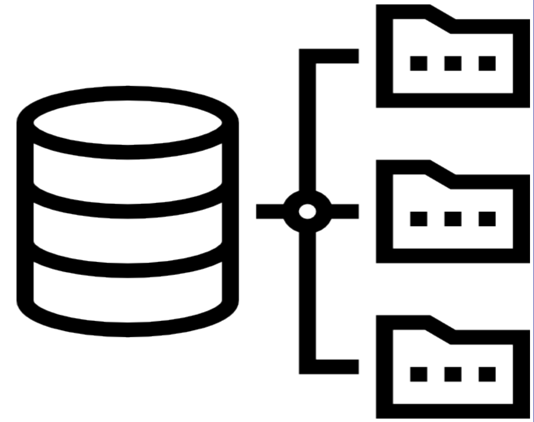
Advanced Database- IS411

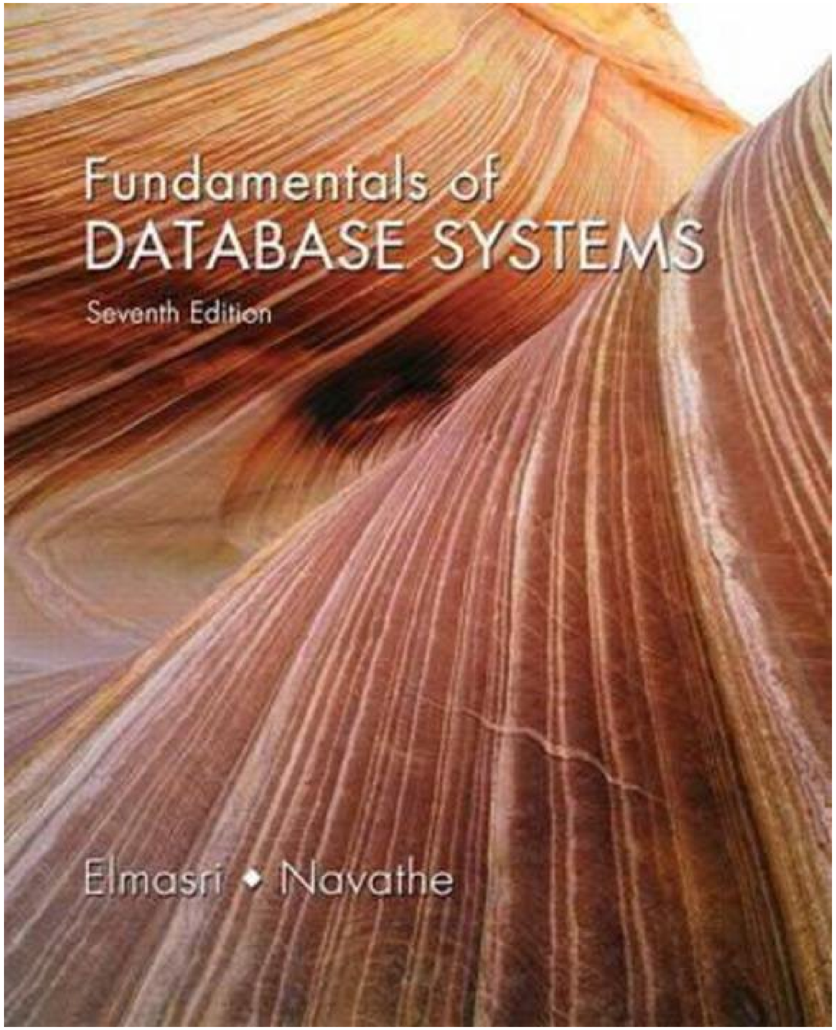


Introduced by

Dr. Ebtsam Adel

Lecturer of Information Systems,
Information Systems department,
Faculty of computers and information,
Damanhour university

Materials


Fundamentals of
DATABASE SYSTEMS
Seventh Edition
Elmasri • Navathe


ADVANCED DATABASE

Topics

- ✓ **Chapter 20** Introduction to Transaction Processing Concepts and Theory
- ✓ **chapter 24** NOSQL Databases and Big Data Storage Systems
- ✓ **chapter 25** Big Data Technologies Based on MapReduce and Hadoop
- ✓ **chapter 27** **Introduction to Information Retrieval and Web Search**
- ✓ **chapter 29** Overview of Data Warehousing and OLAP
- ✓ **chapter 30** Database Security

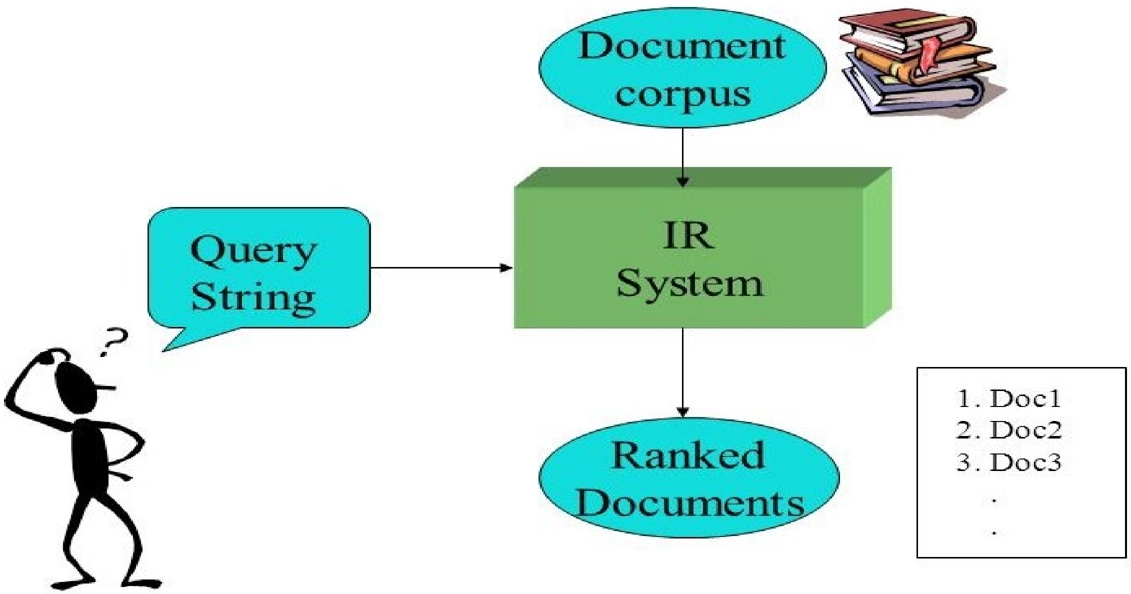
CHAPTER 27

Introduction to Information Retrieval and Web Search

Information Retrieval (IR) Concepts

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).



