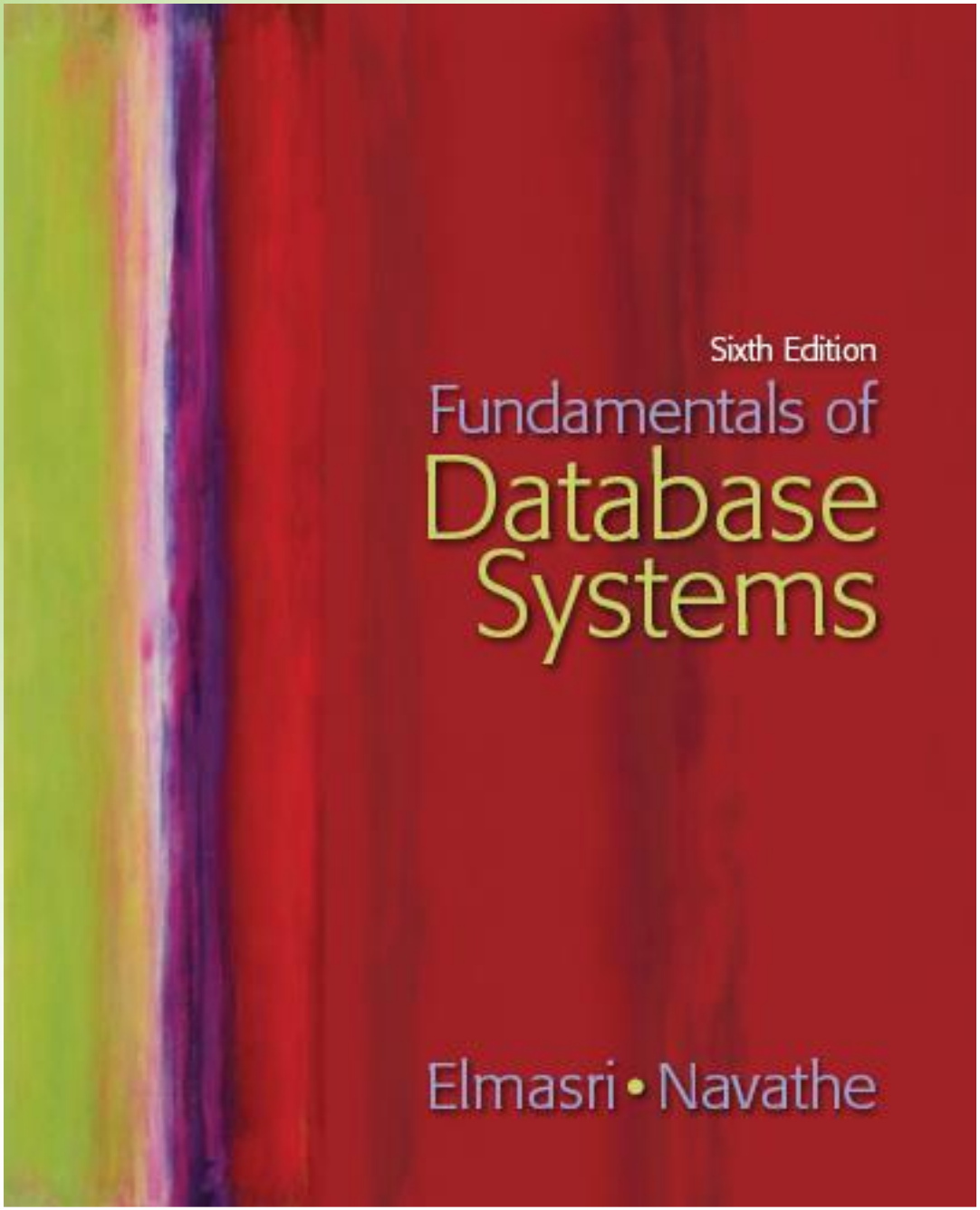
Chapter 27

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
    - Expert users vs. layperson users
  - Types of data
    - Search systems can be tailored to specific types of data for performance purposes (e.g. a customized search system for a specific topic)
  - Types of information needed
    - Navigational search vs. Informational search vs. Transactional search
  - 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	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)	

Brief History of IR

- Inverted file organization
  - Based on keywords and their weights
  - SMART system in 1960s
- Text Retrieval Conference (TREC)
  - Launched by National Institute of Standard and Technology (NIST) in 1992
- **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 (also referred to as keywords)
    - 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

