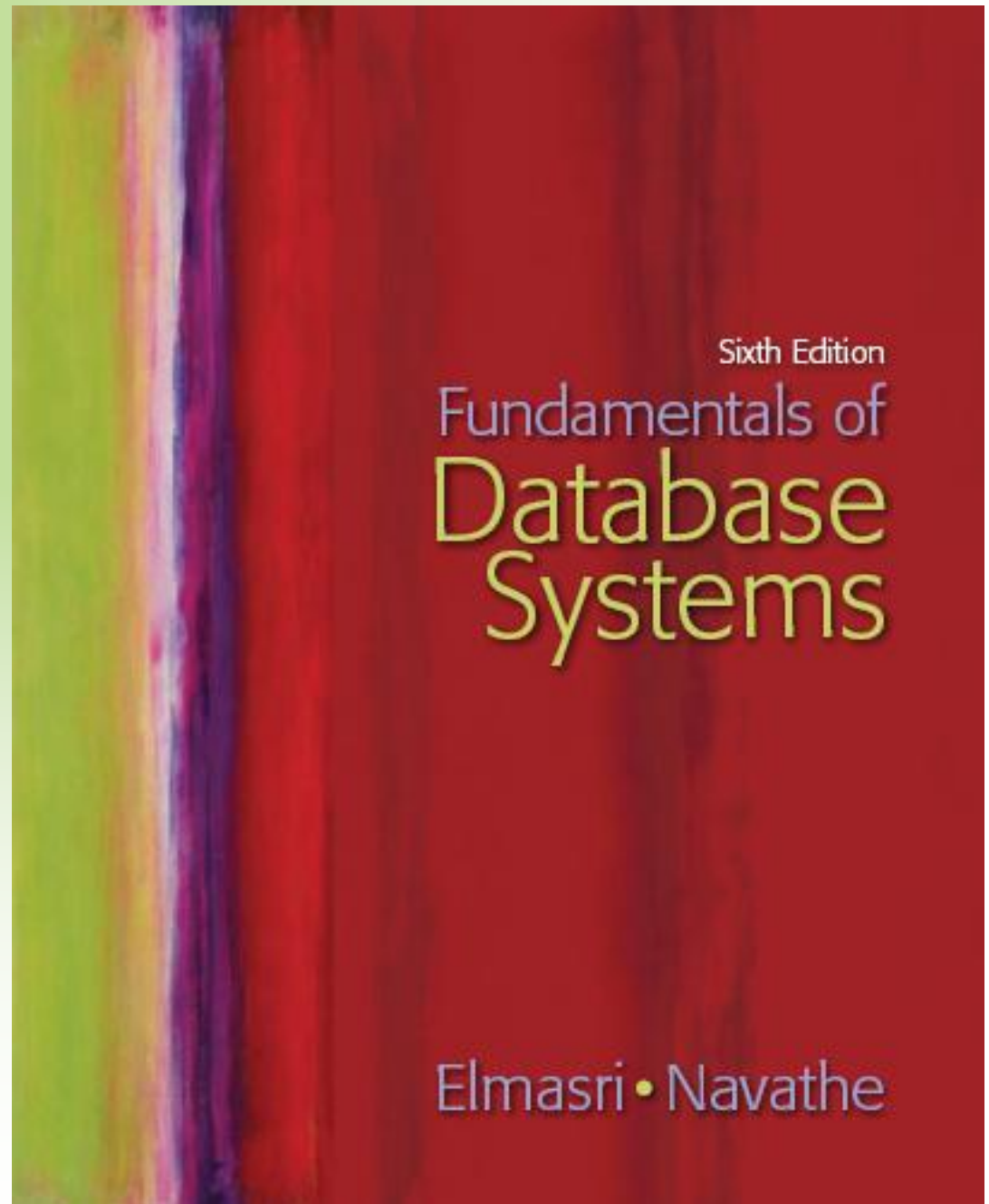


# Chapter 27

## Introduction to Information Retrieval and Web Search



Addison-Wesley  
is an imprint of

PEARSON

# Chapter 27 Outline

- Information Retrieval (IR) Concepts
- Retrieval Models
- Types of Queries in IR Systems
- Text Preprocessing
- Inverted Indexing

# Chapter 27 Outline (cont'd.)

- Evaluation Measures of Search Relevance
- Web Search and Analysis
- Trends in Information Retrieval

# Information Retrieval (IR) Concepts

- **Information retrieval**
  - Process of retrieving documents from a collection in response to a query by a user
- **Introduction to information retrieval**
  - What is the distinction between structured and unstructured data?
  - Information retrieval defined
    - “Discipline that deals with the structure, analysis, organization, storage, searching, and retrieval of information”

