# Information Retrieval

Introduction
IR models and methods

# Search and Information Retrieval

- Search on the Web is a daily activity for many people throughout the world
- Search and communication are the most popular uses of the computer
- Applications involving **search** are **everywhere**

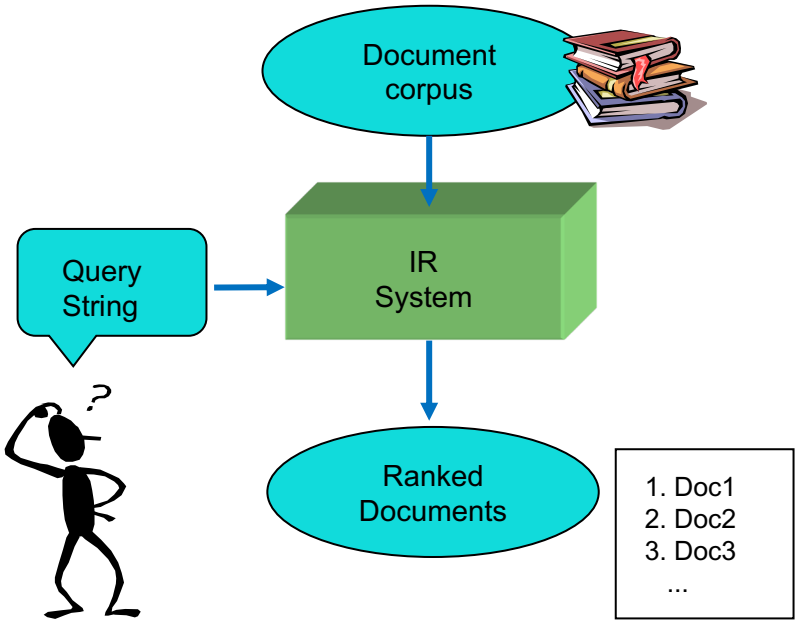
- The field of computer science that is most involved with R&D for search is **information retrieval (IR)**

# 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)."*
  (Manning, et al, 2008)

# Typical IR Task

- Given:
  - A set of documents (corpus)
  - A user query in the form of a textual string
- Find:
  - A ranked set of documents with information that is relevant to the user's information need and helps the user complete a task



# What is a Document?

- Examples:
  - web pages, emails, books, news stories, scholarly papers, text messages, forum postings, patents, social media, etc.
  - TXT, PDF, HTML, JSON, XML, ..etc.

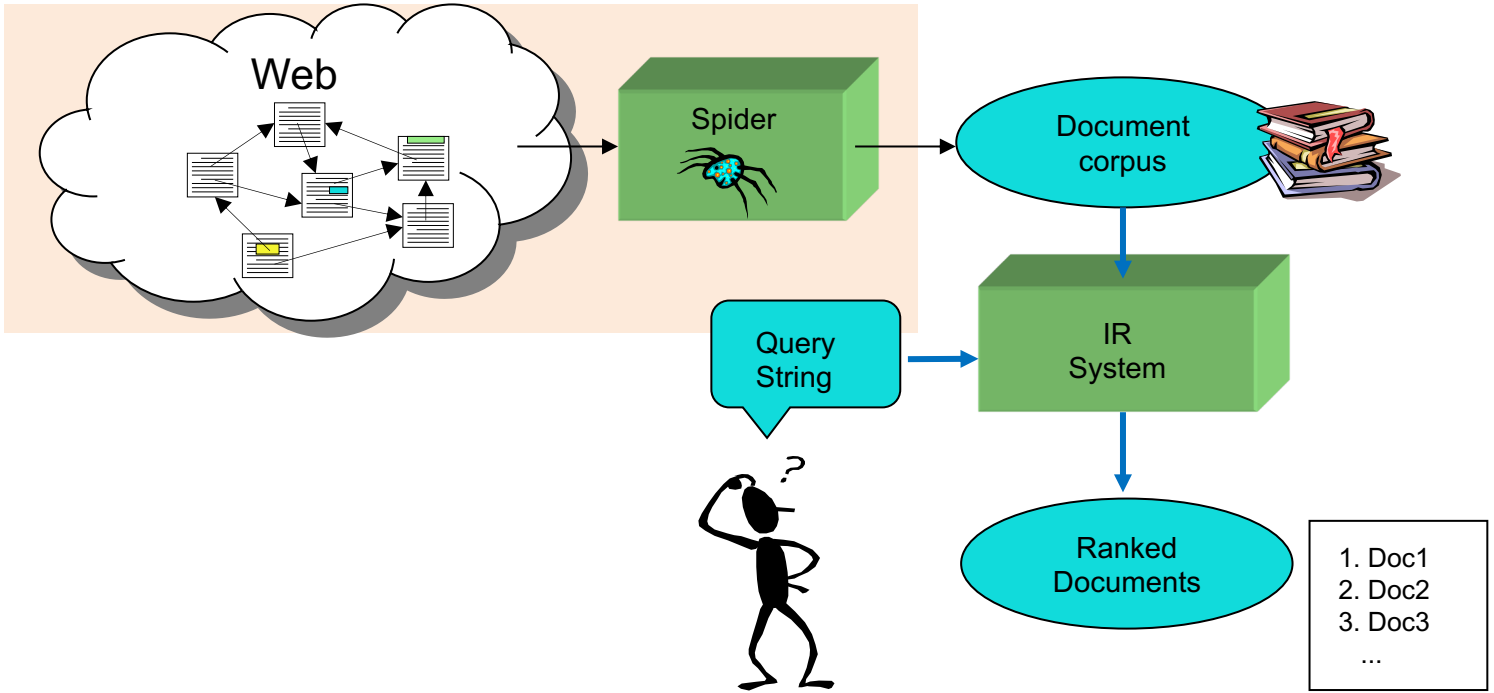
- Common properties
  - Significant text content
  - Some structure
    - e.g.,
    - title, author, and date, for papers
    - subject, sender, and destination, for emails

