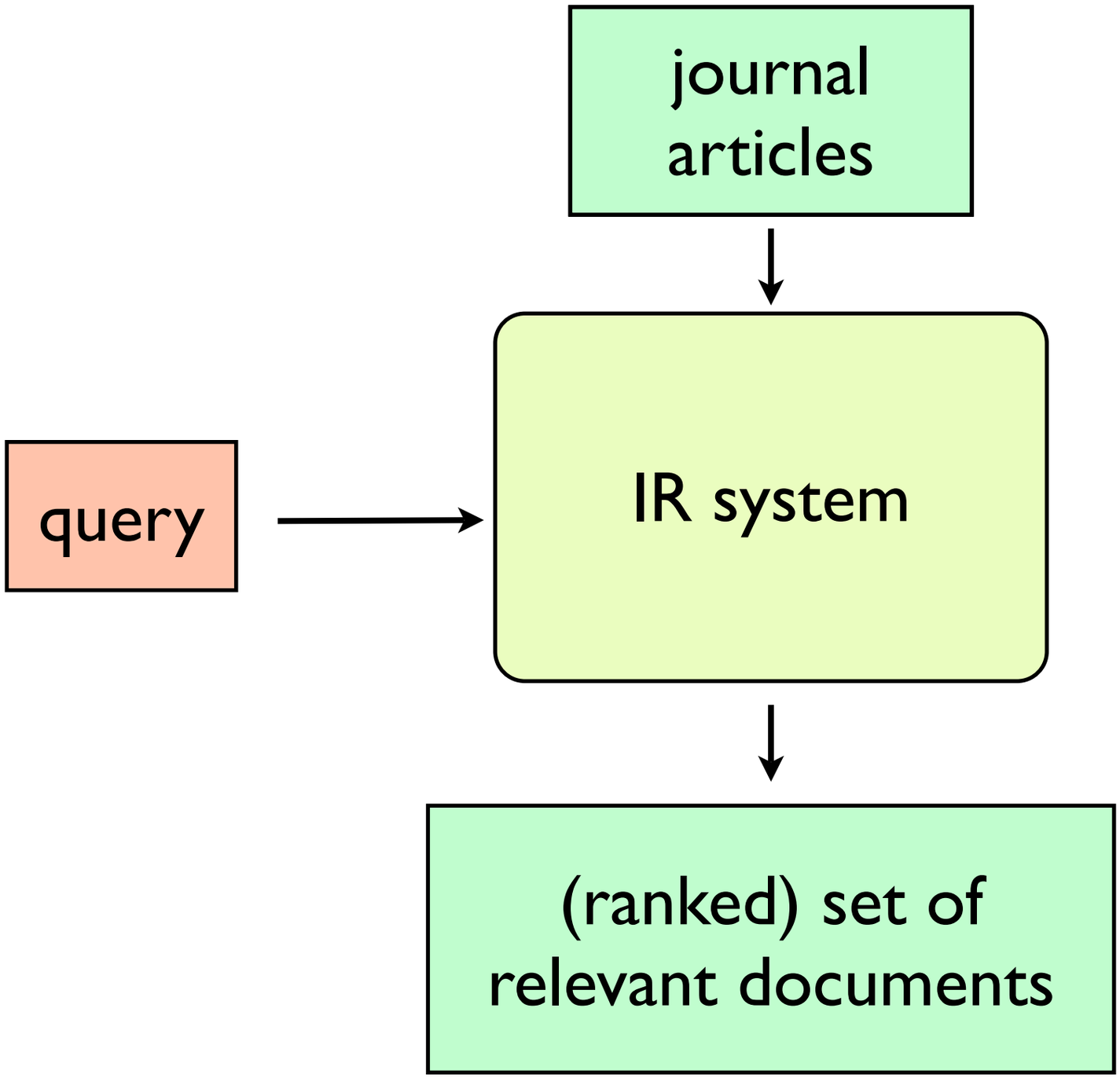
IR basics


document collection
query
IR system
(ranked) set of relevant documents

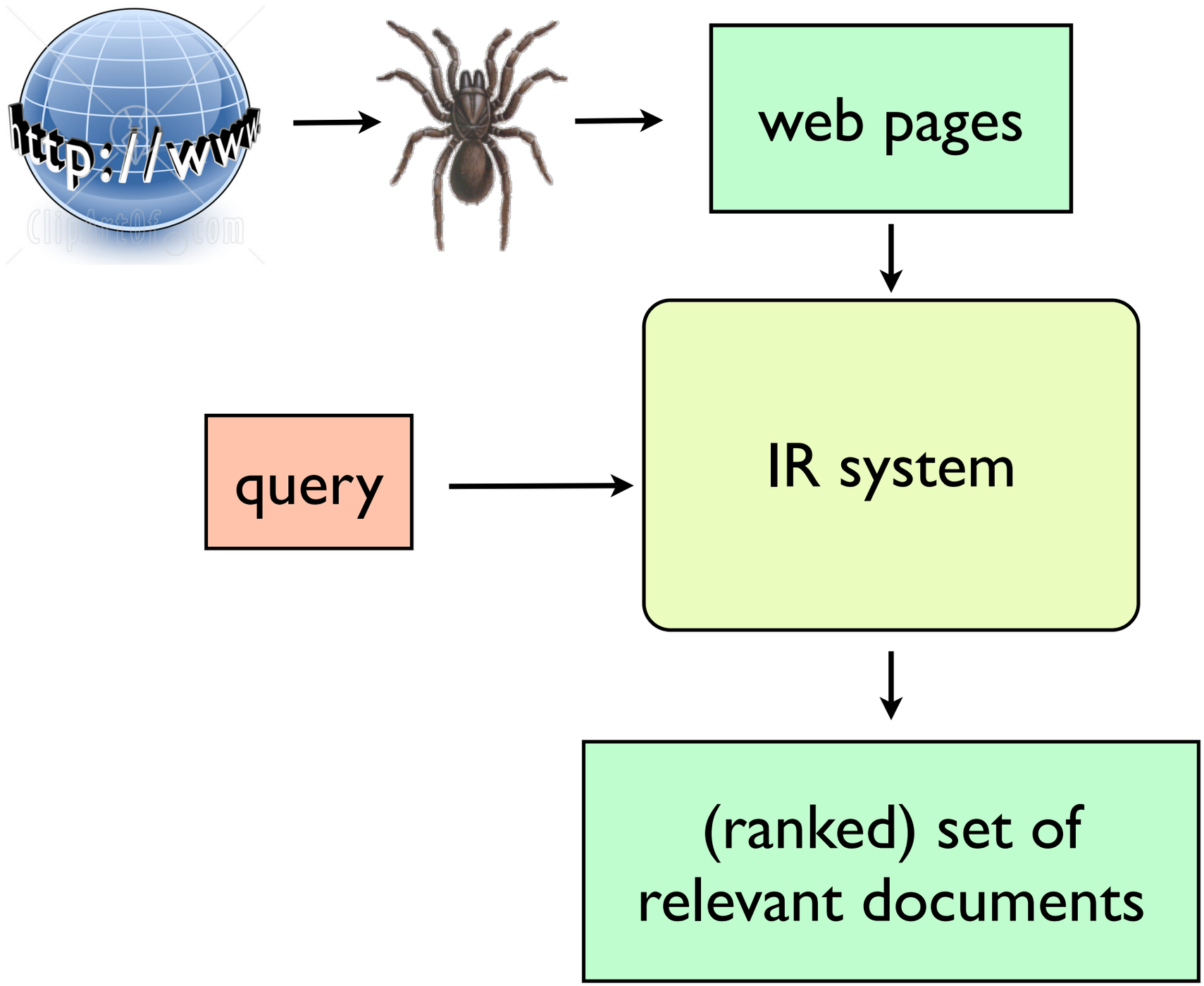
IR basics



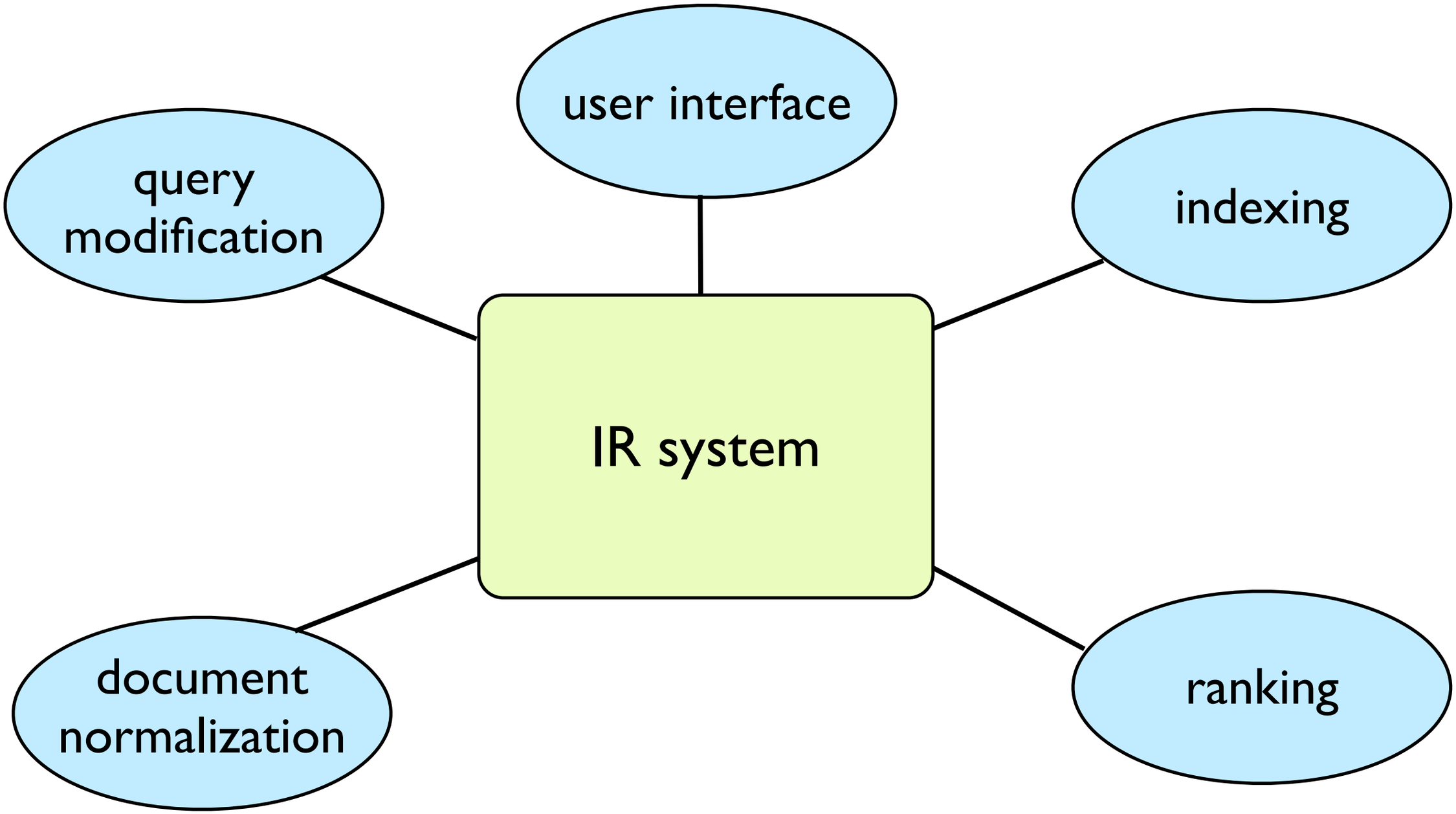
IR basics


journal articles
query
IR system
(ranked) set of relevant documents

IR basics


http://www
web pages
