ADVANCED TOPICS IN INFORMATION RETRIEVAL AND WEB SEARCH

Lecture 2:

Information Retrieval vs. Search Engines

S. M. Vahidipour

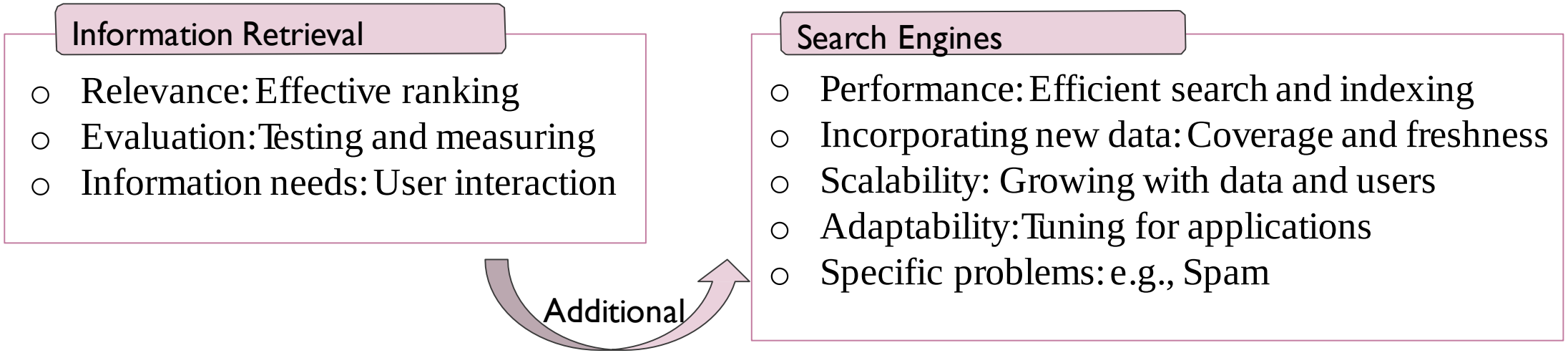
Vahidipour@kashanu.ac.ir

Outline

- **Main issues in IR and SE**
- Search engine architecture
  - Indexing
  - Querying

IR and Search Engines

- A search engine is the practical application of information retrieval techniques to large scale text collections
- Big issues include main IR issues but also some others…



Main Issues in IR

- Relevance
  - A relevant document contains the information a user was looking for when he/she submitted the query.
- Evaluation
  - How well does the ranking meet the expectation of the user.
- Users and information needs
  - Users of a search engine are the ultimate judges of quality.

Relevance

- 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
  - Task at hand, context, novelty, style, serendipity
- Topical relevance vs. user relevance
  - "Storm in Tehran last Sunday" is topically relevant to query "آب و هوا"…
  - … but might not be relevant to user because
    - Read it before
    - Is five years old
    - Is in a foreign language, etc.

Relevance

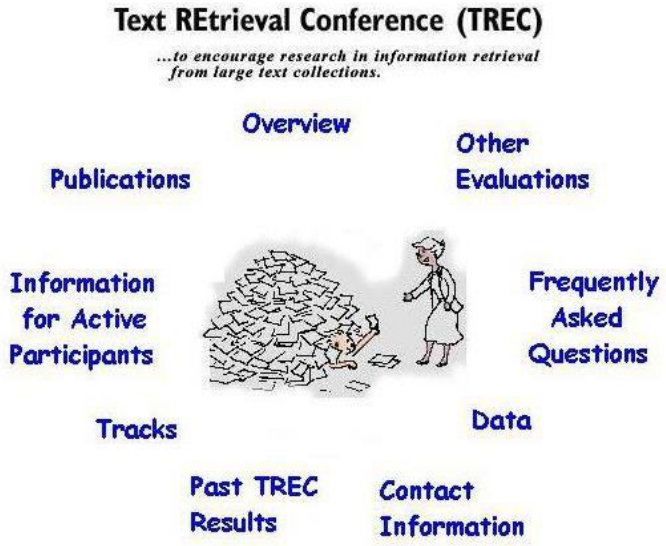
- Retrieval models define a view of relevance
  - Formal representation of the process of matching a query and a document
  - Simple text matching is not sufficient: Vocabulary mismatch problem (synonyms and homonyms)
- Ranking algorithms used in search engines are based on retrieval models
  - Real-world search engines consider topical and user 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
  - Linguistic features can be part of a statistical model

Evaluation

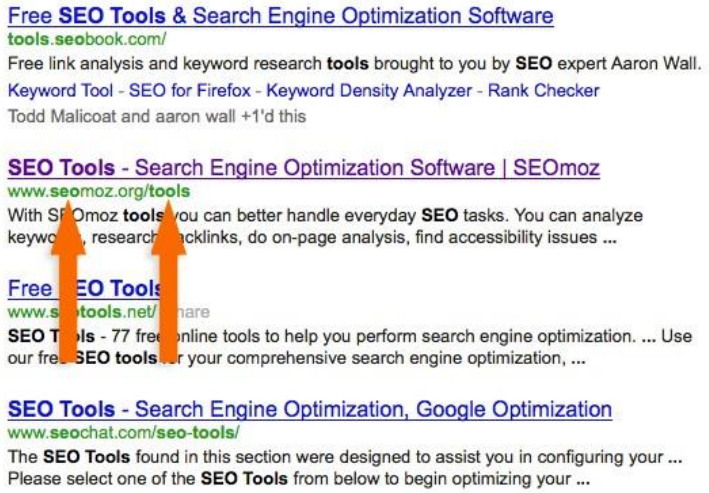
- Experimental procedures and measures for comparing system output with user expectations
  - Originated in Cranfield experiments in the 60s
    - First large scale “benchmark”
- IR evaluation methods now used in many fields
- Recall and precision are two examples of effectiveness measures

Evaluation

- Typically use test collection (corpus) of documents, queries, and relevance judgments
  - Most commonly used are TREC collections (Text REtrieval Conf.)



- Clickthrough data to evaluate retrieval models and search engines.

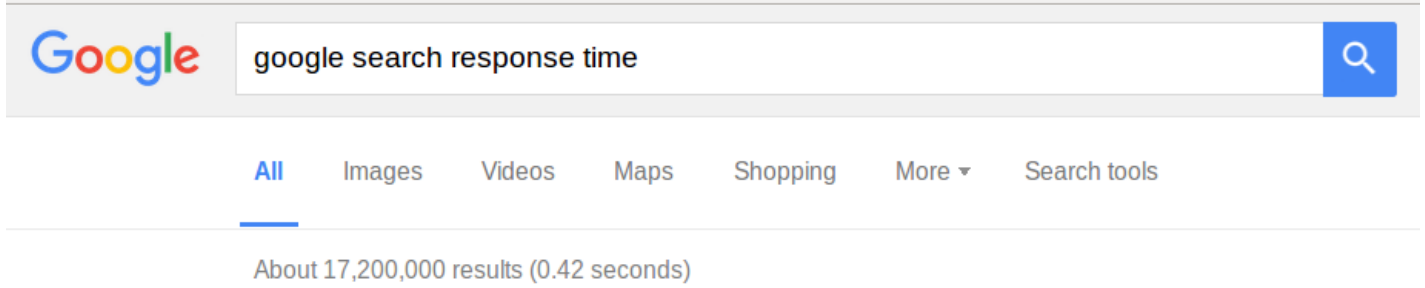


Users and Information Needs

- Search evaluation is user-centered
- Keyword queries are often poor descriptions of actual information needs
  - Query for “cats” could mean places to buy cats or the musical.
  - Search queries (in particular one-word queries) are under-specified.
- Interaction and context are important for understanding user intent
- Query refinement techniques such as
  - Query suggestion
  - Query expansion and relevance feedback
- Improve ranking

Performance

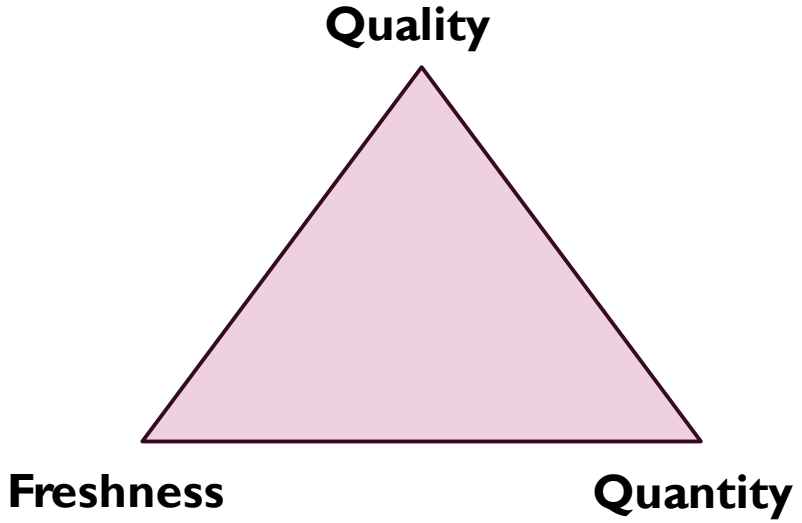
- Measuring and improving the efficiency of search
  - Reduce response time
  - Increase query throughput
  - Increase indexing speed
- Indexes are data structures designed to improve search efficiency.
  - Designing and implementing them are major issues for search engines.
- Same tradeoffs as in DBMS



