Introduction to Data Organization and Retrieval

Virendra Singh
Professor, Indian Institute of Technology Bombay
And
Adjunct Professor, Indian Institute of Technology Jammu
http://www.ee.iitb.ac.in/~viren/
E-mail: viren@ee.iitb.ac.in, virendra.singh@iitjammu.ac.in

CSPL201: Data Organization & Retrieval

Lecture 2 (18 September 2020)

CADSL

Why Information Retrieval?

- Information overload

"… The world produces between 1 and 2 exabytes (10^{18} bytes) of unique information per year, which is roughly 250 megabytes for every man, woman, and child on earth. …" (Lyman & Hal 03)

Information Retrieval

- Information Retrieval (IR) mainly studies unstructured data:
  Merrill Lynch estimates that more than 85 percent of all business information exists as unstructured data - commonly appearing in emails, memos, notes from call centers and support operations, news, user groups, chats, reports, … and Web pages.
  Text in Web pages or emails; image; audio; video; protein sequences..

- Unstructured data:
  No structure: no primary key as in RDBMS
  Semantic meaning unknown: natural language processing systems try to find the meaning in the unstructured text

Databases vs Information Retrieval

Databases	Information Retrieval
Structured Data	Unstructured Data
Schema Driven	No fixed schema; various data models (i.e., vector space model)
Relational (or object, hierarchical, and network) model is predominant	
Structured query model	Free-form query model
Rich mata-data operation	Rich data operation
Query returns data	Search request returns list or pointers to document
Results are based on exact matching (always correct)	Results are based on approximate matching and measures of effectiveness (may be imprecise and ranked)

Search Engines

One widely used example of retrieval system is Web search engine

Search engines have great impact on society

For example: a study was conducted for Vaccination. The search results were altered, i.e. one set of people were shown only pro vaccination results. And the other was shown anti vaccination results.

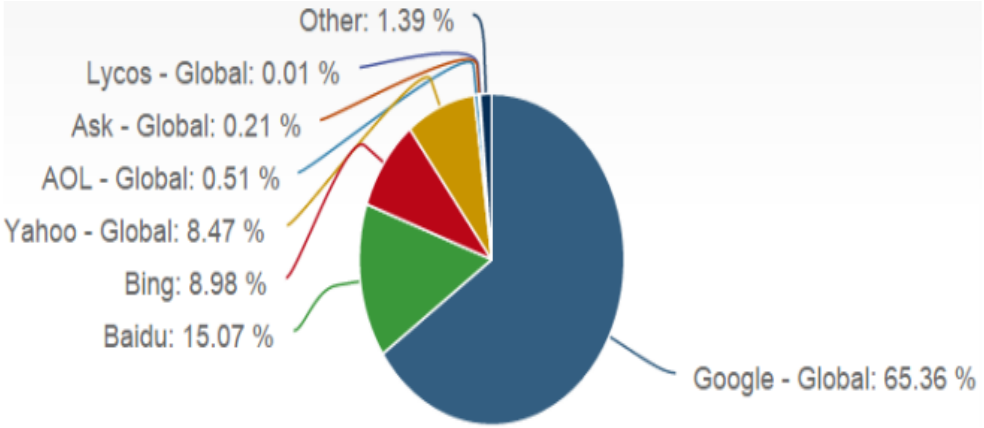
- Pro vaccination group acknowledged the benefit of vaccination.
- The other group first hesitated in believing the anti vaccination side, but after some more exposure they started getting worried about the vaccines.