query
IR system
(ranked) set of
relevant documents

IR system components



Important considerations in IR: Relevance

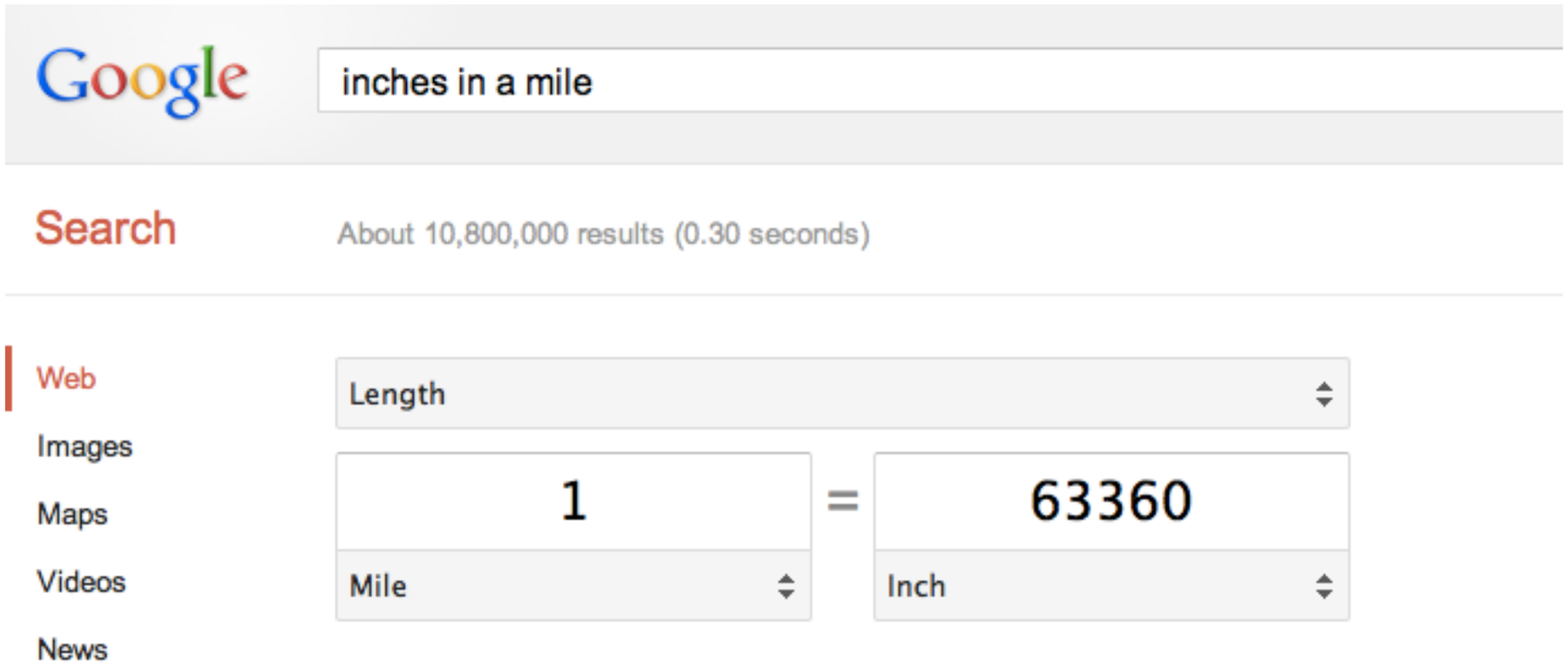
- Are the retrieved documents...
  - about the target subject?
  - timely and up-to-date?
  - from a trusted source?
  - satisfying the user's needs and goals?
- How do we rank documents in terms of these features?
- How do we decide which of these are more important and which are less important?
- More on this in a lecture soon.

Important considerations in IR: Evaluation

- Precision: Proportion of retrieved documents that are relevant.
- Recall: Proportion of relevant documents that are retrieved.
- What is the best balance between the two?
  - Easy to get perfect recall: just retrieve everything.
  - Easy to get good precision: return only the most relevant.
- Best balance may depend on the context:
  - patent search: high recall
  - ad hoc web search: high precision
- Evaluation will be covered in detail in a later lecture.