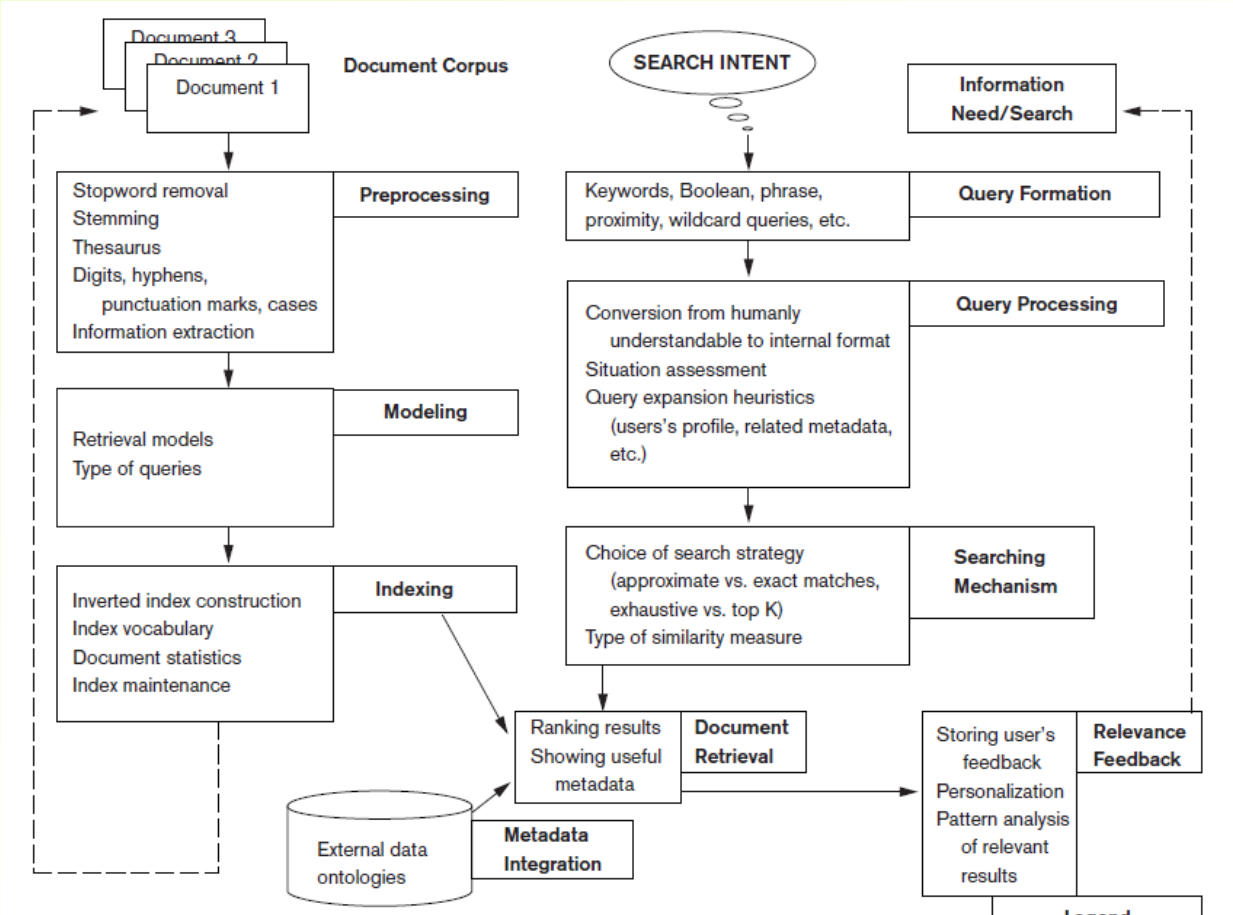


Figure 27.1
Generic IR framework.

Generic IR Pipeline (cont'd.)