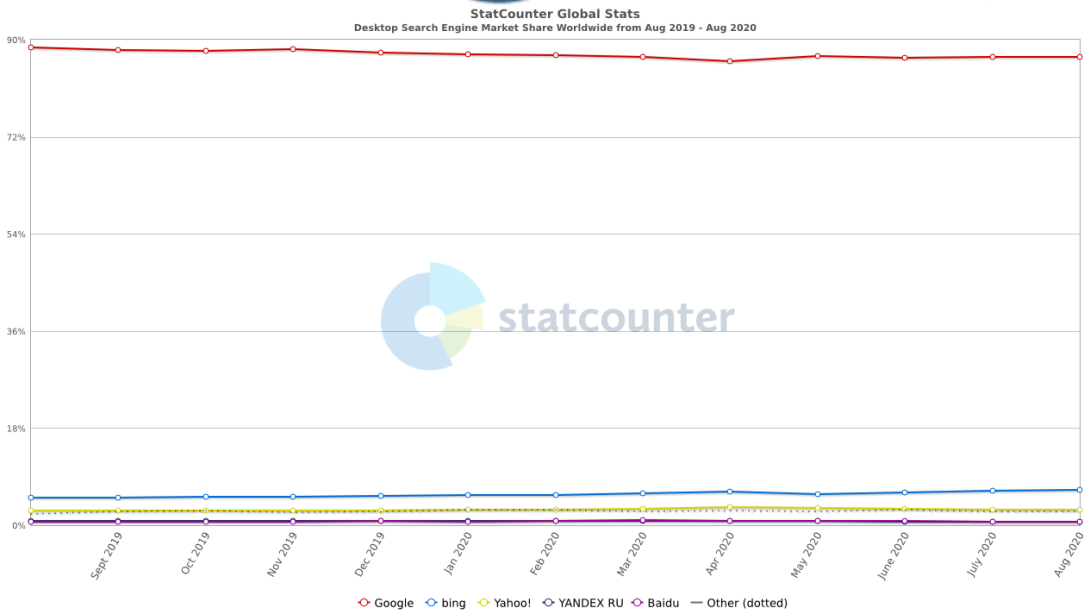
So the experiment shows that as a search engine user, it believe the information provided by search engine is more or less accurate.

In resume search engines, they give high ranking to male candidate. Though the algorithm was not directly differentiating the candidate based on gender but because recruiters were clicking the male candidate profiles more. Hence biased recruiters resulted in biased search engine

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4004139/
https://dl.acm.org/citation.cfm?id=3174225

