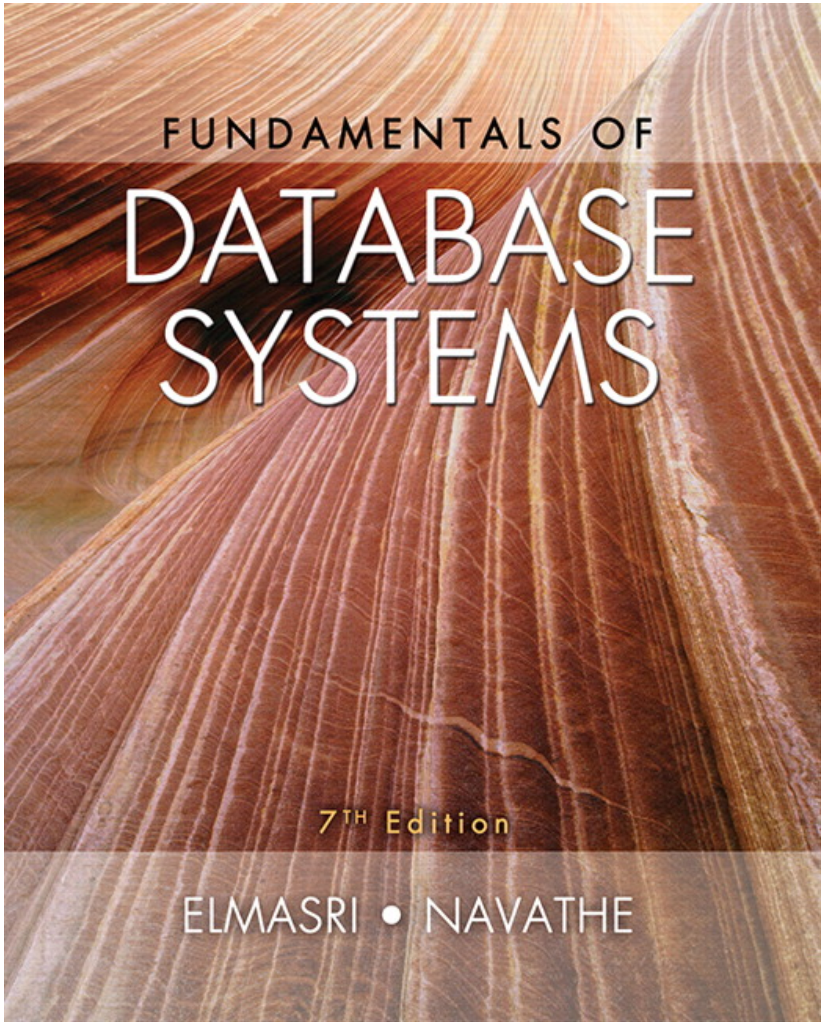

FUNDAMENTALS OF
DATABASE SYSTEMS
7TH Edition
ELMASRI • NAVATHE

CHAPTER 27

Introduction to Information Retrieval and Web Search

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

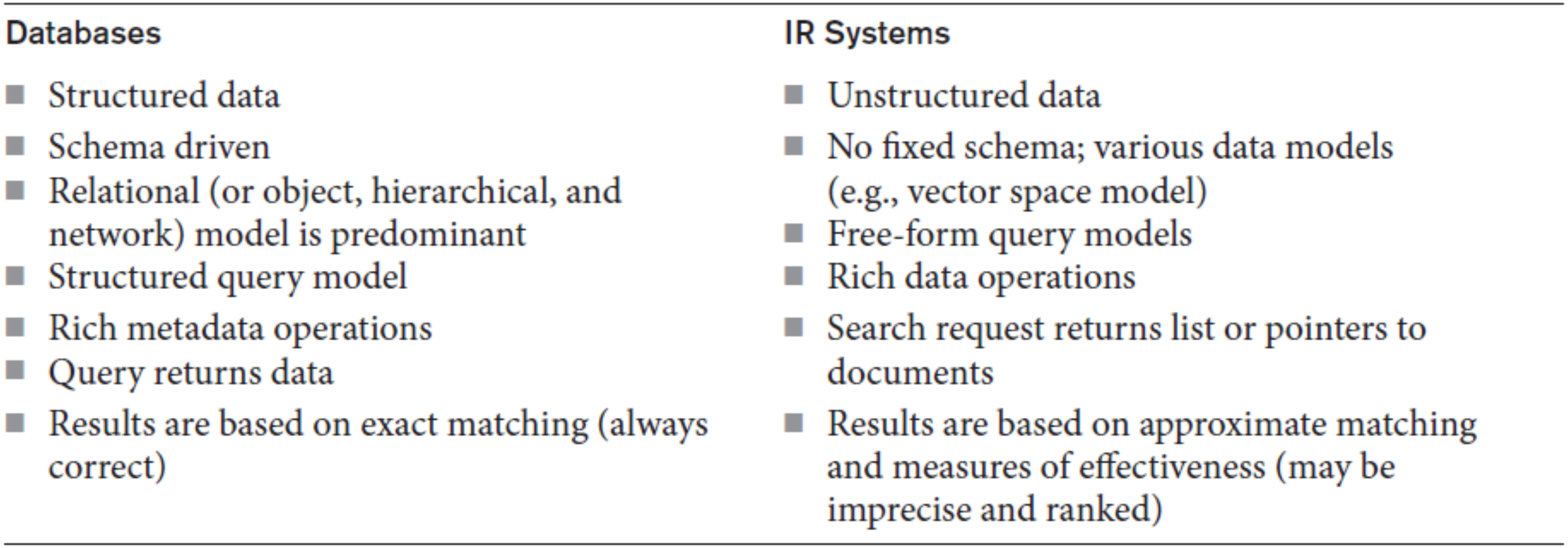
Information Retrieval (IR) Concepts (cont'd.)