# Web Search

- Application of IR to HTML documents

- Differences:
  - Must assemble document corpus by spidering the web
  - Can exploit the structural layout information in HTML
  - Can exploit the link structure of the web
  - Documents change uncontrollably

# Web Search


Web
Spider
Document corpus
Query String
IR System
Ranked Documents
1. Doc1
2. Doc2
3. Doc3
...

# Dimensions of IR

| Content | Applications | Tasks |
|---|---|---|
| Text | Web search | Ad hoc search |
| Images | Vertical search | Filtering |
| Video | Enterprise search | Classification |
| Scanned docs | Desktop search | Question answering |
| Audio | Forum search | |
| Music | P2P search | |
| | Literature search | |

# IR – Related Areas

- Automated document categorization
- Information filtering (spam filtering)
- Automated document clustering
- Recommending information or products
- Information extraction
- Information integration
- Question answering

# IR – Related Areas

- Database Management
  - Easy to compare fields with well-defined semantics to queries in order to find matches
- Library and Information Science
  - Digital libraries
- Artificial Intelligence
  - Web ontologies and intelligent information agents
- Natural Language Processing
  - Analyzing syntax (phrase structure) and semantics
  - Disambiguation
- Machine Learning
  - Classification / Clustering
  - Learning to rank
- Representation Learning and Deep Learning
  - Neural language models
  - Can combine Machine Learning  with Semantics (and syntax)

# 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.
- Also, most RDBMS has some kind of support for full-text indexing and searching
  - A full-text index in MySQL is an index of type FULLTEXT.

# Documents vs. Database Records

- Clearer semantics for both data and queries
  - Database records are typically made up of well-defined fields (or attributes)
  - e.g., bank records with account numbers, balances, names, addresses, social security numbers, dates of birth, …
  - Easy to compare fields with well-defined semantics to queries in order to find matches

- Text is more difficult

# Documents vs. Database Records

- Example bank database query
  - *Find records with balance > $50,000 in branches located in Amherst, MA.*
  - Matches easily found by comparing against field values of records

- Example search engine query
  - *bank scandals in western massachusetts*
  - This text must be compared to the text of entire news stories

# 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
- 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

# Machine Learning

- AI branch 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)

# (Deep) Representation Learning

- Recent advances in deep learning bring together the NLP (semantics, syntax) and Machine Learning
- Neural language models