Dynamic Data

- The "collection" for most real applications is constantly changing in terms of updates, additions, deletions

- Acquiring or "crawling" the documents is a major task
  - Typical measures are
    - Coverage (how much has been indexed)
    - Freshness (how recently was it indexed)



Scalability

- Making everything work with billions of users every day, and many terabytes of documents

- Distributed processing is essential
- But: Large ≠ scalable

Google in 2006	Google in 2008	Google in 2018
> 25 billion pages	• 1 trillion pages (1,000,000,000,000)	• 130 trillion pages

Adaptability

- Changing and tuning search engine components
  - ranking algorithm
  - indexing strategy
  - interface for different applications

- Adapt to different requirements for different applications / users
  - New APIs
  - New uses for search

Spam

- 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 spam
  - e.g., spamdexing or term spam, link spam, "optimization"
  - http://en.wikipedia.org/wiki/Spamdexing
- New subfield called adversarial IR, since spammers are "adversaries" with different goals

Outline

- Main issues in IR and SE
- **Search engine architecture**
  - Indexing
  - Querying

Search EngineArchitecture

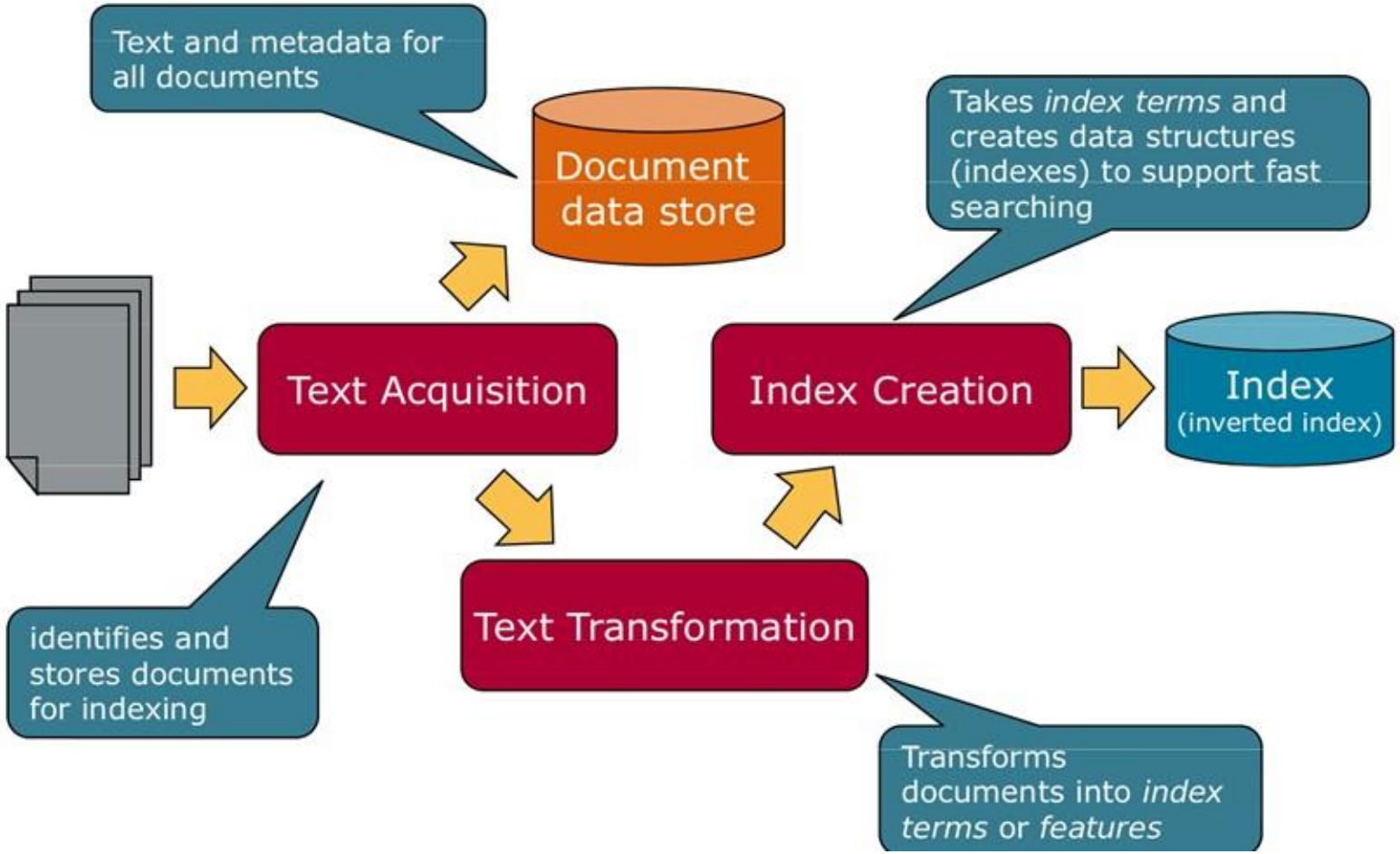
- Basic Building Blocks
- Indexing
  - Text Acquisition
  - Text Transformation
  - Index Creation
- Querying
  - User Interaction
  - Ranking
  - Evaluation

Search EngineArchitecture

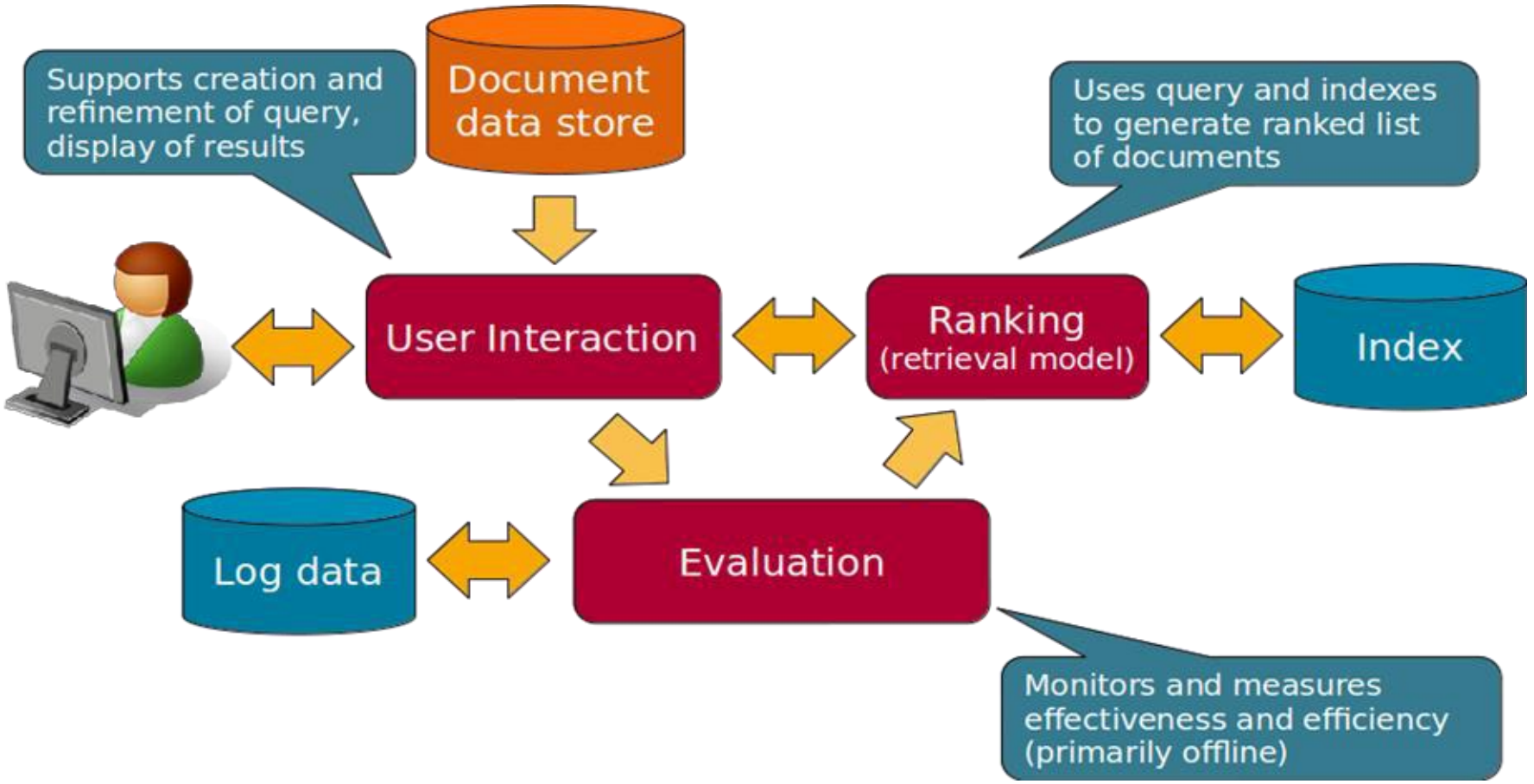
- Determined by two main requirements
  - Effectiveness (quality of results)
    - As good as possible
  - Efficiency (response time and throughput)
    - As quickly as possible