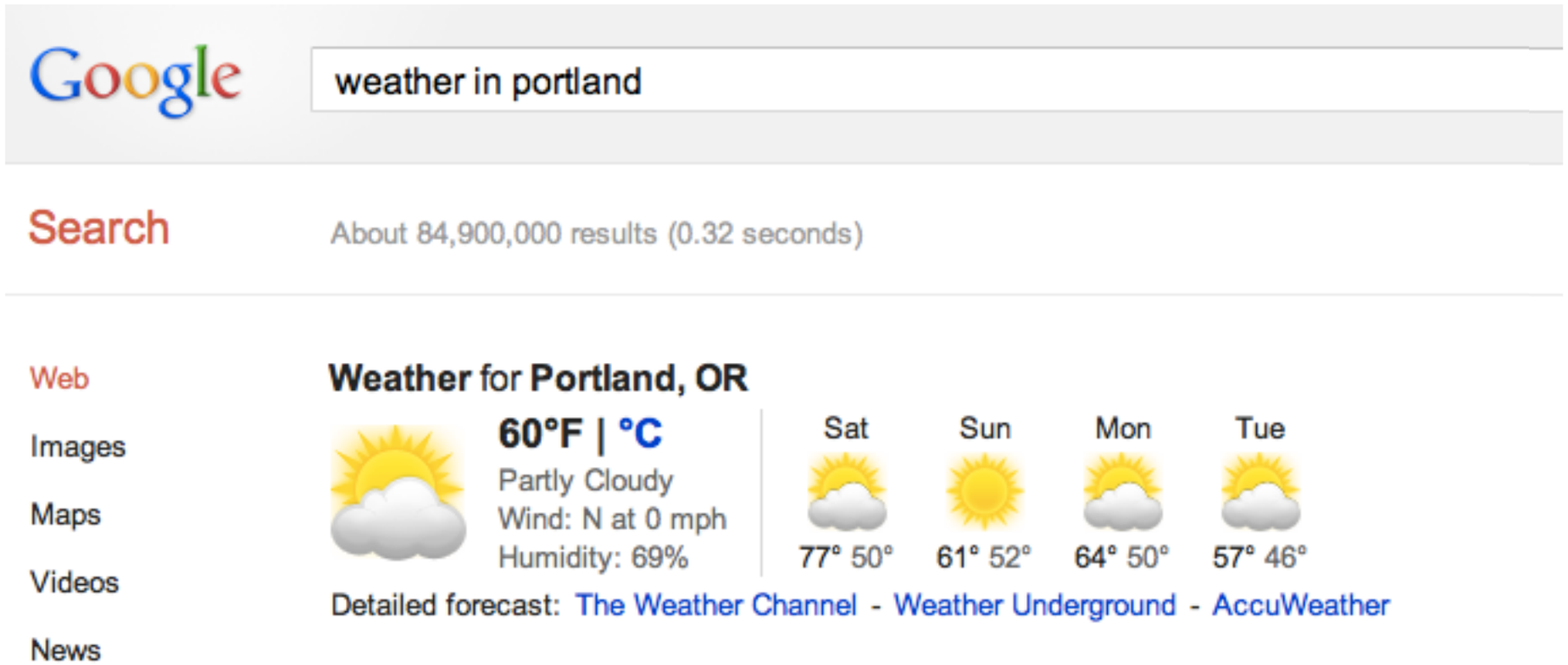
Information needs

- Fact finding



Information needs

- Fact finding



Information needs

- General knowledge

Google imperial measurement

Search About 8,860,000 results (0.31 seconds)

