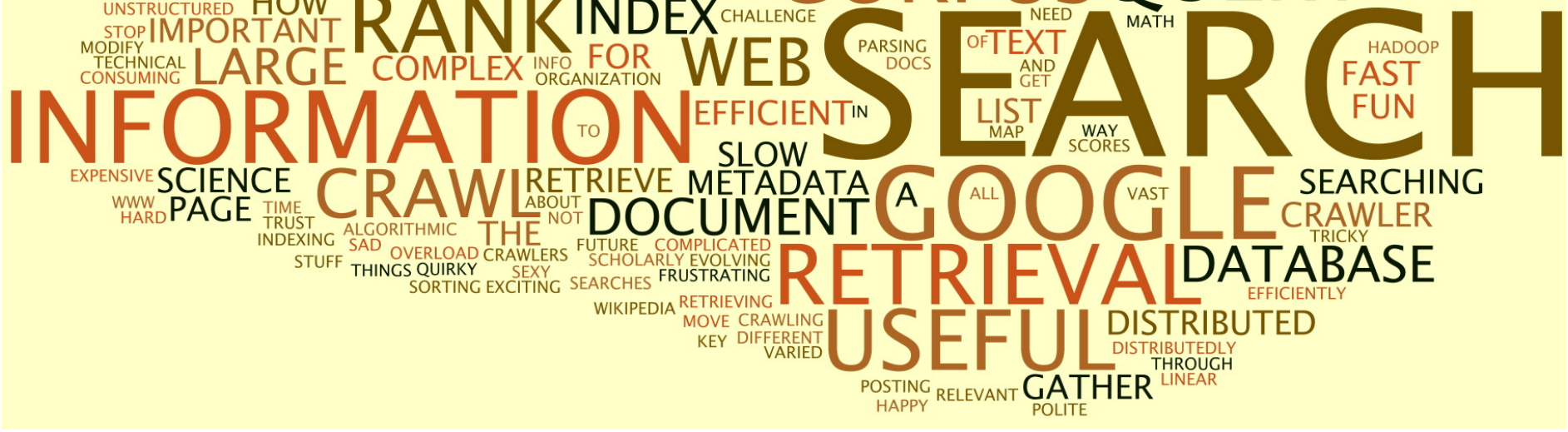


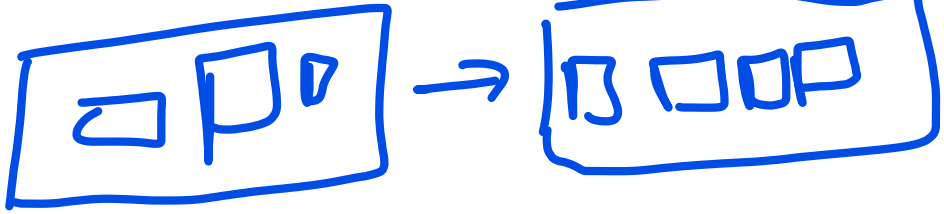
CS4051- Information Retrieval

Introduction

Waqas Ali

13th February 2022

Table of Contents



Why this course?

- Managing Data is one of the primary uses of computers

Why this course?

- Managing Data is one of the primary uses of computers
- Most of the data is not contained in structured databases
  - Therefore, no structured queries

Why this course?

- Managing Data is one of the primary uses of computers
- Most of the data is not contained in structured databases
  - Therefore, no structured queries
- How do we retrieve the required information?

Why this course?

- Managing Data is one of the primary uses of computers
- Most of the data is not contained in structured databases
  - Therefore, no structured queries
- How do we retrieve the required information?

Information Retrieval

Information Retrieval: Challenges

- Data is unstructured
  - Need to guess what is relevant

Information Retrieval: Challenges

- Data is unstructured
  - Need to guess what is relevant
- Query is unstructured
  - Need to guess user intent

Information Retrieval: Challenges

- Data is unstructured
  - Need to guess what is relevant
- Query is unstructured
  - Need to guess user intent
- Computers cannot guess

Inferring relevance and intent from data, query is the science of Information Retrieval

Information Retrieval

- 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

- 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).
  - These days we frequently think first of web search, but there are many other cases:
    - E-mail Search
    - Searching your computer
    - Corporate knowledge bases
    - Legal information retrieval

