Search Engines and Information Retrieval

Unit 1 Syllabus

- **Introduction**
- **Search Engines and Information Retrieval :** What Is Information Retrieval? , The Big Issues, Search Engines, Search Engineers
- **Architecture of a Search Engine:** What Is an Architecture? ,Basic Building Blocks ,Breaking It Down, Text Acquisition Text Transformation , Index Creation, User Interaction, Ranking ,Evaluation, How Does It Really Work?
- **Boolean retrieval:** An example information retrieval problem , A first take at building an inverted index, Processing Boolean queries, The extended Boolean model versus ranked retrieval
- **Vocabulary and postings lists:** Document delineation and character sequence decoding: Obtaining the character sequence in a document, Choosing a document UNIT
- Determining the vocabulary of terms: Tokenization, Dropping common terms: stop words, Normalization (equivalence classing of terms), Stemming and lemmatization, Faster postings list intersection via skip pointers
- Positional postings and phrase queries: Biword indexes, Positional indexes, Combination schemes

What Is Information Retrieval?

- Information retrieval (IR) is a process that facilitates the effective and efficient retrieval of relevant information from large collections of unstructured data.
- Difference between a document and a typical database record.
- Information retrieval involve multimedia documents
- Types of search :

1. Web search (World Wide web search)
2. *Vertical search* - domain of the search is restricted to a particular topic.
3. *Enterprise search* involves finding the required information in the huge variety of computer files scattered across a corporate intranet.
4. *Desktop search* is the personal version of enterprise search, where the information sources are the files stored on an individual computer, including email messages and web pages that have recently been browsed.
5. *Peer-to-peer search* involves finding information in networks of nodes or computers without any centralized control
6. *Ad hoc search* - Search based on a user query

- Other tasks include *filtering*, *classification*, and *question answering*
  - Filtering involves detecting stories of interest based on a person's interests
  - Classification uses a defined set of labels to label the documents
  - Question answering is similar to search but is aimed at more specific questions, such as What is the time now in Tokyo? Who is the Chief minister of Karnataka.

Some dimensions of information retrieval

Examples of Content	Examples of Applications	Examples of Tasks
Text	Web search	Ad hoc search
Images	Vertical search	Filtering
Video	Enterprise search	Classification
Scanned documents	Desktop search	Question answering
Audio	Peer-to-peer search	
Music		

The Big Issues

- Relevance is a fundamental concept in information retrieval.
  - A relevant document contains the information that a person was looking for when she submitted a query to the search engine.
  - It is not simply comparing the text of a query with the text of a document and looking for an exact match (a database system or using the grep utility in Unix)
  - Language can be use to express the same concepts in many different ways, often with very different words. - This is referred to as the *vocabulary mismatch problem* in information retrieval.
  - It is also important to distinguish between *topical relevance* and *user relevance.*
    - *Topical relevance considers the overall subject matter of content rather then just keywords.*
    - *User relevance focuses on creating a connection between users and content – helps attract, engage and retain users.*
  - *Ranking Algorithms used*

- Another core issue for information retrieval is *evaluation*.
  - two of the popular measures used are, *precision* and *recall*
  - Precision is the proportion of retrieved documents that are relevant.
  - Recall is the proportion of relevant documents that are retrieved
- The third core issue for information retrieval is the emphasis on users and their *information needs*.
  - text queries are often poor descriptions
  - A one-word query such as “cats” could be a request for information on where to buy cats or for a description of breeds of cats and so on.
  - Despite their lack of specificity, however, one-word queries are very common in web search.
  - Techniques such as *query suggestion*, *query expansion*, and *relevance feedback* use interaction and context to refine the initial query in order to produce better ranked lists.