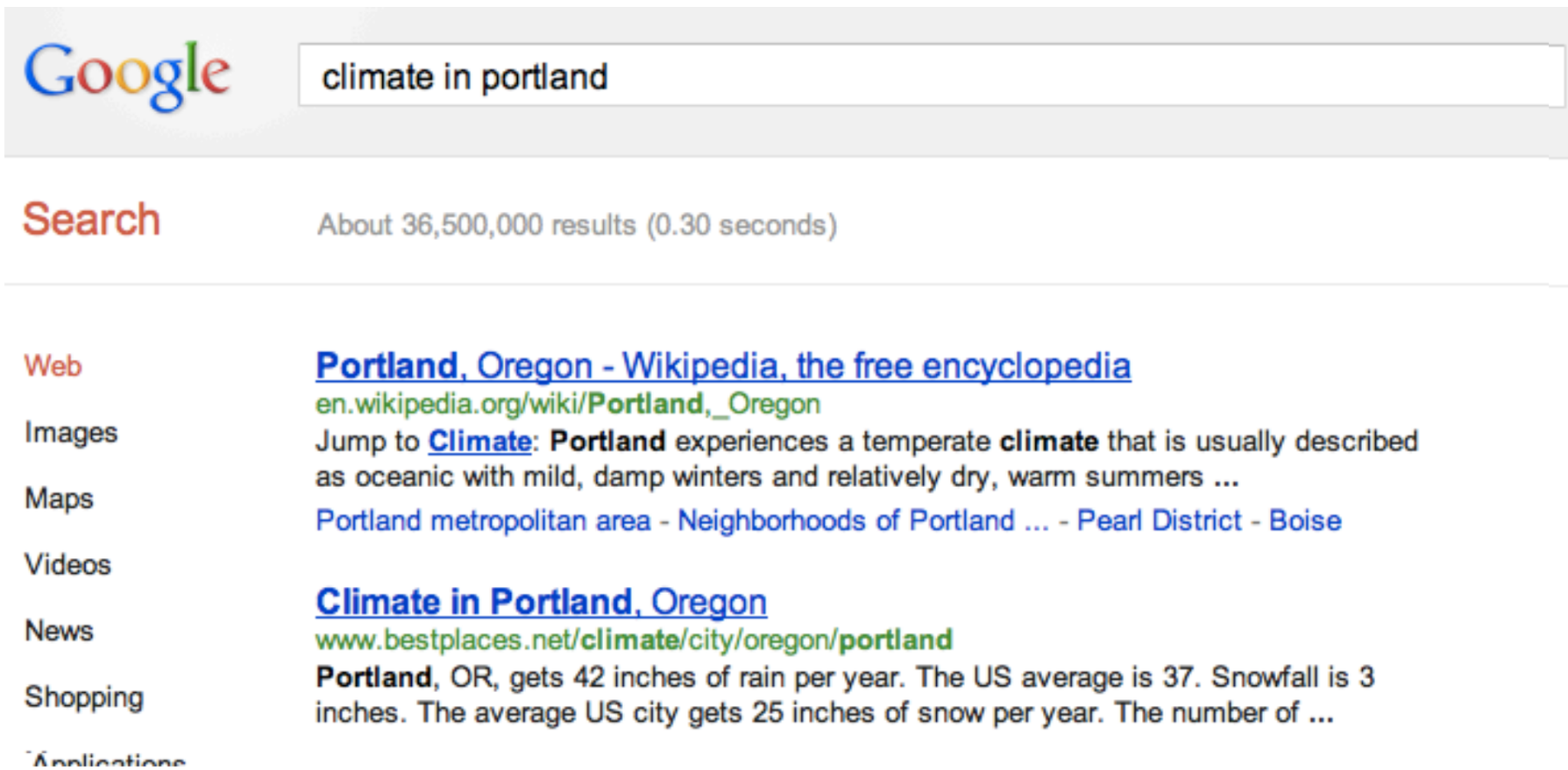
Web
Images
Maps
Videos
News
Shopping
Applications

Imperial units - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Imperial**_units
Metric equivalents in this article usually assume this latest official definition. Before this date, the most precise **measurement** of the **Imperial** Standard Yard was ...
Units - Relation to other systems - Current use of imperial units - See also

Metric and **Imperial** Conversion Charts and Tables
convert.french-property.co.uk/
France uses the metric system of **measurement**. The tables below provide for conversion from/to metric and **imperial** and US **measurement** systems.