Basic assumptions of Information Retrieval

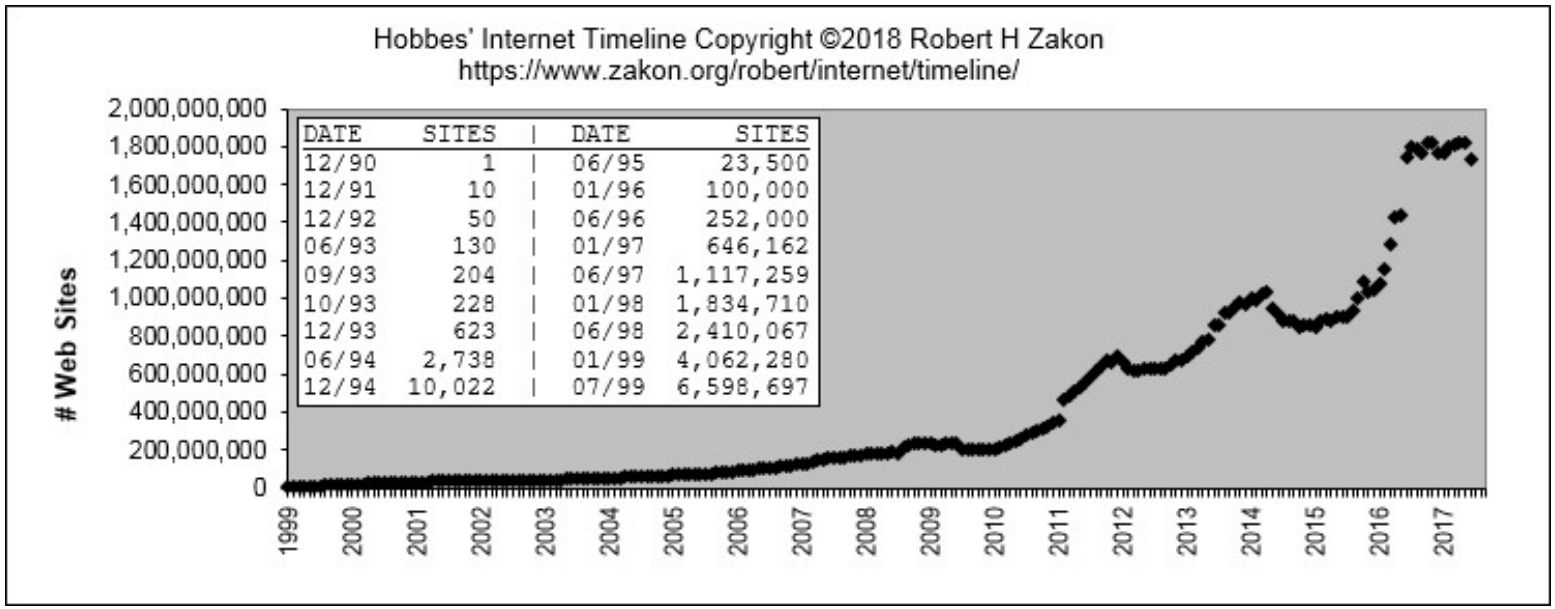
- **Collection:** A set of documents
  - Assume it is a static collection for the moment

Basic assumptions of Information Retrieval

- **Collection:** A set of documents
  - Assume it is a static collection for the moment

- **Goal:** Retrieve documents with information that is relevant to the user's information need and helps the user complete a task.

The growth of WWW



DATE	SITES	DATE	SITES
12/90	1	06/95	23,500
12/91	10	01/96	100,000
12/92	50	06/96	252,000
06/93	130	01/97	646,162
09/93	204	06/97	1,117,259
10/93	228	01/98	1,834,710
12/93	623	06/98	2,410,067
06/94	2,738	01/99	4,062,280
12/94	10,022	07/99	6,598,697