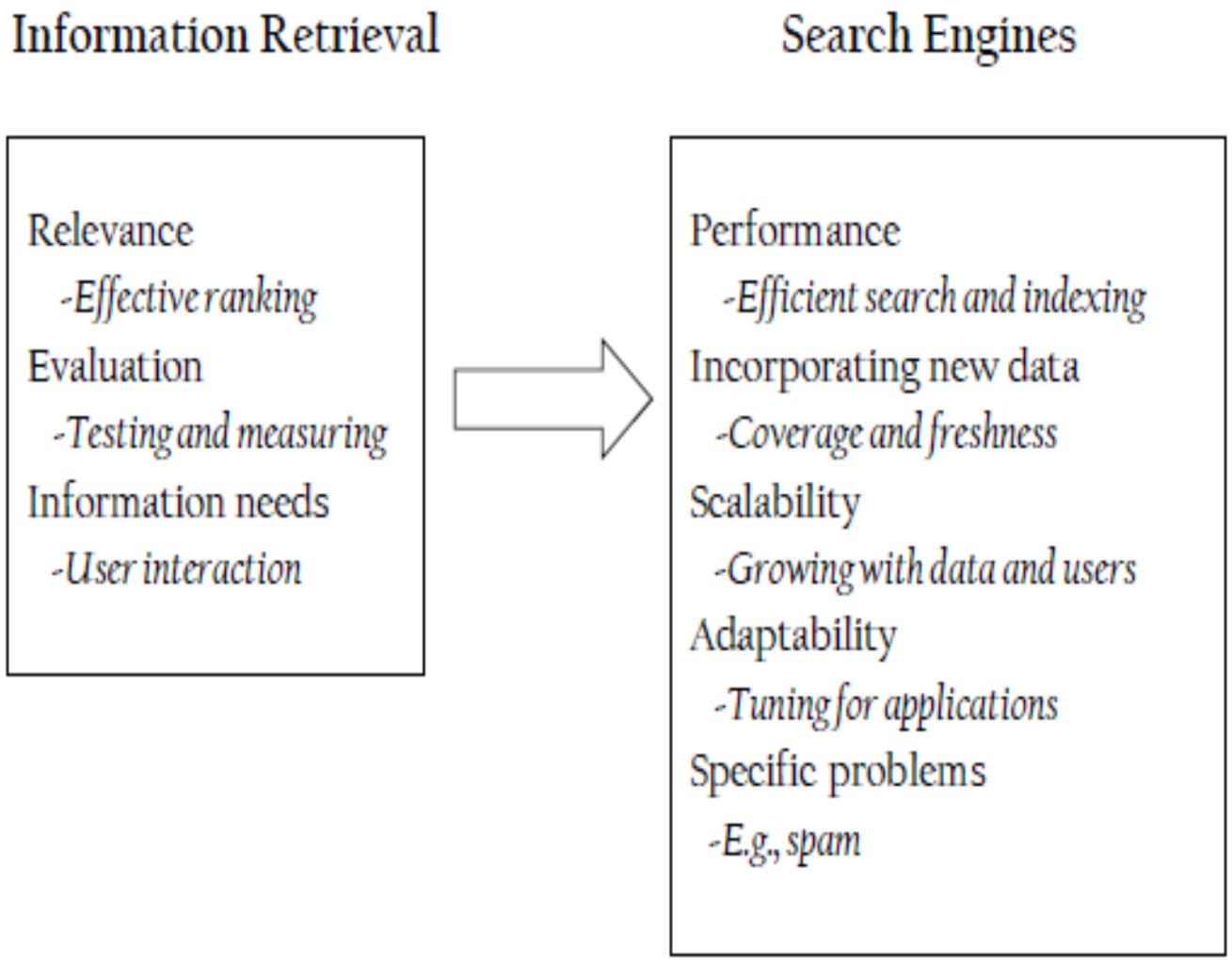
Search Engine

- A search engine is the practical application of information retrieval techniques to large-scale text collections. Eg. A web search engine
- Search engines come in a number of configurations that reflect the applications they are designed for.
- Web search engines, such as Google and Yahoo!, must be able to capture, or *crawl*, many terabytes of data, and then provide subsecond response times to millions of queries submitted every day from around the world.
- Enterprise search engines—for example, Autonomy—must be able to process the large variety of information sources in a company and use company-specific knowledge as part of search and related tasks, such as *data mining*.
- Desktop search engines, such as the Microsoft Vista™ search feature, must be able to rapidly incorporate new documents, web pages, and email as the person creates or looks at them, as well as provide an intuitive interface for searching this very heterogeneous mix of information.

- *Open source* search engines are another important class of systems that have somewhat different design goals than the commercial search engines.
- Three systems of particular interest are Lucene,Lemur and Galago
- The “big issues” in the design of search engines include the ones identified for information retrieval: effective ranking algorithms, evaluation, and user interaction.
- Additional critical features of search engines that result from their deployment in large-scale, operational environments.
  - *Performance* of the search engine in terms of measures such as *response time*, *query throughput*, and *indexing speed*.
  - Response time is the delay between submitting a query and receiving the result list
  - Throughput measures the number of queries that can be processed in a given time, and
  - Indexing speed is the rate at which text documents can be transformed into indexes for searching. An *index* is a data structure that improves the speed of search.
  - Another important performance measure is how fast new data can be incorporated into the indexes. Search applications typically deal with dynamic, constantly changing information.
    - *Coverage* measures how much of the existing information has been indexed and stored in the search engine, and
    - *Recency* or *freshness* measures the “age” of the stored information.

