Information Retrieval and Web Search

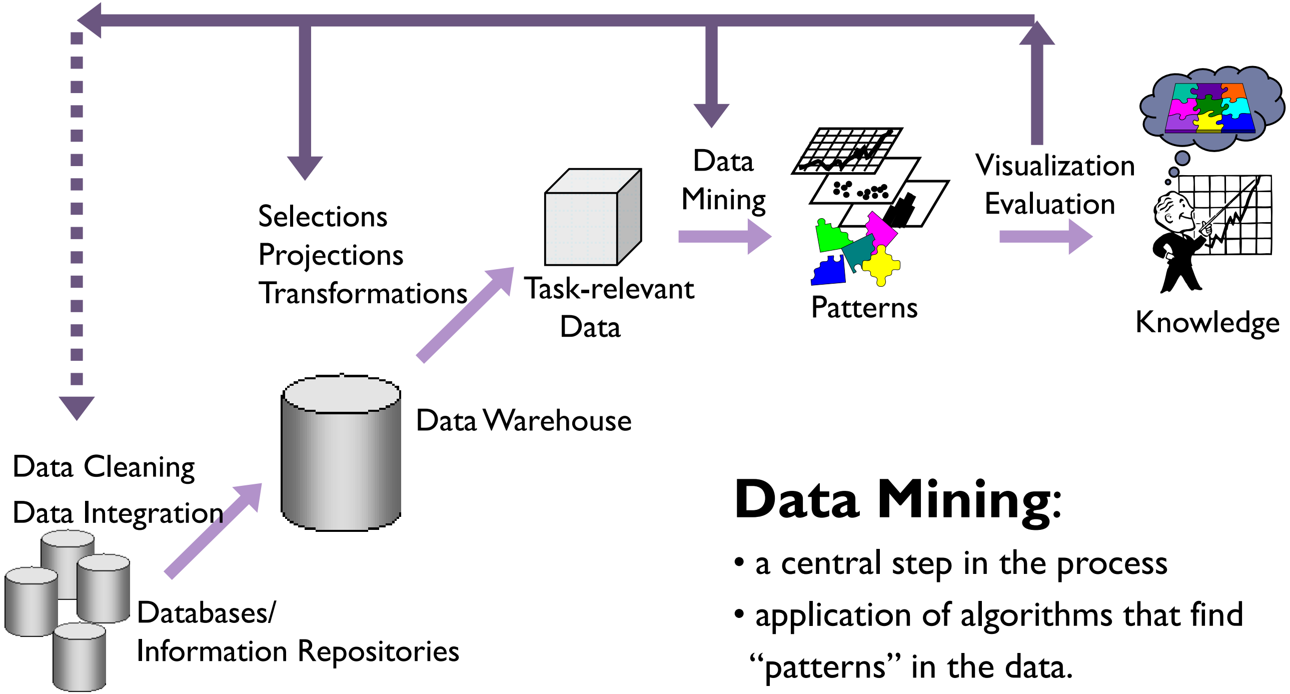
Implementation and Applications of Databases, Spring 2019

Ira Assent

Intended learning outcomes

- Be able to
  - Describe and apply the main information retrieval and web search approaches
  - Evaluate retrieval results

Data Mining recap

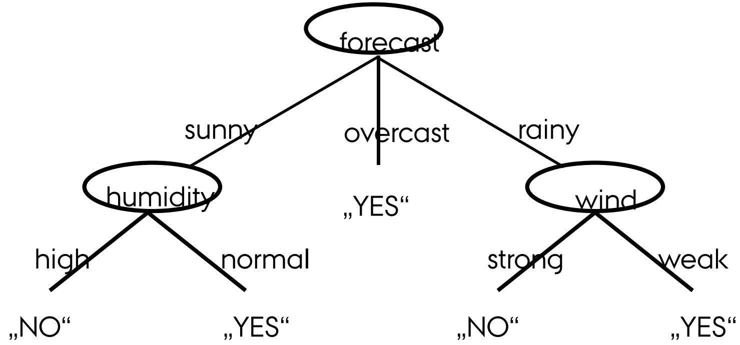


Data Mining:

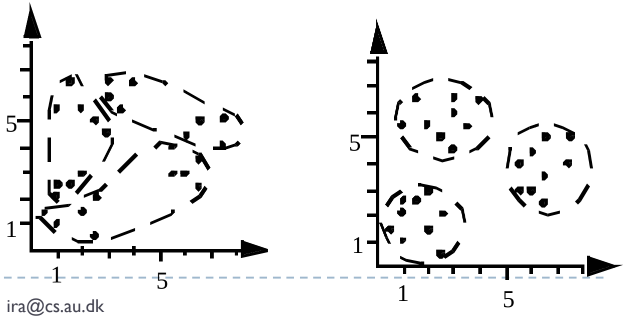
- a central step in the process
- application of algorithms that find "patterns" in the data.

Types of Discovered Knowledge

- Association Rules "Body ⇒ Head [support, confidence]"
  - buys(x, "diapers") ⇒ buys(x, "beers") [0.5%, 60%]
- Classification Models and Predictions



- Clustering: groups of related objects
- …
- Applications
  - Marketing
  - Finance
  - Manufacturing
  - Health
  - …



How many candidates?

- Assume we have 1000 frequent 1-itemsets. How many candidate 2-itemsets does the apriori algorithm generate?

A. 1000

B. 2000

C. 5000

D. 50,000

E. 500,000

Information Retrieval (IR) Concepts

- **Information retrieval**
  - Process of retrieving documents from a collection in response to a query by a user
  - Distinction between structured and unstructured data
    - DBMS highly structured: e.g. relational schema
    - E.g. text documents unstructured: e.g. blog entry
- User's information need expressed as a **free-form search request** (unstructured), also called **keyword search query**
- IR systems characterized by:
  - Types of users
  - Types of data
  - Types of information needed
  - Levels of scale
- 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