- Other requirements fall into these categories
  - Changing documents -> Effectiveness and efficiency
  - Personalization: Effectiveness
  - Spam: Effectiveness and efficiency
  - …

The Indexing Process



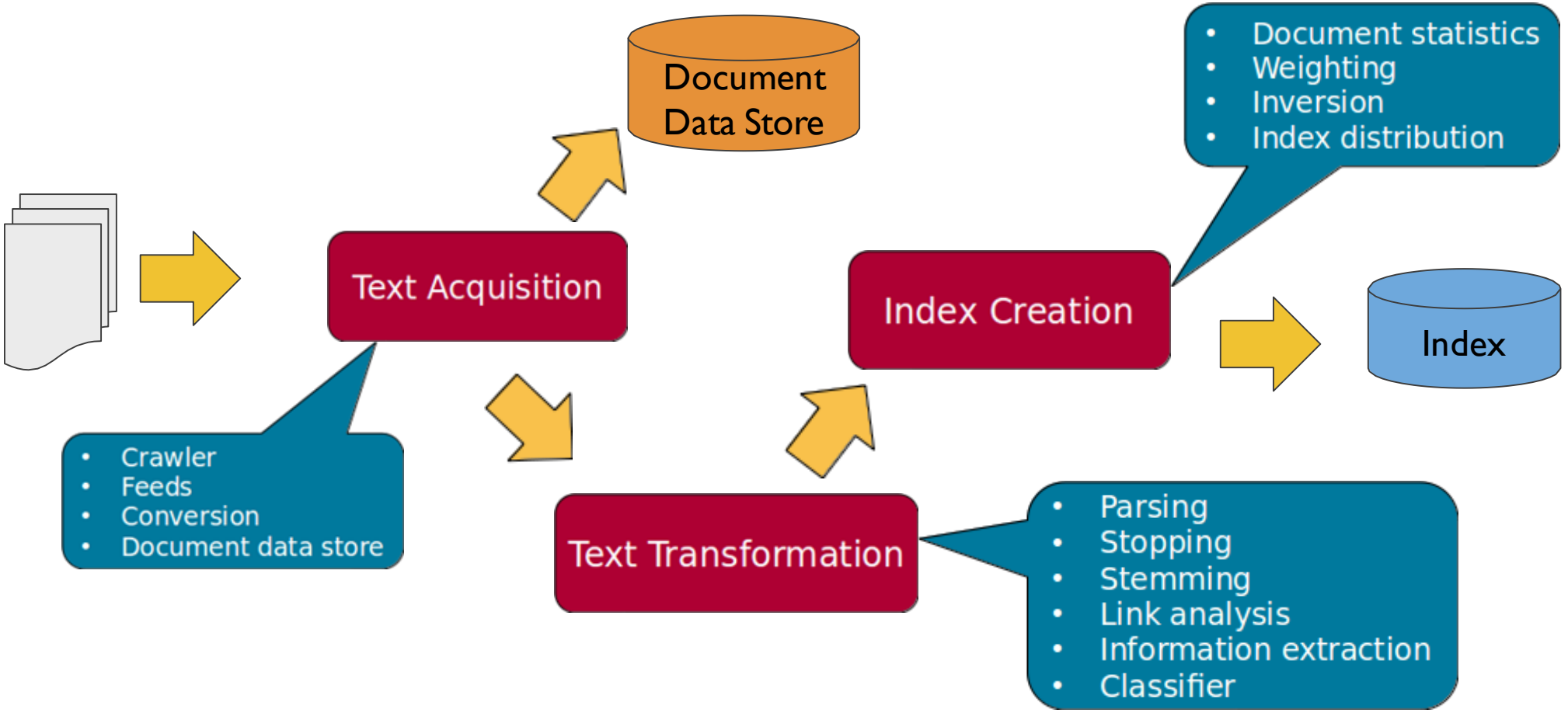
The Query Process



Outline

- Main issues in IR and SE
- **Search engine architecture**
  - **Indexing**
  - Querying

The Indexing Process



Text Acquisition – Crawler

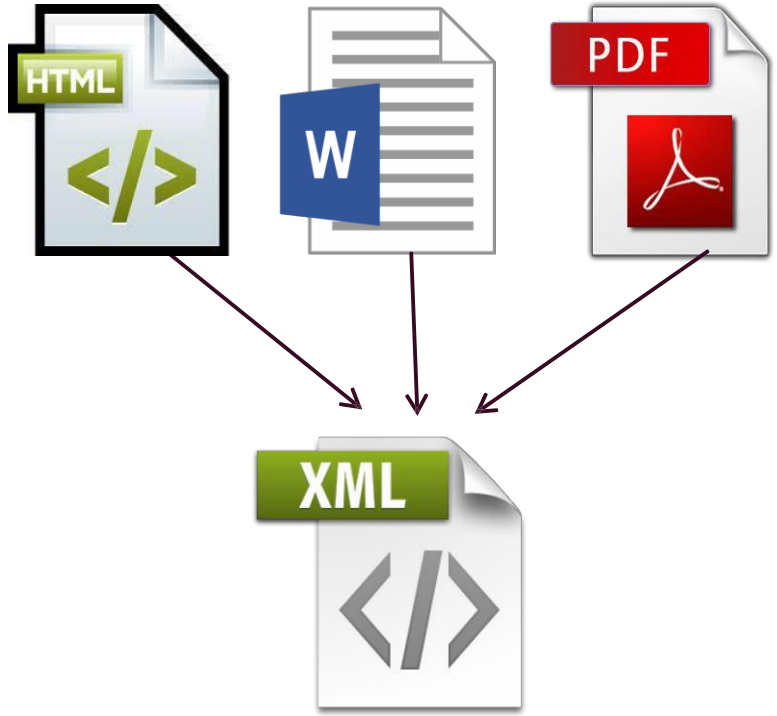
- Identifies and acquires documents for search engine
- Many types
  - Web, enterprise, desktop
- Web crawlers follow links to find documents
  - Must efficiently find huge numbers of web pages (coverage) and keep them up-to-date (freshness)
  - Single site crawlers for site search
  - Topical or focused crawlers for vertical search
- Document crawlers for enterprise and desktop search
  - Follow links and scan directories

Text Acquisition – Feeds

- Real-time streams of documents
  - Web feeds for news, blogs, video, radio, TV
- RSS is common standard
  - RSS "reader" can provide new XML documents to search engine

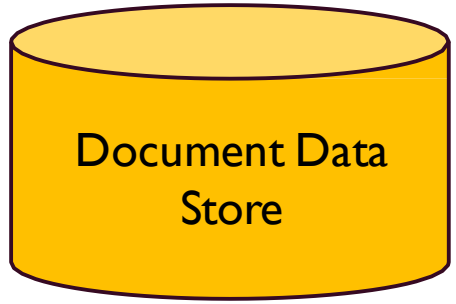
TextAcquisition – Conversion

- Convert variety of document formats into a consistent text-plus-metadata format
  - e.g., HTML, XML, Word, PDF, etc. → XML
- Convert text encoding for different languages
  - Using a standard like UTF-8
  - Be consistent throughout application
- Non-content data (tags, metadata) is either removed or stored as metadata.
- First step towards text transformation