- *Scalability* is clearly an important issue for search engine design. Designs that work for a given application should continue to work as the amount of data and the number of users grow.
- Search Engines have to be *customizable* or *adaptable*. This means that many different aspects of the search engine, such as the ranking algorithm, the interface, or the indexing strategy, must be able to be tuned and adapted to the requirements of the application.
- Spam is an unwanted email, but more generally it could be defined as misleading, inappropriate, or non-relevant information in a document that is designed for some commercial benefit.
  - Search engines must handle spam words put into a document to cause it to be retrieved in response to popular queries.
  - The practice of “spamdexing” can significantly degrade the quality of a search engine’s ranking, and web search engine designers have to develop techniques to identify the spam and remove those documents.

Search engine design and the core information retrieval issues



Search Engineers

- Search engineers are primarily people trained in computer science, mostly with a systems or database background. Role of such engineers include
  - designing and implementing new search engines
  - modify, extend, maintain, or tune existing search engines for a wide range of commercial applications.
  - design or “optimize” content for search engines
  - implement techniques to deal with spam.
- They primarily use open source and enterprise search engines for application development, but also get the most out of desktop and web search engines.

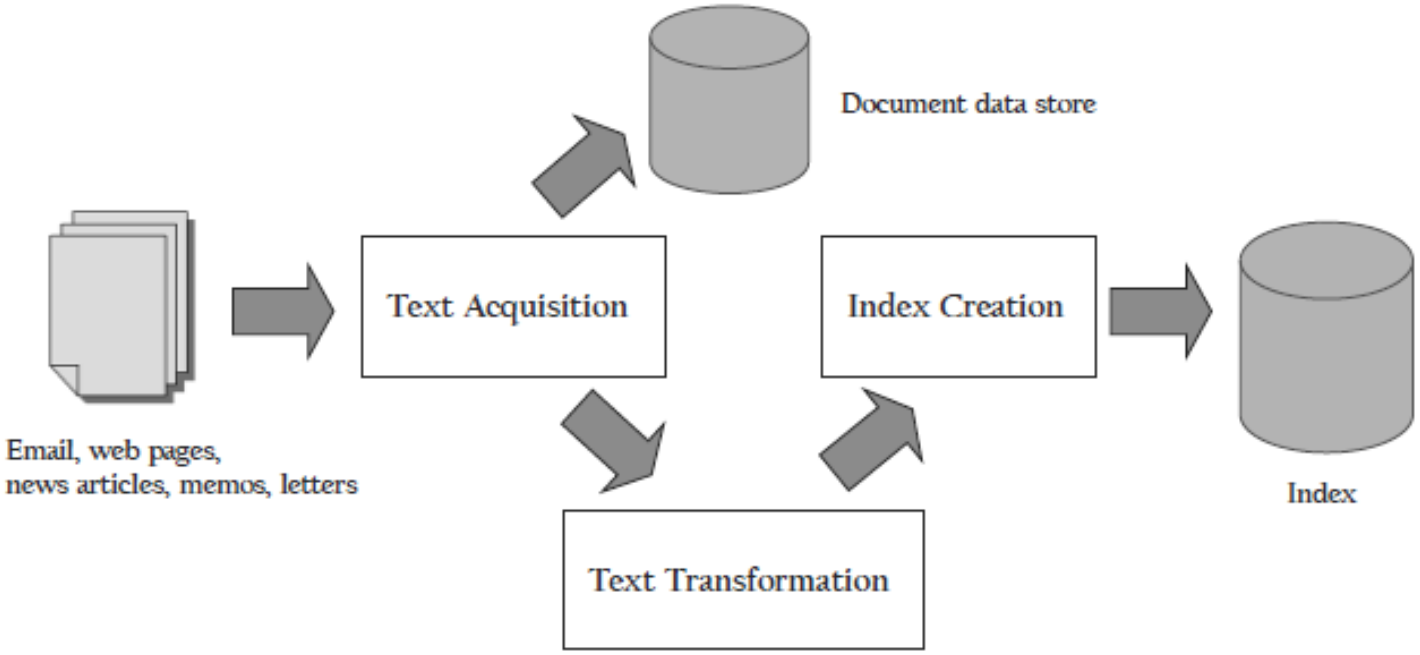
Architecture of a Search Engine

- What is Architecture?
- A search engine architecture is used to present high-level descriptions of the important components of the system and the relationships between them.
- The two primary goals of a search engine are:
  - Effectiveness (quality): We want to be able to retrieve the most relevant set of documents possible for a query.
  - Efficiency (speed): We want to process queries from users as quickly as possible.
- Making sure that the search engine immediately reacts to changes in documents is both an effectiveness issue and an efficiency issue.
- For an efficient system, search engines employ specialized data structures that are optimized for fast retrieval.
- Because we want high-quality results, search engines carefully process text and store text statistics that help improve the relevance of results.