27.1 Information Retrieval (IR) Concepts

- Information retrieval
  - Process of retrieving documents from a collection in response to a query (search request)
  - Deals mainly with unstructured data
    - Example: homebuying contract documents
- Unstructured information
  - Does not have a well-defined formal model
  - Based on an understanding of natural language
  - Stored in a wide variety of standard formats

Information Retrieval (IR) Concepts (cont'd.)

- Information retrieval field predates database field
  - Academic programs in Library and Information Science
- **RDBMS** vendors providing new capabilities to support various data types
  - Extended RDBMSs or object-relational database management systems.
- User's information need expressed as **free-form** search request
  - Keyword search query

Information Retrieval (IR) Concepts (cont'd.)

Characterizing an IR system

- Types of users
  - Expert
  - Layperson
- Types of data
  - Domain-specific
- Types of information needs
  - Navigational search: locations
  - Informational search: information- wiki
  - Transactional search: transactions

Information Retrieval (IR) Concepts (cont'd.)

Types of Information Need. In the context of Web search, users' information needs may be defined as navigational, informational, or transactional.[3] **Navigational search** refers to finding a particular piece of information (such as the Georgia Tech University Web site) that a user needs quickly. The purpose of **informational search** is to find current information about a topic (such as research activities in the college of computing at Georgia Tech—this is the classic IR system task). The goal of **transactional search** is to reach a site where further interaction happens resulting in some transactional event (such as joining a social network, shopping for products, making online reservations, accessing databases, and so on).

Information Retrieval (IR) Concepts (cont'd.)

- Enterprise search systems
  - Limited to an intranet
- Desktop search engines
  - Searches an individual computer system
- Databases have fixed schemas
  - IR system has no fixed data model

Comparing Databases and IR Systems

Databases	IR Systems
■ Structured data	■ Unstructured data
■ Schema driven	■ No fixed schema; various data models (e.g., vector space model)
■ Relational (or object, hierarchical, and network) model is predominant	■ Free-form query models
■ Structured query model	■ Rich data operations
■ Rich metadata operations	■ Search request returns list or pointers to documents
■ Query returns data	■ Results are based on approximate matching and measures of effectiveness (may be imprecise and ranked)
■ Results are based on exact matching (always correct)	

Table 27.1 A comparison of databases and IR systems

Generic IR approaches

1. **Statistical approach**

2. **Semantic approaches**

Generic IR approaches

- **Statistical approach**
  - Documents analyzed and broken down into **chunks** of text.
  - Each word or phrase is counted, weighted, and measured for relevance or importance
- Types of statistical approaches
  1. Boolean
  2. Vector space
  3. Probabilistic

Generic IR Pipeline (cont'd.)

- **Semantic approaches**
  - Use knowledge-based retrieval techniques.
  - Rely on syntactic, lexical, sentential, discourse-based, and pragmatic levels of knowledge understanding.
  - Also apply some form of statistical analysis