Global market: Desktop





In 2020

Google: 86.81%

Bing: 6.52%

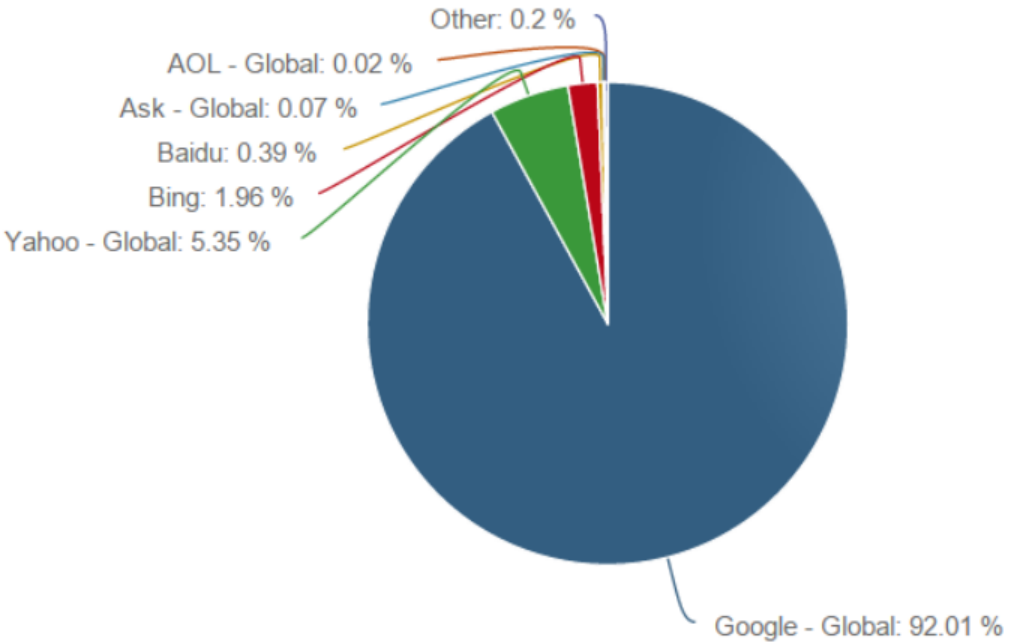
Yahoo: 2.89%

Duckduckgo: 0.64%

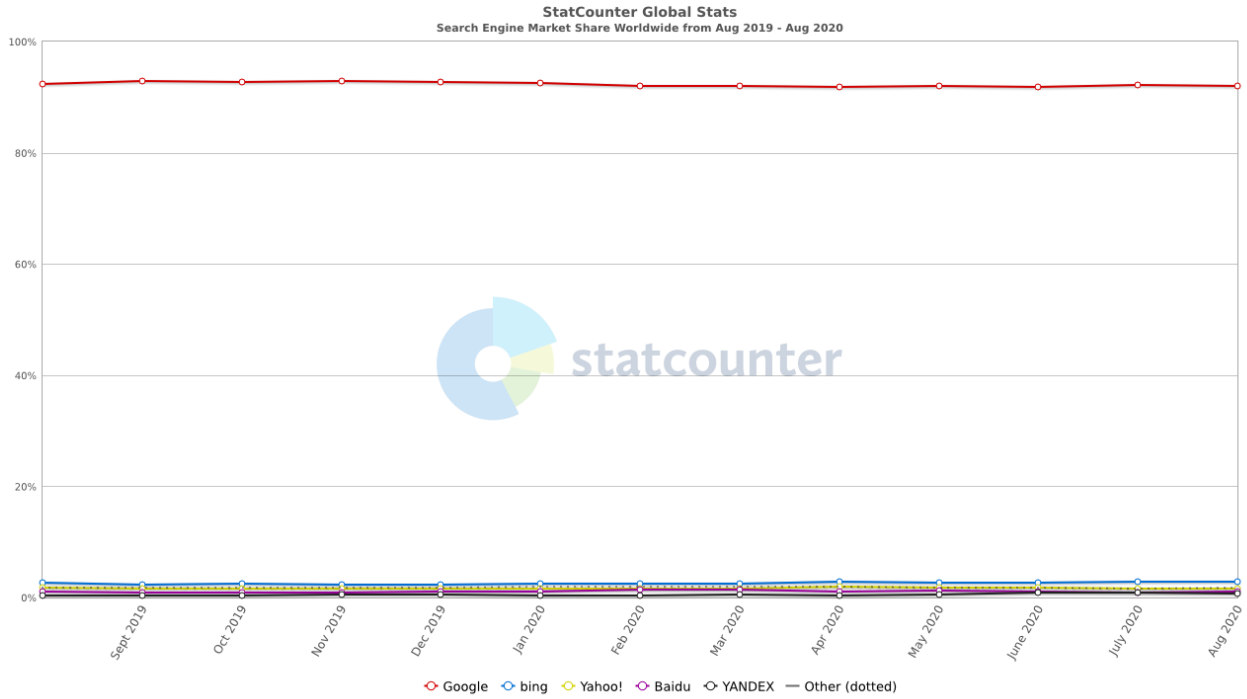
Baidu: 0.66%

Yandex: 0.62%

Global market: Mobile



Global market: Overall



In 2020

Google: 92.05%

Bing: 2.83%

Yahoo: 1.65%

Baidu: 0.98%

Yandex: 0.76%

Duckduckgo: 0.53%

Information Retrieval (1955-1992)

- Primary Users
  - Law Clerks
  - Reference Librarians
  - (Some) News organizations, product research, congressional committees, medical/chemical abstract searches
- Primary Search Models
  - Boolean keyword searches on Abstract, Title, keyword
- Vendors
  - Mead Data Central(Lexis – Nexis)
  - Dialog
  - Westlaw
  - Total searchable online data : O(10 terabytes)

Information Retrieval (1993+)

- Primary users
  - 1st time computer users
  - novices
- Primary search modes
  - Still Boolean keyword searches with limited probabilistic models
  - But FULL TEXT Retrieval
- Vendors
  - Lycos, Infoseek, Yahoo, Excite, AltaVista, Google
  - Total online data : ???
- Searching FTPable documents on internet
  - Archie
  - WAIS

