Information Retrieval

Introduction
IR models and methods

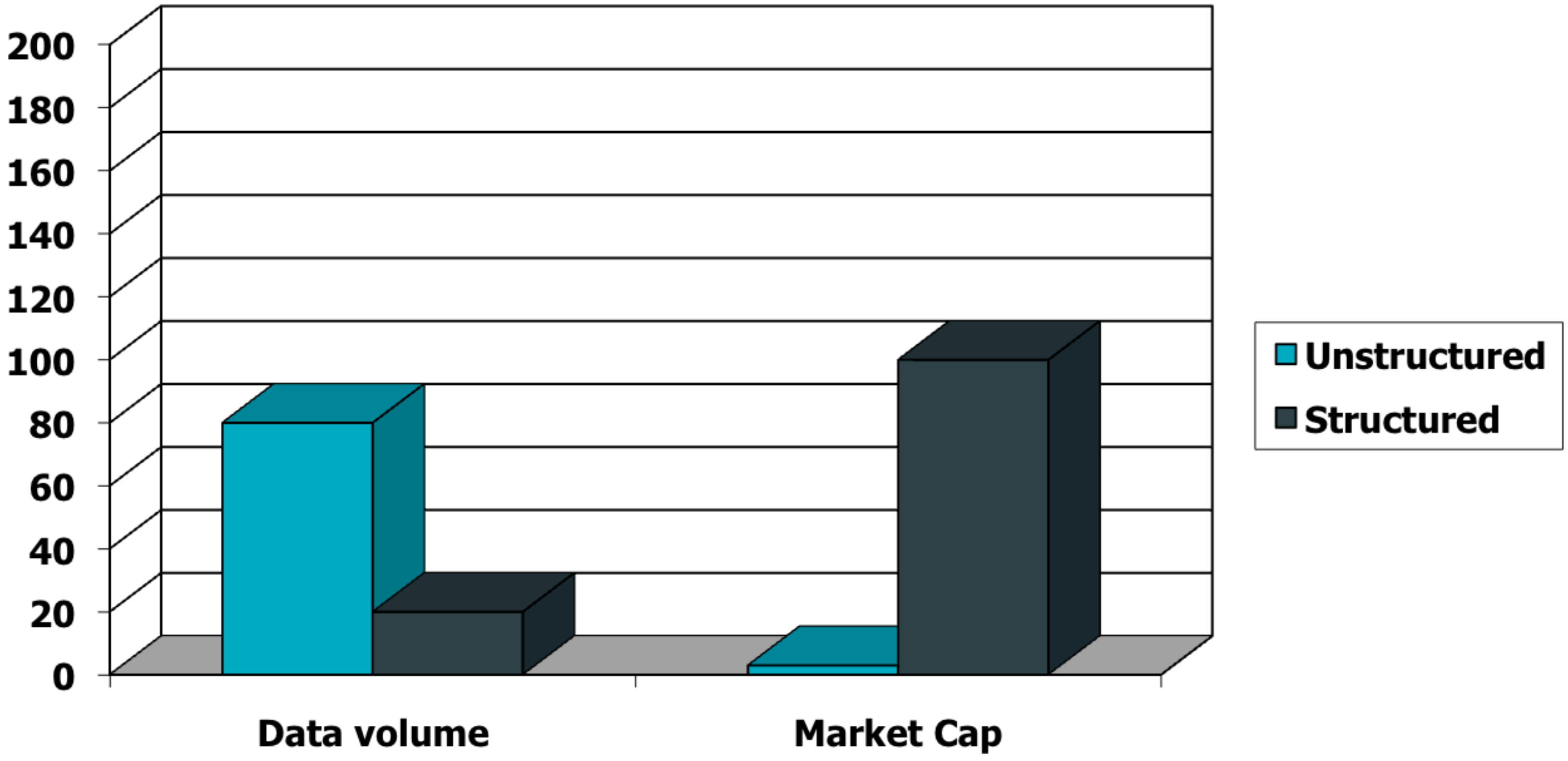
(Adapted from several IR courses: Croft's "Search Engines" book, Christopher Manning and P. Raghavan (at Stanford), and R Mooney at UT Austin)

Information Retrieval

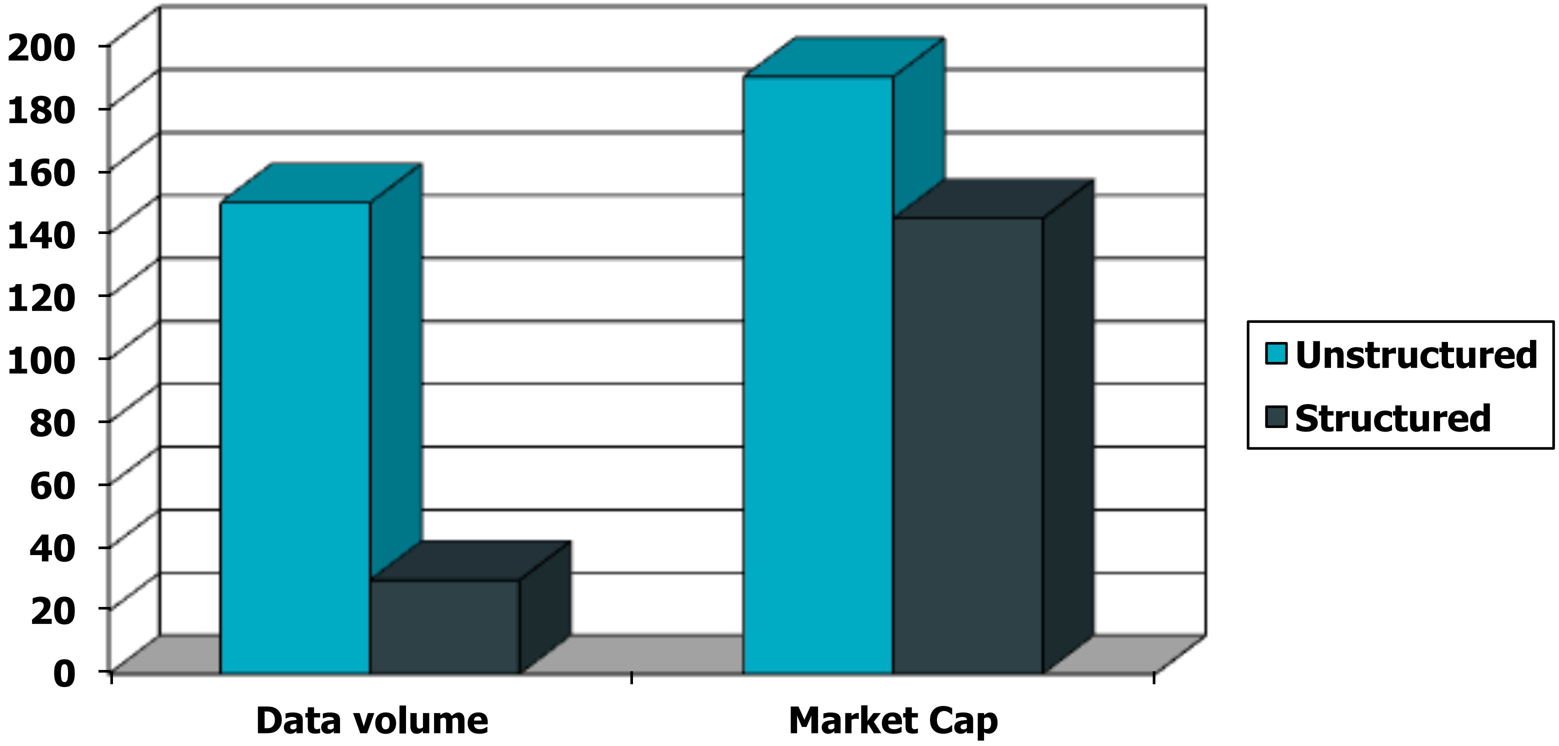
- *"Information Retrieval (IR) is finding material (usually documents) of an unstructured* nature *(usually text) that satisfies an information need from within large collections (usually stored on computers)."* (Manning, et al, 2008)