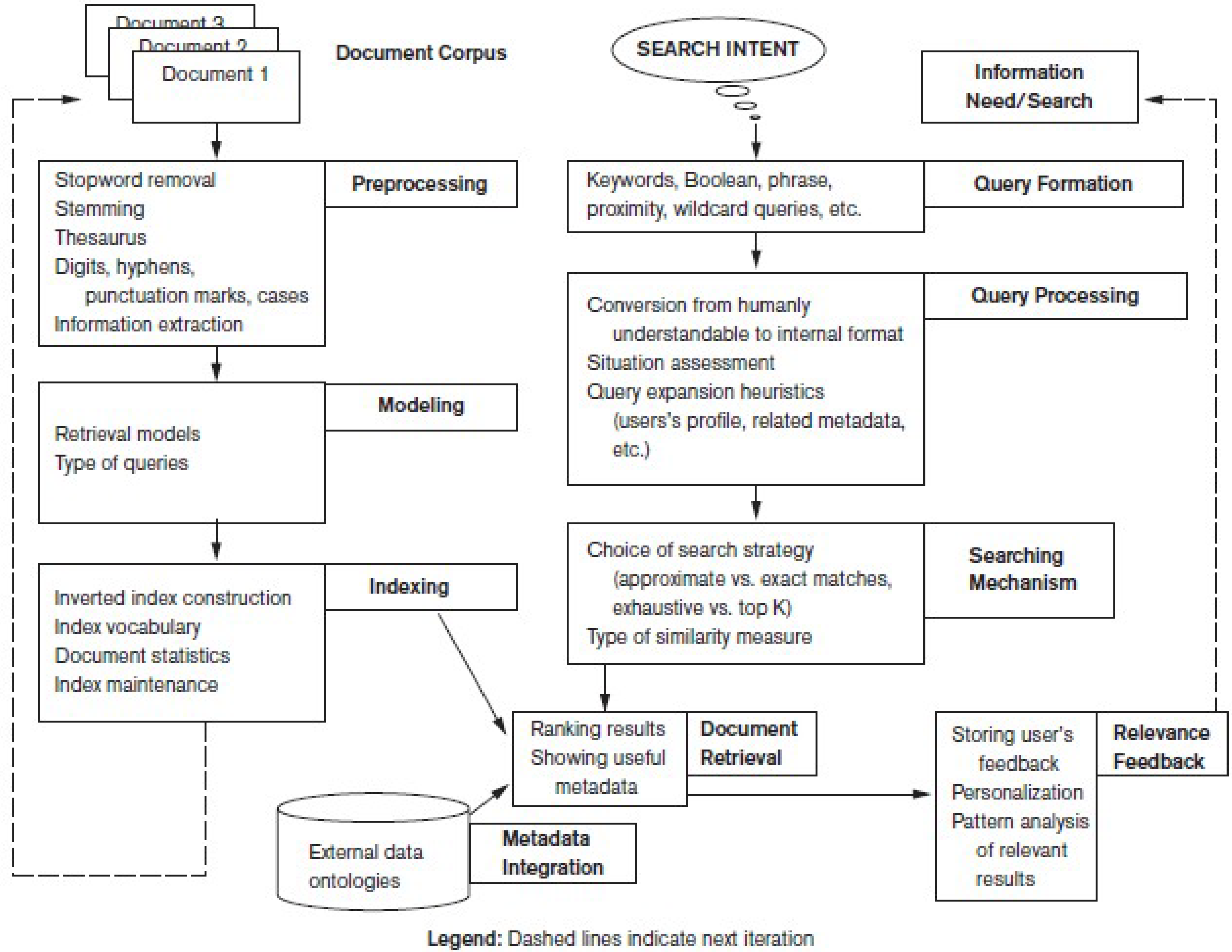


Figure 27.1 Generic IR framework

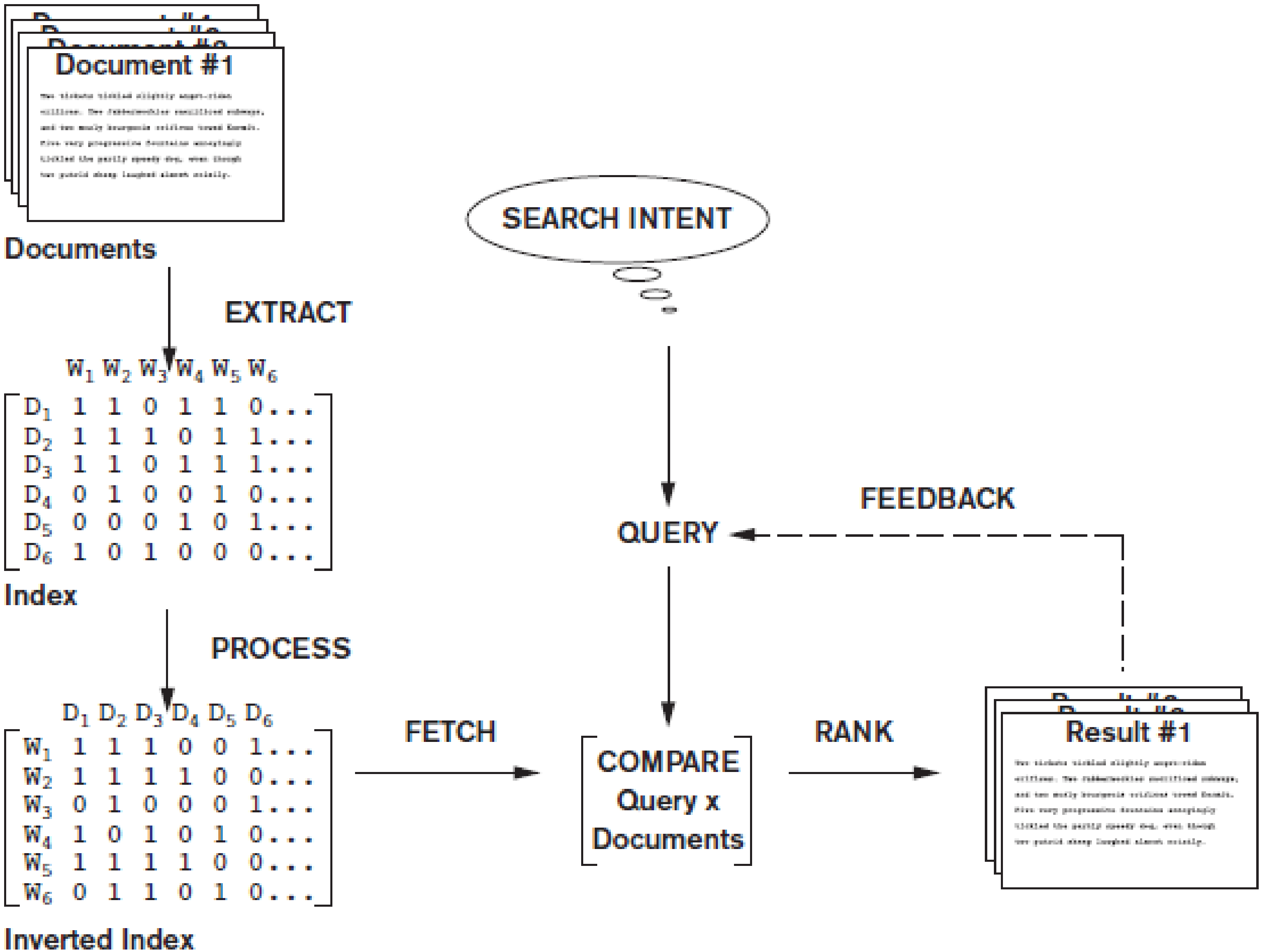


Figure 27.2 Simplified IR process pipeline

Retrieval Models

27.2 Retrieval Models

- Boolean model
  - One of earliest and simplest IR models
  - Documents represented as a set of terms
  - Queries formulated using AND, OR, and NOT
  - Retrieved documents are an **exact match**
    - No notion of ranking of documents
  - Easy to associate metadata information and write queries that match contents of documents.

27.2 Retrieval Models - Boolean model

plays

words

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth	...
Antony	1	1	0	0	0	1	
Brutus	1	1	0	1	0	0	
Caesar	1	1	0	1	1	1	
Calpurnia	0	1	0	0	0	0	
Cleopatra	1	0	0	0	0	0	
mercy	1	0	1	1	1	1	
worser	1	0	1	1	1	0	
...							

27.2 Retrieval Models - Boolean model

- a vector for each term.

a) Brutus: 110100

b) Caesar: 110111

c) NOT Calpurnia: (complemented (**1's** complement) Calpurnia) 101111

- To answer the query Brutus AND Caesar AND NOT Calpurnia, we take the vectors for Brutus, Caesar and Calpurnia, complement the last, and then do a bitwise AND:

110100 **AND** 110111 **AND** 101111 = 100100.