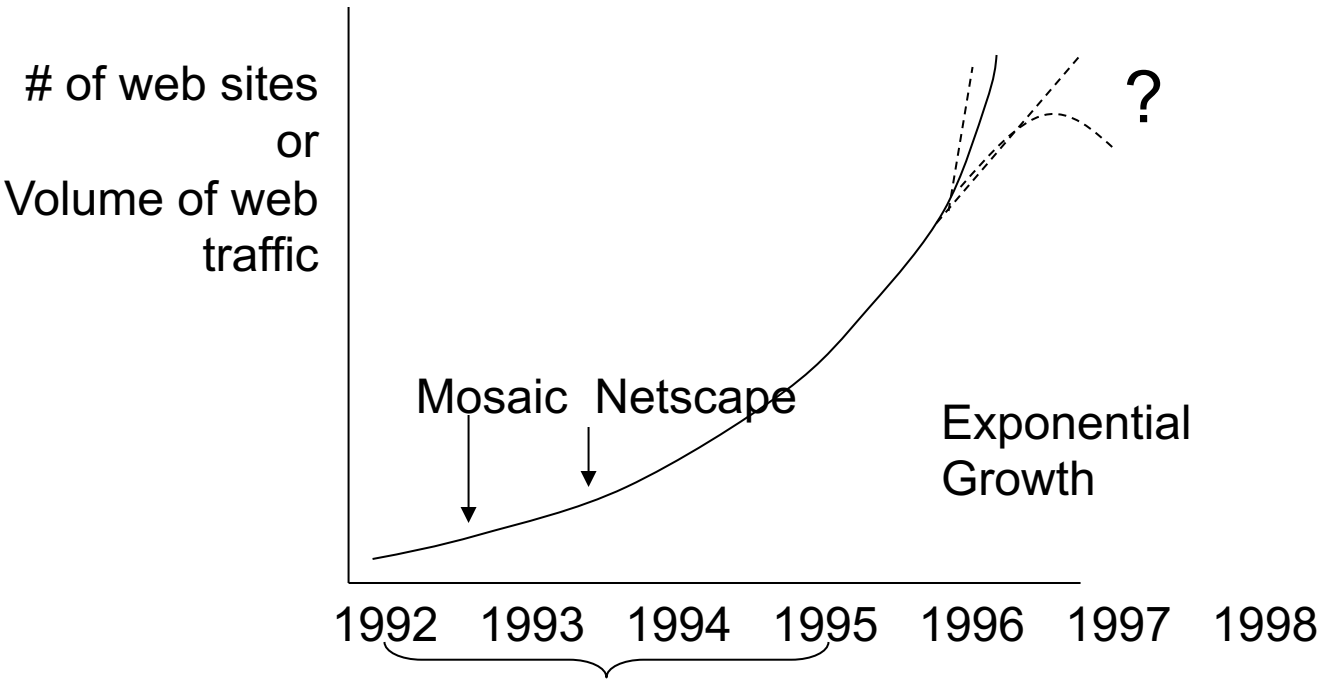
Information Retrieval (2000's)

- Multimedia IR
  - Image
  - Video
  - Audio and music
- Cross-Language IR
  - DARPA Tides
- Document Summarization
- Learning to Rank

Information Retrieval (2010's)

- Intelligent Personal Assistants
  - Siri
  - Cortana
  - Google Now
  - Alexa
- Complex Question Answering
  - IBM Watson
- Distributional Semantics
- Deep Learning

Growth of the Web



Information Retrieval

- Early IR system basically extended library catalog systems, allowing
  - Keyword searches,
  - Limited abstract searches
    in addition to Author/Title/Subject and including Boolean combination functionality

- IR was seen as reference retrieval
    (full documents still had to be ordered/delivered by hand)

Information Retrieval View

Old View

Function of IR :

Map queries to relevant documents

New View

Satisfy user's information need

Infer goals/information need from:

- query itself
- past user query history
- User profiling
- Collective analysis of other user feedback on similar queries

Information Retrieval View

Returns information in a format useful/intelligible to the user

- weighted orderings
- clustering of documents by different attributes
- visualization tools