  - Various uses:
    - **web search**
    - E-mail search
    - Searching your laptop
    - Corporate knowledge bases
    - Legal information retrieval

Unstructured (text) vs. structured (database) data in the mid-nineties

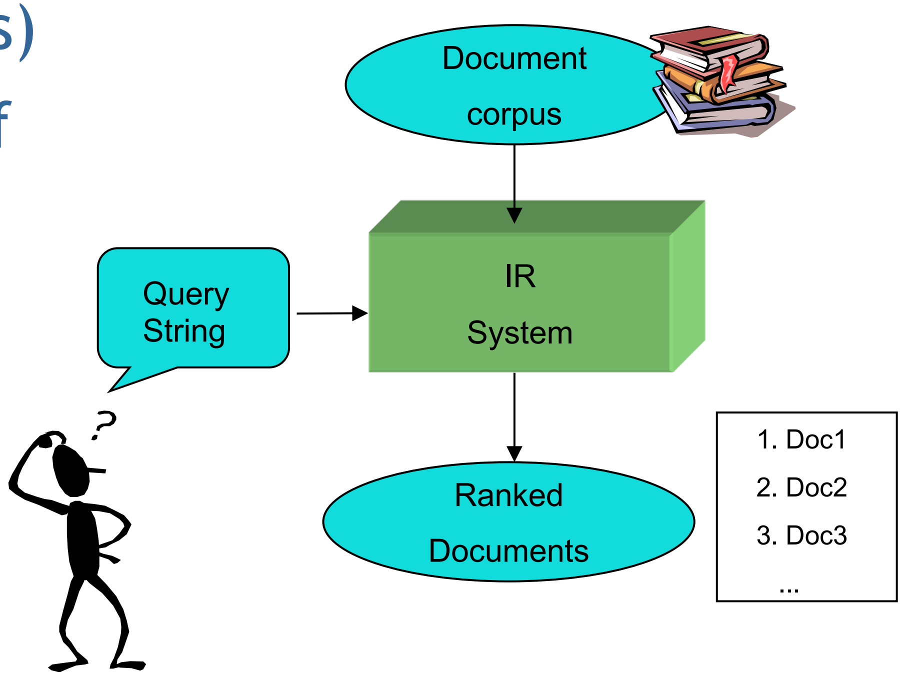


Unstructured (text) vs. structured (database) data today



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?

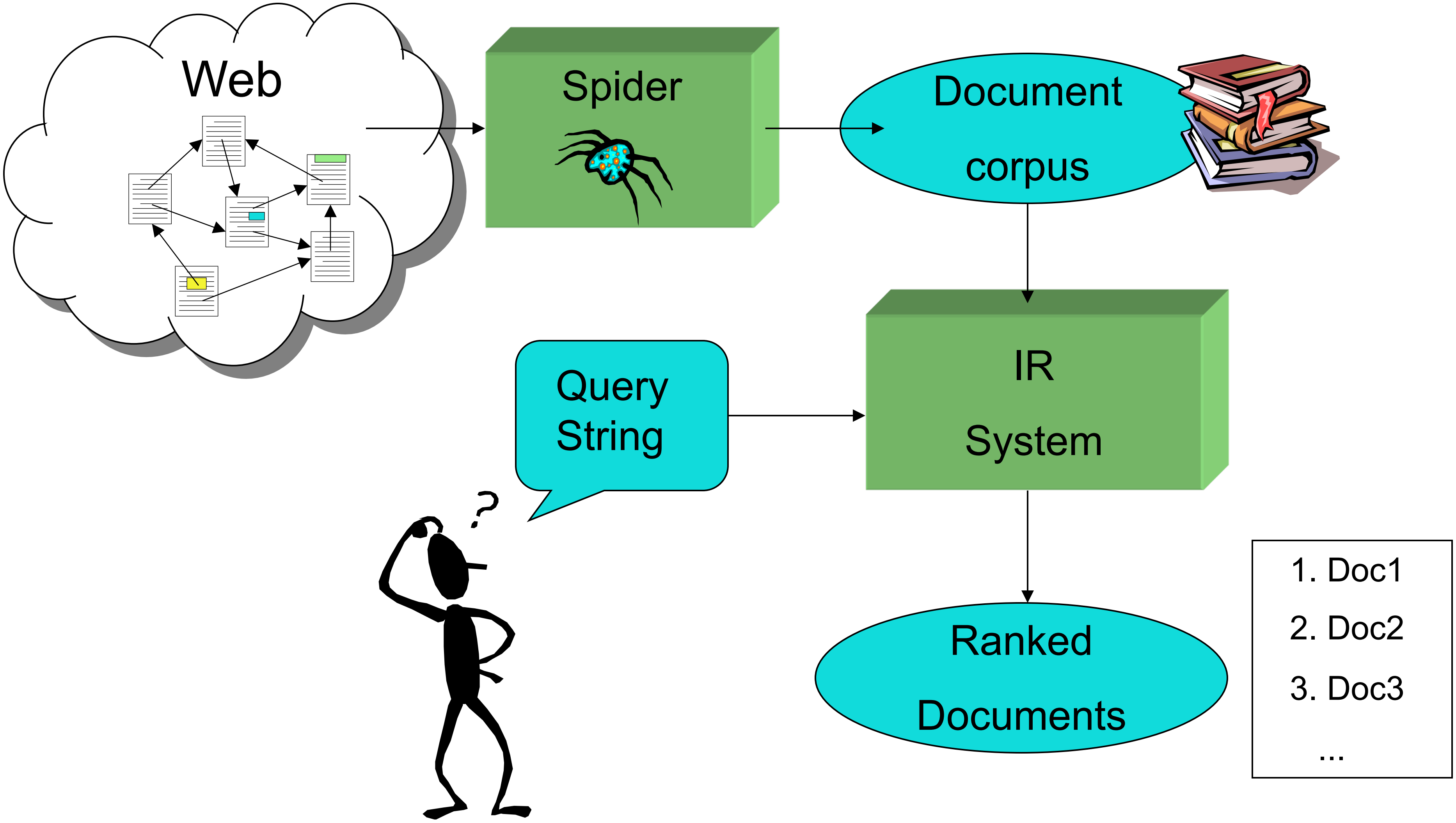
- Examples:
  - web pages, email, books, news stories, scholarly papers, text messages, Word™, Powerpoint™, PDF, forum postings, patents, IM sessions, etc.
- Common properties
  - Significant text content
  - Some structure (e.g., title, author, date for papers; subject, sender, destination for email)

