INFORMATION RETRIEVAL

- Text is the primary way that human knowledge is stored after speech.
- Techniques for storing and searching for textual documents are nearly as old as written language itself.
- In past, information retrieval means going to town's library and asking the librarian for help.
- The librarian usually knew all the books in the possession and could give one a define answer.
- As the number of books grew, it became impossible. Then tools for information retrieval had to be devised.
- One of the most important tools is indexing.
- Index is a terms with pointer to places where information about them can be found.
- The terms can be subject matter, author names, etc.
- Oliver Wendell Holmes wrote in 1872, "It is the province of knowledge to speak and it is the privilege of wisdom to listen".
- In future, "It is the province of knowledge to write and it is the privilege of wisdom to query."
- The field of computer science that deals with the automated storage and retrieval of a document is called information retrieval.
- Requires:
    - o Algorithm – For manipulating natural language.
    - o Data Structures – To efficiently store and process data.

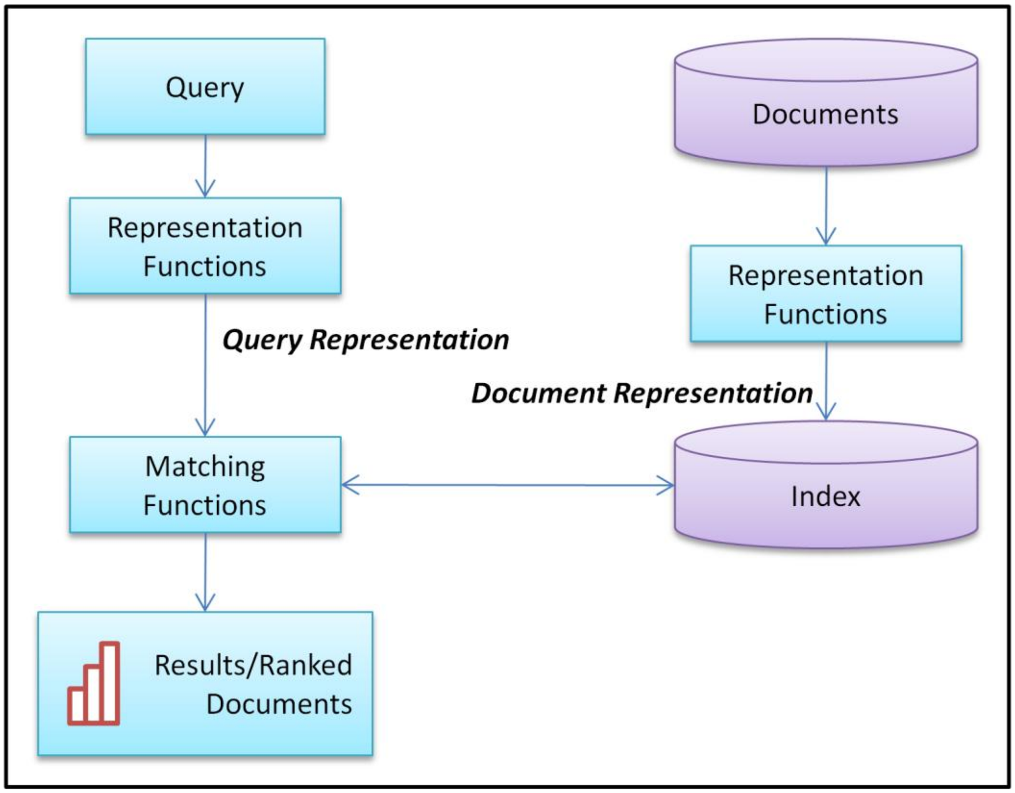
WHAT MAKES IR A HARD PROBLEM?

1. Under good circumstances
    - Text is unstructured.
    - Requires understanding of semantics. For example: restaurant → café, PRC → China, fast automobiles → fast cars.
    - Human language presents distinct problems like ambiguity. For example: bat (mammal or baseball), apple (company or fruit), bit (unit of data or act of eating), etc.

2. Under hard circumstances
   - Web pages change rapidly.
   - Many pages lie about their content.
   - New pages are not linked to.