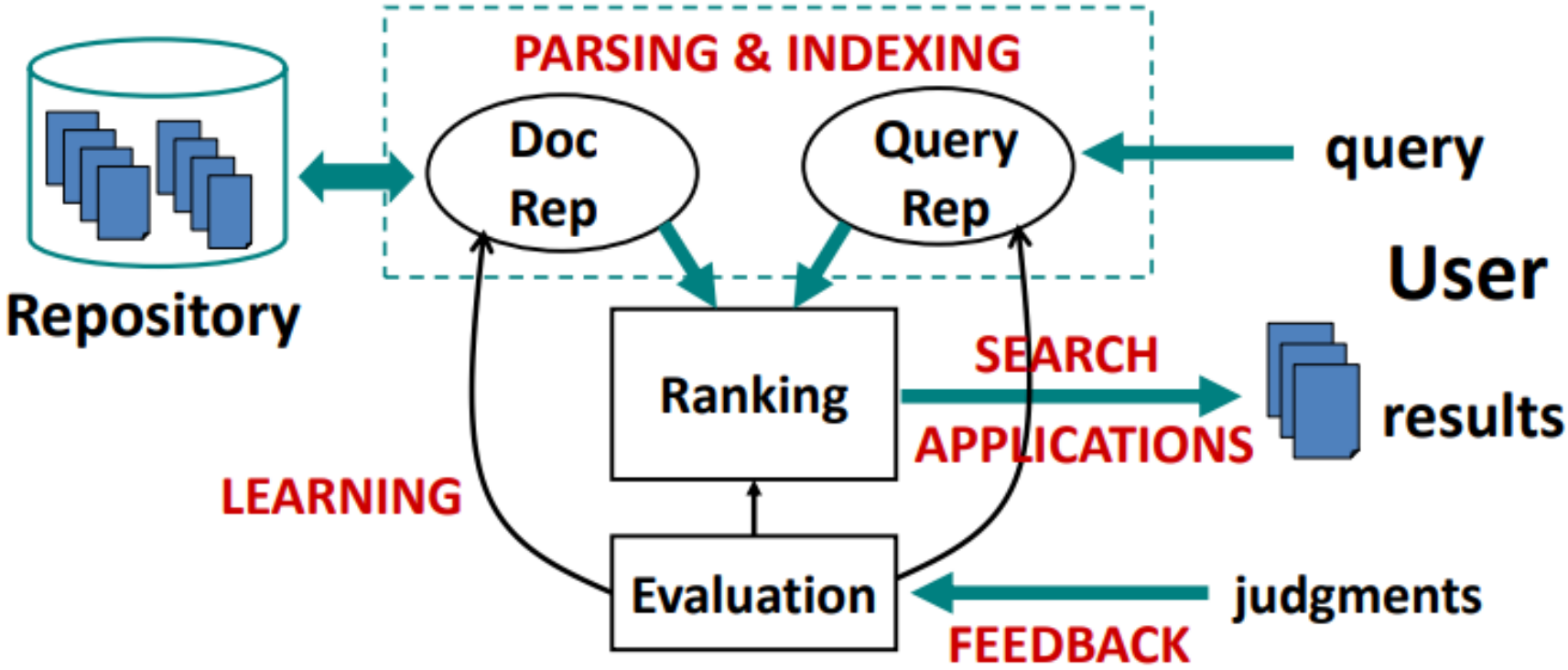
Text Understanding techniques to extract answer to questions or at least subregion of text

Who is the current mayor of Jammu?

⟹ don't need full article on city/state,

just the answer(and available source for proof)