110100
110111
101111

100100

27.2 Retrieval Models - Boolean model

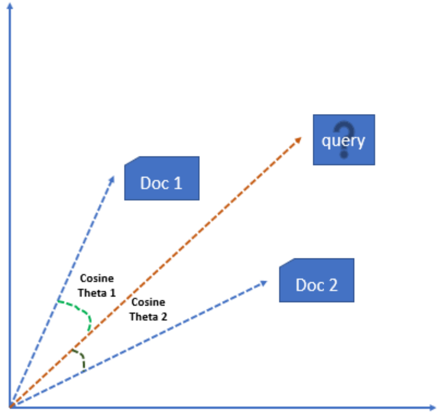
	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth	...
Antony	1	1	0	0	0	1	
Brutus	1	1	0	1	0	0	
Caesar	1	1	0	1	1	1	
Calpurnia	0	1	0	0	0	0	
Cleopatra	1	0	0	0	0	0	
mercy	1	0	1	1	1	1	
worser	1	0	1	1	1	0	
...							
Result	**1**	0	0	**1**	0	0	

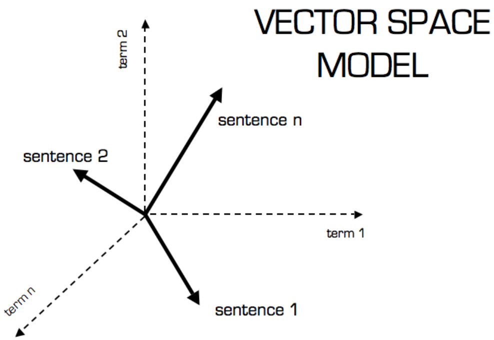
Antony and Cleopatra , **and** *Hamlet*

Vector space model

Retrieval Models (cont'd.)

- Vector space model
  - Weighting, ranking, and determining relevance are possible
  - Uses individual terms as dimensions
  - Each document represented by an n-dimensional vector of values.
  - **Features**
    - Subset of **terms** in a document set that are considered most relevant to an IR search for the document set.

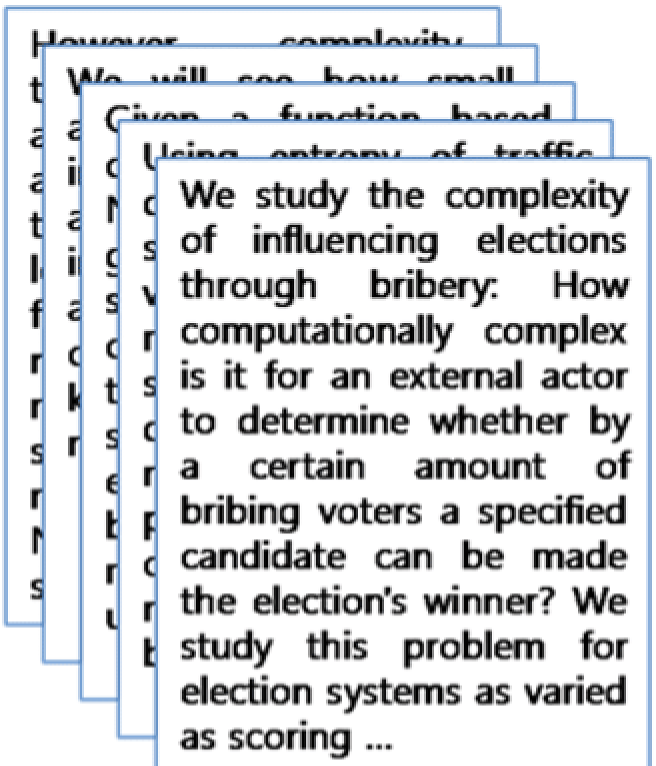




Retrieval Models (cont'd.)-

Vector space model

Documents



Vector-space representation

	D1	D2	D3	D4	D5
complexity	2		3	2	3
algorithm	3			4	4
entropy	1			2	
traffic		2	3		
network		1	4		

Term-document matrix

Retrieval Models (cont’d.)- Vector space model

- Vector space model
  - Different similarity assessment functions can be used.
- Term frequency-inverse document frequency (TF-IDF)
  - Statistical weight measure used to evaluate the importance of a document word in a collection of documents.

Vector space model - Term Frequency

How TF-IDF Works?

TF-IDF combines two components: Term Frequency (TF) and Inverse Document Frequency (IDF).

Term Frequency (TF): Measures how often a word appears in a document.

- A higher frequency suggests greater importance. If a term appears frequently in a document, it is likely relevant to the document's content.

$$TF(t, d) = \frac{\text{Number of times term t appears in document d}}{\text{Total number of terms in document d}}$$

Vector space model - **Inverse Document Frequency**

Limitations of TF Alone:

•TF does not account for the global importance of a term across the entire **corpus**.

•Common words like **"the"** or **"and"** may have high TF scores but are not meaningful in distinguishing "تمييز" documents.

Inverse Document Frequency (IDF): Reduces the weight of common words across multiple documents while increasing the weight of rare words. If a term appears in fewer documents, it is more likely to be meaningful and specific.

$$\mathrm{IDF}(t, D) = \log \frac{\text{Total number of documents in corpus D}}{\text{Number of documents containing term t}}$$

Vector space model

- Term frequency-inverse document frequency (TF-IDF)

IDF computation. The following formulas can be used:

$$TF_{ij} = f_{ij} \Big/ \sum_{i=1 \text{ to } |V|} f_{ij}$$

$$IDF_i = \log\left(N / n_i\right)$$

In these formulas, the meaning of the symbols is:

- TF_{ij} is the normalized term frequency of term i in document D_j.
- f_{ij} is the number of occurrences of term i in document D_j.
- IDF_i is the inverse document frequency weight for term i.
- N is the number of documents in the collection.
- n_i is the number of documents in which term i occurs.

Probabilistic model

Retrieval Models (cont'd.)