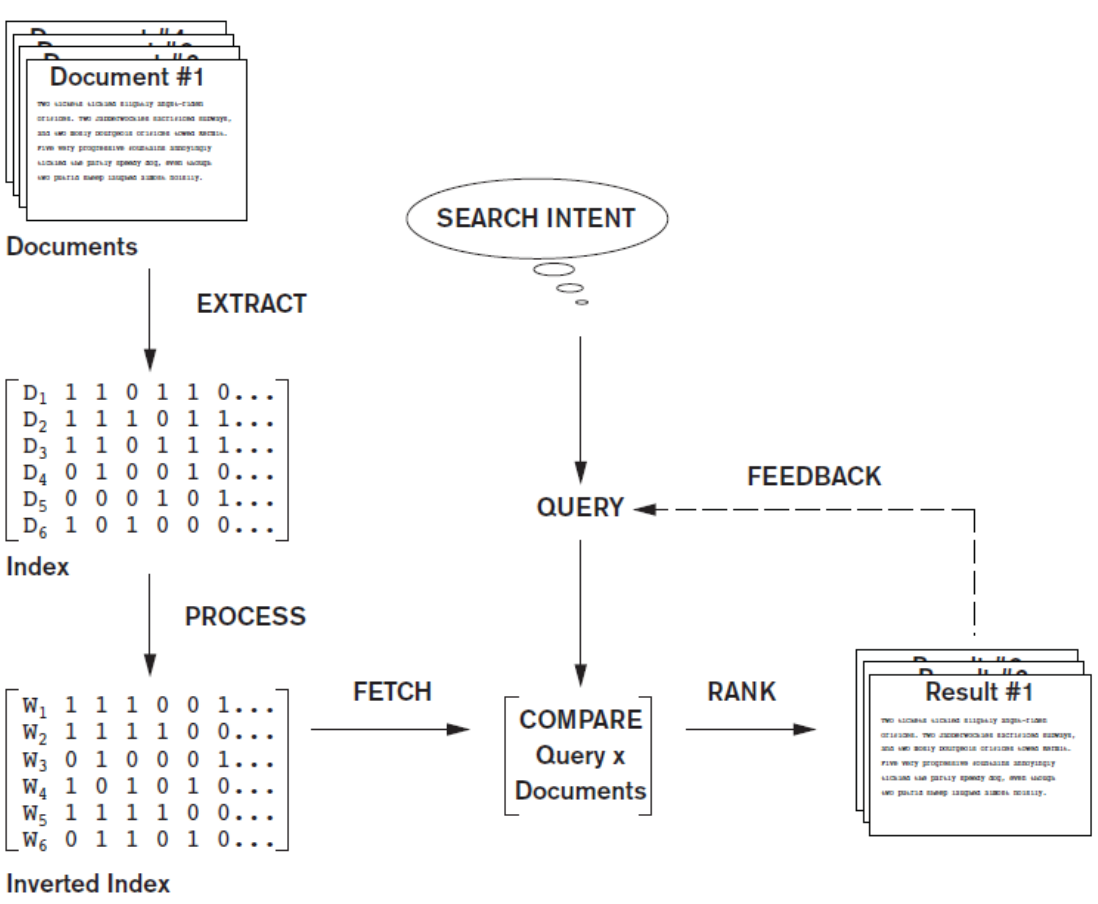


Figure 27.2
Simplified IR process pipeline.

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
  - Features are a subset of the terms in a set of documents that are deemed most relevant to an IR search for this particular set of documents.
- 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 (Term Frequency-Inverse Document Frequency)**
  - 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
  - Assume a predefined relevant set and nonrelevant set exist for the query
  - Calculate the probability that the document D belongs to the relevant set (R) and compare that with the probability that the document D belongs to the nonrelevant set (NR)
  - The likelihood ratio P(D|R)/P(D/NR) of a document D is used for ranking D (assumption: a highly ranked document will have a high likelihood of belonging to the relevant set)

Semantic Model

- Include different levels of analysis
  - **Morphological**
    - **Determines the part of the speech (nouns, verbs, adjectives, etc.) of the words**
  - **Syntactic**
    - **Parses and analyzes complete phrases in documents**
  - **Semantic**
    - **Resolves word ambiguities and/or generates relevant synonyms based on the semantic relationships between levels of structural entities in documents (words, paragraphs, pages, or entire documents)**
- Knowledge-based IR systems
  - Based on semantic models
  - Cyc knowledge base
    - Commonsense knowledge about assertions (over 2.5 million facts and rules) interrelating more than 155,000 concepts for reasoning about objects and events of everyday life.
  - WordNet
    - Extensive thesaurus (over 115,000 concepts)

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 (with relative positions of word occurrences in documents)
- 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
  - Example: “comput” is the stem word for computer, computing and computation
- Reduces different forms of the word formed by inflection (due to plurals or tenses)
- A stemming algorithm can be applied to reduce a word to its stem.
- Most famous stemming algorithm:
  - Martin Porter’s stemming algorithm
- Using stemming for preprocessing data results in a decrease in the size of the indexing structure and an increase in recall, possibly at the cost of precision.