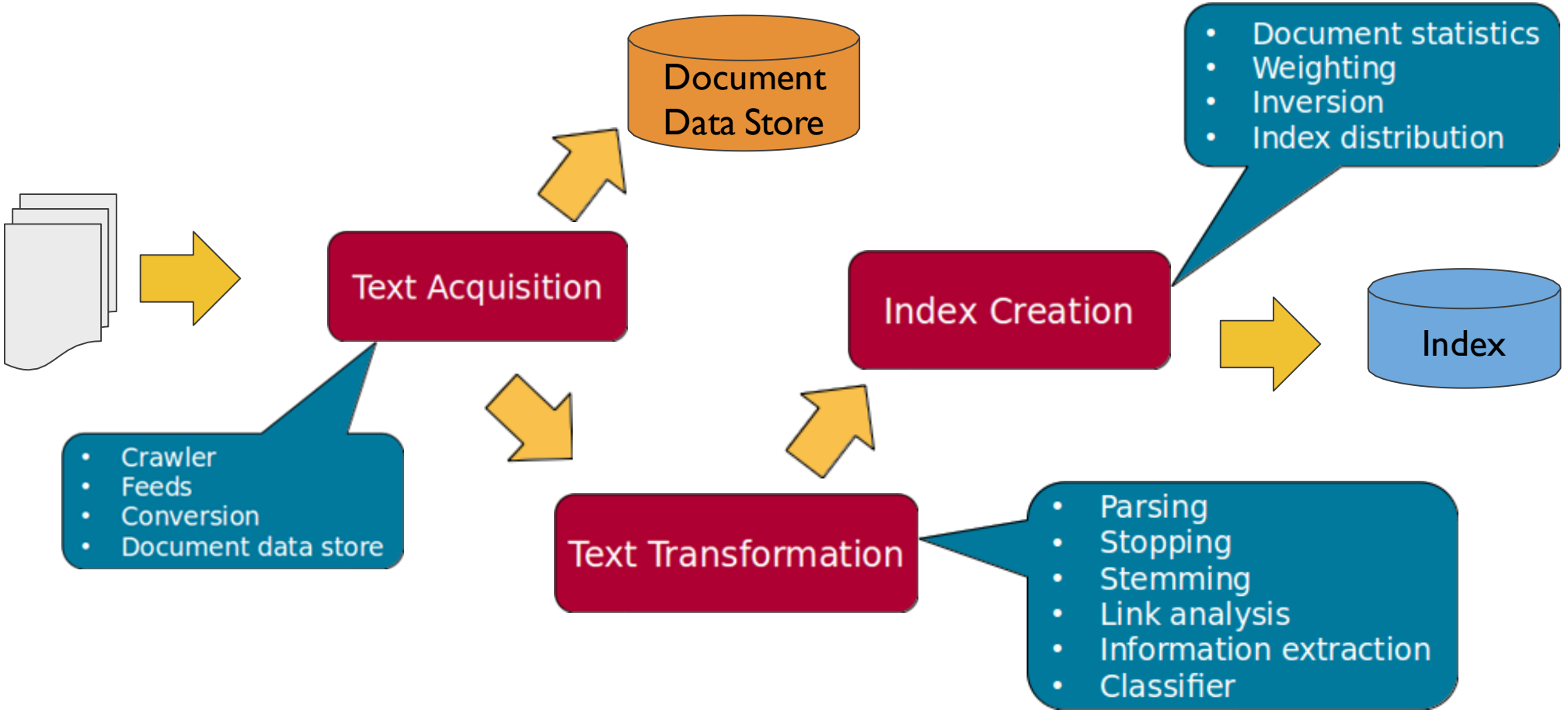


Text Acquisition – Document Data Store

- Two parts
  - Unstructured text
  - Structured metadata
- Stores text, metadata, and other related content for documents
  - Metadata is information about document
    - Type, creation date, …
  - Other content includes links, anchor text
- Why store documents? They are available on the Web anyway…
  - Provide fast access to document contents for search engine components
    - Result list generation, document summary, snippet



The Indexing Process



Text Transformation – Parsing

- Processing the sequence of text tokens in the document to recognize structural elements
  - Titles, links, headings, etc.

- Markup languages such as HTML, XML often used to specify structure
  - Tags used to specify document elements
    - E.g., <h2> Overview </h2>
  - Document parser uses syntax of markup language (or other formatting) to identify structure
    - E.g. email format, MS Word metadata etc.

Text Transformation – Parsing

Input: Friends, Romans, Countrymen, lend me your ears;

Output: | Friends | Romans | Countrymen | lend | me | your | ears |

- Tokenizer recognizes "words" in the text.
  - Must consider issues like capitalization, hyphens, apostrophes, non-alpha characters, separators

neill
oneill
o'neill
o'	neill
o	neill	?

aren't
arent
are	n't
aren	t	?

朝鲜外务省发言人11月1日在平壤宣布，朝鲜将重返六方会谈，但前提条件是朝鲜与美国在 六方会谈框架内讨论解除美国对朝鲜[illegible]钱问题。

针对朝鲜方面[illegible]对表示欢迎。

Where are the words?

美联社11月1日报道说："长期以来一直拒绝与平壤进行直接对话的美国总统布什认为， 各方达成一致、同意恢复六方会谈应归功于中国的斡旋。

Text Transformation – Stopping

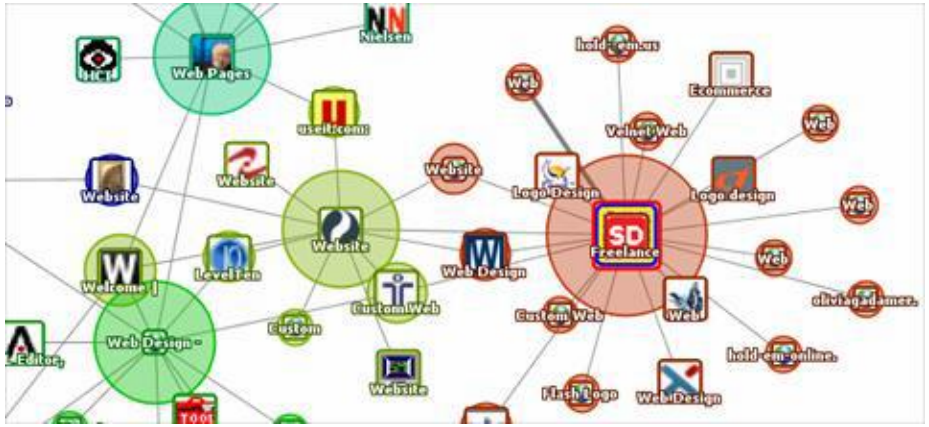
- Remove common words
  - “and”, “or”, “the”, “in”, …
- Some impact on efficiency and effectiveness
- Can be a problem for some queries
  - “To be or not to be”

Text Transformation – Stemming

- Group words derived from a common stem
  - “computer”, “computers”, “computing”, “compute”
  - “fish”, “fishing”, “fisherman”
- Usually effective, but not for all queries
- Benefits vary for different languages
  - Arabic: Very complicated morphology
  - Chinese: Few word variations anyway

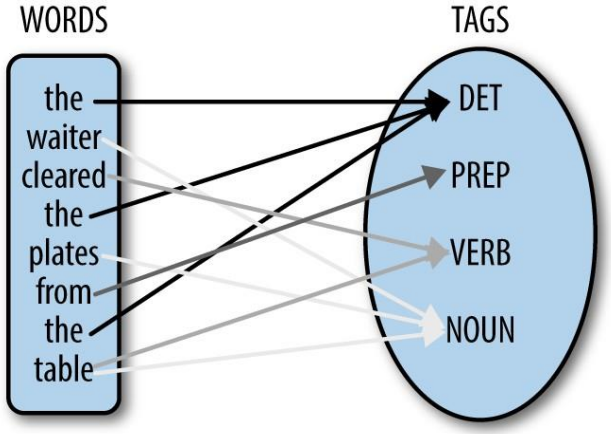
Text Transformation – Link Analysis

- Makes use of links and anchor text in web pages.
  - Stored and indexed separately
  - University of Kashan
- Link analysis identifies popularity and community information
  - e.g., PageRank
- Anchor text can significantly enhance the representation of pages pointed to by links
- Significant impact on web search
  - Less importance in other applications



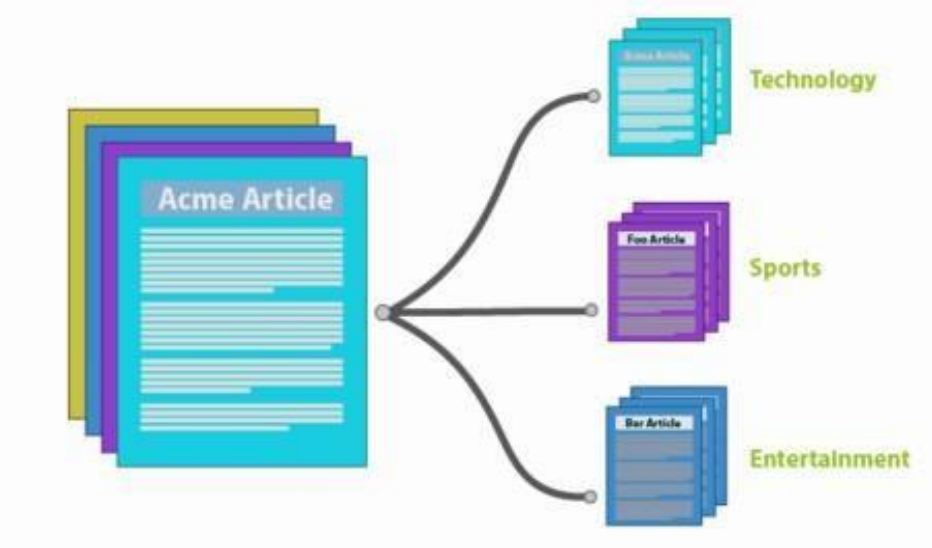
Text Transformation – Information Extraction

- Identify classes of index terms that are important for some applications
- Simple: Bold-face, heading, title
- Part of speech tagging
- Named entity recognizers (NER) identify classes such as
  - People
  - Locations
  - Companies
  - Dates
  - etc.