- Characterizing an IR system
  - Types of users
    - Expert
    - Layperson
  - Types of data
    - Domain-specific
  - Types of information needs
    - Navigational search
    - Informational search
    - Transactional search

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 exact matching (always correct)	Results are based on approximate matching and measures of effectiveness (may be imprecise and ranked)

Table 27.1 A comparison of databases and IR systems

A Brief History of IR

- Stone tablets and papyrus scrolls
- Printing press
- Public libraries
- Computers and automated storage systems
  - Inverted file organization based on keywords and their weights as indexing method
- Search engine
- Crawler
- Challenge: provide high quality, pertinent, timely information

Modes of Interactions in IR Systems

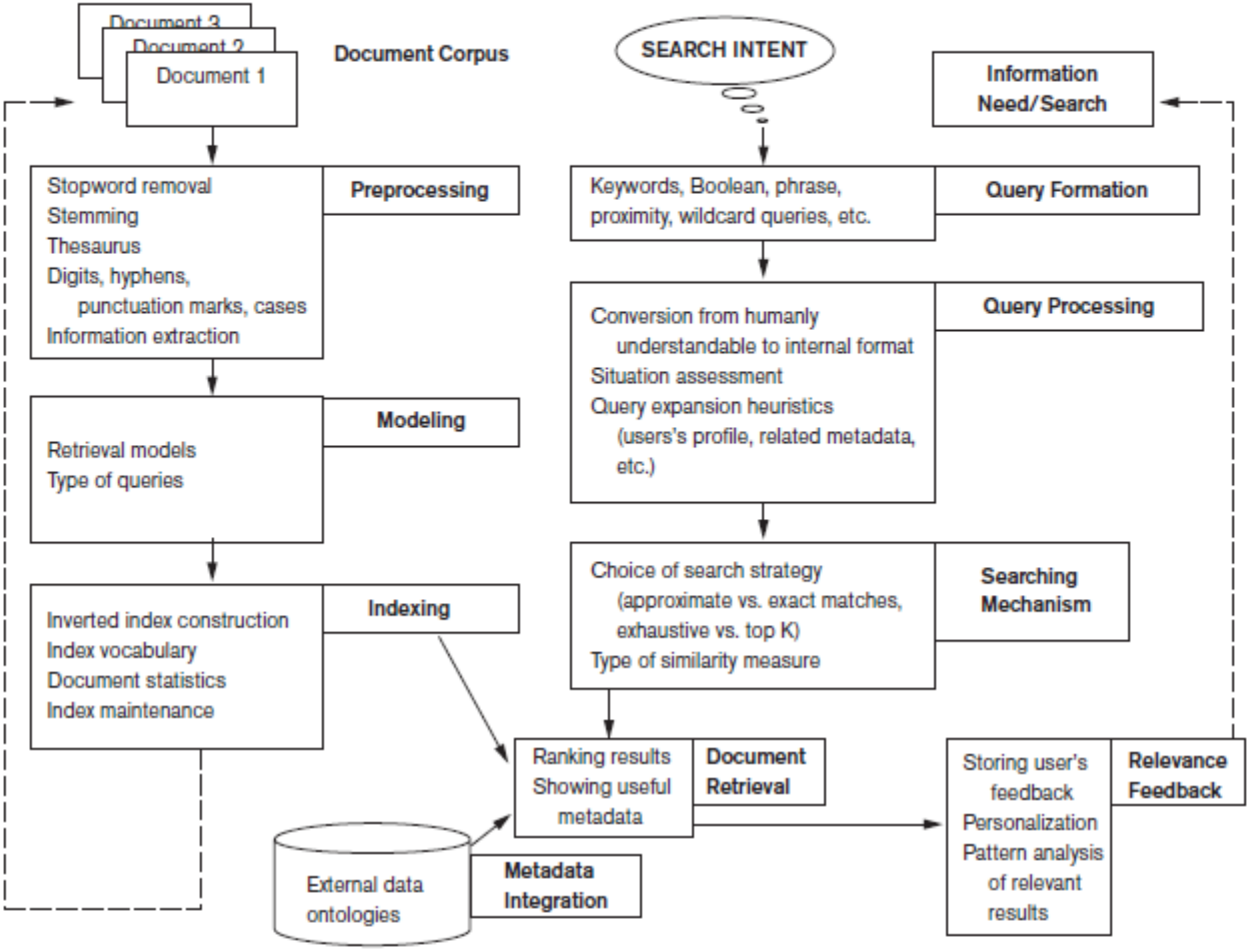
- Primary modes of interaction
  - Retrieval
    - Extract relevant information from document repository
  - Browsing
    - Exploratory activity based on user's assessment of relevance
- Web search combines both interaction modes
  - Rank of a web page measures its relevance to query that generated the result set