IR vs RDBMS

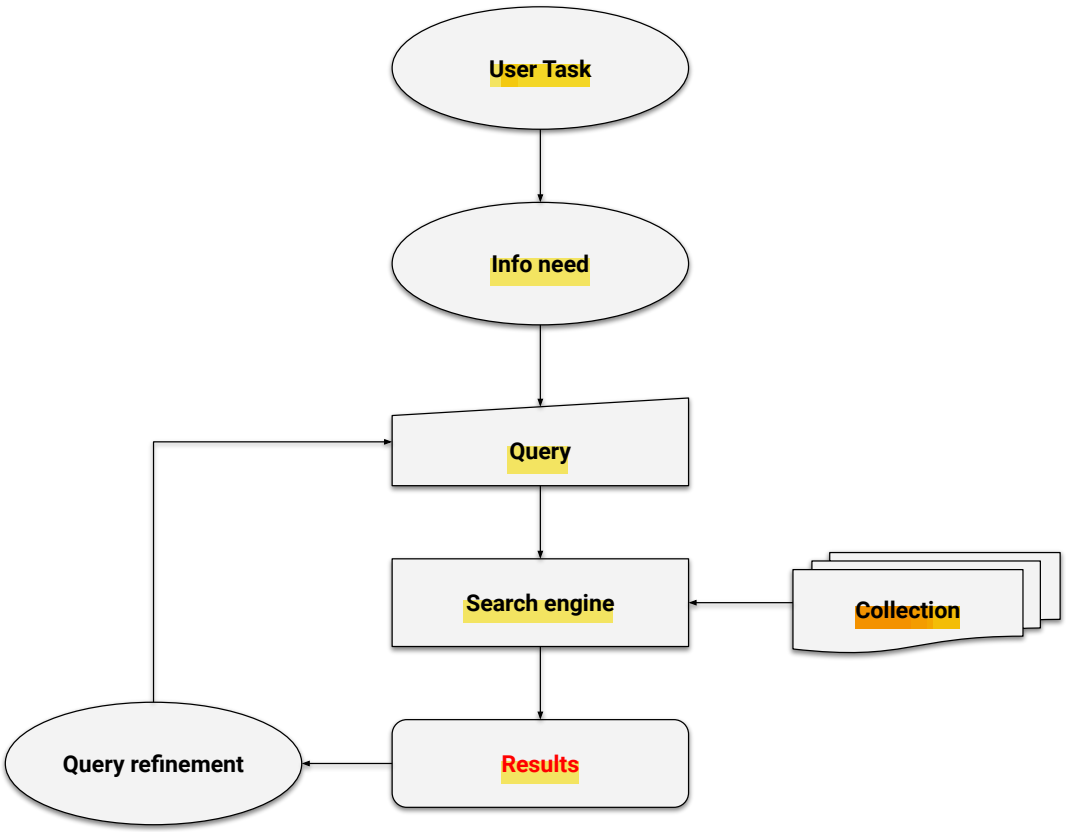
- Relational Database Management Systems (RDBMS)
  - Semantics of each object are well defined
  - Complex query languages (e.g., SQL)
  - Exact retrieval for what you ask
  - Emphasis on efficiency

IR vs RDBMS

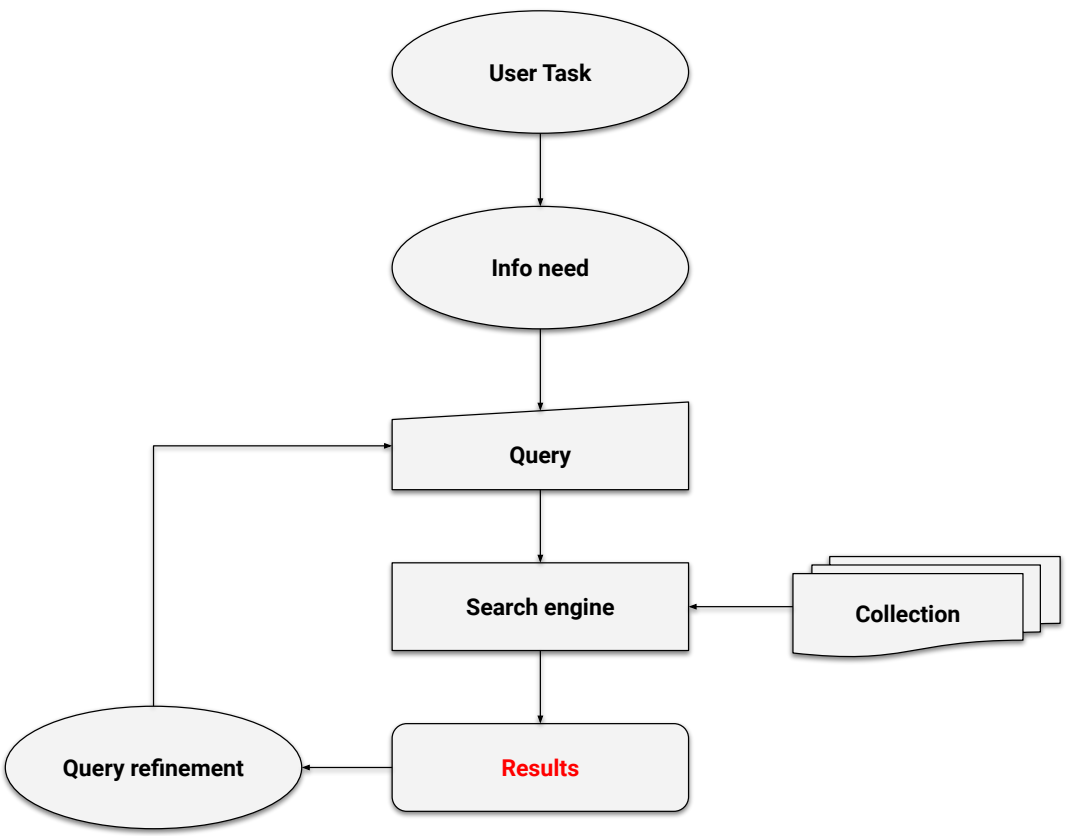
- Relational Database Management Systems (RDBMS)
  - Semantics of each object are well defined
  - Complex query languages (e.g., SQL)
  - Exact retrieval for what you ask
  - Emphasis on efficiency