Information Retrieval System



Information Retrieval System

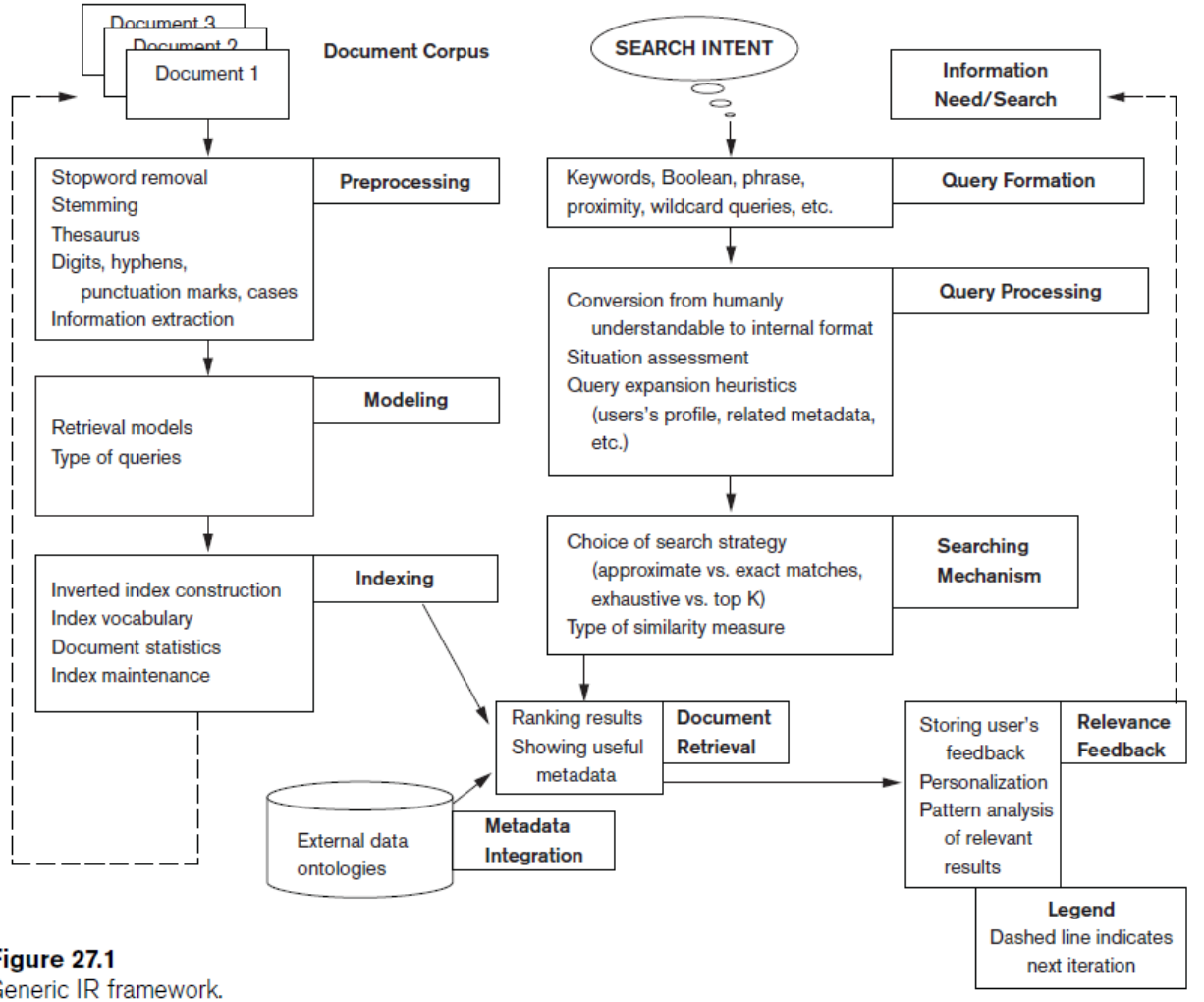


Figure 27.1
Generic IR framework.