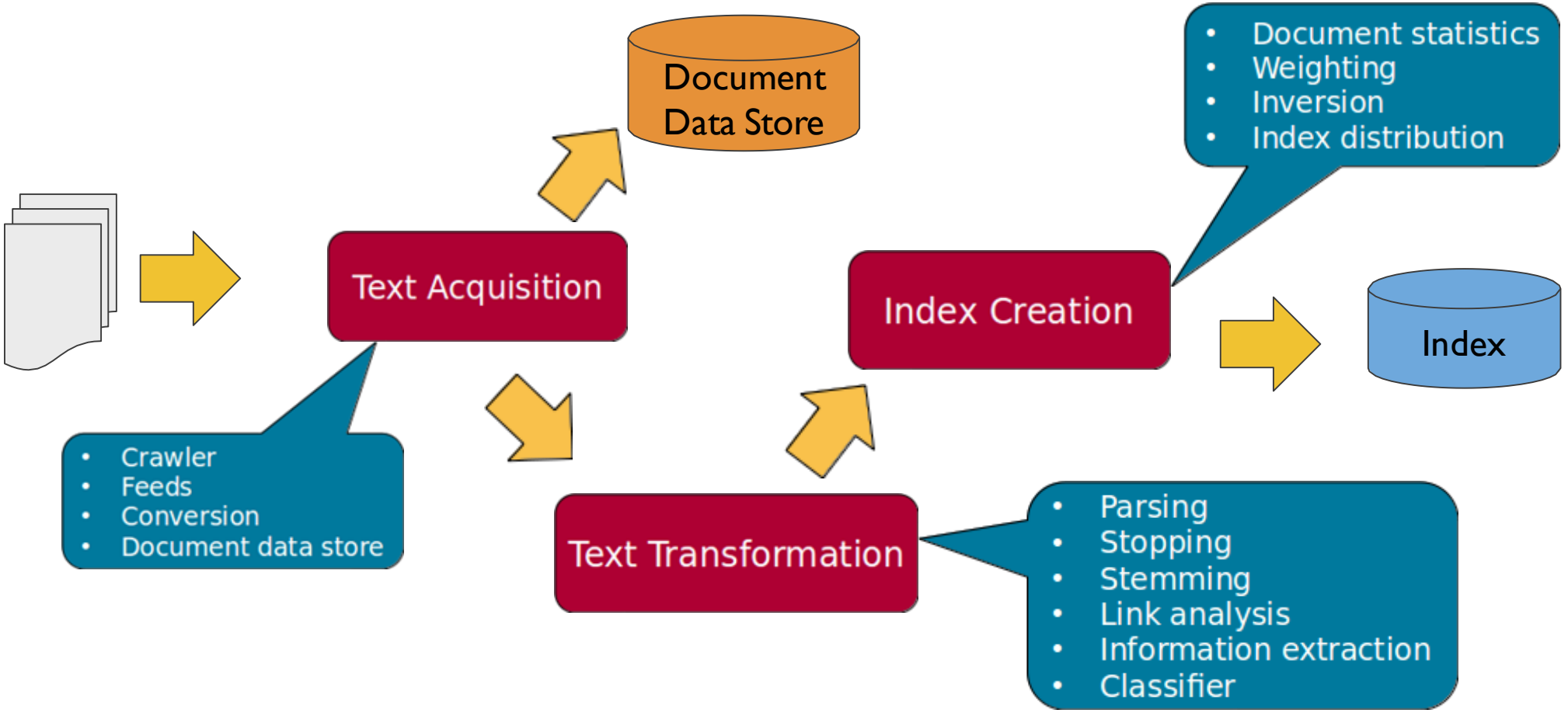


Text Transformation – Classifier

- Identifies class-related metadata for documents
  - i.e., assigns labels to documents
  - e.g., topics, reading levels, sentiment, genre
  - Spam!
  - Advertisements in documents
- Use depends on application



The Indexing Process



Index Creation - Document Statistics

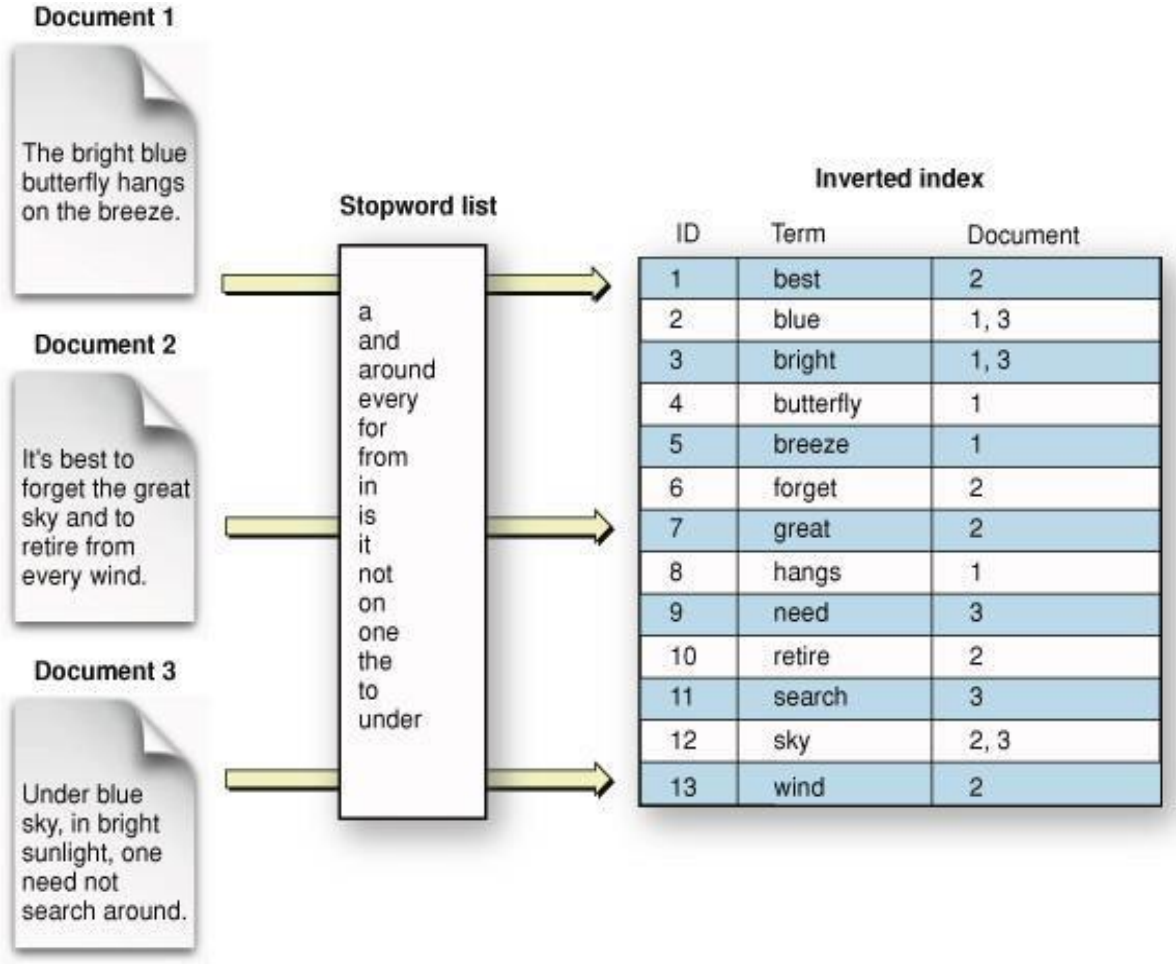
- Statistical information about words, features and documents
- Gathers counts and positions of words and other features
  - Within a document
  - Across groups of documents
  - Across all documents
- Used in ranking algorithm

Index Creation –Weighting

- Computes weights for index terms
  - Relative importance of words in documents
  - Used in ranking algorithm
- Global weight
  - Query-dependent weight
- TF.IDF weight
  - Combination of term frequency in document
  - and inverse document frequency in the collection

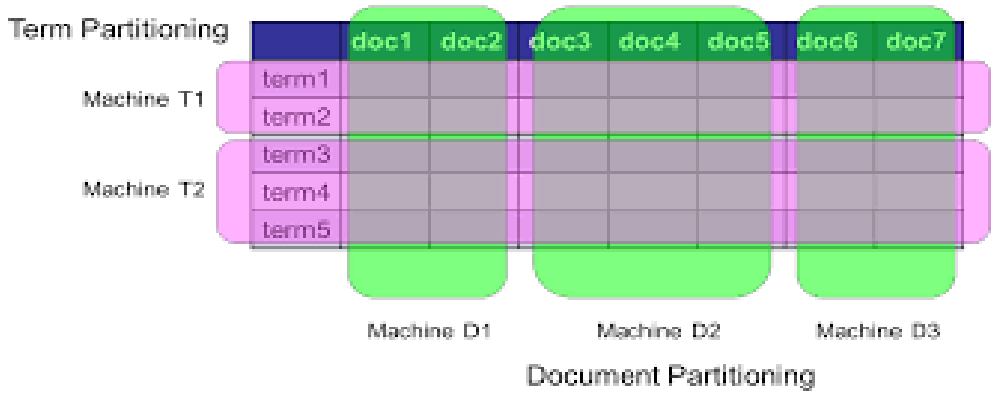
Index Creation – Inversion

- Core of indexing process
- Converts document-term information to term-document for indexing
  - Difficult for very large numbers of documents
  - Classical Map/Reduce use case
- Format of inverted file is designed for fast query processing
  - Must also handle updates
  - Compression used for efficiency