# Information Retrieval (IR)

## Concepts (cont'd.)

- User's information need expressed as a **free-form search request**
  - **Keyword search query**
  - **Query**
- IR systems characterized by:
  - Types of users
  - Types of data
  - Types of information needed
  - Levels of scale

# Information Retrieval (IR)

## Concepts (cont'd.)

- High noise-to-signal ratio
- **Enterprise search systems**
  - IR solutions for searching different entities in an enterprise's intranet
- **Desktop search engines**
  - Retrieve files, folders, and different kinds of entities stored on the computer

# Databases and IR Systems: A Comparison

**Table 27.1** A Comparison of Databases and IR Systems

---

## Databases

- Structured data
- Schema driven
- Relational (or object, hierarchical, and network) model is predominant
- Structured query model
- Rich metadata operations
- Query returns data
- Results are based on exact matching (always correct)

## IR Systems

- Unstructured data
  - No fixed schema; various data models (e.g., vector space model)
  - Free-form query models
  - Rich data operations
  - Search request returns list or pointers to documents
  - Results are based on approximate matching and measures of effectiveness (may be imprecise and ranked)
-

# Brief History of IR

- Inverted file organization
  - Based on keywords and their weights
  - SMART system in 1960s
- Text Retrieval Conference (TREC)
- **Search engine**
  - Application of information retrieval to large-scale document collections
  - **Crawler**
    - Responsible for discovering, analyzing, and indexing new documents



# Modes of Interaction in IR Systems

- **Query**
  - Set of terms
    - Used by searcher to specify information need
- Main modes of interaction with IR systems:
  - **Retrieval**
    - Extraction of information from a repository of documents through an IR query
  - **Browsing**
    - User visiting or navigating through similar or related documents