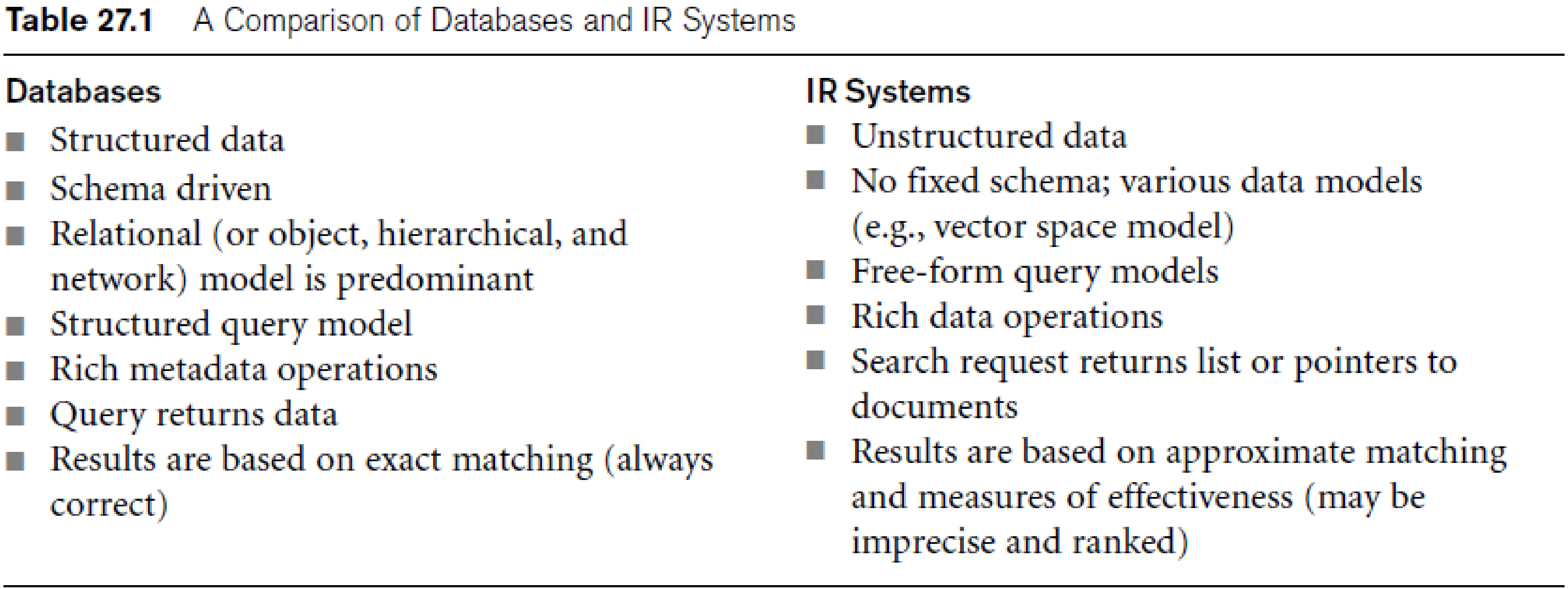


Table 27.1 A Comparison of Databases and IR Systems

Databases	**IR Systems**
■ Structured data ■ 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)	■ 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)