# Big Issues in IR

- Relevance
- Evaluation
- Information needs

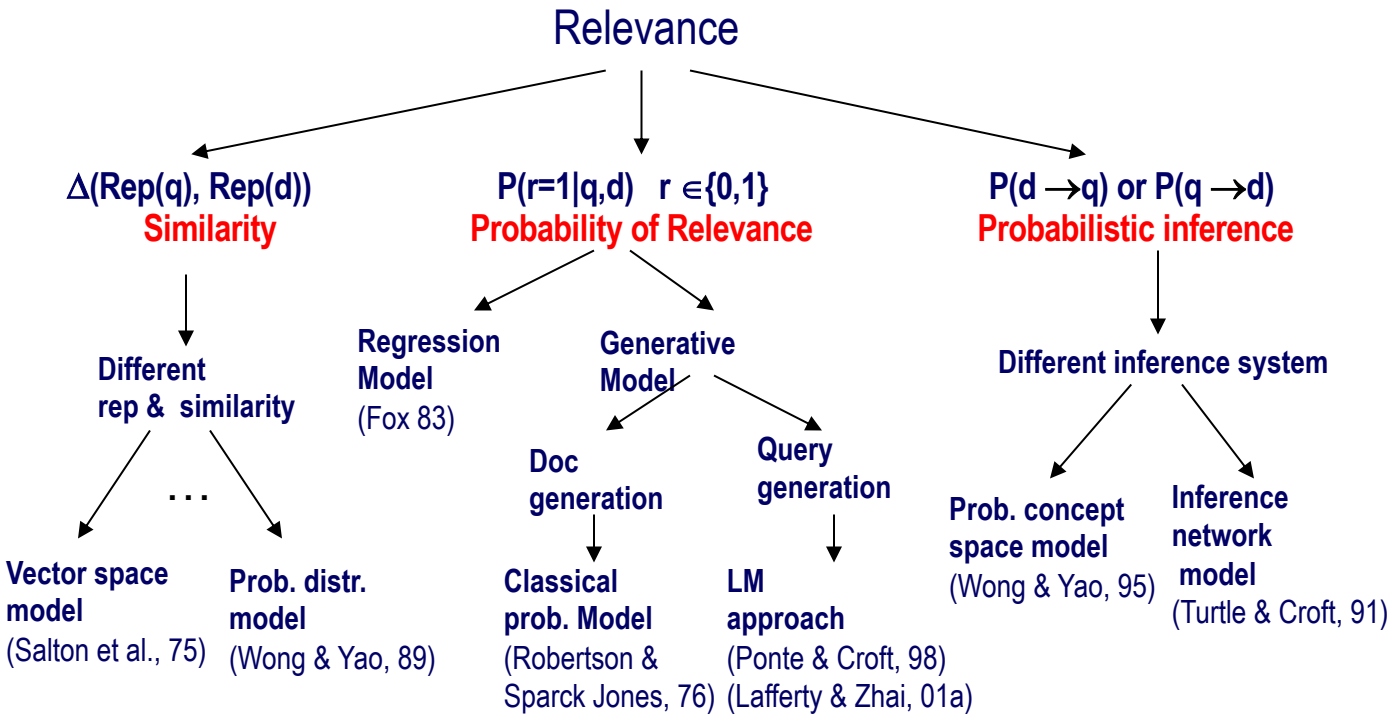
# Big Issues in IR - Relevance

- What is it?
  - Simple (and simplistic) definition: A relevant document contains the information that a person was looking for when they submitted a query to the search engine
- Many factors influence a person's decision about what is relevant: e.g., task, context, novelty, style
- Topical relevance (same topic) vs. user relevance

# Big Issues in IR - Relevance

- Ranking algorithms used in search engines are based on retrieval models
- Each retrieval model defines a view of relevance
- Most models describe statistical properties of text rather than linguistic
  - i.e. counting simple text features such as words instead of parsing and analyzing the sentences
  - Statistical approach to text processing started with Luhn in the 50s
  - Linguistic features can be part of a statistical model

# Notion of relevance –> IR Models



# Notion of relevance –> IR Models

- Boolean model
- Vector-space model
- Probabilistic model
- Language-based model
- Neural model

# Big Issues in IR - Evaluation

- Experimental procedures and measures for comparing system output with user expectations
  - Originated in Cranfield experiments in the 60s
- IR evaluation methods now used in many fields
- Typically use test collection of documents, queries, and relevance judgments
  - Most commonly used are TREC collections
- Recall and precision are two examples of effectiveness measures