# Modes of Interaction in IR Systems (cont'd.)

- **Hyperlinks**

- Used to interconnect Web pages
- Mainly used for browsing

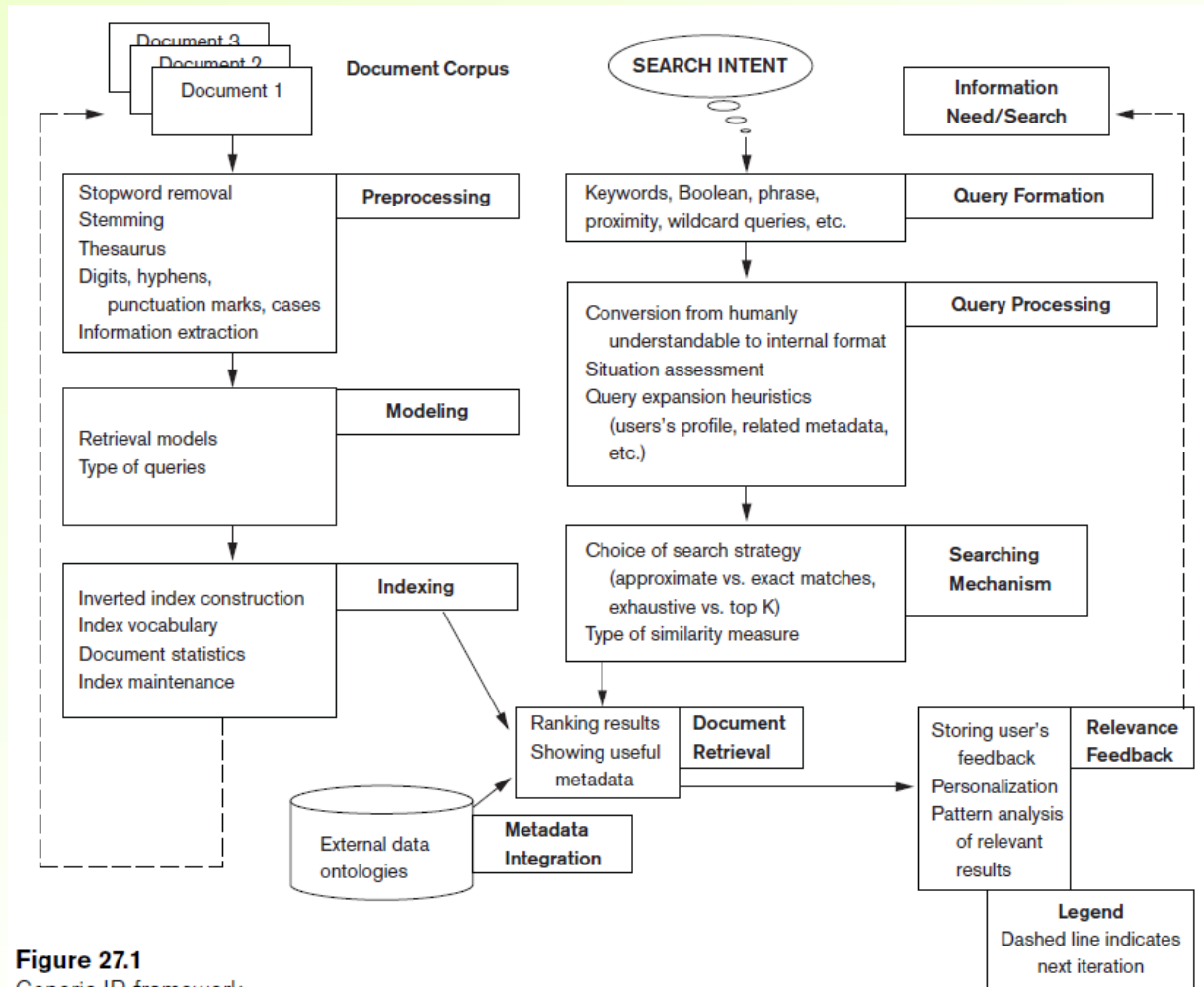
- **Anchor texts**

- Text phrases within documents used to label hyperlinks
- Very relevant to browsing

# Modes of Interaction in IR Systems (cont'd.)

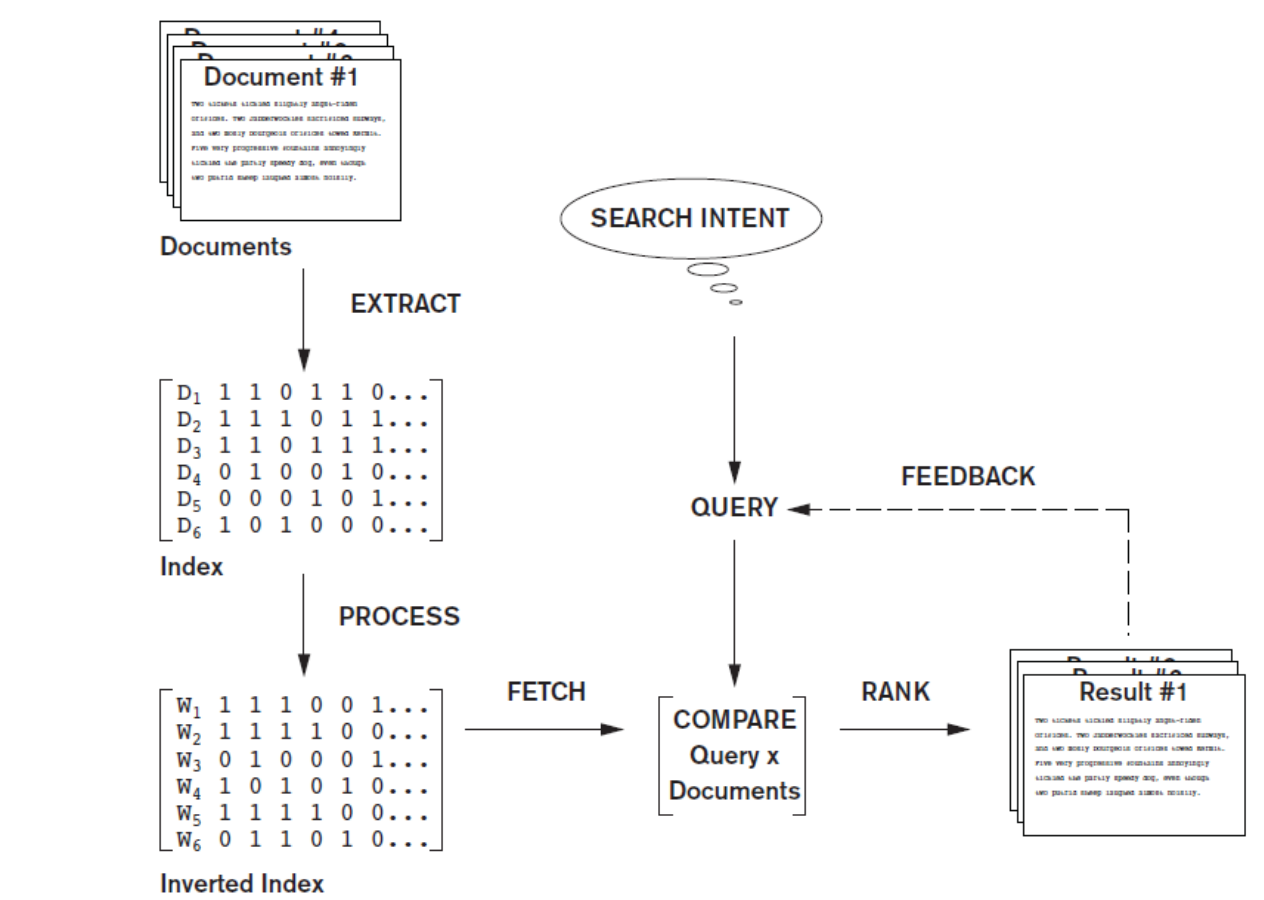
- **Web search**
  - Combines browsing and retrieval
- **Rank of a Webpage**
  - Measure of relevance to query that generated result set