3. Multimedia information
   - Hard to store (size), represent and compare.

IR SYSTEM



INFORMATION RETRIEVAL SYSTEM

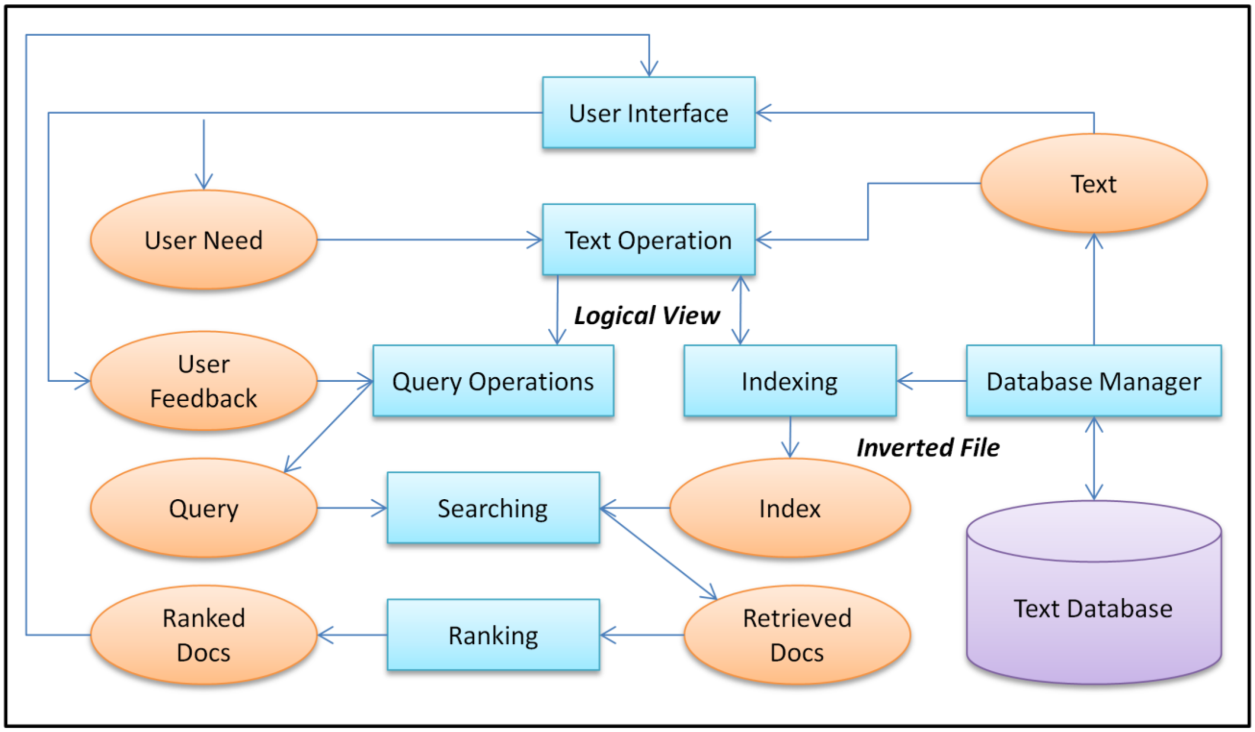
- Searching for pages on the World Wide Web is the most recent killer application.
- IR concerns firstly with retrieving relevant documents as a query.

- Relevance is a subjective judgment and may include:
  1. Being on the proper subject.
  2. Being timely (recent information).
  3. Being authoritative (from a trusted source).
  4. Satisfy the goals of the user.

TYPICAL IR

1. Given
   - A corpus of textual natural language documents.
   - A user query in the form of a textual string.
2. Find
   - A ranked set of documents that is relevant to the query.