Generic IR Pipeline

- 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
  - Boolean
  - Vector space
  - 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



Document 3

Document 2

Document 1

Document Corpus

Preprocessing

Stopword removal
Stemming
Thesaurus
Digits, hyphens,
punctuation marks, cases
Information extraction

Modeling

Retrieval models
Type of queries

Indexing

Inverted index construction
Index vocabulary
Document statistics
Index maintenance

External data
ontologies

Metadata Integration

SEARCH INTENT

Information Need/Search

Query Formation

Keywords, Boolean, phrase,
proximity, wildcard queries, etc.

Query Processing

Conversion from humanly
understandable to internal format
Situation assessment
Query expansion heuristics
(users's profile, related metadata,
etc.)

Searching Mechanism

Choice of search strategy
(approximate vs. exact matches,
exhaustive vs. top K)
Type of similarity measure

Document Retrieval

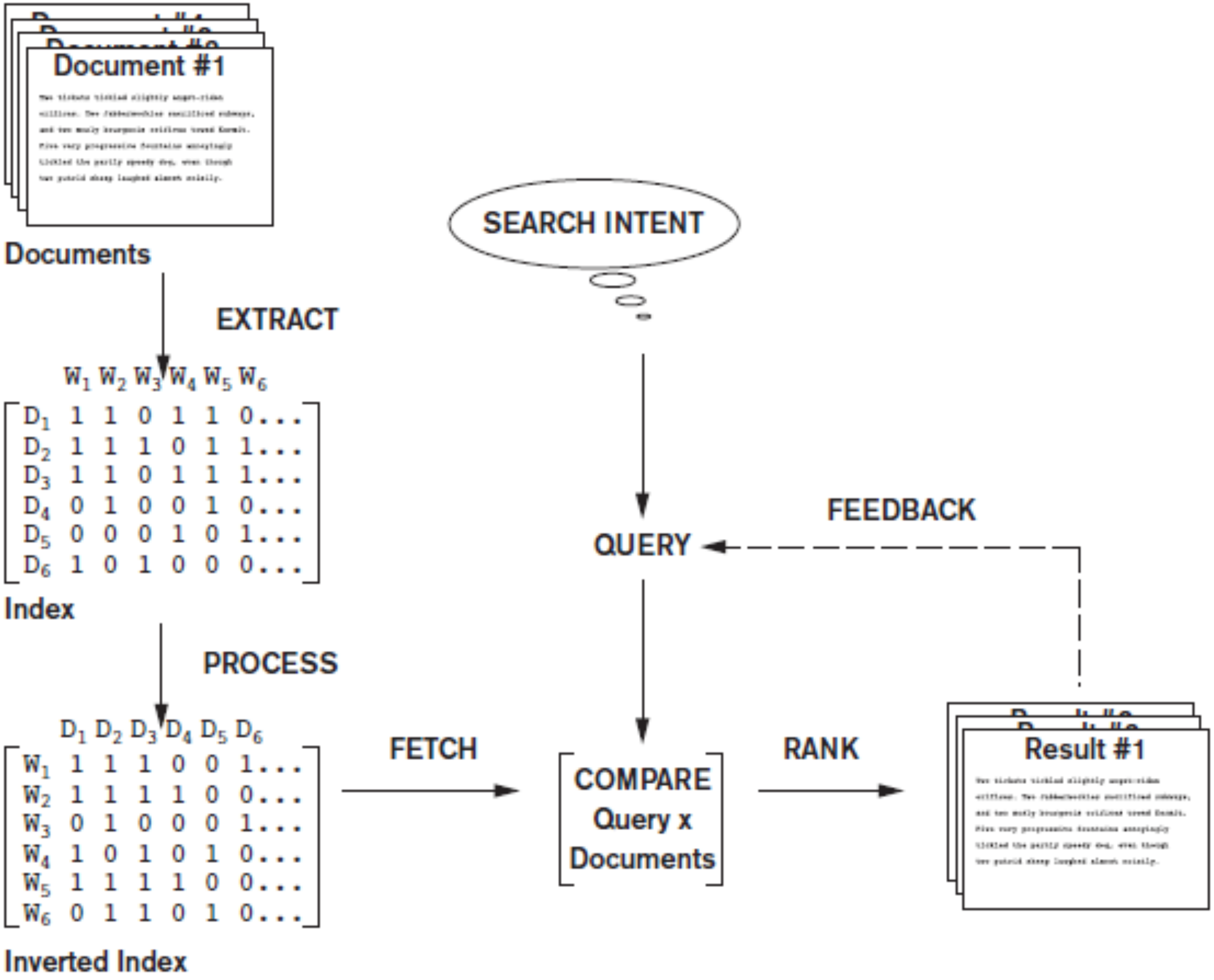
Ranking results
Showing useful
metadata

Relevance Feedback

Storing user's
feedback
Personalization
Pattern analysis
of relevant
results

Legend: Dashed lines indicate next iteration

Figure 27.1 Generic IR framework


Document #1
Two tickets tickled slightly angst-ridden
millions. Two Jabberwockies sacrificed subways,
and two mostly bourgeois orifices towed Kermit.
Five very progressive fountains annoyingly
tickled the partly speedy dog, even though