# Big Issues in IR - Evaluation

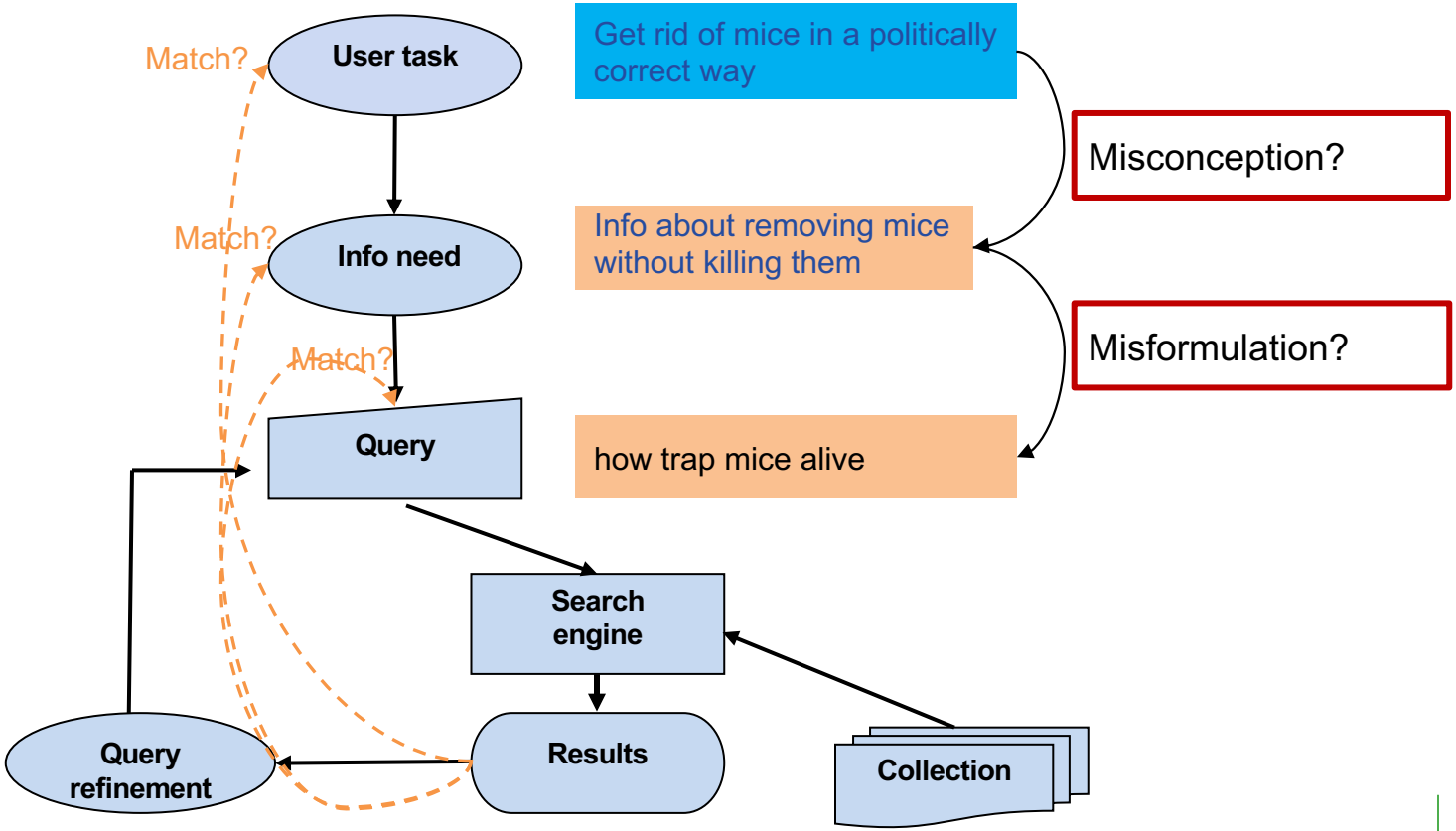
- How good are the retrieved docs?
  - *Precision* : Fraction of retrieved docs that are relevant to the user's information need
  - *Recall* : Fraction of relevant docs in collection that are retrieved

  - We will look at more precise definitions and measurements later

# Big Issues in IR

- Users and Information Needs
  - Search evaluation should be user-centered
  - Keyword queries are often poor descriptions of actual information needs
  - Interaction and context are important for understanding user intent
  - Query refinement techniques such as *query expansion*, *query suggestion*, *relevance feedback* improve ranking

# The classic search model



# IR and Search Engines

- A search engine is the practical application of information retrieval techniques to large scale text collections

- Web search engines are best-known examples, but many others (e.g. enterprise search)
  - *Open source* search engines are important for research and development (e.g., Lucene,Solr, Elasticsearch, Sphinx, Nutch, …)

# Search Engine Issues

- Performance
  - Measuring and improving the efficiency of search
    - e.g., reducing response time, increasing query throughput, increasing indexing speed
  - Indexes are data structures designed to improve search efficiency
    - designing and implementing them are major issues for search engines
- Dynamic data
  - The "collection" for most real applications is constantly changing in terms of updates, additions, deletions
    - e.g., web pages
  - Acquiring or "crawling" the documents is a major task
    - Typical measures are coverage (how much has been indexed) and freshness (how recently was it indexed)
  - Updating the indexes while processing queries is also a design issue