# Generic IR Pipeline



**Figure 27.1**  
Generic IR framework.

# Generic IR Pipeline (cont'd.)



**Figure 27.2**  
Simplified IR process pipeline.

# Retrieval Models

- Three main statistical models
  - Boolean
  - Vector space
  - Probabilistic
- Semantic model

# Boolean Model

- Documents represented as a set of terms
- Form queries using standard Boolean logic set-theoretic operators
  - AND, OR and NOT
- Retrieval and relevance
  - Binary concepts
- Lacks sophisticated ranking algorithms

# Vector Space Model

- Documents
  - Represented as features and weights in an  $n$ -dimensional vector space
- Query
  - Specified as a terms vector
  - Compared to the document vectors for similarity/relevance assessment



# Vector Space Model (cont'd.)

- Different similarity functions can be used
  - Cosine of the angle between the query and document vector commonly used
- **TF-IDF**
  - Statistical weight measure
  - Used to evaluate the importance of a document word in a collection of documents
- Rocchio algorithm
  - Well-known relevance feedback algorithm

# Probabilistic Model

- Probability ranking principle
  - Decide whether the document belongs to the **relevant** set or the **nonrelevant** set for a query
- Conditional probabilities calculated using Bayes' Rule
- **BM25** (Best Match 25)
  - Popular probabilistic ranking algorithm
- **Okapi** system

# Semantic Model

- Include different levels of analysis
  - **Morphological**
  - **Syntactic**
  - **Semantic**
- Knowledge-based IR systems
  - Based on semantic models
  - Cyc knowledge base
  - WordNet

# Types of Queries in IR Systems

- **Keywords**
  - Consist of words, phrases, and other characterizations of documents
  - Used by IR system to build inverted index
- **Queries compared to set of index keywords**
- **Most IR systems**
  - Allow use of Boolean and other operators to build a complex query

# Keyword Queries

- Simplest and most commonly used forms of IR queries
- Keywords implicitly connected by a logical AND operator
- Remove **stopwords**
  - Most commonly occurring words
    - a, the, of
- IR systems do not pay attention to the ordering of these words in the query

# Boolean Queries

- AND: both terms must be found
- OR: either term found
- NOT: record containing keyword omitted
- ( ): used for nesting
- +: equivalent to and
- – Boolean operators: equivalent to AND NOT
- Document retrieved if query logically true as exact match in document

# Phrase Queries

- Phrases encoded in inverted index or implemented differently
- Phrase generally enclosed within double quotes
- More restricted and specific version of proximity searching

# Proximity Queries

- Accounts for how close within a record multiple terms should be to each other
- Common option requires terms to be in the exact order
- Various operator names
  - NEAR, ADJ(adjacent), or AFTER
- Computationally expensive



# Wildcard Queries

- Support regular expressions and pattern matching-based searching
  - ‘Data\*’ would retrieve data, database, datapoint, dataset
- Involves preprocessing overhead
- Not considered worth the cost by many Web search engines today
- Retrieval models do not directly provide support for this query type

# Natural Language Queries

- Few natural language search engines
- Active area of research
- Easier to answer questions
  - Definition and factoid questions

# Text Preprocessing

- Commonly used text preprocessing techniques
- Part of text processing task

# Stopword Removal

- **Stopwords**
  - Very commonly used words in a language
  - Expected to occur in 80 percent or more of the documents
  - the, of, to, a, and, in, said, for, that, was, on, he, is, with, at, by, and it
- Removal must be performed before indexing
- Queries can be preprocessed for stopword removal

# Stemming

- **Stem**
  - Word obtained after trimming the suffix and prefix of an original word
- Reduces different forms of the word formed by inflection
- Most famous stemming algorithm:
  - Martin Porter's stemming algorithm

# Utilizing a Thesaurus

## ■ Thesaurus

- Precompiled list of important concepts and the main word that describes each
- Synonym converted to its matching concept during preprocessing
- Examples:
  - **UMLS**
    - Large biomedical thesaurus of concepts/meta concepts/relationships
  - **WordNet**
    - Manually constructed thesaurus that groups words into strict synonym sets