Information needs

- General knowledge



A few models of information seeking

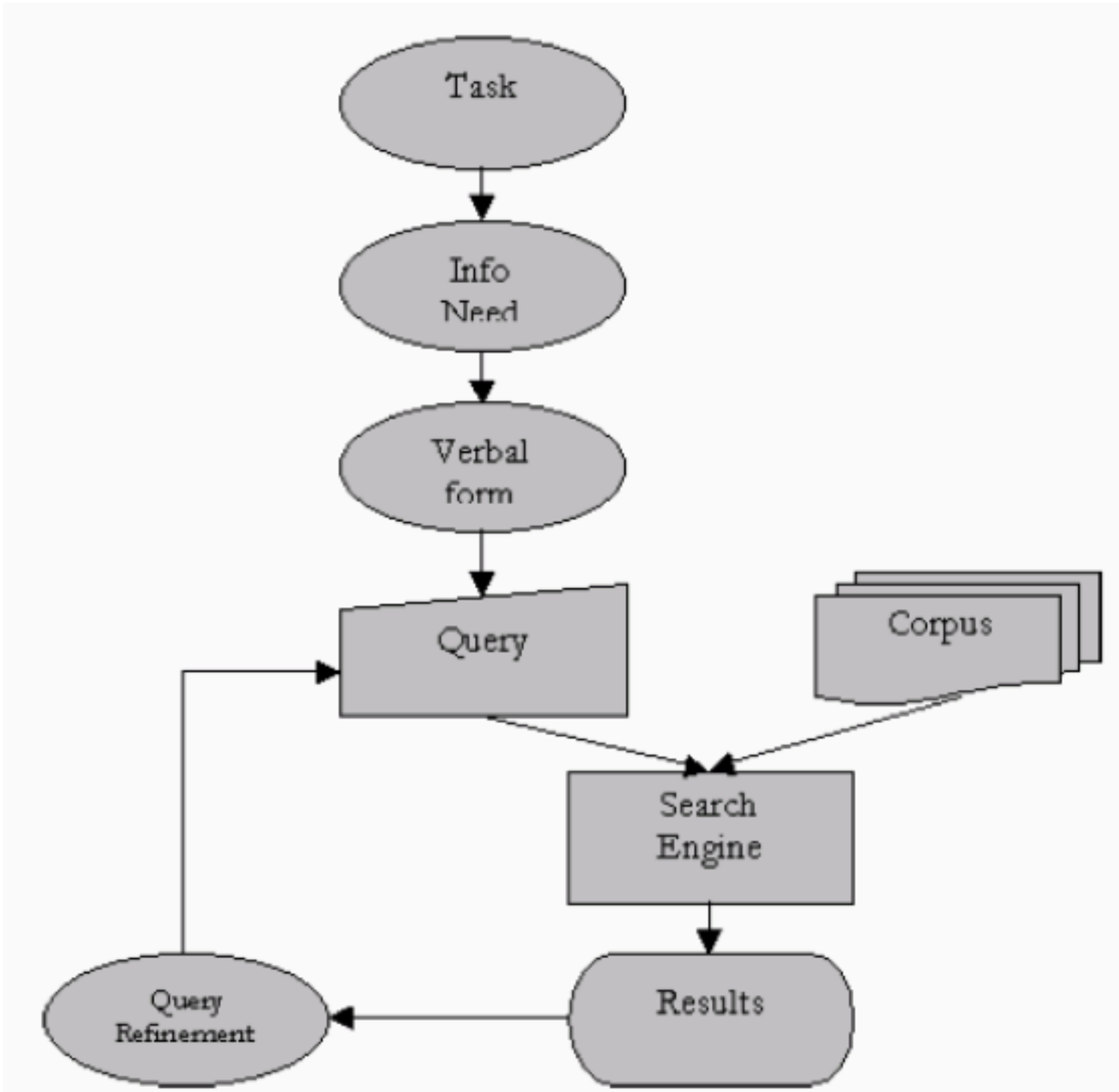
- Standard model
- Dynamic (berry-picking) model
- Search as a strategic process
- Sensemaking

Standard model of information seeking

- Recognizing a need for information,
- Accepting the challenge to take action to fulfill the need,
- Formulating the problem,
- Expressing the information need in a search system,
- Examination of the results,
- Reformulation of the problem and its expression,
- Admitting you are powerless over the internet, and
- Use of the results.

Slightly modified summary of information-seeking process outlined in Marchionini and White (2008), described in M. Hearst (2009) *Search User Interfaces*.