- Probabilistic model
  - Involves ranking documents by their estimated probability of relevance with respect to the query and the document.
  - IR system must decide whether a document belongs to the relevant set or nonrelevant set for a query
    - Calculate probability that document belongs to the relevant set
  - BM25: a popular ranking algorithm

Retrieval Models - **Semantic model**

Semantic model

- Morphological analysis
  - Analyze roots and affixes to determine parts of speech of search words
- Syntactic analysis
  - Parse and analyze complete phrases in documents
- Semantic analysis
  - Resolve word ambiguities and z

➢ Uses techniques from artificial intelligence and expert systems.

Types of Queries in IR Systems

27.3 Types of Queries in IR Systems

- Keyword queries
  - Simplest and most commonly used
  - Keyword terms implicitly connected by logical AND
- Boolean queries
  - Allow use of AND, OR, NOT, and other operators
  - Exact matches returned
    - No ranking possible

Types of Queries in IR Systems (cont'd.)

- Phrase queries
  - Sequence of words that make up a phrase
  - Phrase enclosed in double quotes
  - Each retrieved document must contain at least one instance of the exact phrase
- Proximity queries
  - How **close** within a record multiple search terms are to each other
  - Phrase search is most commonly used proximity query

Types of Queries in IR Systems (cont'd.)

- Proximity queries (cont'd.)
  - Specify **order** of search terms
  - NEAR, ADJ (adjacent), or AFTER operators
  - Sequence of words with maximum allowed distance between them
  - Computationally expensive
    - Suitable for smaller document collections rather than the Web.

Types of Queries in IR Systems (cont'd.)

- Wildcard queries
  - Supports regular expressions and pattern-based matching
    - Example 'data*' would retrieve data, database, dataset, etc.
  - Not generally implemented by Web search engines.
- Natural language queries
  - Definitions of textual terms or common facts
  - **Semantic models** can support

Text Preprocessing

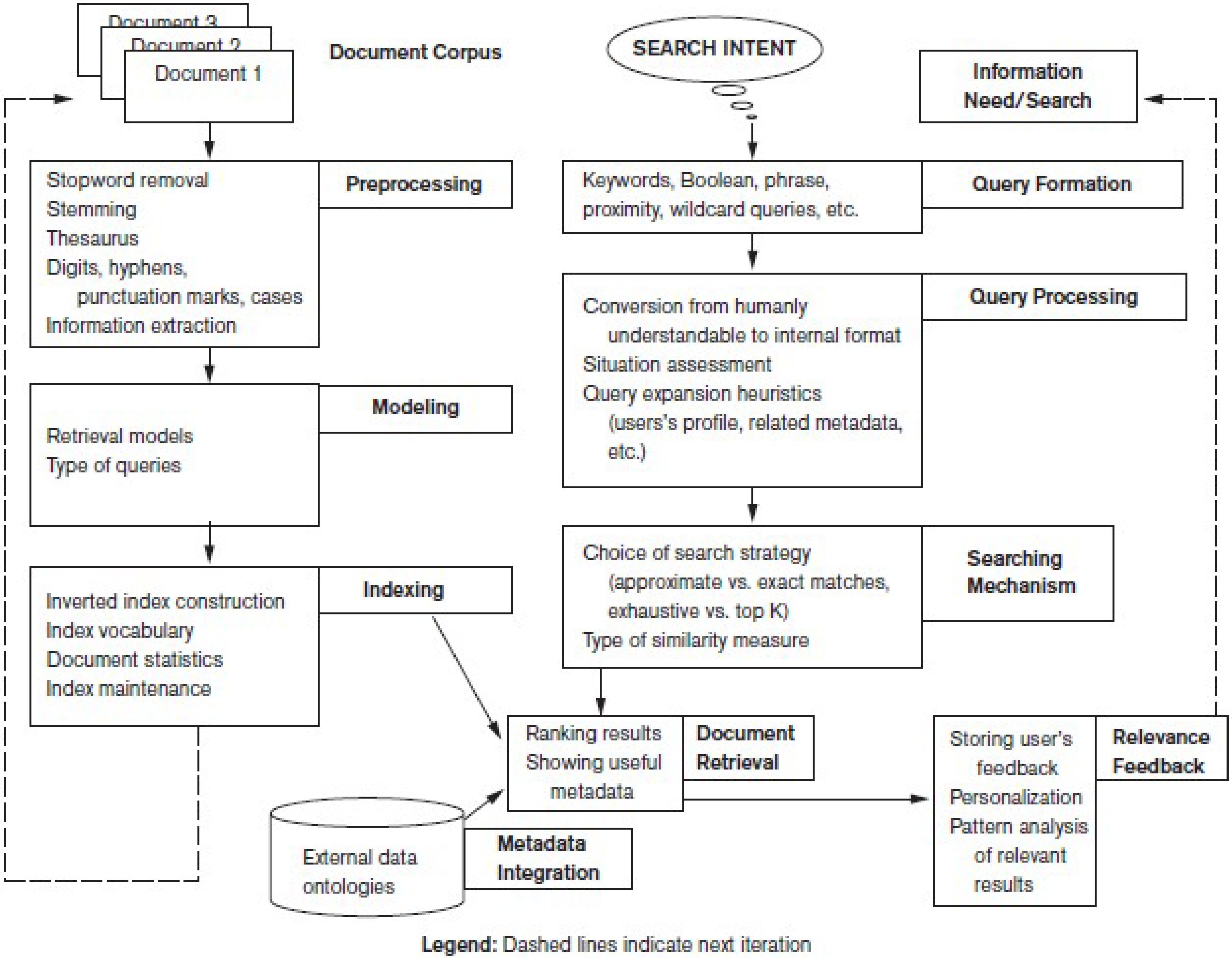


Figure 27.1 Generic IR framework

27.4 Text Preprocessing

- Stopword removal must be performed before indexing
- Stopwords
  - Words that are expected to occur in 80% or more of the documents of a collection
    - Examples: the, of, to, a, and, said, for, that
  - Do not contribute much to relevance
- Queries preprocessed for stopword removal before retrieval process
  - *Many search engines do not remove stopwords*

Text Preprocessing (cont'd.)