Utilizing a Thesaurus

- **Thesaurus**
  - Precompiled list of important concepts and the main word that describes each concept for a particular domain of knowledge
  - For each concept in the list, a set of synonyms and related words is also compiled
  - Synonym can be 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 are language specific (such as involving accents and diacritics)

Information Extraction (IE)

- A generic term used for extracting structured content from text
- Examples of IE tasks
  - Identifying noun phrases, facts, events, people, places, and relationships
- Mostly used to identify contextually relevant features that involve text analysis, matching, and categorization for improving the relevance of search systems

Inverted Indexing

- **Vocabulary**
  - Set of distinct query terms in the document set
  - The simplest form of vocabulary terms consists of words or individual tokens of the documents
  - In some cases, the vocabulary terms also consist of phrases, n-grams, links, names, dates or manually assigned descriptor terms from documents and/or Web pages
- **Inverted index**
  - Data structure that attaches distinct terms with a list of all documents that contains the terms
- **Steps involved in inverted index construction**
  - Brake the documents into vocabulary terms by tokenizing, cleansing, stopword removal, stemming, and/or use of an additional thesaurus as vocabulary
  - Collect document statistics (e.g counts of vocabulary terms in individual documents as well as different collections, their positions of occurrence within the documents, and the lengths of the documents) and store the statistics in a document lookup table
  - Invert the document-term stream into a term-document stream along with 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 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 (r)**
  - Number of relevant documents retrieved by a search / Total number of existing relevant documents
- **Precision (p)**
  - 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 (or F_1-score)
  - Single measure that combines precision and recall to compare different result sets
  - $F = 2pr/(p+r)$

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**
    - Deals with extracting useful info/knowledge from Web page contents
  - **Web structure analysis**
    - Discovers knowledge from hyperlinks representing the structure of the Web
  - **Web usage analysis**
    - Mines user access patterns from usage logs

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
  - Developed by Google's co-founders, Larry Page and Sergey Brin in 1998
  - Used by Google
  - Highly linked pages are more important (have greater authority) than pages with fewer links
  - Not all backlinks (inbound links) are important: a backlink to a page from a credible source is more important than a link from some arbitrary page => A page has a high rank if the sum of the ranks of its backlinks is high
  - PageRank of page A:
    - *PR(A) = (1-d)/N + d * (PR(T1)/C(T1)+…+PR(Tn)/C(Tn))*
  - PR(X) is PageRank of page X, d is a damping factor in the range 0<d<1 (usually d is set to 0.85); N is the total number of pages in the collection; T1,…, Tn are the pages that point to page A, and C(X) is number of forward links (outbound links) from page X.
  - PageRank is a measure of query-independent importance of a page
  - PageRank forms a probability distribution over Web pages, so the sum of all Web pages' PageRanks is 1.

Web Content Analysis

- Structured data extraction
  - Example: a structured table on the Web showing the airline flight schedules
  - Several approaches:
    - writing a **wrapper:** a program that looks for different structural characteristics of the info on the page and extracts the right content.
    - **manual extraction**: manually write an extraction program for each Website based on the observed format patterns of the site
    - **wrapper induction/learning**: user first manually labels a set of training Web pages, and the learning system generate rules based on the training Web pages to extract target items from other Web pages
    - **wrapper generation:** finds patterns/grammars from the Web pages and then uses wrapper generation to produce a wrapper to extract data automatically
- Web information integration
  - **Web query interface integration:** enabling querying multiple Web databases that are not visible to external interfaces and are hidden in the "deep Web"
  - **Schema matching:** integrating directories and catalogs to come up with a global schema for applications (e.g. combine a personal health record of an individual by matching and collecting data from multiple sources)
- Ontology-based information integration
  - **Single ontology**: uses one global ontology that provides a shared vocabulary for the specifications of the semantics
  - **Multiple ontology:** each info source is described by its own ontology
  - **Hybrid ontology:** each info source is described by its own ontology, but to make the source ontologies comparable to each other, they are built on one global shared vocabulary.

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**
    - Software agents that search for relevant info using characteristics of a particular app domain to organize and interpret the discovered info.
  - **Information filtering/categorization**
    - Web agents that use methods from IR, and semantic info based on the links among various documents to organize documents into a concept hierarchy
  - **Personalized Web agents**
    - Web agents that utilize the personal preferences of users to organize search results, or to discover info and documents that could be of value for a particular user.
- Database-based approach
  - Infer the structure of the Website or transform a Web site to organize it as a database so that better info management and querying on the Web become possible

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
  - Patterns that identify sequences of Web accesses
- Dependency modeling
  - Model significant dependencies among the various variables in the Web domain (e.g. building a model representing the different stages a visitor undergoes while shopping in an online store based on the actions chosen)
- Pattern analysis
  - Filter out those rules or patterns that are considered to be not of interest from the discovered patterns

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

- **Social search**
  - User cooperation (co-located or remotely) during Web-based search
- **Conversational search (CS)**
  - Interactive and collaborative information finding interaction
  - Participants engage in a conversation and perform a social search activity aided by intelligent agents
  - The collaborative search helps the agent learn about conversations with interactions and feedback from participants.
  - The agent uses the semantic retrieval model with natural language understanding to provide the users with faster and relevant search results.

Summary

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