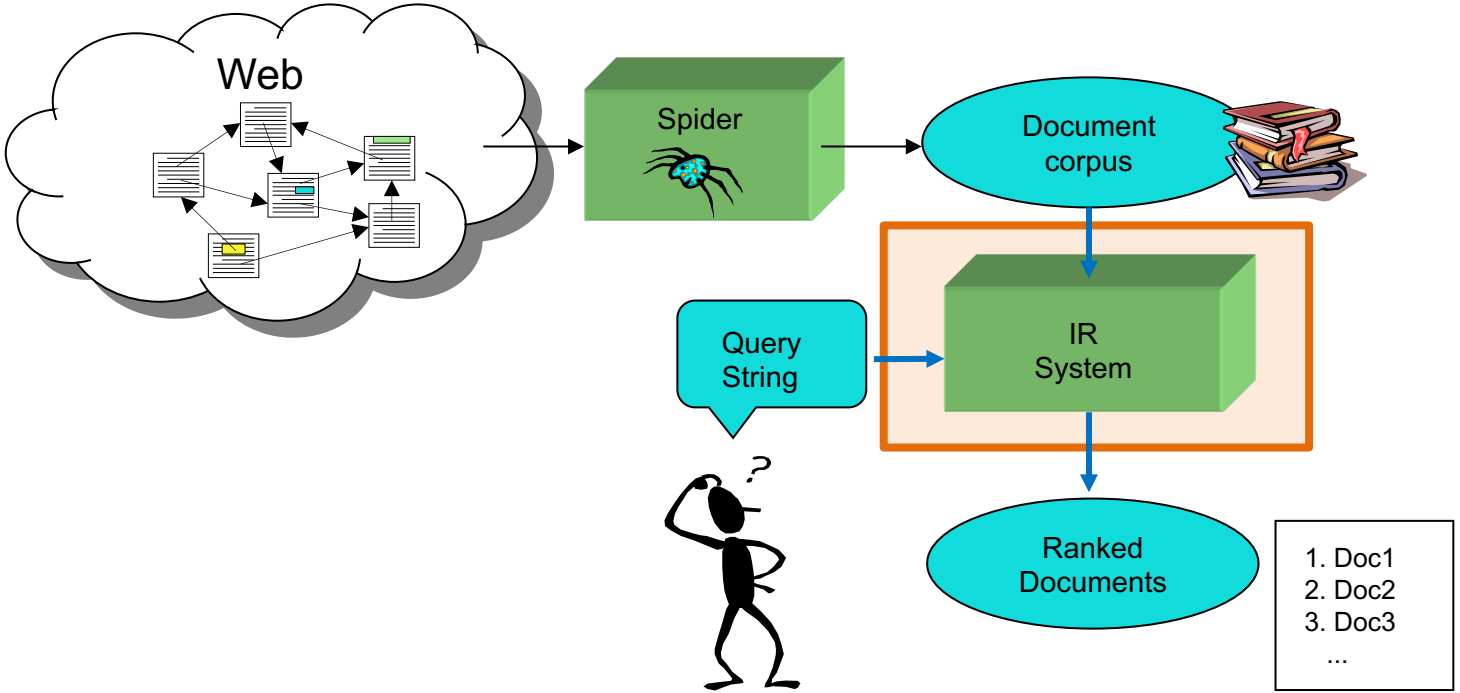
# Search Engine Issues

- Scalability
  - Making everything work with millions of users every day, and many terabytes of documents
  - Distributed processing is essential
- Adaptability
  - Changing and tuning search engine components such as ranking algorithm, indexing strategy, interface for different applications