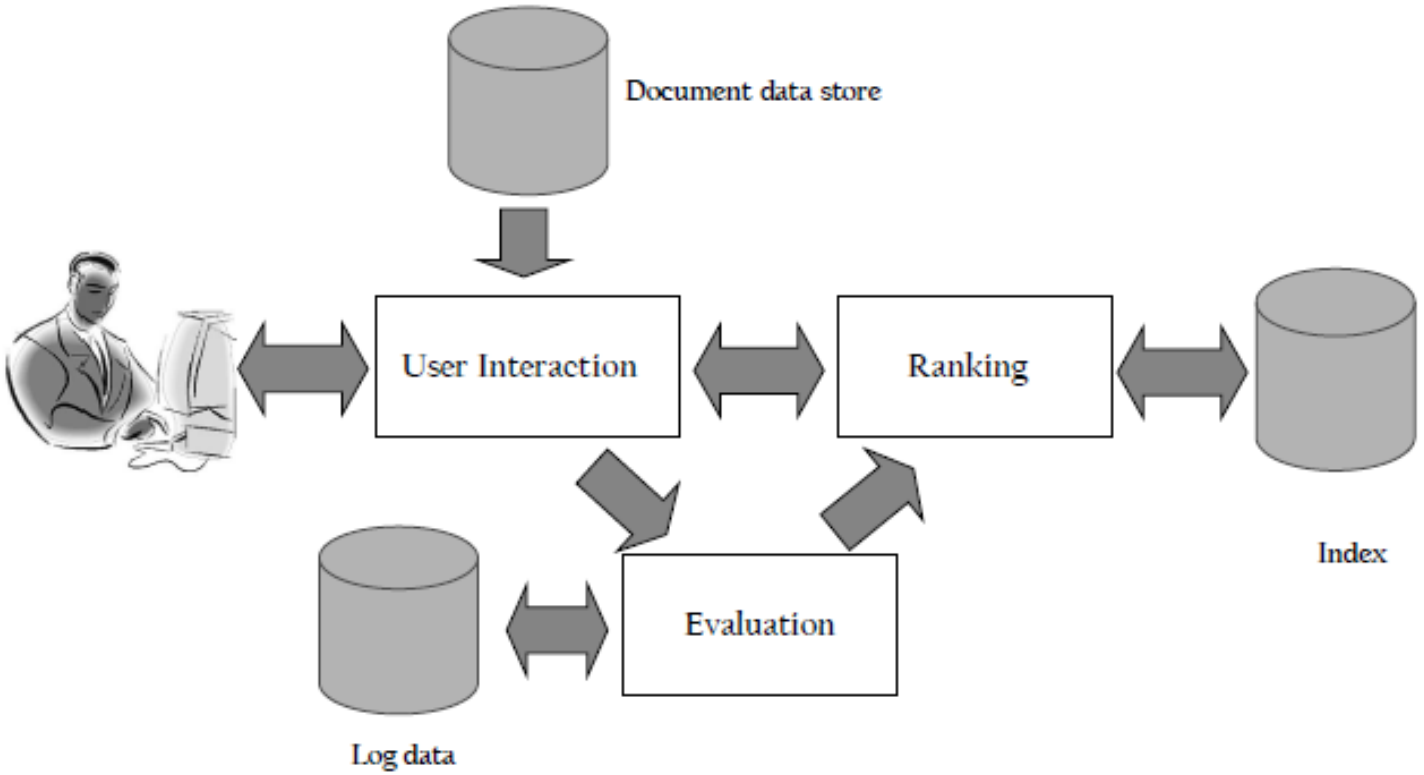
Basic Building Blocks

- Search engine components support two major functions, which we call the *indexing process* and the *query process*.
- The indexing process builds the structures that enable searching, and
- the query process uses those structures and a person's query to produce a ranked list of documents.

The indexing process



The query process



- The indexing process builds the structures that enable searching, and
- the query process uses those structures and a person's query to produce a ranked list of documents
- These major components of indexing process are text acquisition, text transformation, and index creation.
- Text Acquisition
  - Crawler
  - Feeds
  - Conversion
  - Document data store

- Text Transformation
  - Parser
  - Stopping
  - Stemming
  - Link Analysis
  - Information Extraction
  - Classifier
- Index Creation
  - Document Statistics
  - Weighting
  - Inversion
  - Distribution

- The major components of the query process are user interaction, ranking, and evaluation.
- User Interaction
  - Query input
  - Query transformation
  - Results Output
- Ranking
  - Scoring
  - Optimization
  - Distribution
- Evaluation
  - Logging
  - Ranking Analysis
  - Performance Analysis