# Other Preprocessing Steps: Digits, Hyphens, Punctuation Marks, Cases

- Digits, dates, phone numbers, e-mail addresses, and URLs may or may not be removed during preprocessing
- Hyphens and punctuation marks
  - May be handled in different ways
- Most information retrieval systems perform case-insensitive search
- Text preprocessing steps language specific

# Information Extraction

- Generic term
- Extracting structured content from text
- Examples of IE tasks
- Mostly used to identify contextually relevant features that involve text analysis, matching, and categorization



# Inverted Indexing

- Vocabulary
  - Set of distinct query terms in the document set
- **Inverted index**
  - Data structure that attaches distinct terms with a list of all documents that contains term
- Steps involved in inverted index construction

### 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 2

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.

# Evaluation Measures of Search Relevance

- **Topical relevance**
  - Measures extent to which topic of a result matches topic of query
- **User relevance**
  - Describes “goodness” of a retrieved result with regard to user’s information need
- **Web information retrieval**
  - Must evaluate document ranking order

# Recall and Precision

- **Recall**

- Number of relevant documents retrieved by a search / Total number of existing relevant documents

- **Precision**

- Number of relevant documents retrieved by a search / Total number of documents retrieved by that search

# Recall and Precision (cont'd.)

- Average precision
  - Useful for computing a single precision value to compare different retrieval algorithms
- Recall/precision curve
  - Usually has a negative slope indicating inverse relationship between precision and recall
- F-score
  - Single measure that combines precision and recall to compare different result sets

# Web Search and Analysis

- **Vertical search engines**
  - Topic-specific search engines
- **Metasearch engines**
  - Query different search engines simultaneously
- **Digital libraries**
  - Collections of electronic resources and services

# Web Analysis and Its Relationship to IR

- Goals of Web analysis:
  - Improve and personalize search results relevance
  - Identify trends
- Classify Web analysis:
  - **Web content analysis**
  - **Web structure analysis**
  - **Web usage analysis**

# Searching the Web

- **Hyperlink** components
  - **Destination page**
  - **Anchor text**
- **Hub**
  - Web page or a Website that links to a collection of prominent sites (**authorities**) on a common topic



# Analyzing the Link Structure of Web Pages

- The **PageRank** ranking algorithm
  - Used by Google
  - Highly linked pages are more important (have greater authority) than pages with fewer links
  - Measure of query-independent importance of a page/node
- **HITS** Ranking Algorithm
  - Contains two main steps: a sampling component and a weight-propagation component

# Web Content Analysis

- Structured data extraction
  - Several approaches: writing a **wrapper**, manual extraction, **wrapper induction**, **wrapper generation**
- Web information integration
  - **Web query interface integration** and **schema matching**
- Ontology-based information integration
  - **Single**, **multiple**, and **hybrid**

# Web Content Analysis (cont'd.)

- Building **concept hierarchies**
  - Documents in a search result are organized into groups in a hierarchical fashion
- Segmenting Web pages and detecting noise
  - Eliminate superfluous information such as ads and navigation

# Approaches to Web Content Analysis

- Agent-based approach categories
  - **Intelligent Web agents**
  - **Information filtering/categorization**
  - **Personalized Web agents**
- Database-based approach
  - Infer the structure of the Website or to transform a Web site to organize it as a database

# Web Usage Analysis

- Typically consists of three main phases:
  - Preprocessing, pattern discovery, and pattern analysis
- Pattern discovery techniques:
  - Statistical analysis
  - Association rules
  - **Clustering of users**
    - Establish groups of users exhibiting similar browsing patterns

# Web Usage Analysis (cont'd.)

- **Clustering of pages**
  - Pages with similar contents are grouped together
- Sequential patterns
- Dependency modeling
- Pattern modeling

# 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 results returned by search engines
- **Web security**
- Alternate uses for **Web crawlers**

# Trends in Information Retrieval

## ■ Faceted search

- Allows users to explore by filtering available information
- **Facet**
  - Defines properties or characteristics of a class of objects

## ■ Social search

- New phenomenon facilitated by recent Web technologies: **collaborative social search, guided participation**



# Trends in Information Retrieval (cont'd.)

- **Conversational search (CS)**
  - Interactive and collaborative information finding interaction
  - Aided by intelligent agents

# Summary

- IR introduction
  - Basic terminology, query and browsing modes, semantics, retrieval modes
- Web search analysis
  - Content, structure, usage
  - Algorithms
  - Current trends