Web Search

- Application of IR to HTML documents on the World Wide Web

- 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 System



Web

Spider

Document corpus

Query String

IR System

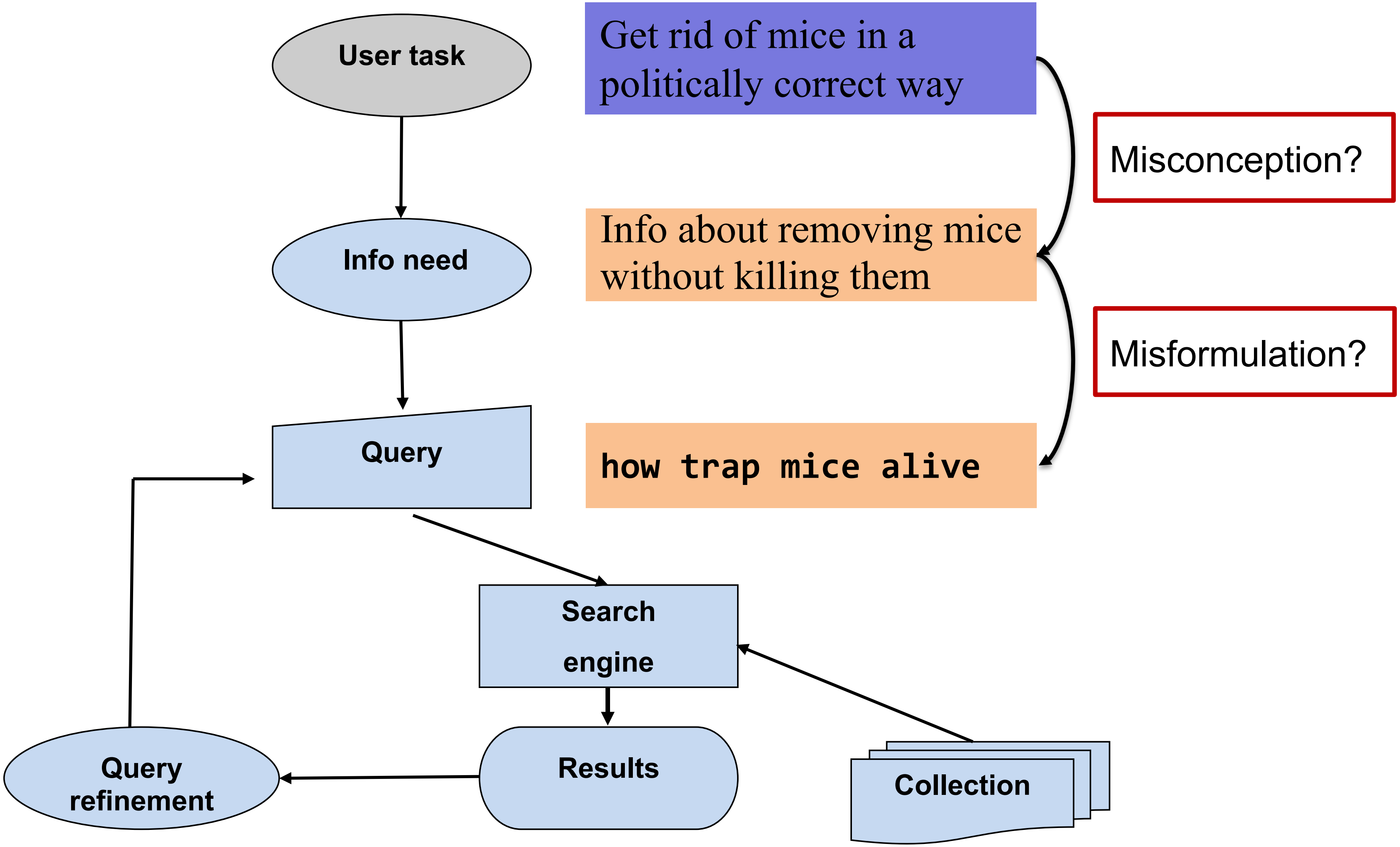
Ranked Documents

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	

The classic search model



User task

Info need

Query

Search engine

Results

Query refinement

Collection

Get rid of mice in a politically correct way

Misconception?

Info about removing mice without killing them

Misformulation?

`how trap mice alive`

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

- Relevance
- Evaluation
- Information needs

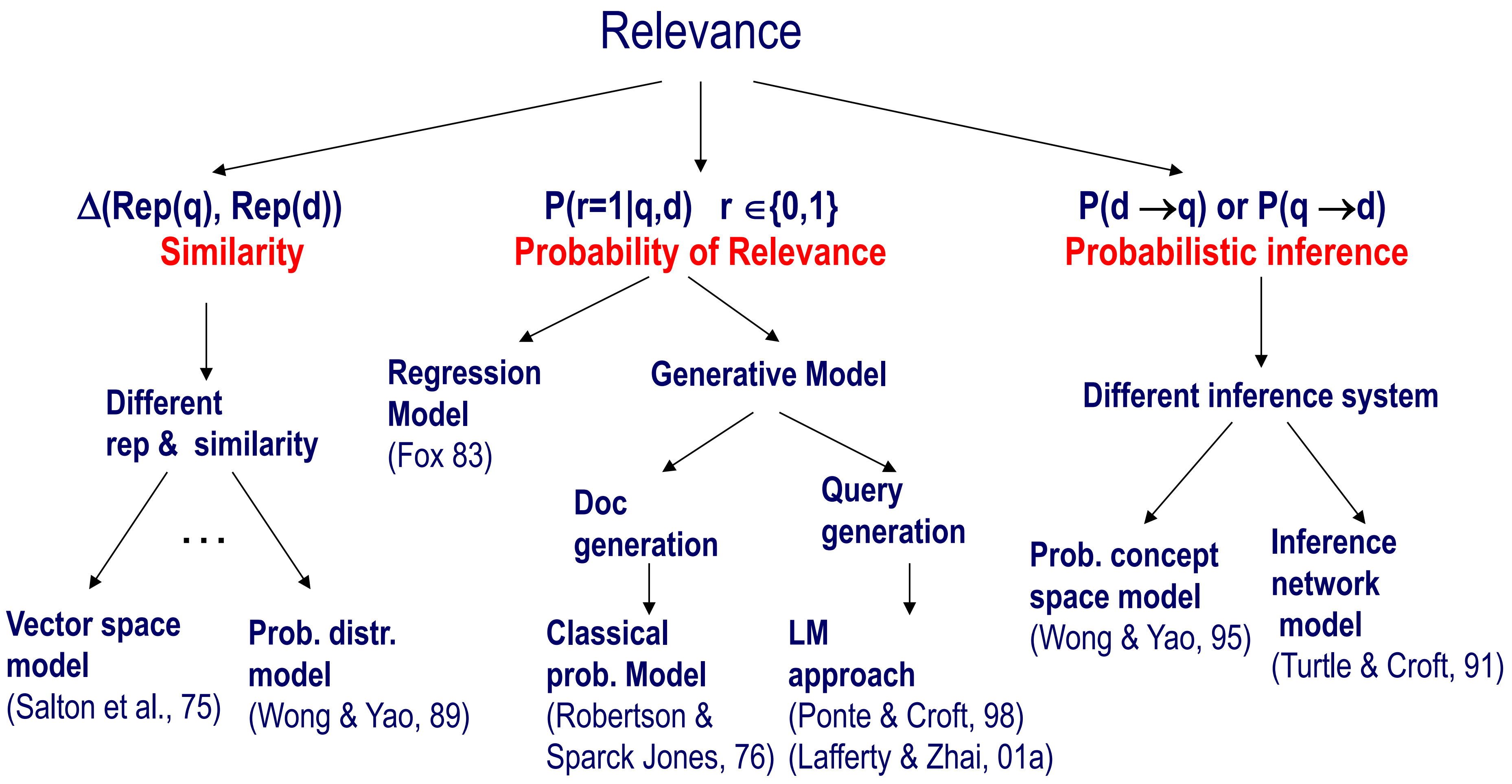
Big Issues in IR - Relevance

- 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



Relevance

$\Delta(Rep(q), Rep(d))$
Similarity

$P(r=1|q,d) \quad r \in \{0,1\}$
Probability of Relevance

$P(d \rightarrow q)$ or $P(q \rightarrow d)$
Probabilistic inference

Different
rep & similarity

...