- Stemming
  - Trims suffix and prefix
  - Reduces the different forms of the word to a common stem
  - Martin Porter's stemming algorithm
- Utilizing a thesaurus
  - Important concepts and main words that describe each concept for a particular knowledge domain
  - Collection of synonyms
  - UMLS

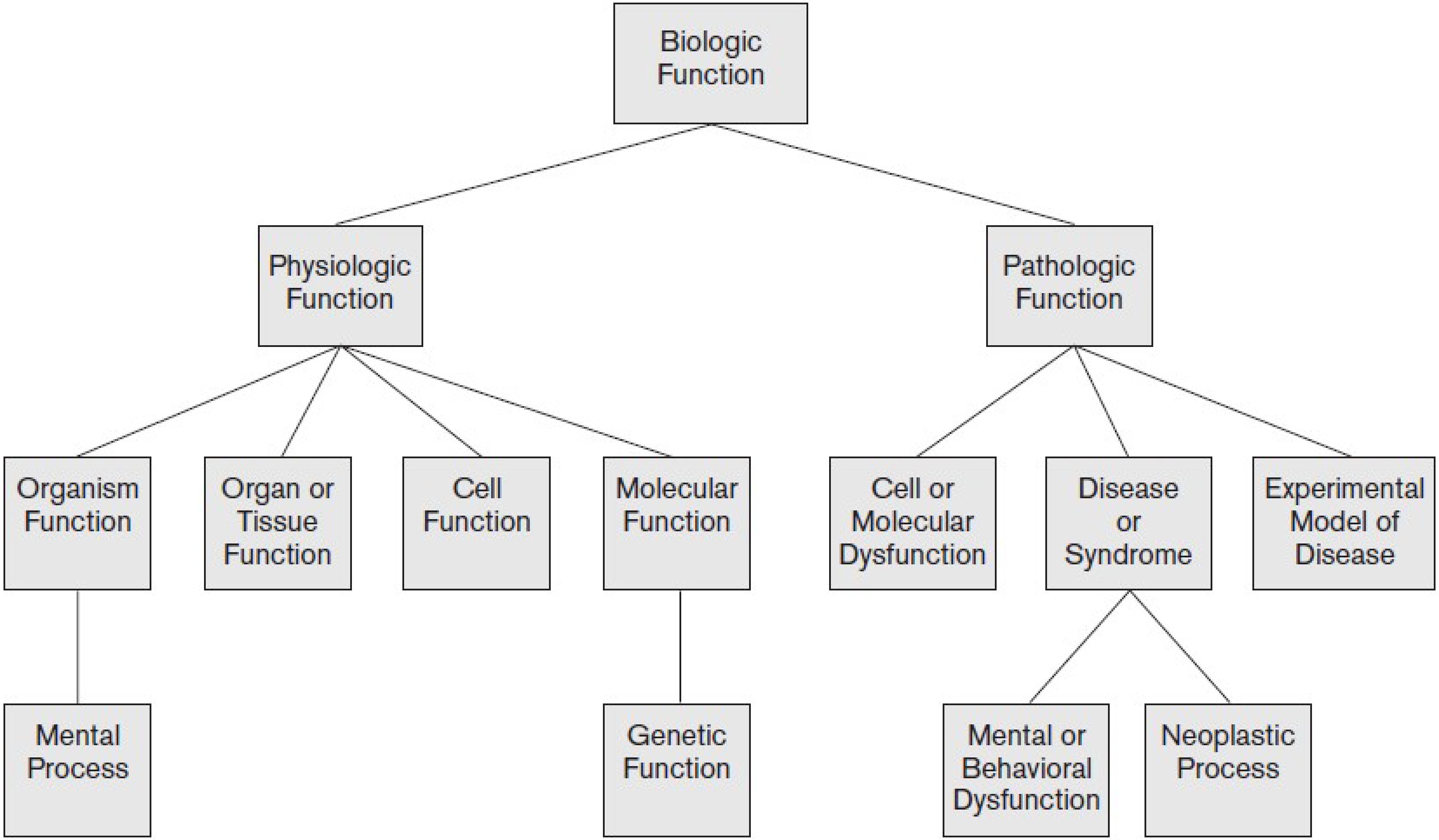


Figure 27.3 A portion of the UMLS Semantic Network: "Biologic Function" Hierarchy
Source: UMLS Reference Manual, National Library of **Medicine**

Text Preprocessing (cont'd.)

- Other preprocessing steps
  - Digits
    - May or may not be removed during preprocessing
  - Hyphens and punctuation marks
    - Handled in different ways
  - Cases
    - Most search engines use case-insensitive search
- Information extraction tasks
  - Identifying noun phrases, facts, events, people, places, and relationships.

Inverted Indexing

27.5 Inverted Indexing

- Inverted index structure
  - Vocabulary information
    - Set of distinct query terms in the document set
  - Document information: term frequency

- **Inverted index** : Data structure that attaches distinct terms with a list of all documents that contain the term