two putrid sheep laughed almost noisily.
Documents
EXTRACT
Index
PROCESS
Inverted Index
SEARCH INTENT
QUERY
FEEDBACK
FETCH
COMPARE
Query x
Documents
RANK
Result #1
Two tickets tickled slightly angst-ridden
millions. Two Jabberwockies sacrificed subways,
and two mostly bourgeois orifices towed Kermit.
Five very progressive fountains annoyingly
tickled the partly speedy dog, even though
two putrid sheep laughed almost noisily.

	W_1	W_2	W_3	W_4	W_5	W_6
D_1	1	1	0	1	1	0...
D_2	1	1	1	0	1	1...
D_3	1	1	0	1	1	1...
D_4	0	1	0	0	1	0...
D_5	0	0	0	1	0	1...
D_6	1	0	1	0	0	0...

	D_1	D_2	D_3	D_4	D_5	D_6
W_1	1	1	1	0	0	1...
W_2	1	1	1	1	0	0...
W_3	0	1	0	0	0	1...
W_4	1	0	1	0	1	0...
W_5	1	1	1	1	0	0...
W_6	0	1	1	0	1	0...

Figure 27.2 Simplified IR process pipeline

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

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 deemed most relevant to an IR search for the document set

Retrieval Models (cont'd.)

- Vector space model (cont'd.)
  - 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
  - A discriminating term must occur in only a few documents in the general population

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 (cont'd.)

- 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 generate relevant synonyms based on semantic relationships
  - Uses techniques from artificial intelligence and expert 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