Vector space
model
(Salton et al., 75)

Prob. distr.
model
(Wong & Yao, 89)

Regression
Model
(Fox 83)

Generative Model

Doc
generation

Classical
prob. Model
(Robertson &
Sparck Jones, 76)

Query
generation

LM
approach
(Ponte & Croft, 98)
(Lafferty & Zhai, 01a)

Different inference system

Prob. concept
space model
(Wong & Yao, 95)

Inference
network
model
(Turtle & Croft, 91)

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

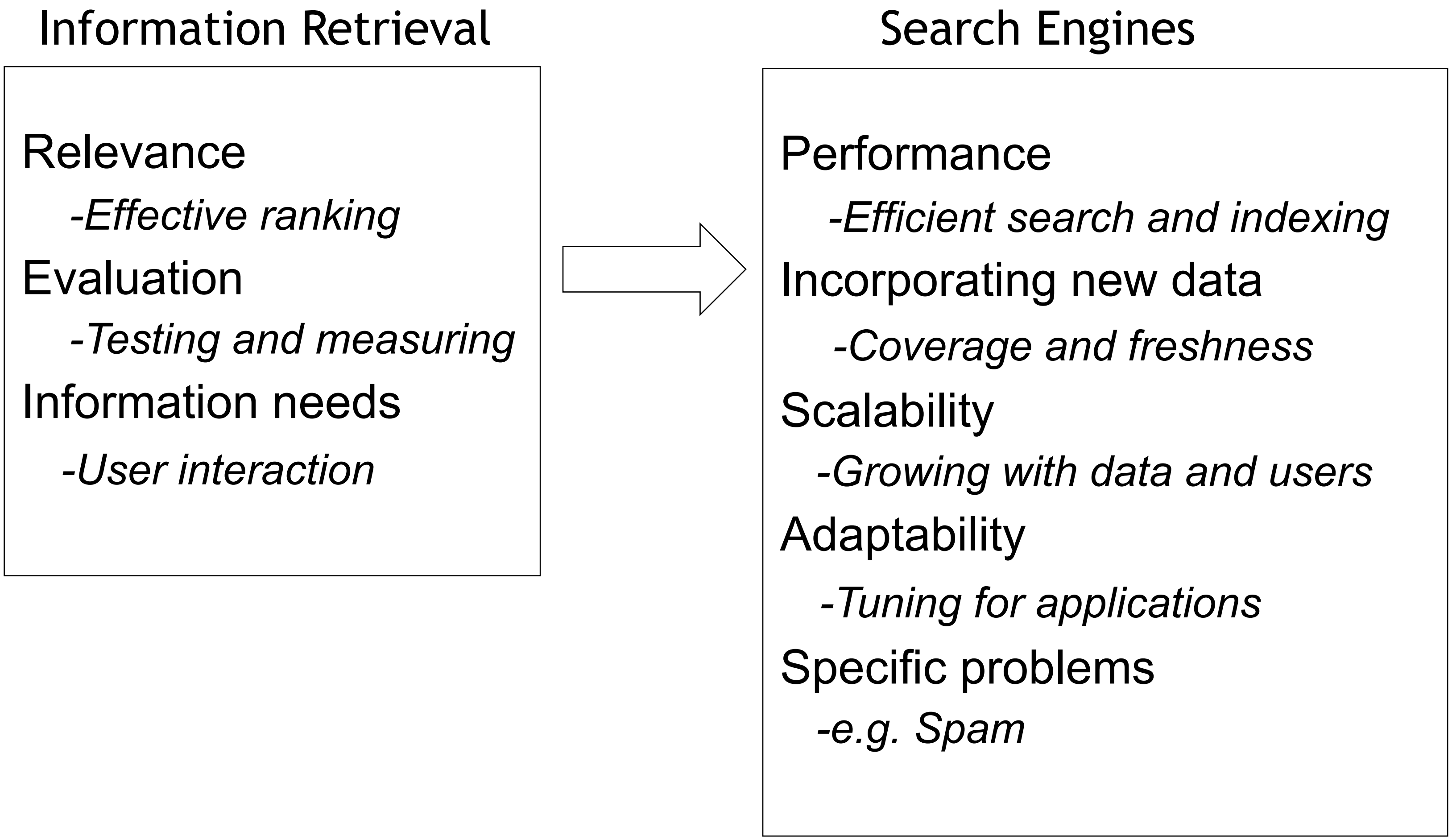
- Users and Information Needs
  - Search evaluation is user-centered
  - Keyword queries are often poor descriptions of actual information needs
  - Interaction and context are important for understanding user intent
  - Query refinement techniques such as *query expansion*, *query suggestion*, *relevance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

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, …)