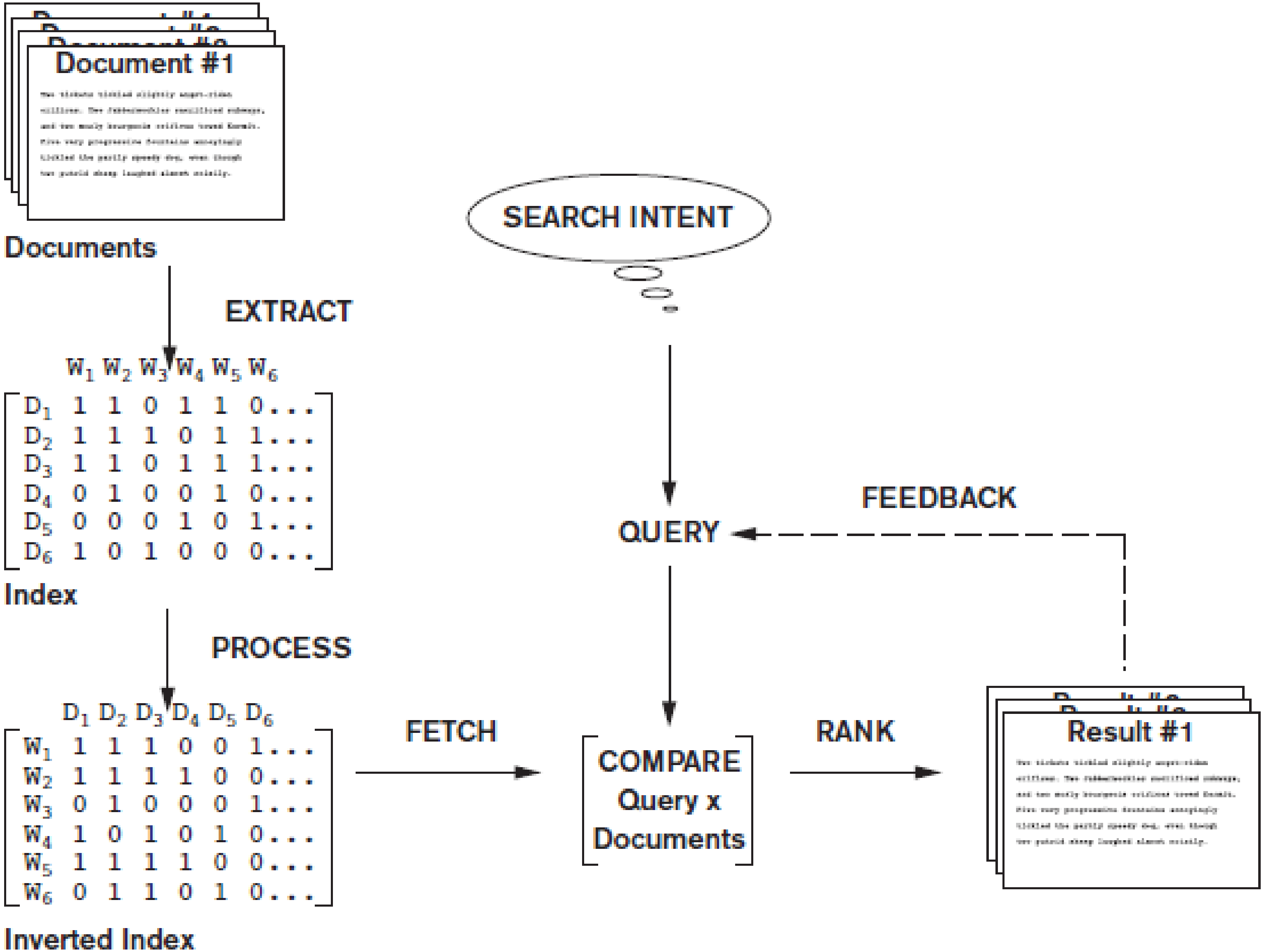


Figure 27.2 Simplified IR process pipeline

Inverted Indexing (cont'd.)

- Construction of an inverted index
  - Break documents into vocabulary terms
    - Tokenizing, removing stopwords, stemming, and/or using a thesaurus
  - Collect document statistics
    - Store statistics in **document lookup** table
  - Invert the document-term stream into a term-document stream
    - Add additional information such as term frequencies, term positions, and term weights

Document 1

This example shows an example of an inverted index.

Document 2

Inverted index is a data structure for associating terms to documents.

Document 3

Stock market index is used for capturing the sentiments of the financial market.

ID	Term	Document: position
1.	example	1:2, 1:5
2.	inverted	1:8, 2:1
3.	index	1:9, 2:2, 3:3
4.	market	3:2, 3:13

Figure 27.4 Example of an inverted index

Inverted Indexing (cont'd.)

- Searching for relevant documents from an inverted index
  - Vocabulary search
  - Document information retrieval
  - Manipulation of retrieved information

Introduction to Lucene

- Lucene: open source indexing/search engine
  - Indexing is primary focus
- Document composed of set of fields
  - Chunks of untokenized text
  - Series of processed lexical units called **token streams**
    - Created by tokenization and filtering algorithms
- Highly-configurable search API
- Ease of indexing large, unstructured document collections

Evaluation Measures of Search Relevance

27.6 Evaluation Measures of Search Relevance

- Topical relevance
  - Measures result topic match to query topic
- User relevance
  - Describes ‘goodness’ of retrieved result with regard to user’s information need
- Web information retrieval
  - No binary classification made for relevance or nonrelevance
  - Ranking of documents

Evaluation Measures of Search Relevance (cont'd.)

- Recall
  - Number of relevant documents retrieved by a search divided by the total number of actually relevant documents existing in the database
- Precision
  - Number of relevant documents retrieved by a search divided by total number of documents retrieved by that search

Retrieved Versus Relevant Search Results

- TP: true positive
- FP: false positive
- TN: true negative
- FN: false negative

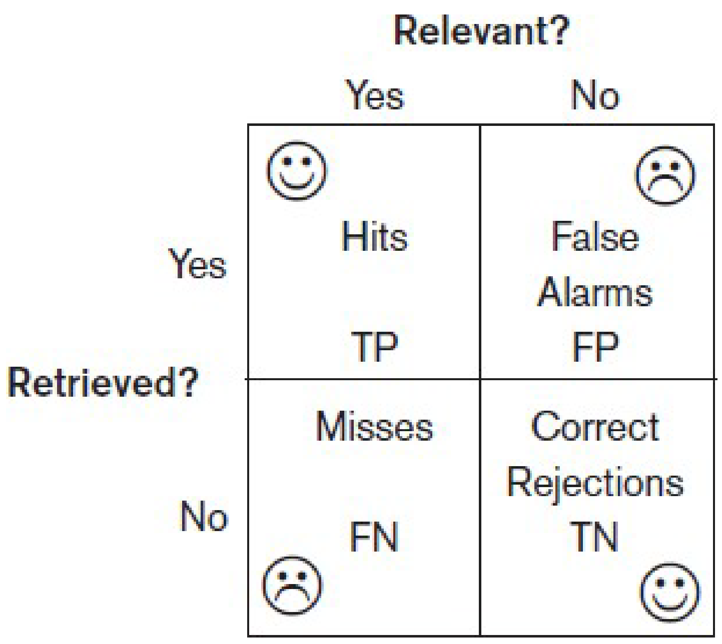


Figure 27.5 Retrieved versus relevant search results

Information retrieval system evaluation

- The standard approach to information retrieval system evaluation revolves around the notion of *relevant* and non-relevant documents.