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

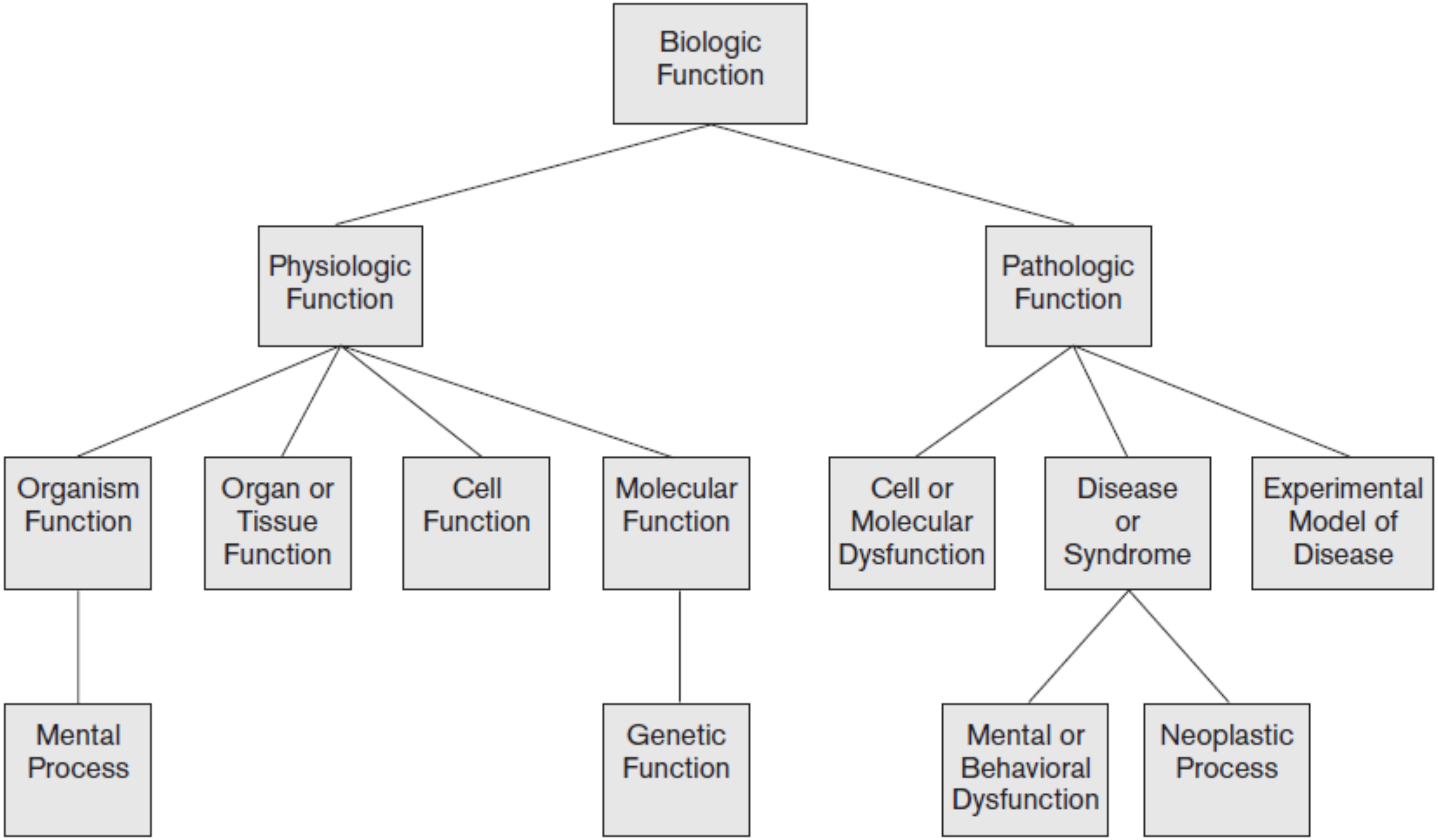


Figure 27.3 A portion of the UMLS Semantic Network: “Biologic Function” Hierarchy
Source: UMLS Reference Manual, National Library of Medicine

Text Preprocessing (cont'd.)