IR and Search Engines

- Big issues include main IR issues but also some others



Information Retrieval

Relevance

-Effective ranking

Evaluation

-Testing and measuring

Information needs

-User interaction

Search Engines

Performance

-Efficient search and indexing

Incorporating new data

-Coverage and freshness

Scalability

-Growing with data and users

Adaptability

-Tuning for applications

Specific problems

-e.g. Spam

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**

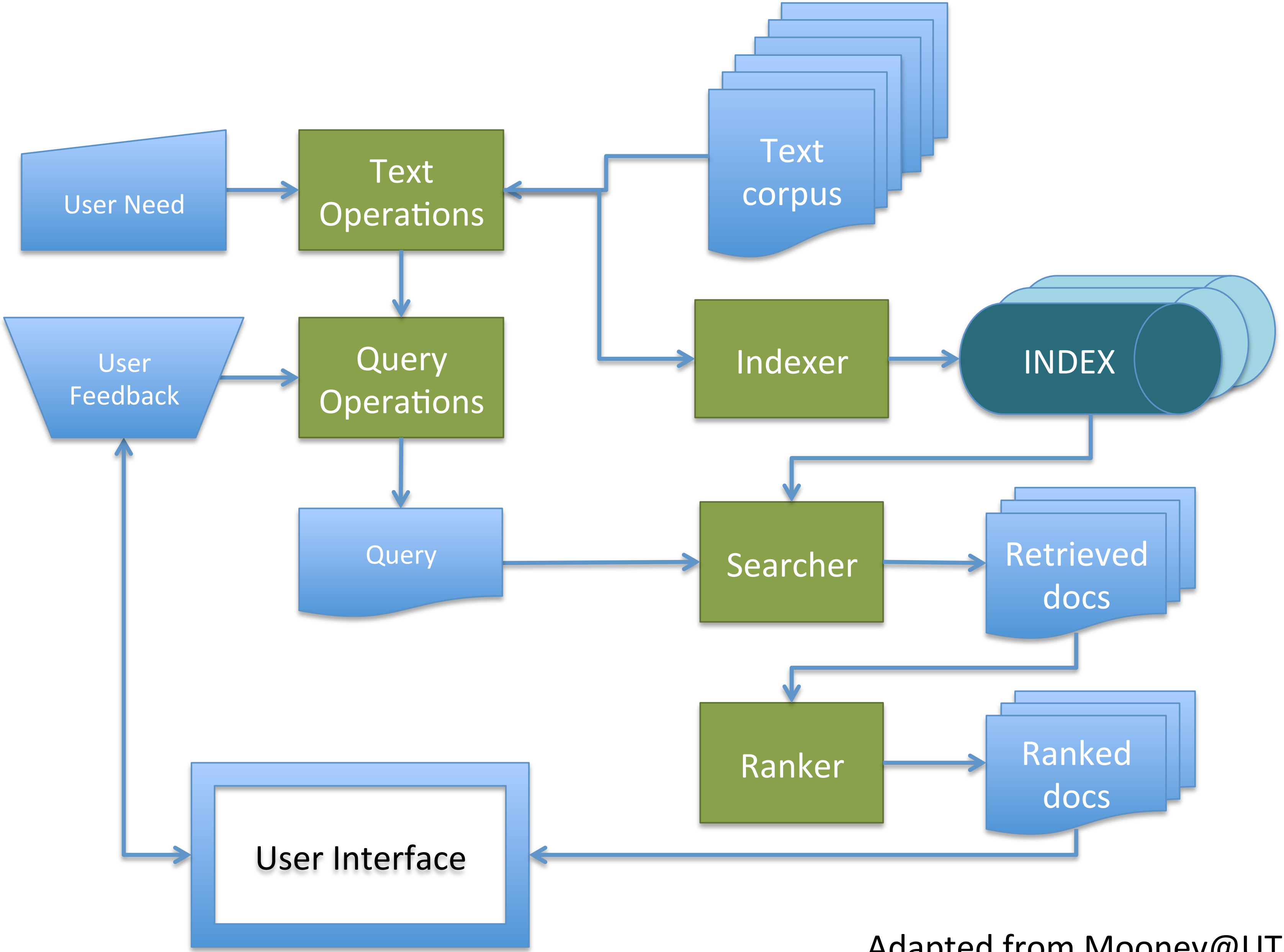
Search Engine Issues

- 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

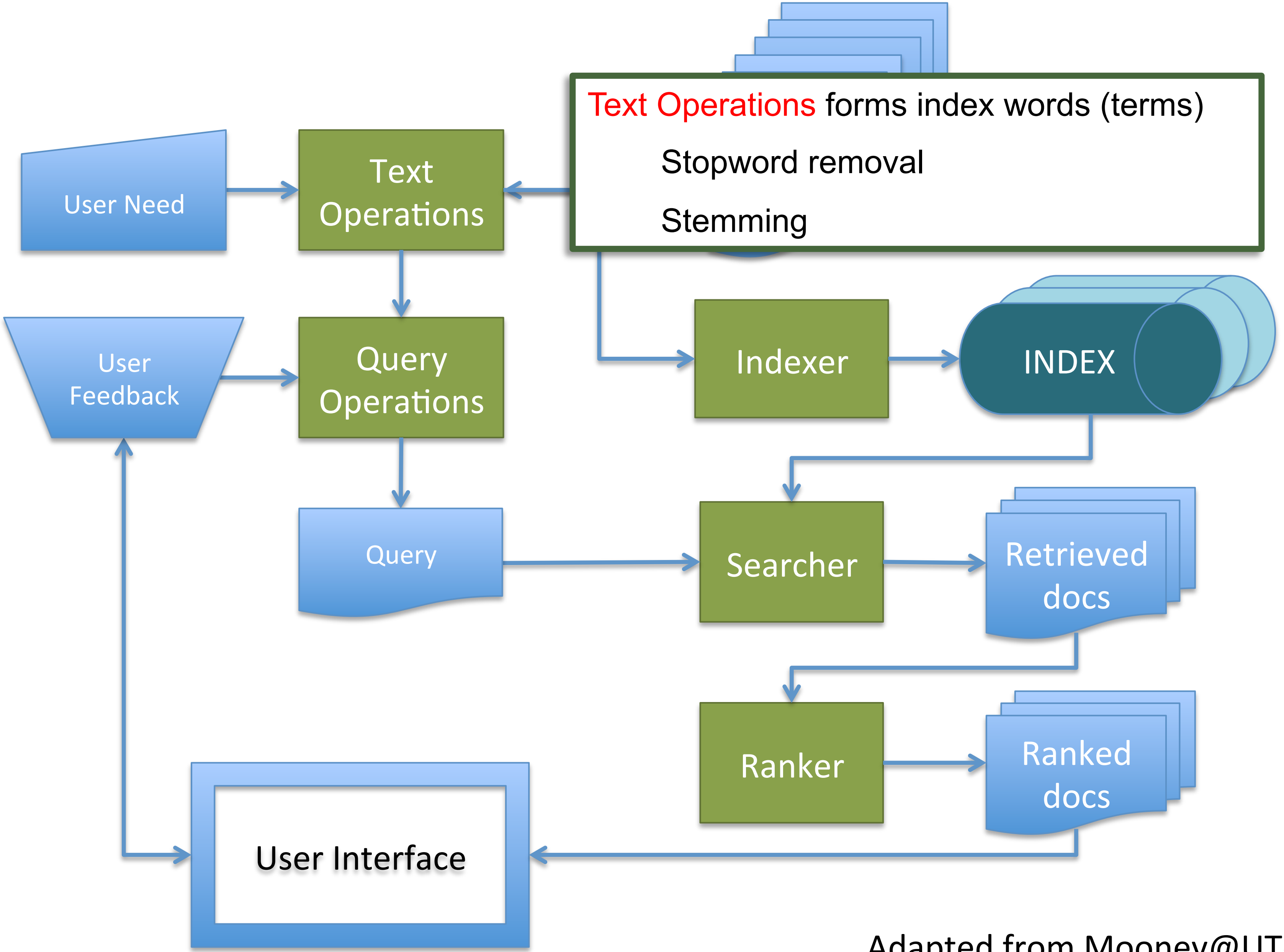
- Scalability
  - Making everything work with millions of users every day, and many terabytes of documents
  - Distributed processing is essential
- Adaptability
  - Changing and tuning search engine components such as ranking algorithm, indexing strategy, interface for different applications

IR System Architecture



Adapted from Mooney@UTexas

IR System Components



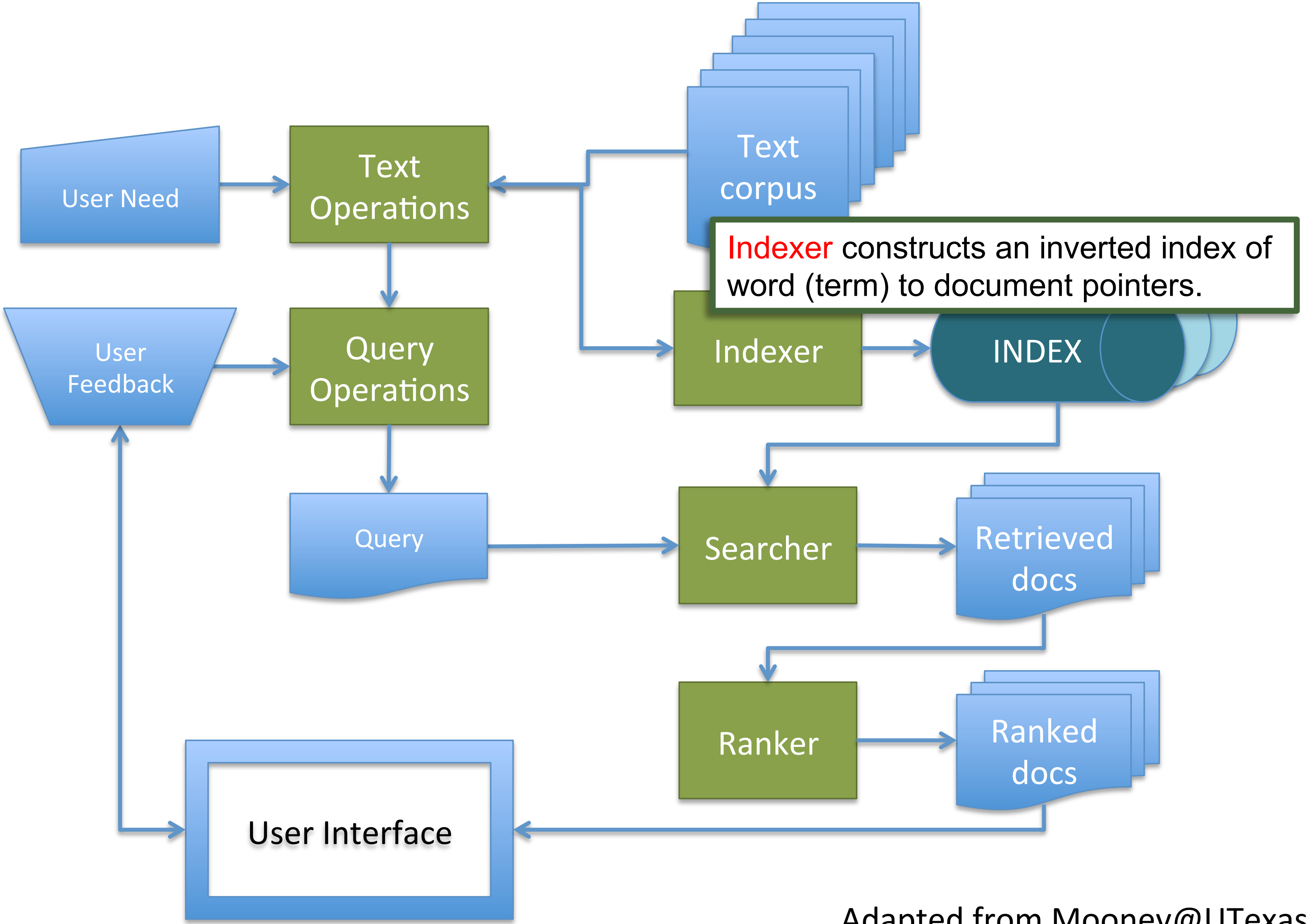
Text Operations forms index words (terms)