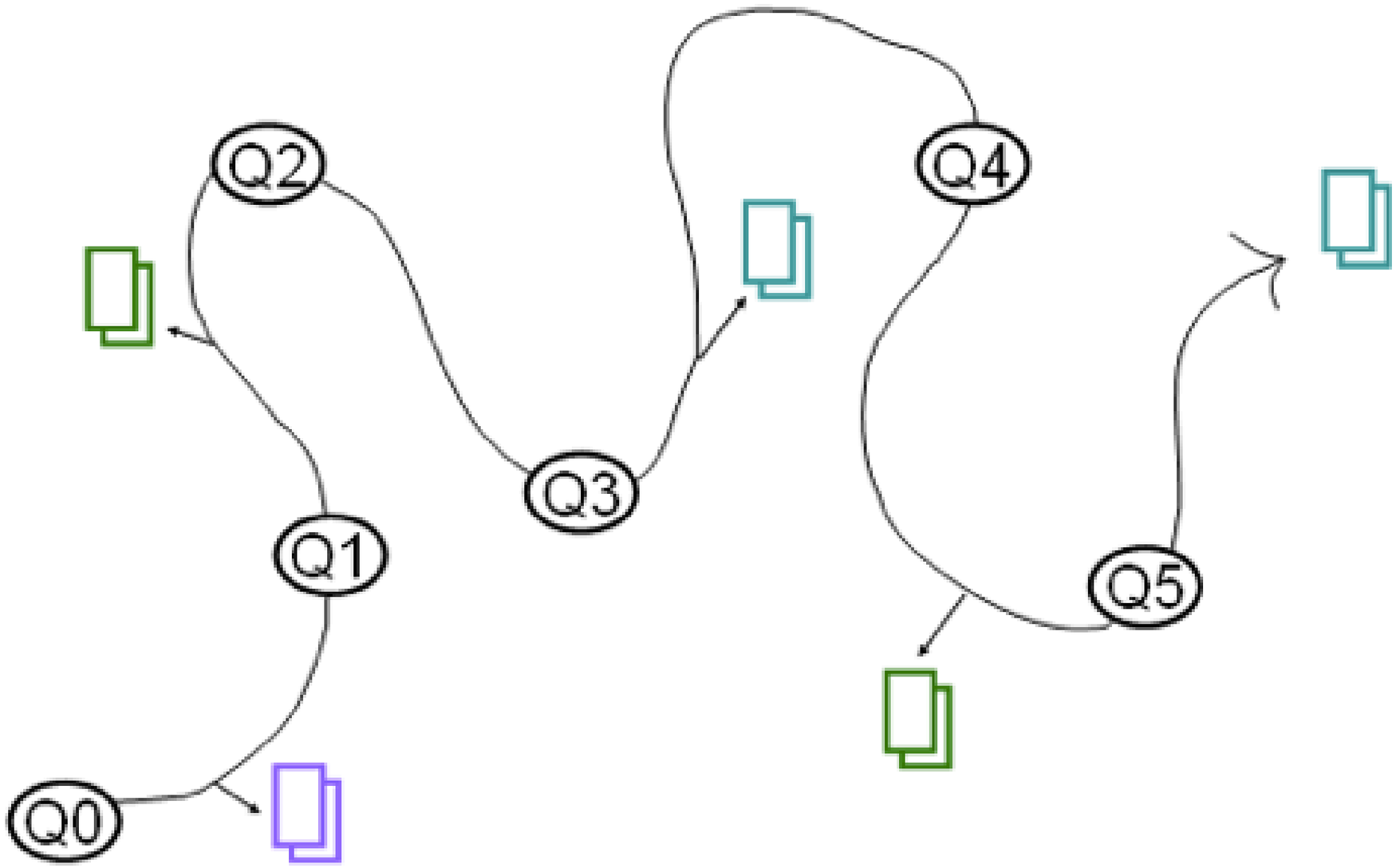
Standard model of information seeking



“Berry-picking” model of info seeing

- General model assumes a user’s information is static.
- Turns out that information needs change as results are retrieved.
- As an information seeker gets results, he learns more about what he’s looking for.
  - Reformulated queries may veer in a different direction.
  - One search goal may take precedence over another.
  - Results may lead to entirely new search needs.
- For some information seekers, the value of the search is in the total acquisition of knowledge rather than in the final set of search results.

“Berry-picking” model of info seeing



Information seeking as a strategic process

- Querying/searching vs. browsing/navigating.
  - search queries → ad hoc collections of documents
  - browsing → following links to predefined collections
- Less work required to browse than formulate a query, but...
  - if too many links, takes to long to find what you want
  - desired links may not be available or labeled appropriately
- Some strategies incorporate both querying and navigating.
  - To find web-page for this class, you might Google Brian Roark, and then follow the link to the class page from Brian's personal web page.

Sensemaking

- "an iterative process of formulating a conceptual representation from of a large volume of information"
  - information retrieval via searching, browsing
  - analysis and synthesis of results (analyzing and saving data, organizing documents)
- Process involved in scientific research, legal discovery.
- Interestingly, persistence on the part of the information seeker (i.e., time spent searching and browsing) results in more relevant documents retrieved -- not the kinds of queries issued.

Next lecture

- Information retrieval basics:
  - Term-document matrix
  - Inverted indices
  - Boolean retrieval, index intersection
- Additional topics on terms and postings
  - Faster intersection of posting lists
  - Positional indices
  - Tokenization and normalization