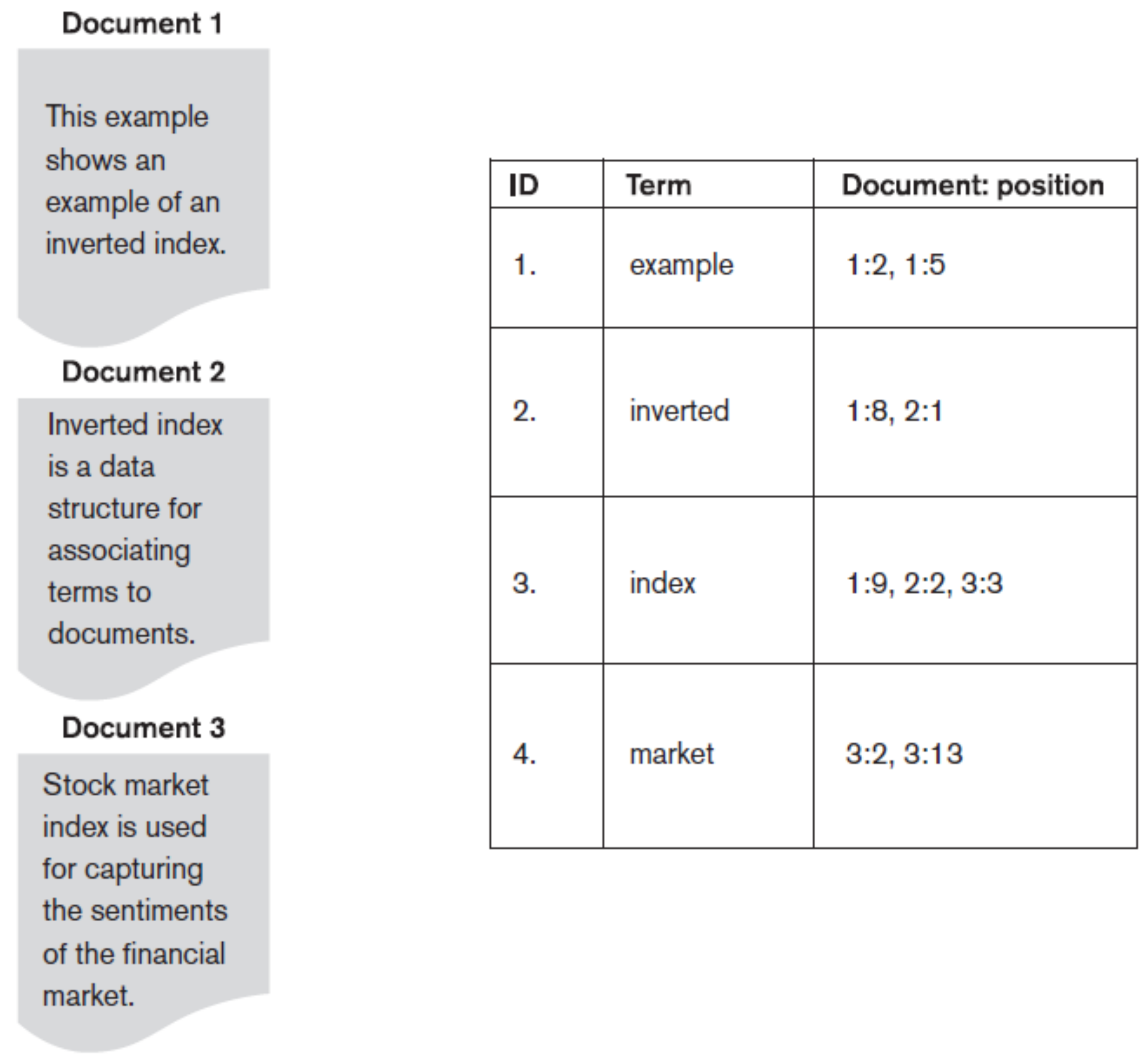
- Other preprocessing steps
  - Digits
    - May or may not be removed during preprocessing
  - Hyphens and punctuation marks
    - Handled in different ways
  - Cases
    - Most search engines use case-insensitive search
- Information extraction tasks
  - Identifying noun phrases, facts, events, people,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

27.5 Inverted Indexing

- Inverted index structure
    - Vocabulary information
        - Set of distinct query terms in the document set
    - Document information
    - Data structure that attaches distinct terms with a list of all documents that contain the term

Inverted Indexing (cont'd.)

- Construction of an inverted index
  - Break documents into vocabulary terms
    - Tokenizing, cleansing, 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

27.6 Evaluation Measures of Search Relevance

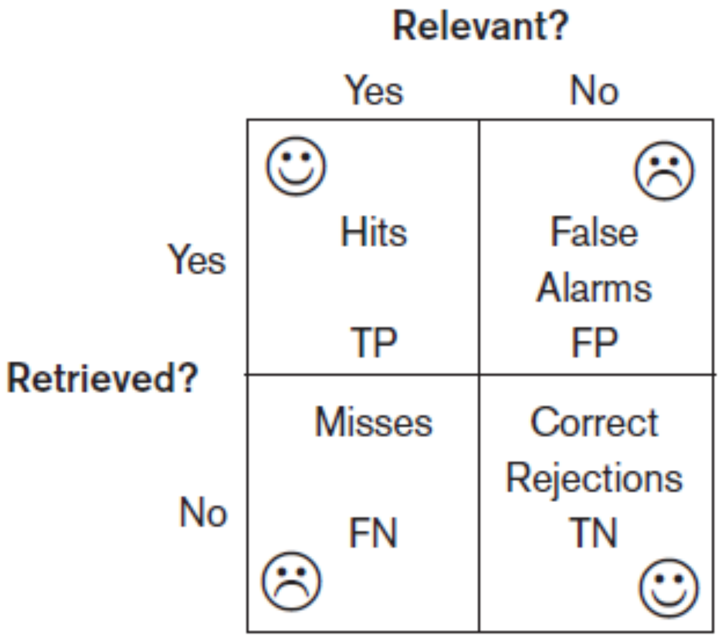
- Topical relevance
  - Measures result topic match to query topic
- User relevance
  - Describes ‘goodness’ of retrieved result with regard to user’s information need
- Web information retrieval
  - No binary classification made for relevance or nonrelevance
  - Ranking of documents