Stopword removal

Stemming

Adapted from Mooney@UTexas

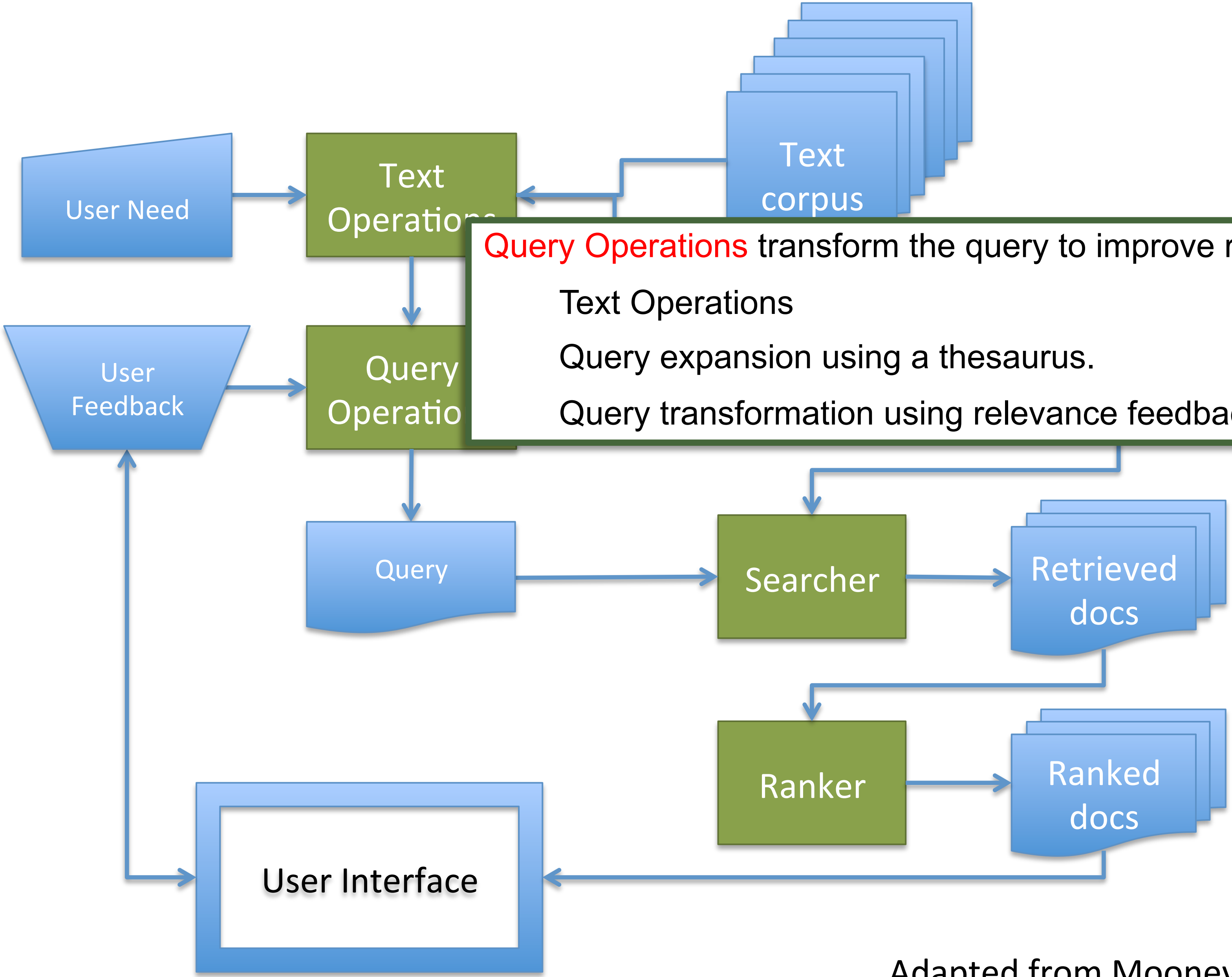
IR System Components



Indexer constructs an inverted index of word (term) to document pointers.

Adapted from Mooney@UTexas

IR System Components



Query Operations transform the query to improve retrieval:

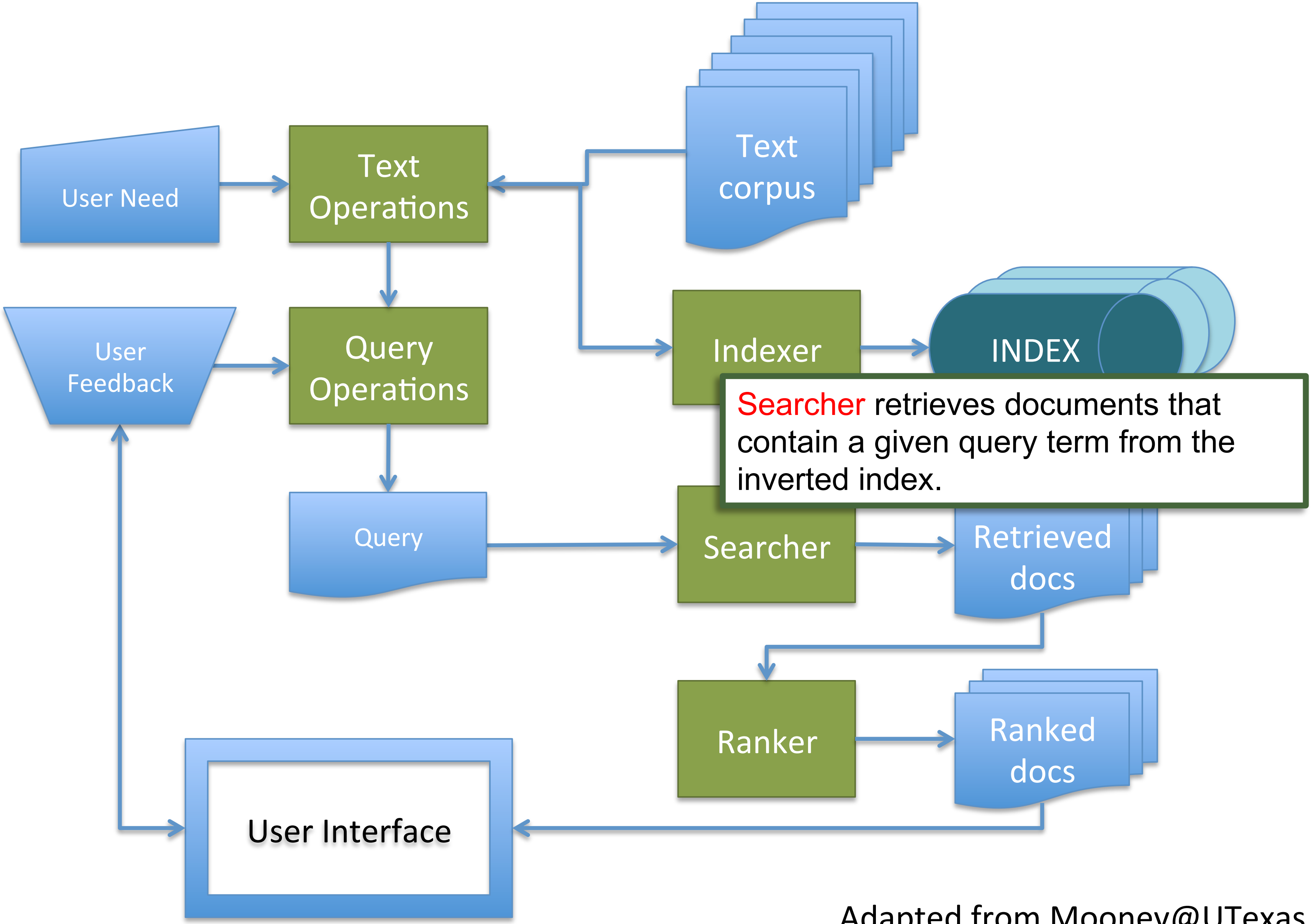
Text Operations

Query expansion using a thesaurus.

Query transformation using relevance feedback.