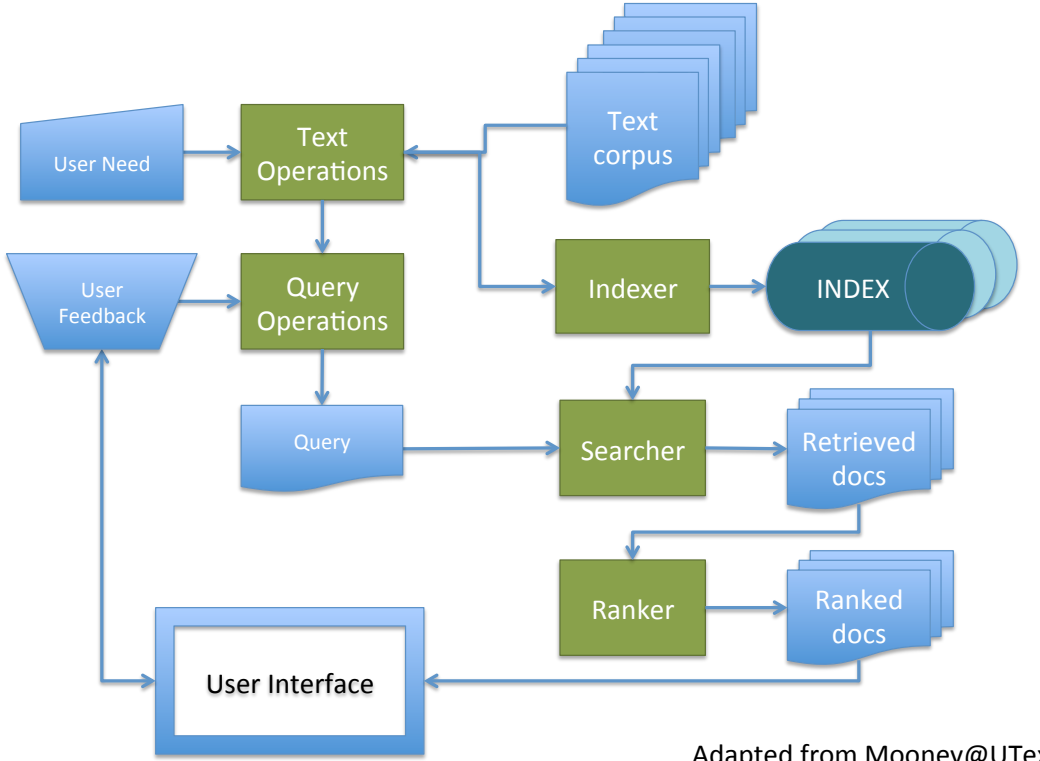
# Spamdexing

- For Web search, spam in all its forms is one of the major issues
- Affects the efficiency of search engines and, more seriously, the effectiveness of the results
- Many types of spamdexing
  - Content spam
    - keyword stuffing, hidden text, meta-tag stuffing, doorway pages,…
  - Link spam
    - link-building software, link farms, hidden links, …