Index Creation – Index Distribution

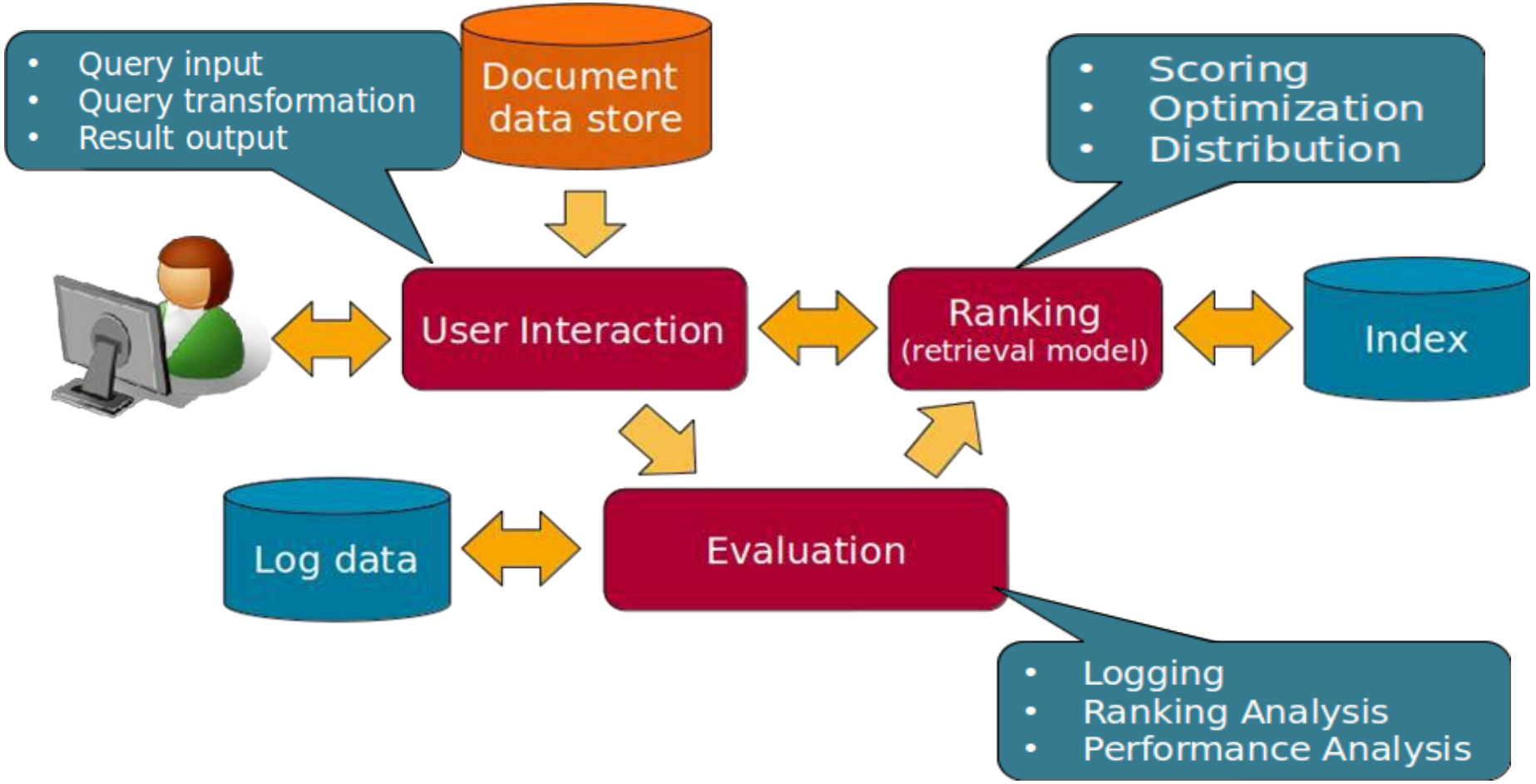
- Distribute indexes
  - across multiple computers
  - and/or multiple sites
- Essential for fast query processing with large numbers of documents
- Many variations
  - Document distribution: Distribute index for subsets of documents
  - Term distribution: Distribute index for subset of terms
  - Replication



Outline

- Main issues in IR and SE
- **Search engine architecture**
  - Indexing
  - **Querying**

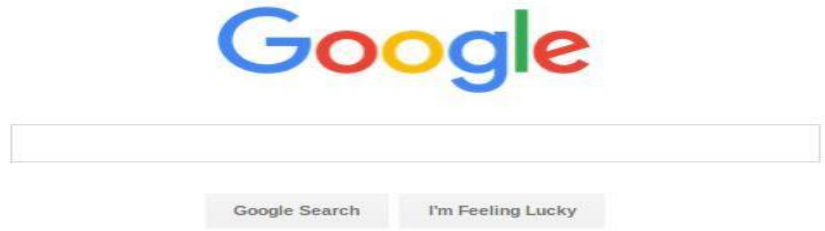
The Query Process



User Interaction - Query input

- Provides interface and parser for internal query language
- Most web queries are very simple
  - Other applications may use forms
- Query language used to describe more complex queries and results of query transformation
  - +, -, " ", ~, site:, AND, OR, …
  - Similar to SQL language used in database applications
    - Not for "end users"
  - IR query languages also allow structure specifications, but focus on content

User Interaction - Query input



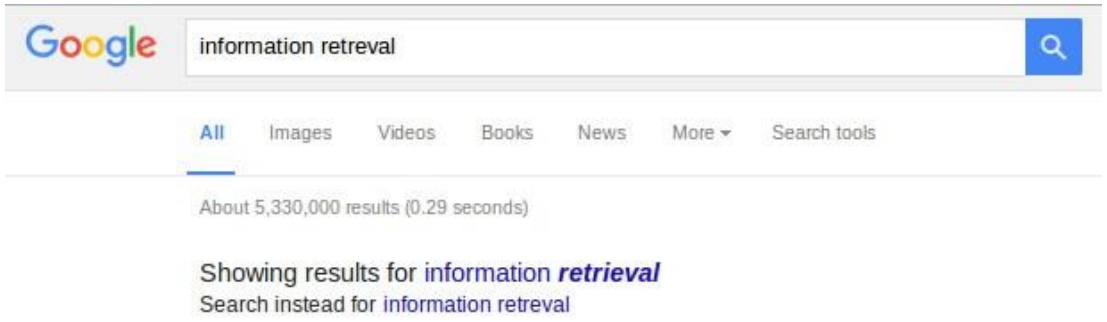
Google

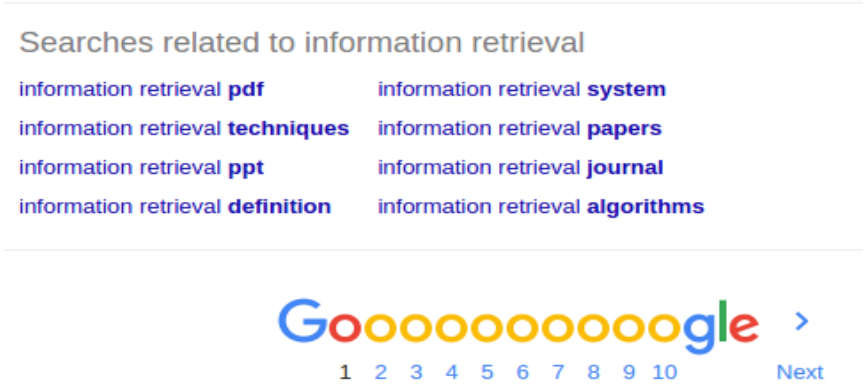
Advanced Search

Find pages with...		To do this in the search box.
all these words:		Type the important words: `tri-colour rat terrier`
this exact word or phrase:		Put exact words in quotes: `"rat terrier"`
any of these words:		Type OR between all the words you want: `miniature OR standard`
none of these words:		Put a minus sign just before words that you don't want: `-rodent, -"Jack Russell"`
numbers ranging from:	to	Put two full stops between the numbers and add a unit of measurement: `10..35 kg, £300..£500, 2010..2011`

User Interaction – Query transformation

- Improves initial query
  - both before and after initial search
- Includes same text transformation techniques used for documents
  - Tokenization, stemming, stopping
- Spell checking and query suggestion provide alternatives to original query
  - Based on query logs
- Query expansion and relevance feedback modify the original query with additional terms