Adapted from Mooney@UTexas

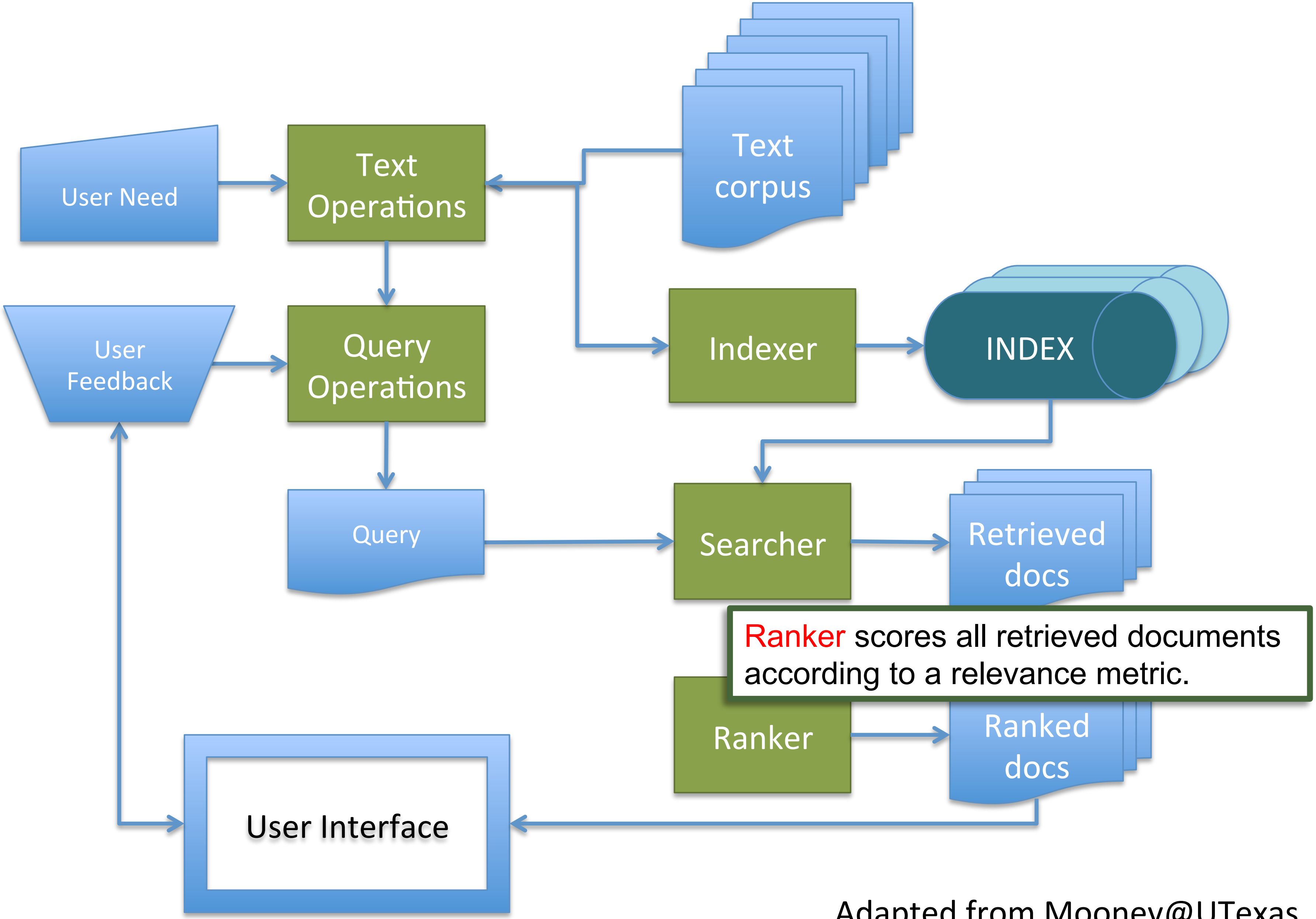
IR System Components



Searcher retrieves documents that contain a given query term from the inverted index.

Adapted from Mooney@UTexas

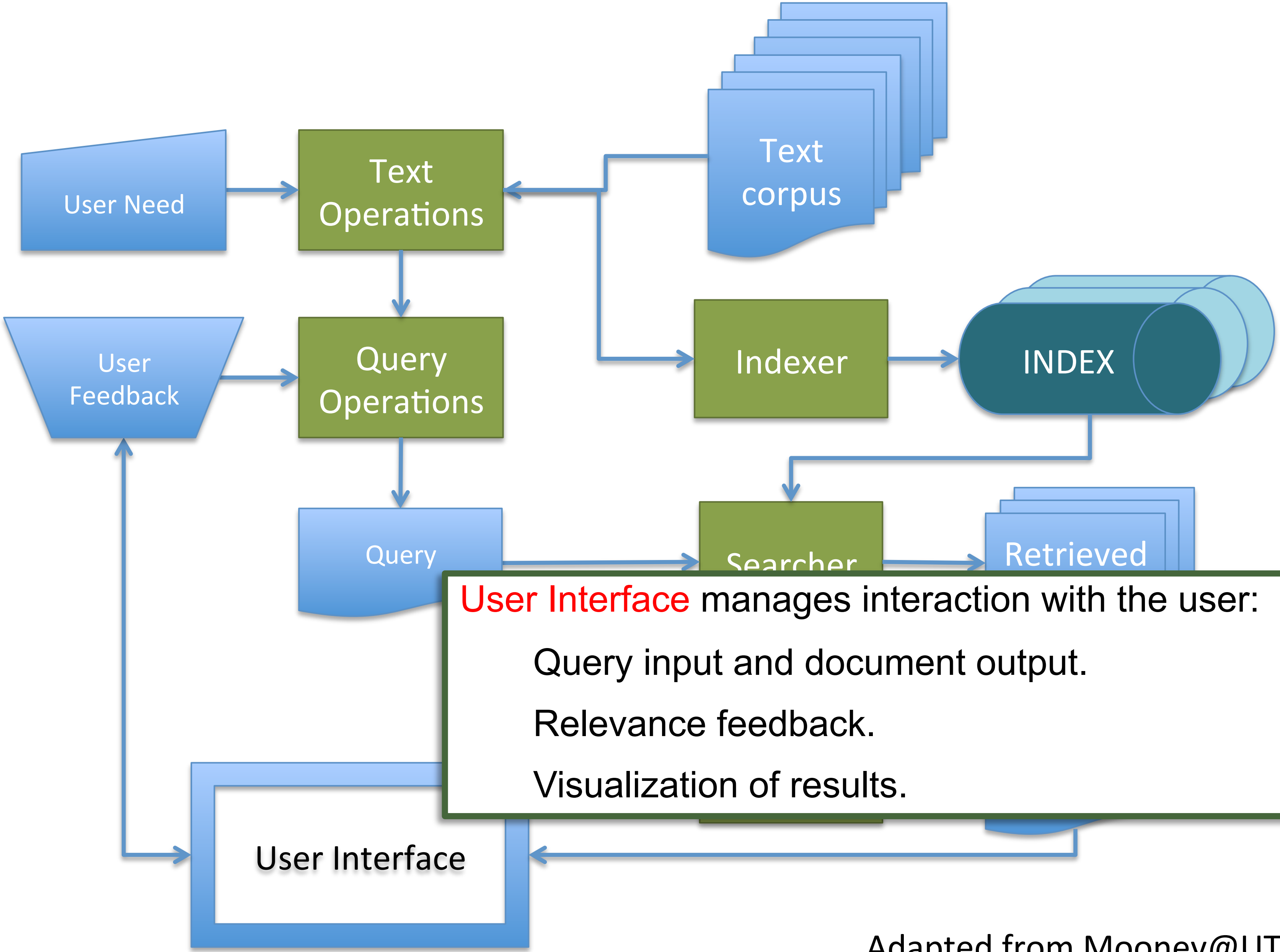
IR System Components



Ranker scores all retrieved documents according to a relevance metric.