IR SYSTEM ARCHITECTURE

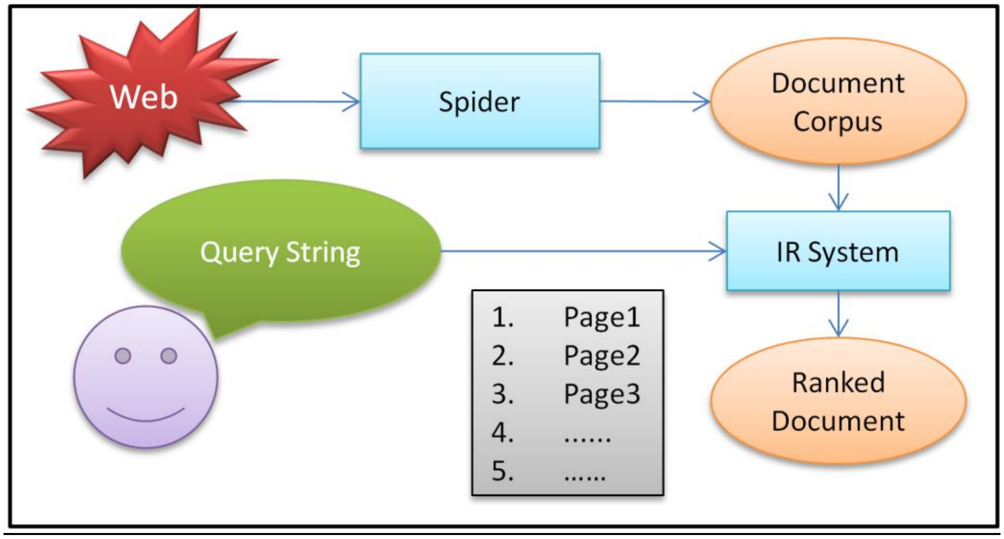


INFORMATION RETRIEVAL ARCHITECTURE

IR SYSTEM COMPONENTS

1. Text Operations
   - Forms index words (tokens) by stop-word removal and stemming.
2. Indexing
   - Constructs an inverted index of word to document pointers.
3. Searching
   - Retrieves documents that contain a given query token from the inverted index.
4. Ranking
   - Scores all retrieved documents according to relevance matrix.
5. User Interface
   - Manage interaction with the user.
   - Query input and document output.
   - Relevance feedback.
   - Visualization of results.
6. Query Operations
   - Transform the query to improve retrieval.
   - Query expansion using thesaurus.

WEB SEARCH AND IR (Application of IR to HTML documents on the World Wide Web)



WEB CHALLENGES OF IR

1. Distributed Data
   - Documents spread over millions of different web servers.
2. Volatile Data
   - Many documents change or disappear rapidly. For example: dead link.
3. Large Volume
   - Billions of separate documents.
4. Unstructured and Redundant Data
   - No Uniform Structure.
   - Up to 30% (near) are duplicate documents
5. Quality of Data
   - No editorial control.
   - False information.
   - Poor quality writing.
6. Heterogeneous Data
   - Multiple media types (image, video)
   - Languages.

AREAS OF AI FOR IR

1. Natural Language Processing
   - Focused on syntactic, semantic and pragmatic analysis of natural language text.
   - Retrieval based should be focused on semantic.
   - Methods for determining the sense of ambiguous word based on context.
   - Question answering.
2. Machine Learning
   - Focused on the development of computational system that improves their performance with experience.
   - Automated classification of examples based on learning concepts from labeled training.

- For example: supervised learning.
- Automated methods for clustering unlabeled examples into meaningful groups (unsupervised).
- Text categorization (For example: spam filtering).
- Text clustering (clustering of IR query results).
- Text mining.

3. Knowledge Representation
   - Expert system
4. Reasoning Under Uncertainty
   - Bayesian network
5. Cognitive Theory

STEPS IN IR PROCESS (RETRIEVAL PROCESS)

1. Indexing (Creating document representation)
   - Indexing is the manual or automated process of making statements about a document, lesson, and person and so on.
   - For example: author wise, subject wise, text wise, etc.
   - Index can be:
     i. Document oriented: – the indexer accesses the document relevance to subjects and other features of interests to user.
     ii. Request oriented: – the indexer accesses the document relevance to subjects and other features of interests to user.
   - Automated indexing begins with feature extraction such as extracting all words from a text, followed by refinements such as eliminating stop words (a, an, the, of), stemming (walking → walk), counting the most frequent words, mapping the concepts using thesaurus (tube → pipe).

2. Query Formulation (Creating query representation)
   - Retrieval means using the available evidence to predict the degree to which a document is relevant or useful for a given user need as described in a free form query description.

- A query can specify text words or phrase, the system should look for.
- The query description is transformed manually or automatically into a formed query representation, ready to match with document representation.

3. Matching the Query Representation With Entity Representation
    - The match uses the features specified in the query to predict document relevance.
    - Exact match (0 or 1).
    - Synonym expansion (pipe → tube).
    - Hierarchical expansion (pipe → capillary).
    - The system ranks the result.

4. Selection
    - User examines the results and selects the relevant items.

5. Relevance Feedback & Interactive Retrieval
    - The system can assist the user in improving the query by showing a list of features (option) found in many relevant items.