- *Precision* (*P*) is the fraction of retrieved documents that are relevant.
- *Recall* (*R*) is the fraction of relevant documents that are retrieved.

Precision and recall

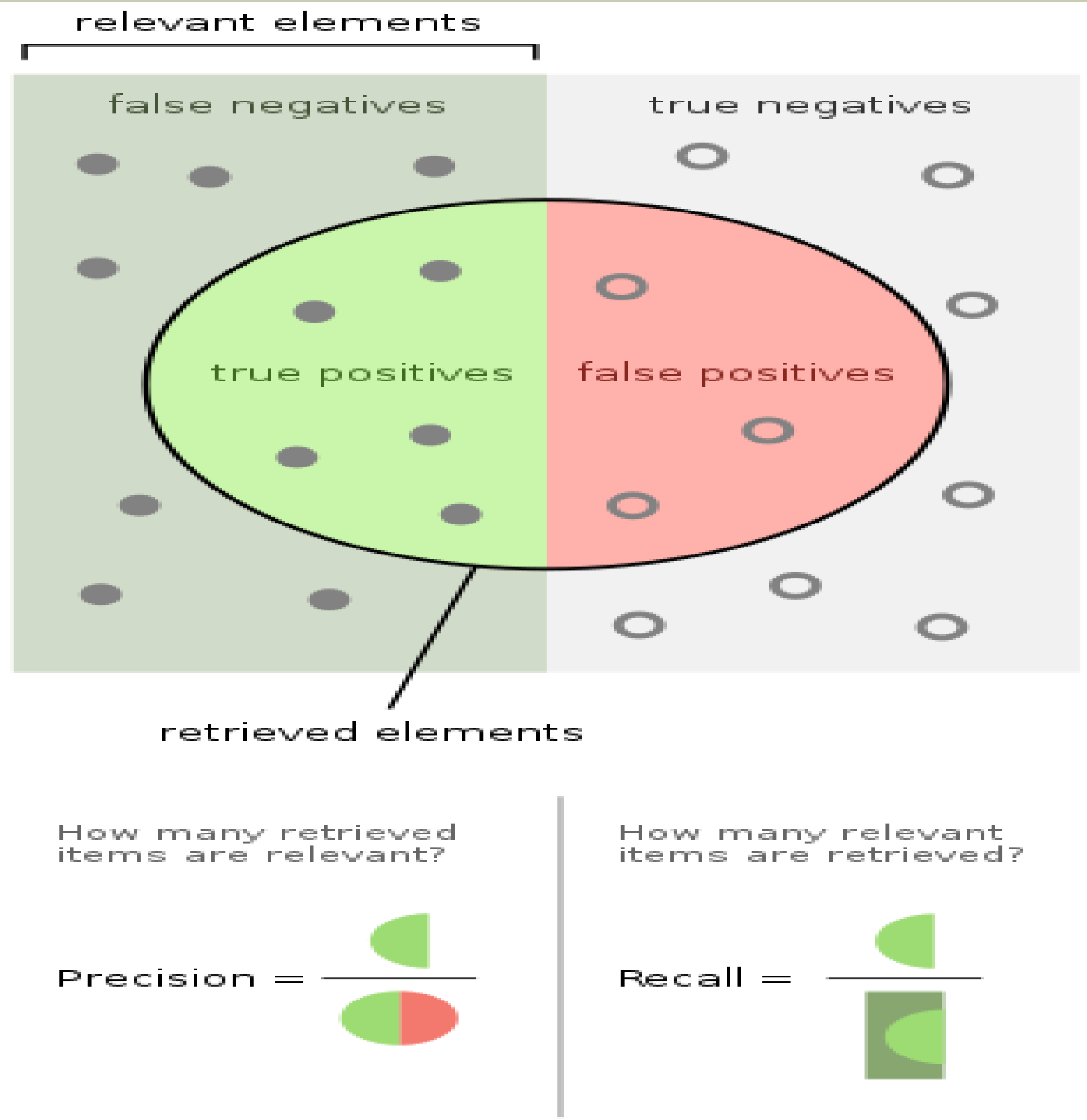
- Precision (P) is the fraction of retrieved documents that are relevant.

$$\text{Precision}=\frac{=\#(\text{relevant items retrieved})}{\#(\text{retrieved items})}= P(\text{relevant}|\text{retrieved})$$

- Recall (R) is the fraction of relevant documents that are Retrieved.

$$\text{Recall}=\frac{\#(\text{relevant items retrieved})}{\#(\text{relevant items})}= = P(\text{retrieved}|\text{relevant})$$

Positive=retrieved
Negative=Not Retrieved



Precision and recall

THE TRUTH

	Relevant	Not Relevant
Retrieved	true positives (TP)	false positives (FP)
Not retrieved	false negatives (FN)	true negatives (TN)

$$P = \frac{TP}{(TP+FP)}$$

$$R = \frac{TP}{(TP+FN)}$$

Evaluation Measures of Search Relevance (cont'd.)

- Recall can be increased by presenting more results to the user
  - May decrease the **precision**

Doc. No.	Rank Position i	Relevant	Precision(i)	Recall(i)
10	1	Yes	1/1 = 100%	1/10 = 10%
2	2	Yes	2/2 = 100%	2/10 = 20%
3	3	Yes	3/3 = 100%	3/10 = 30%
5	4	No	3/4 = 75%	3/10 = 30%
17	5	No	3/5 = 60%	3/10 = 30%
34	6	No	3/6 = 50%	3/10 = 30%
215	7	Yes	4/7 = 57.1%	4/10 = 40%
33	8	Yes	5/8 = 62.5%	5/10 = 50%
45	9	No	5/9 = 55.5%	5/10 = 50%
16	10	Yes	6/10 = 60%	6/10 = 60%

Table 27.2 Precision and recall for ranked retrieval

Evaluation Measures of Search Relevance (cont'd.)

- Average precision
  - Computed based on the precision at each relevant document in the ranking
- Recall/precision curve
  - Based on the recall and precision values at each rank position
    - x-axis is recall and y-axis is precision
- F-score
  - Harmonic mean of the precision (p) and recall (r) values


Thank You !
Any questions?