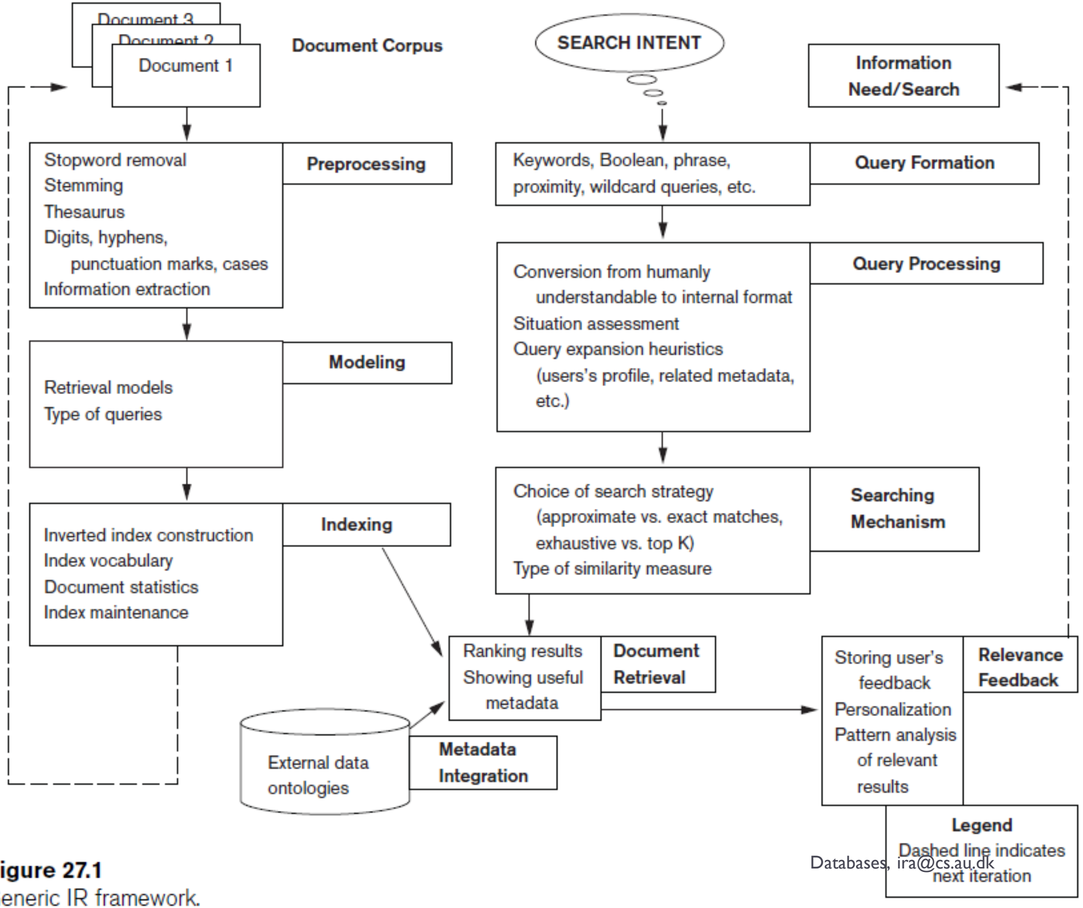
Brief History of IR

- Inverted file organization
  - Based on keywords and their weights
  - SMART system in 1960s
- Text Retrieval Conference (TREC) since 1992
- **Search engine**
  - **Rapid growth in search engine technology due to the Web**
  - 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
- **Hyperlinks**
  - Used to interconnect Web pages
  - Mainly used for browsing
- **Anchor texts**
  - Text phrases within documents used to label hyperlinks
  - Very relevant to browsing
- **Web search**
  - Combines browsing and retrieval
- **Rank of a Webpage**
  - Measure of relevance to query that generated result set

Generic IR Pipeline



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

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

Metadata Integration

External data ontologies

Relevance Feedback

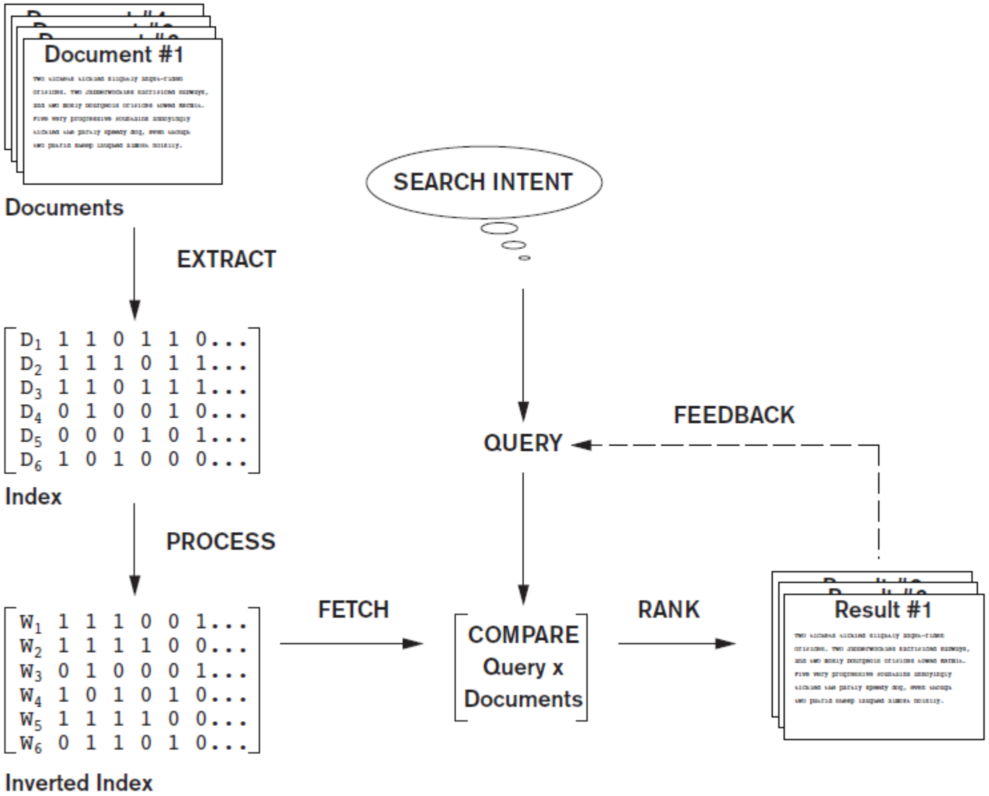
Storing user's feedback
Personalization
Pattern analysis of relevant results

Legend

Dashed line indicates next iteration

Figure 27.1
Generic IR framework.