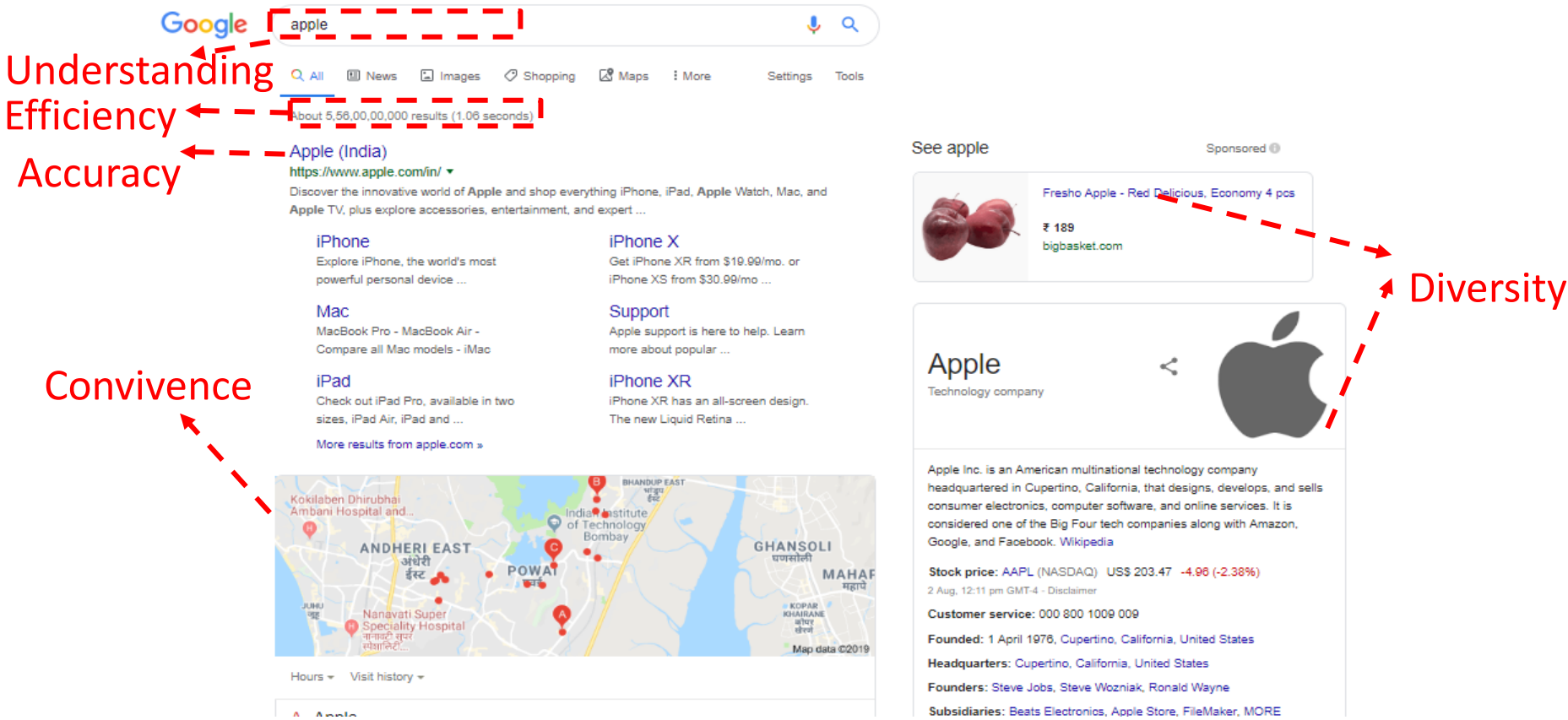
Major Topics in IR

- Indexing : for retrieving faster
- Ranking: to present most relevant results at the top
- Compression: to store efficiently
- Error tolerating search
- Machine learning: when fixed rule-based approaches fail
- Knowledge bases: how to organize the structured data
- Evaluation
- NLP

Modern Information Retrieval Systems

- Demand of accuracy
- Demand of efficiency
- Demand of understanding
- Demand of diversity
- Demand of convivence

Modern Information Retrieval Systems



Relevance

- Relevance is a subjective judgment and may include:
  - Being on the proper subject.
  - Being timely (recent information).
  - Being authoritative (from a trusted source).
  - Satisfying the goals of the user and his/her intended use of the information (*information need*).

More than Web Search

- IR is more than web search

Example

- Recommendation systems
- Question answering systems
- Text mining
- Online advertising

Related Areas

- Database Management
- Library and Information Science
- Artificial Intelligence
- Natural Language Processing
- Machine Learning

Database Management

- Focused on *structured* data stored in relational tables rather than free-form text
- Focused on efficient processing of well-defined queries in a formal language (SQL)
- Clearer semantics for both data and queries
- Recent move towards *semi-structured* data (XML) brings it closer to IR

Library and Information Science

- Focused on the human user aspects of information retrieval (human-computer interaction, user interface, visualization)
- Concerned with effective categorization of human knowledge
- Concerned with citation analysis and *bibliometrics* (structure of information)
- Recent work on *digital libraries* brings it closer to CS & IR

Artificial Intelligence

- Focused on the representation of knowledge, reasoning, and intelligent action
- Formalisms for representing knowledge and queries
  - First-order Predicate Logic
  - Bayesian Networks
- Recent work on web ontologies and intelligent information agents brings it closer to IR

Natural Language Processing

- Focused on the syntactic, semantic, and pragmatic analysis of natural language text and discourse
- Ability to analyze syntax (phrase structure) and semantics could allow retrieval based on *meaning* rather than keywords

Natural Language Processing: IR Directions

- Methods for determining the sense of an ambiguous word based on context (*word sense disambiguation*)
- Methods for identifying specific pieces of information in a document (*information extraction*)
- Methods for answering specific NL questions from document corpora or structured data like FreeBase or Google's Knowledge Graph.