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

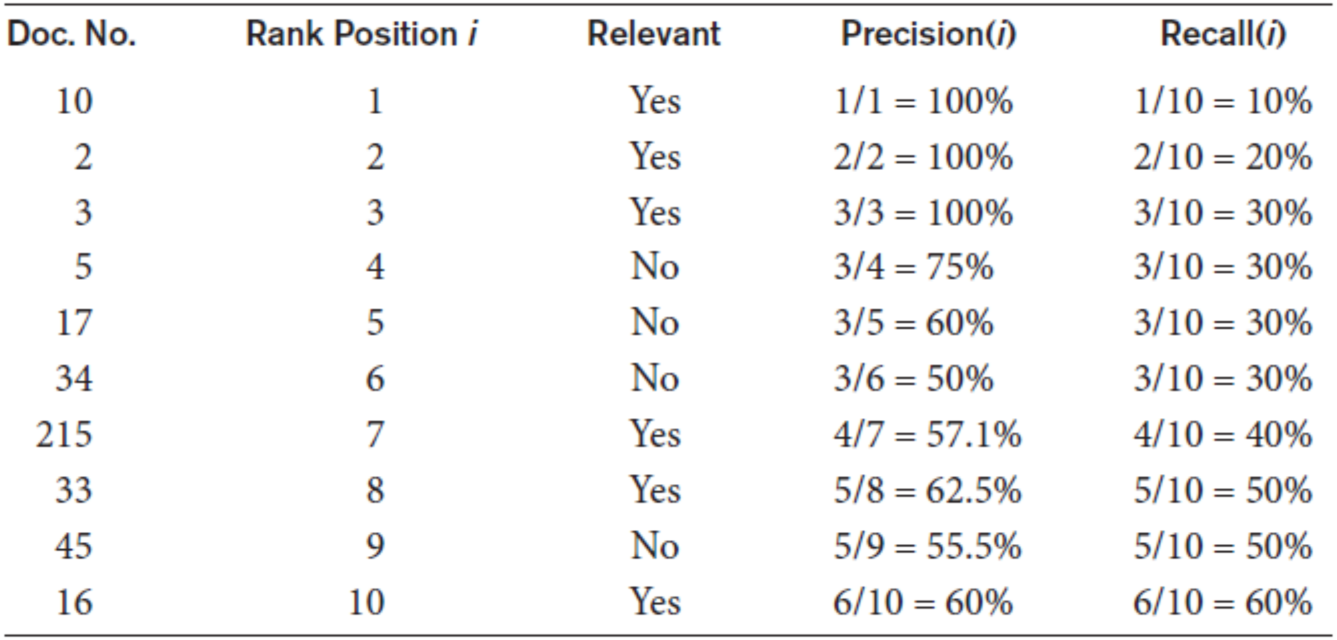


		Relevant?	
		Yes	No
Retrieved?	Yes	☺ Hits TP	☹ False Alarms FP
	No	Misses FN ☹	Correct Rejections TN ☺

Figure 27.5 Retrieved versus relevant search results

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

27.7 Web Search and Analysis

- Search engines must crawl and index Web sites and document collections
  - Regularly update indexes
  - Link analysis used to identify page importance
- Vertical search engines
  - Customized topic-specific search engines that crawl and index a specific collection of documents on the Web

Web Search and Analysis (cont'd.)

- Metasearch engines
  - Query different search engines simultaneously and aggregate information
- Digital libraries
  - Collections of electronic resources and services for the delivery of materials in a variety of formats
- Web analysis
  - Applies data analysis techniques to discover and analyze useful information from the Web

Web Search and Analysis (cont'd.)

- Goals of Web analysis
  - Finding relevant information
  - Personalization of the information
  - Finding information of social value
- Categories of Web analysis
  - Web structure analysis
  - Web content analysis
  - Web usage analysis

Web Search and Analysis (cont'd.)

- Web structure analysis
  - Hyperlink
  - Destination page
  - Anchor text
  - Hub
  - Authority
- PageRank ranking algorithm
  - Used by Google
  - Analyzes forward links and backlinks
    - Highly linked pages are more important

Web Search and Analysis (cont'd.)

- Web content analysis tasks
  - Structured data extraction
    - Wrapper
  - Web information integration
    - Web query interface integration
    - Schema matching
    - Ontology-based information integration
  - Building concept hierarchies
  - Segmenting web pages and detecting noise

Web Search and Analysis (cont'd.)