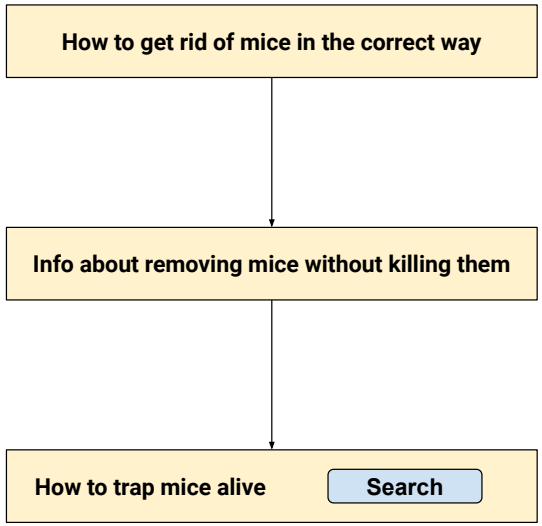
- Information Retrieval (IR)
  - Semantics of object are subjective, not well defined
  - Usually simple query languages (e.g., natural language query)
  - You should get what you want, even the query is bad
  - Effectiveness is primary issue, although efficiency is important

The classic search model

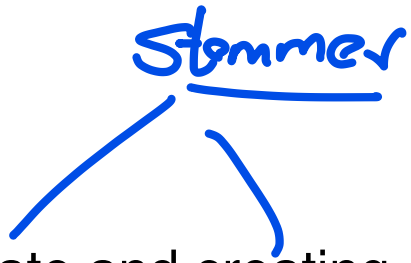


The classic search model

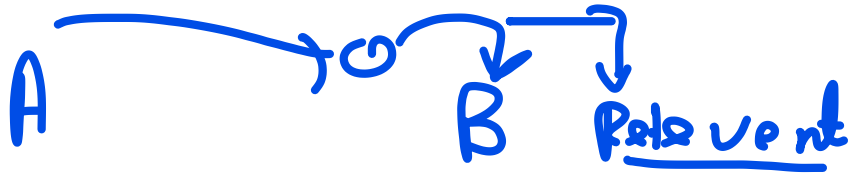




Core Concepts of IR



- Query Representation
  - Bridge lexical gap: system and systems; create and creating (stemmer)
  - Bridge semantic gap: car and automobile (feedback)
- Document Representation
  - Internal representation of document contents: a list of documents that contain specific word (inverted document list)
  - Representation of document structure: different fields (e.g., title, body)
- Retrieval Model
  - Algorithms that best match meaning of user query and available documents. (e.g., vector space model and statistical language modeling)



How good are the retrieved documents?

- **Precision:** Fraction of the retrieved documents that are relevant to the user's information need
- **Recall:** Fraction of relevant documents in collection that are retrieved
  - More precise definitions and measurements to follow later