IR Process Pipeline



Document #1

Two tickets tickled slightly angst-ridden orifices. 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

$$\begin{bmatrix} D_1 & 1 & 1 & 0 & 1 & 1 & 0\dots \\ D_2 & 1 & 1 & 1 & 0 & 1 & 1\dots \\ D_3 & 1 & 1 & 0 & 1 & 1 & 1\dots \\ D_4 & 0 & 1 & 0 & 0 & 1 & 0\dots \\ D_5 & 0 & 0 & 0 & 1 & 0 & 1\dots \\ D_6 & 1 & 0 & 1 & 0 & 0 & 0\dots \end{bmatrix}$$

Index

PROCESS

$$\begin{bmatrix} W_1 & 1 & 1 & 1 & 0 & 0 & 1\dots \\ W_2 & 1 & 1 & 1 & 1 & 0 & 0\dots \\ W_3 & 0 & 1 & 0 & 0 & 0 & 1\dots \\ W_4 & 1 & 0 & 1 & 0 & 1 & 0\dots \\ W_5 & 1 & 1 & 1 & 1 & 0 & 0\dots \\ W_6 & 0 & 1 & 1 & 0 & 1 & 0\dots \end{bmatrix}$$

Inverted Index

SEARCH INTENT

QUERY

FEEDBACK

FETCH

COMPARE Query x Documents

RANK

Result #1

Two tickets tickled slightly angst-ridden orifices. 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.

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
  - E.g. tags / metadata to describe news articles
- Form queries using standard Boolean logic set-theoretic operators
  - AND, OR and NOT
- Retrieval and relevance
  - Binary concepts
- Lacks sophisticated ranking algorithms
  - All "matching" documents considered equally important
    - Are all documents tagged with "databases" equally important?

Vector Space Model

- Documents
  - Represented as features and weights in an *n*-dimensional vector space, where each term represents one dimension (VERY high-dimensional)
- Query
  - Specified as a terms vector
  - Compared to the document vectors for similarity/relevance assessment
- 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

Given the query term ”data” and four documents…

…which one is the most relevant?

- A long one with one term occurrence ”data”
- A long one with three occurrences ”databases”, ”datalogi”, ”dada”
- A short one with one term occurrence ”data”
- A short one with two occurrences ”databases”, ”datalogi”

TF-IDF

TF-IDF (Term frequency * Inverse Document Frequency)

f_{ij} = frequency of term ***i*** in document ***j***

$$TF_{ij} = \frac{f_{ij}}{\max_k f_{kj}}$$

Note: we normalize TF to discount for "longer" documents (other variants for normalization exist e.g. textbook uses sum)

n_i = number of docs that mention term ***i***

N = total number of docs

$$IDF_i = \log \frac{N}{n_i}$$

TF-IDF score: $w_{ij} = TF_{ij} \times IDF_i$

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
  - Builds a score based on frequency of term in document, frequency of term in query, total number of documents, total number of documents that contain the term, document length, and average document length

Semantic Model

- Idea: capture "meaning" of words, instead of only counting terms
- Include different levels of analysis
  - **Morphological**
    - **Parts of speech (noun, adjective, verb,…)**
  - **Syntactic**
    - **Phrases of text**
  - **Semantic**
    - **Resolve word ambiguities, identify synonyms,…**
- 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 and Boolean Queries

- Simplest and most commonly used forms of IR queries
- Keywords implicitly connected by a logical AND operator
- Remove **stopwords**
  - Most commonly occurring words such as the, of, to, a, and, in, said, for, that, was, on, he, is, with, at, by, it
    - Very commonly used words in a language (expected to occur in >80% of documents)
  - Removal must be performed before indexing
- 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
  - Document retrieved if query logically true as exact match in document

Phrase Queries and Proximity Queries

- Phrases often encoded in inverted index
- Phrase generally enclosed within double quotes
  - “President of the United States”,“make my day”
- 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 and Natural Language Queries