Machine Learning

- Focused on the development of computational systems that improve their performance with experience.
- Automated classification of examples based on learning concepts from labeled training examples (*supervised learning*).
- Automated methods for clustering unlabeled examples into meaningful groups (*unsupervised learning*).

Machine Learning: IR Directions

- Text Categorization
  - Automatic hierarchical classification (Yahoo).
  - Adaptive filtering/routing/recommending.
  - Automated spam filtering.
- Text Clustering
  - Clustering of IR query results.
  - Automatic formation of hierarchies (Yahoo).
- Learning for Information Extraction
- Text Mining
- Learning to Rank

Generic IR Pipeline

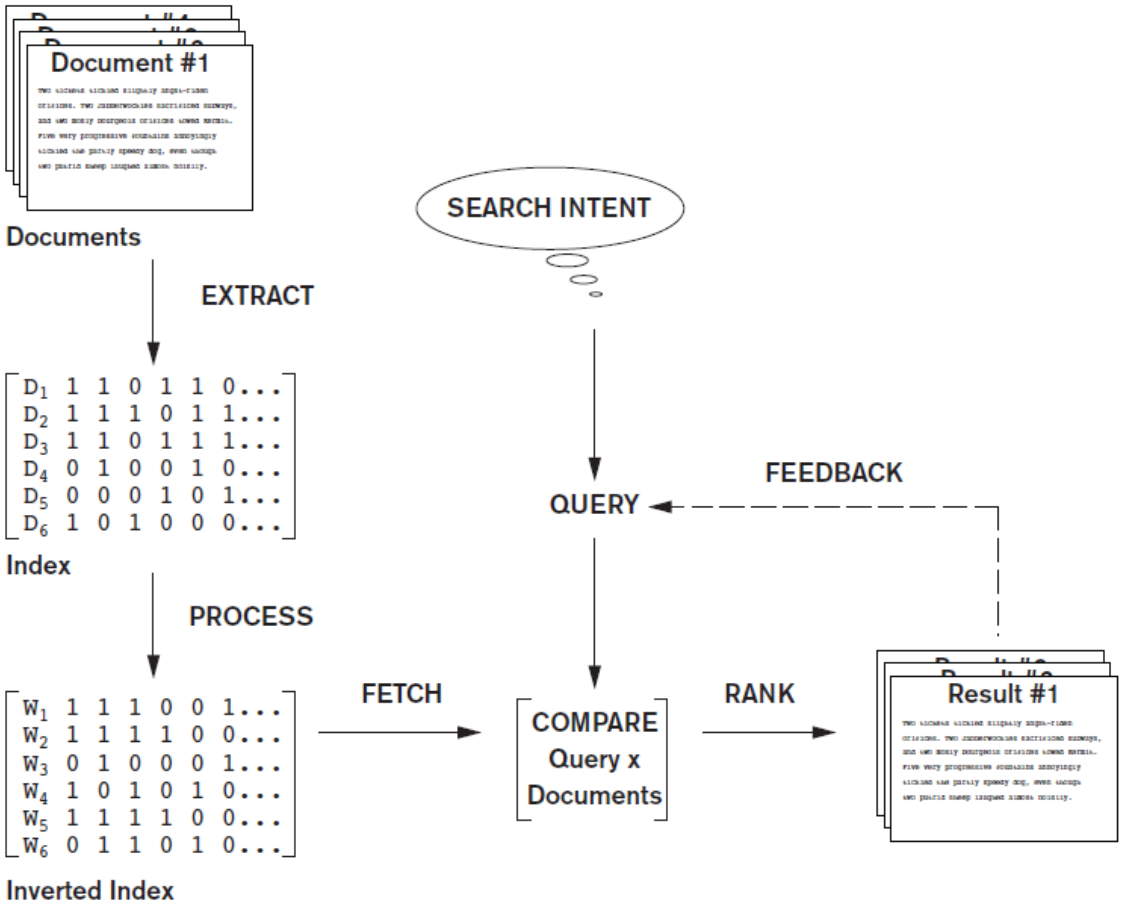


Figure 27.2
Simplified IR process pipeline.

More Information

http://www.ee.iitb.ac.in/~viren/Courses/2020/DOR.htm

Thank You