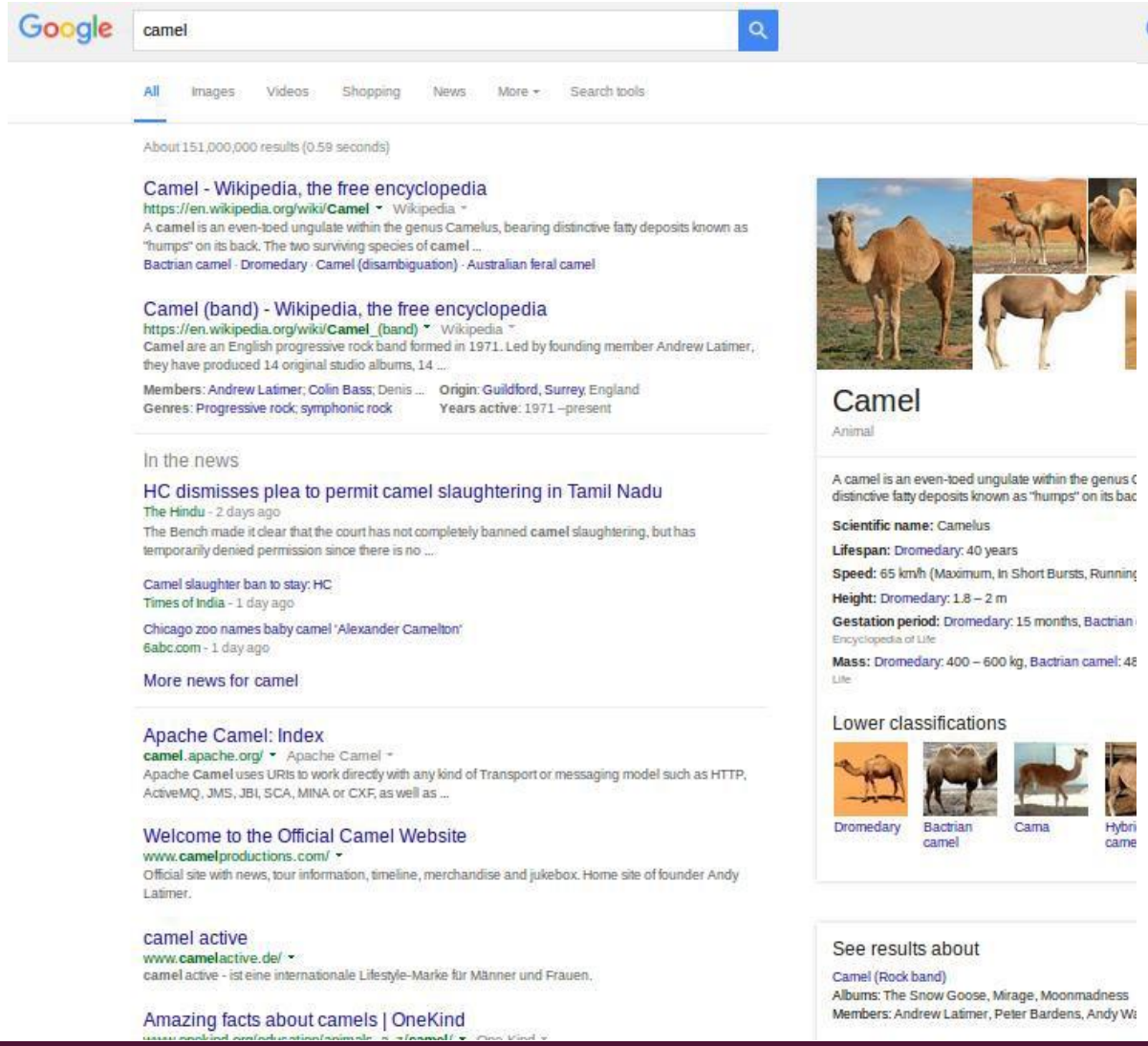


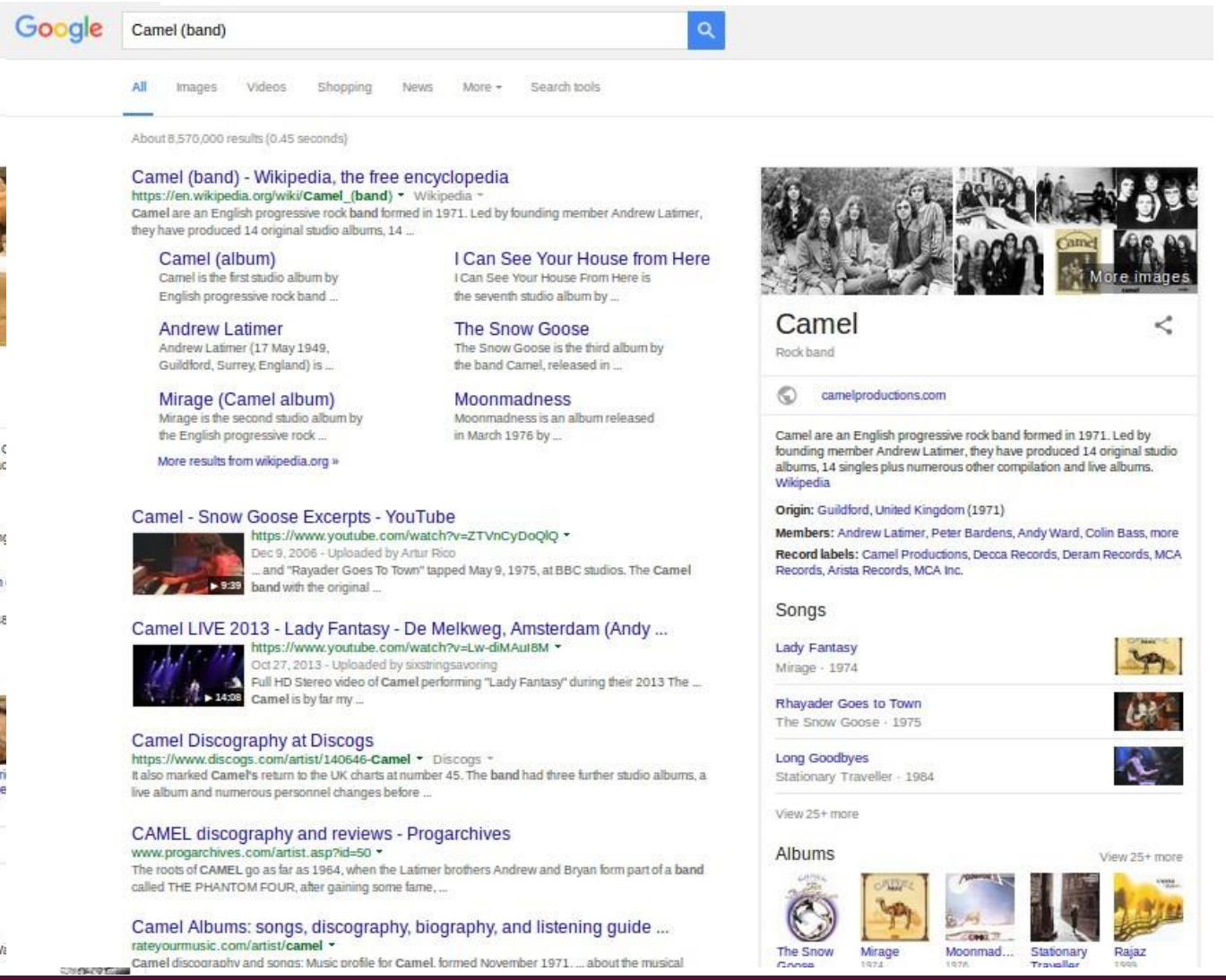


User Interaction – Results output

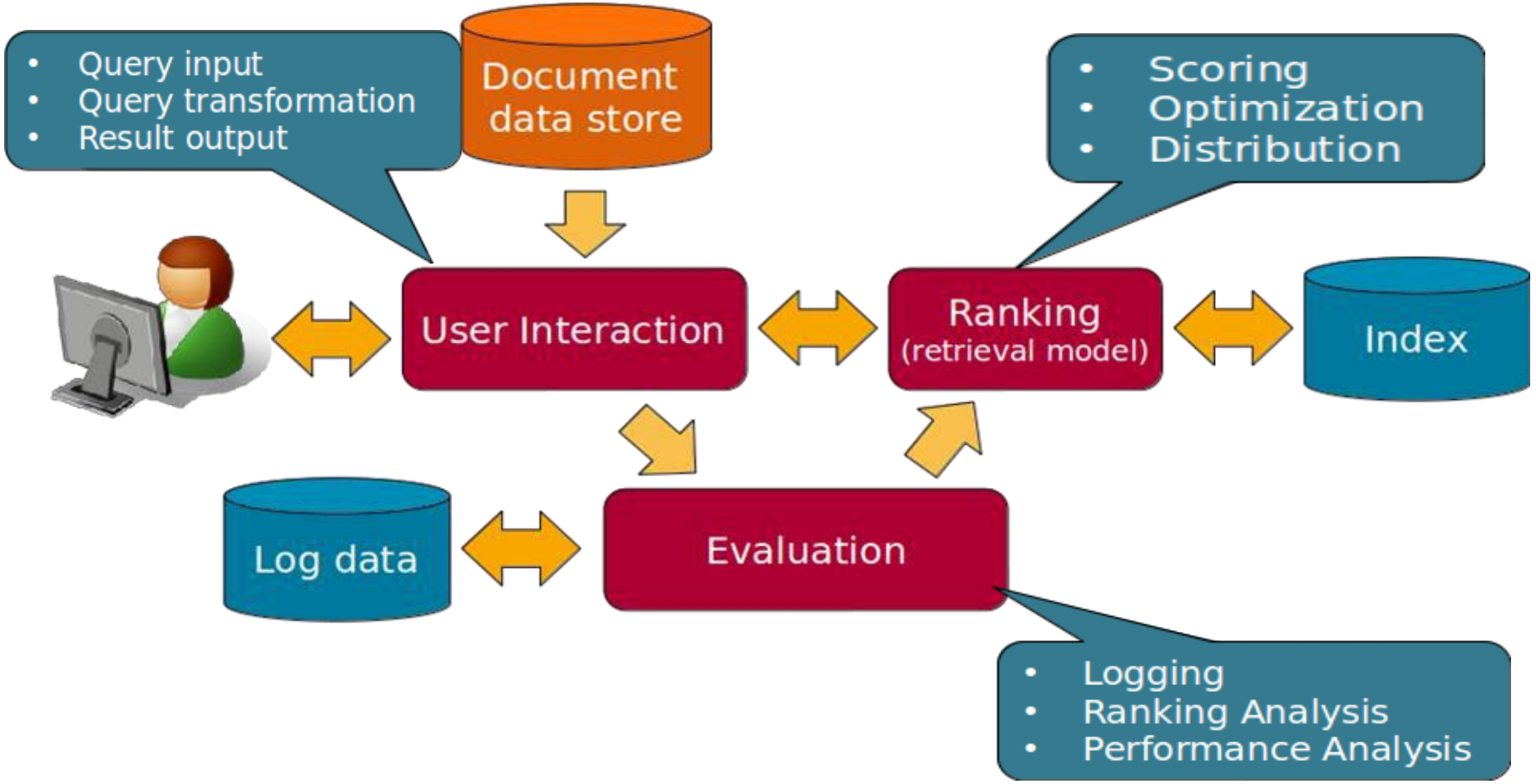
- Constructs the display of ranked documents for a query
- Generates snippets to show how queries match documents
- Highlights important words and passages
- Retrieves appropriate advertising in many applications
- May provide clustering and other visualization tools
- May translate results from foreign languages