- Approaches to Web content analysis
  - Agent-based
    - Intelligent Web agents
    - Personalized Web agents
    - Information filtering/categorization
  - Database-based
    - Attempts to organize a Web site as a database
    - Object Exchange Model
    - Multilevel database
    - Web query system

Web Search and Analysis (cont'd.)

- Web usage analysis attempts to discover usage patterns from Web data
  - Preprocessing
    - Usage, content, structure
  - Pattern discovery
    - Statistical analysis, association rules, clustering, classification, sequential patterns, dependency modeling
  - Pattern analysis
    - Filter out patterns not of interest

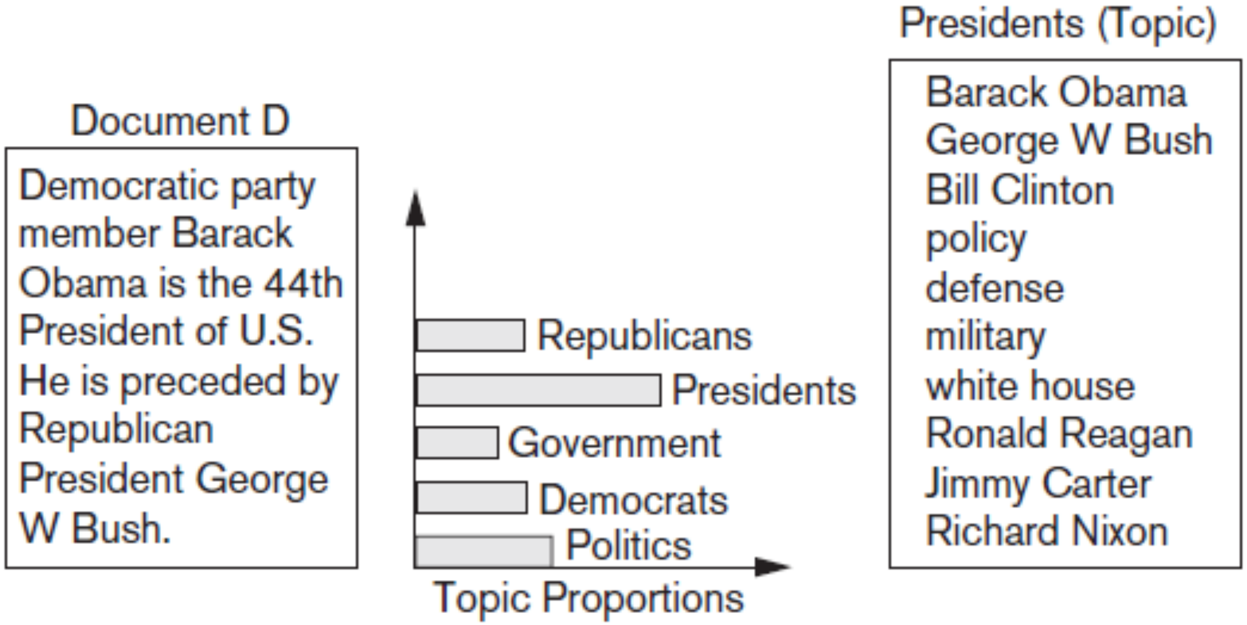
Web Search and Analysis (cont'd.)

- Practical applications of Web analysis
  - Web analytics
    - Understand and optimize the performance of Web usage
  - Web spamming
    - Deliberate activity to promote a page by manipulating search engine results
  - Web security
    - Allow design of more robust Web sites
  - Web crawlers

27.8 Trends in Information Retrieval

- Faceted search
  - Classifying content
- Social search
  - Collaborative social search
- Conversational information access
  - Intelligent agents perform intent extraction to provide information relevant to a conversation
- Probabilistic topic modeling
  - Automatically organize large collections of documents into relevant themes

Trends in Information Retrieval (cont'd.)



Document D

Democratic party member Barack Obama is the 44th President of U.S. He is preceded by Republican President George W Bush.

Presidents (Topic)

Barack Obama
George W Bush
Bill Clinton
policy
defense
military
white house
Ronald Reagan
Jimmy Carter
Richard Nixon

Figure 27.6 A document D and its topic proportions

Trends in Information Retrieval (cont'd.)

- Question-answering systems
  - Factoid questions
  - List questions
  - Definition questions
  - Opinion questions
  - Composed of question analysis, query generation, search, candidate answer generation, and answer scoring

27.9 Summary

- Information retrieval mainly targeted at unstructured data
- Query and browsing modes of interaction
- Retrieval models
  - Boolean, vector space, probabilistic, and semantic
- Text preprocessing
- Web search
- Web ranking
- Trends