Adapted from Mooney@UTexas

IR System Components



User Interface manages interaction with the user:

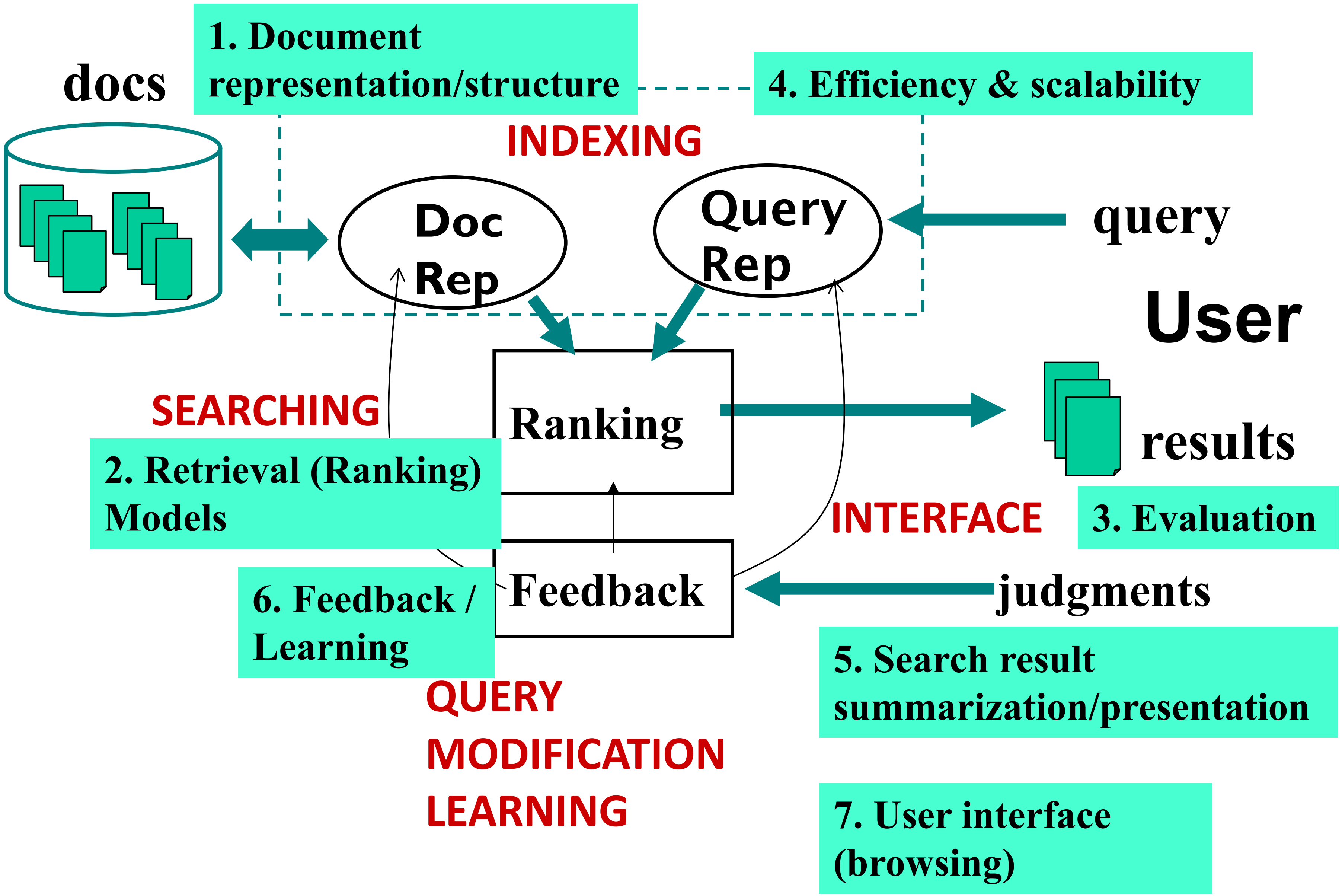
Query input and document output.

Relevance feedback.

Visualization of results.

Adapted from Mooney@UTexas

IR Topics (narrow view)



1. Document representation/structure

4. Efficiency & scalability

2. Retrieval (Ranking) Models

3. Evaluation

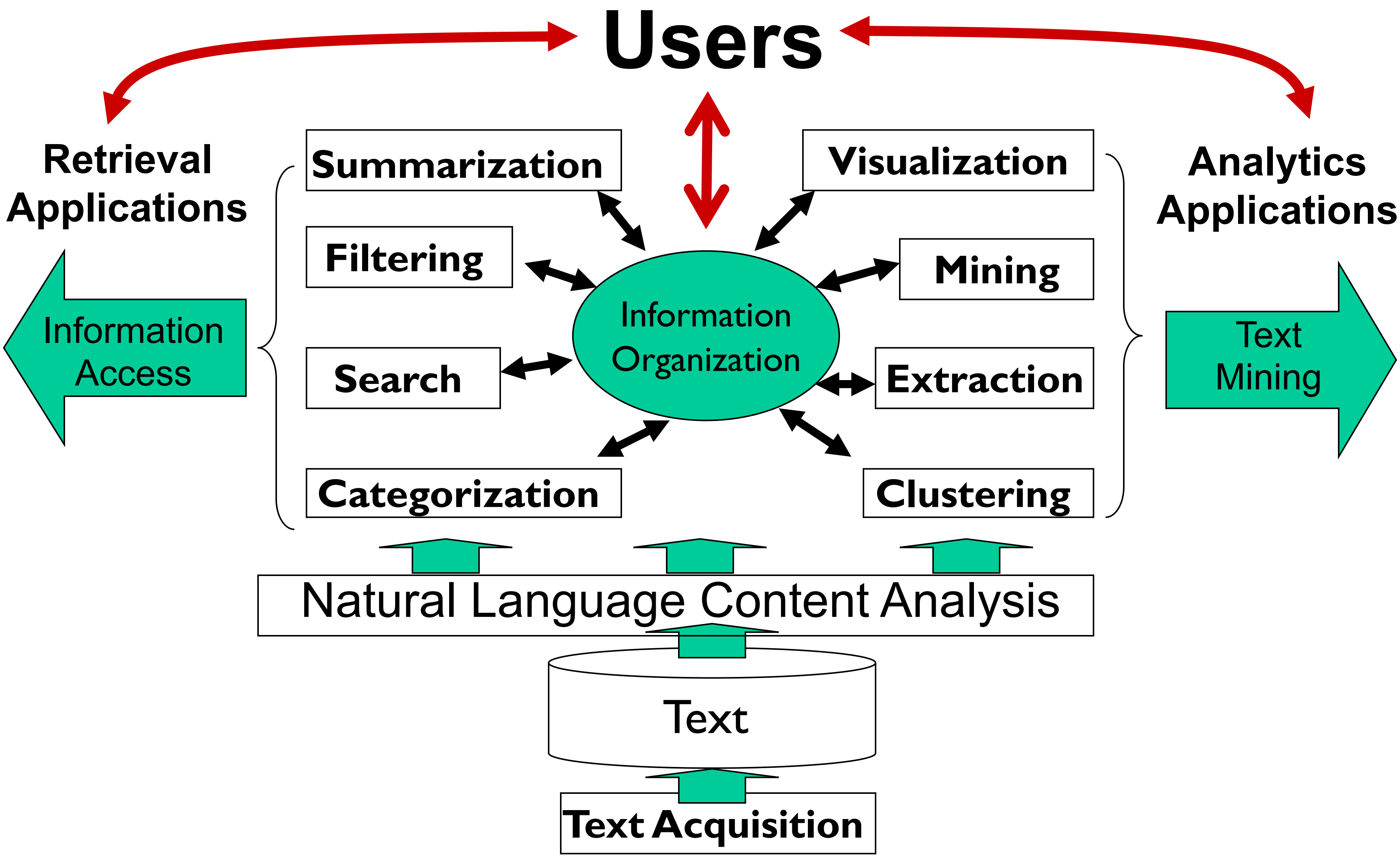
6. Feedback / Learning

5. Search result summarization/presentation

7. User interface (browsing)

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

IR Research Topics (Broad View)



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

IR Directions: NLP

- 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
  - Question answering

IR Directions: Machine Learning

- 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

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