User Interaction – Results output





The Query Process



Ranking – Scoring

- ≈ database query processing
- Calculates scores for documents using a ranking algorithm
  - Based on retrieval model
- Core component of search engine
- Many variations of ranking algorithms and retrieval models
- Key requirement: Fast execution!

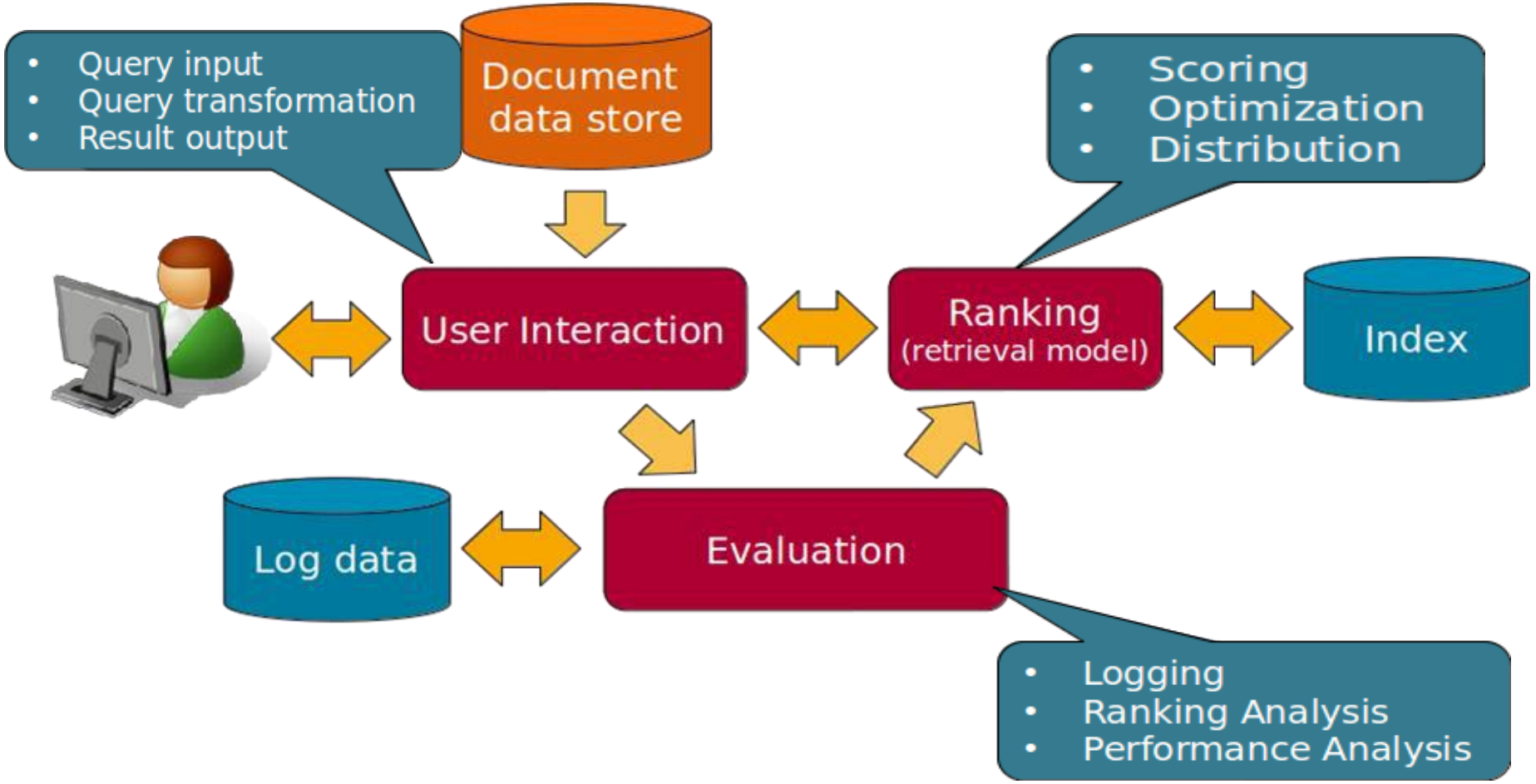
Ranking – Performance optimization

- Designing ranking algorithms for efficient processing
  - Term-at-a time vs. document-at-a-time processing
  - Safe vs. unsafe optimizations
    - Trade-off between speed and quality

Ranking – Distribution

- Processing queries in a distributed environment
- Query broker distributes queries and assembles results

The Query Process

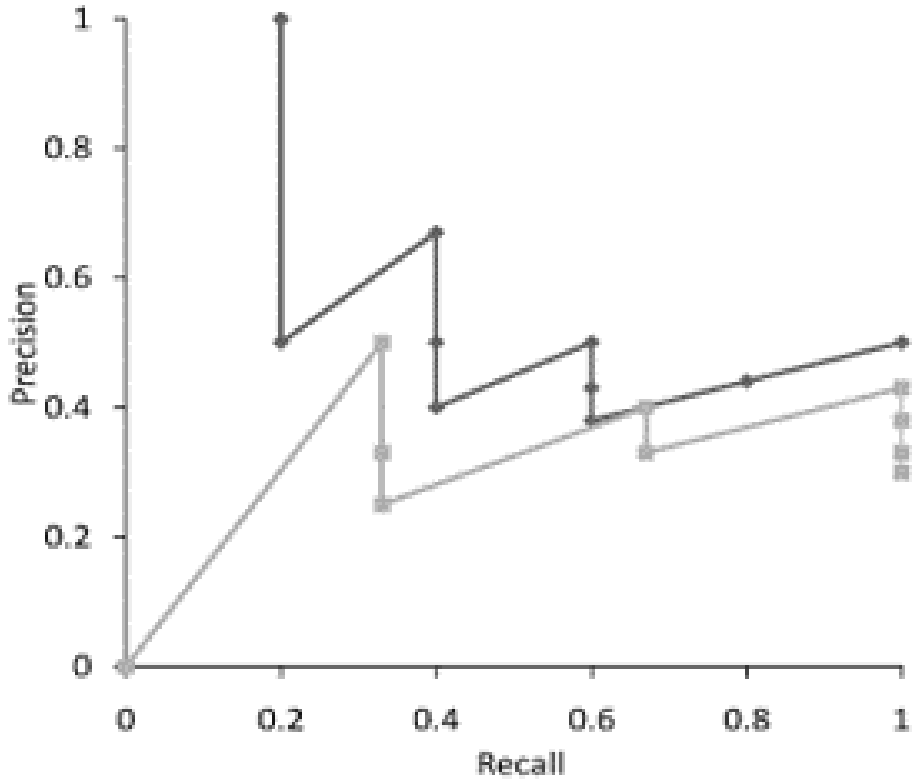


Evaluation – Logging

- Logging user queries and interaction is crucial for improving search effectiveness and efficiency.
- Query logs and clickthrough data (& dwell time) used for
  - Query suggestion
  - Spell checking
  - Query caching
  - Ranking
  - Advertising search
  - …
- Assumption: Pages clicked on are relevant to query.

Evaluation – RankingAnalysis

- Measuring and tuning ranking effectiveness
- Variety of measures



Evaluation – Performance Analysis

- Measuring and tuning system efficiency
- Response time, throughput
- Simulation

How Does It ReallyWork?

- This course explains the components of a search engine in more detail.
- Often many possible approaches and techniques for a given component
  - Focus is on the most important alternatives
    - Explain a small number of approaches in detail rather than many approaches
  - "Importance" based on research results and use in actual search engines
  - Alternatives described in references (see book)

Questions?