# IR System Architecture

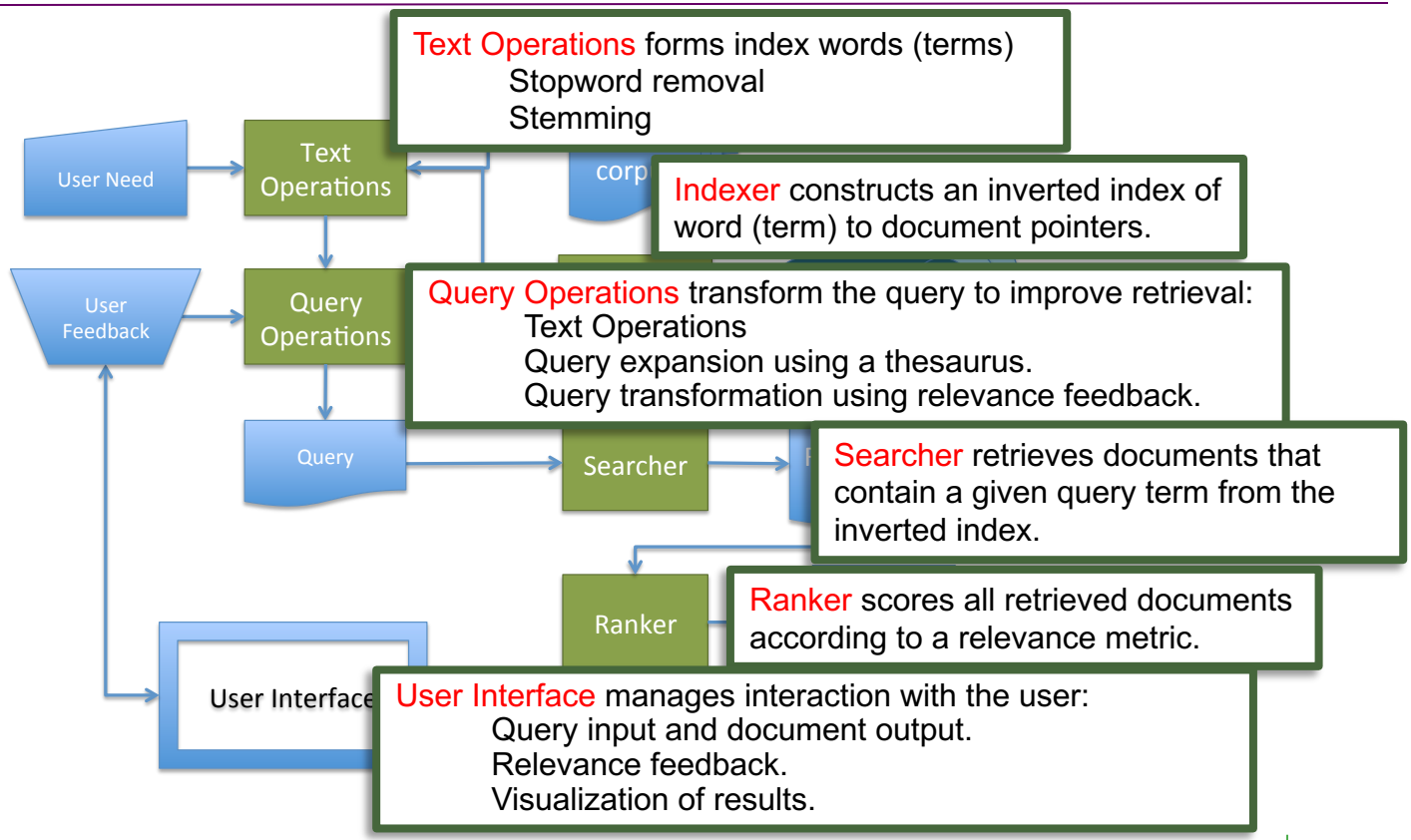


# IR System Architecture

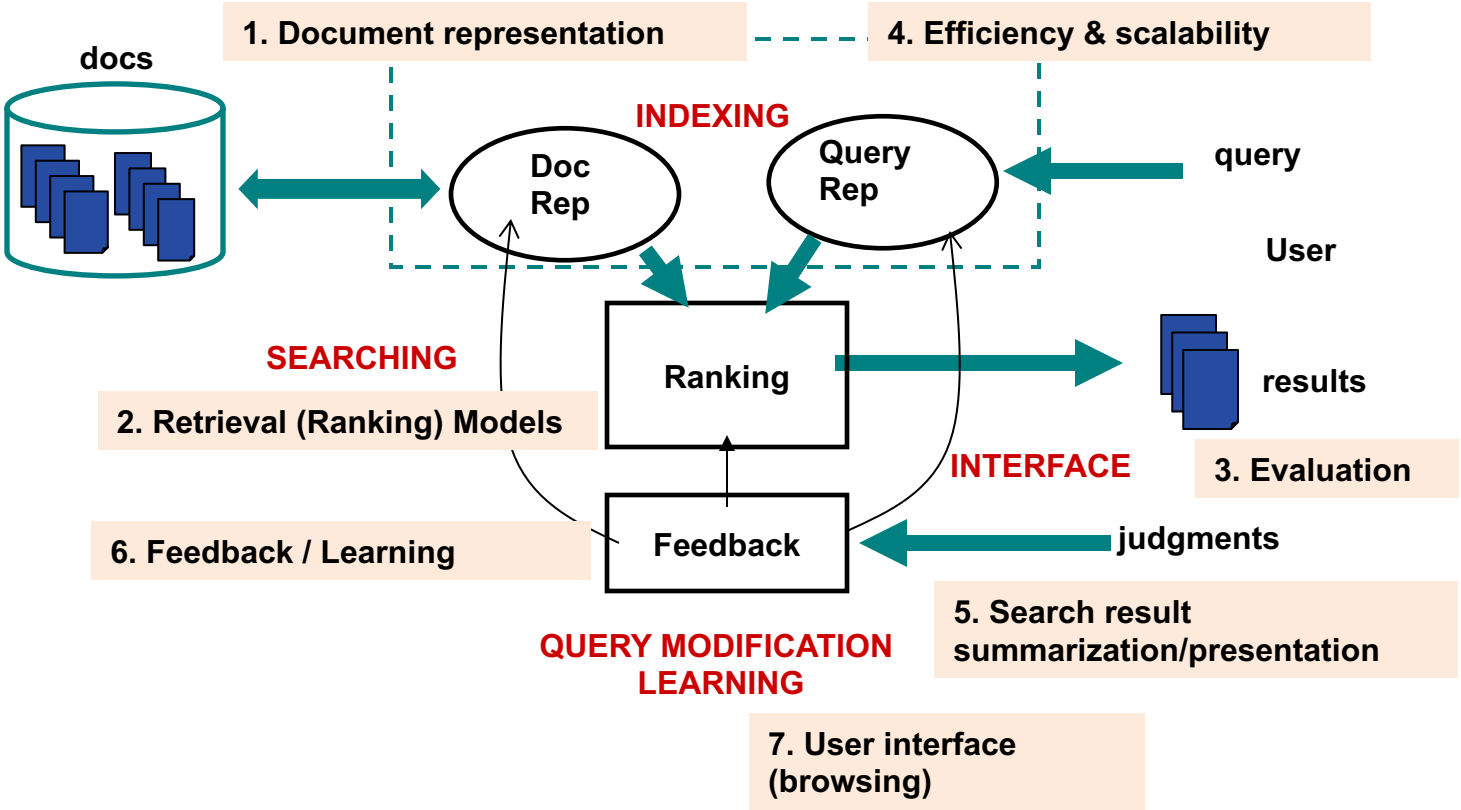



Adapted from Mooney@UTexas

# IR System Architecture

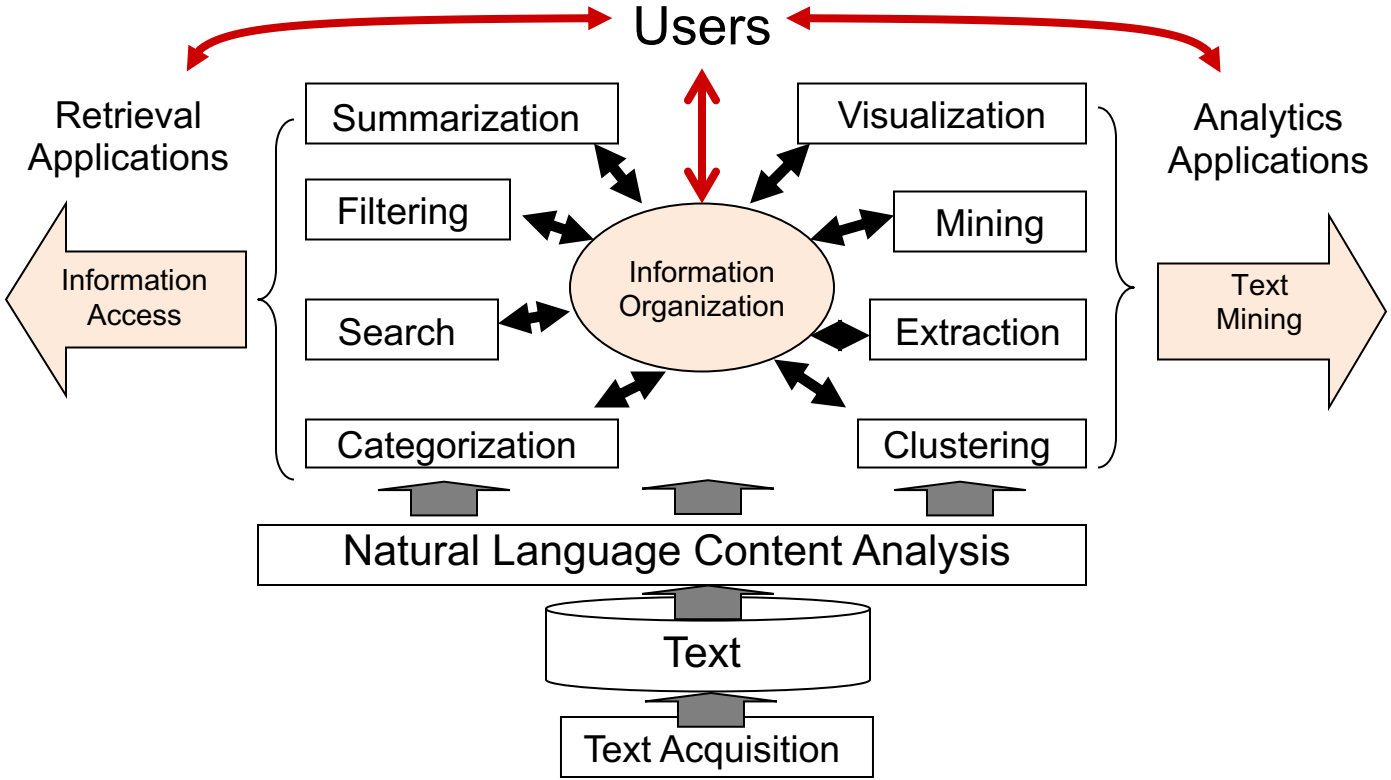


# IR Research Topics (narrow view)




Slide from: http://times.cs.uiuc.edu/course/598f16

# IR Research Topics (Broad View)




Slide from: http://times.cs.uiuc.edu/course/598f16

# Key Terms Used in IR

- **QUERY**
  - a representation of what the user is looking for - can be a list of words or a phrase.
- **DOCUMENT**
  - an information entity that the user wants to retrieve
- **COLLECTION**
  - a set of documents
- **INDEX**
  - a representation of information that makes querying easier
- **TERM**
  - word or concept that appears in a document or a query

# Other Important Terms

- Classification
- Cluster
- Similarity
- Information Extraction
- Term Frequency
- Inverse Document Frequency
- Precision
- Recall
- Inverted File
- Query Expansion
- Relevance
- Relevance Feedback
- Stemming
- Stopword
- Vector Space Model
- Weighting
- TREC/TIPSTER/MUC