- 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

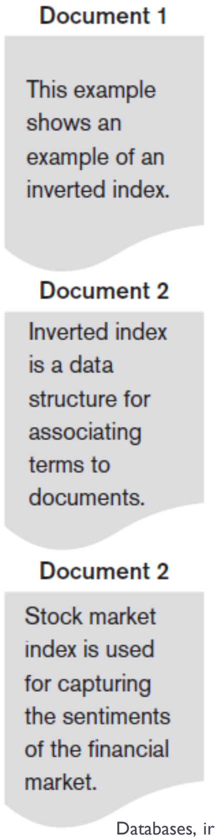
- Stopword removal
- **Stemming**
  - Word obtained after trimming the suffix and prefix of an original word
- Reduces different forms of the word formed by inflection
- E.g. Porter's stemming algorithm
- **Thesaurus**
  - Precompiled list of important concepts and the main word that describes each
  - Synonym converted to its matching concept during preprocessing
  - E.g. **WordNet:** Manually constructed thesaurus that groups words into strict synonym sets
- Digits, dates, phone numbers, e-mail addresses, and URLs may or may not be removed
- 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 unstructured text
  - E.g. find facts, events, people, relationships
  - Named entity recognition taks
  - Rule-based approaches
- Mostly used to identify contextually relevant features that involve text analysis, matching, and categorization
  - May improve the relevance of search results

Indexing occurrences of terms

- Vocabulary
  - Set of distinct query terms in the document set
- **Inverted index**
  - attaches distinct terms with a list of all documents that contains term
    - Seen before, most recently in connection with MapReduce
- Steps involved in inverted index construction
  1. break the documents into vocabulary terms by tokenizing, cleaning, stopword removal, stemming, …
  2. derive frequency counts and other statistics
  3. invert document-term stream into a term-document stream



Document 1

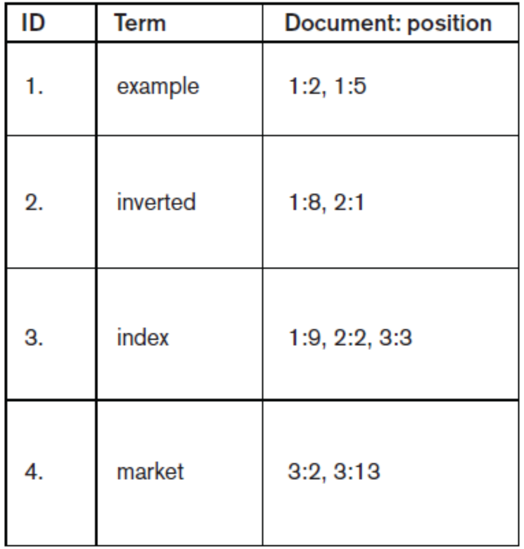
This example shows an example of an inverted index.

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

Document search

1. Vocabulary search
   - Treat query terms independently
   - Typically via index structures such as B+-tree or hashes
2. Retrieve document information for each term
3. Apply query logic
   - E.g. range, context, proximity queries

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

What is best?

A search engine that returns

A. 5 documents, 2 relevant

B. 15 documents, 3 relevant

C. 1 document, 1 relevant

D. 300 documents, 99 relevant

Recall and Precision

- **Recall**
  - Number of relevant documents retrieved by a search (Hits) / Total number of existing relevant documents (TP+FN)
- **Precision**
  - Number of relevant documents retrieved by a search / Total number of documents retrieved by that search (TP+FP)
- 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 (y-axis) and recall (x-axis)
- F-score
  - Single measure that combines precision and recall (harmonic mean) to compare different result sets: 2pr/(p+r)

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

Web Search and Analysis

- **Vertical search engines**
  - Topic-specific search engines
    - Specialized search
      - e.g. Shopzilla - A shopping search engine that allows users to browse retail categories or search for specific items. Merchants pay for inclusion
- **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
    - Sampling component finds small collection of pages that is considered authoritative
    - Weight component assigns scores to documents based on authority

What do you imagine is an issue for PageRank?

- Artificial booster pages with lots of links
- Lots of artificial booster pages with the same link
- Pages linking to the same document are hard to distinguish
- Documents with few links are hard to distinguish

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

Intended learning outcomes

- Be able to
  - Describe and apply the main information retrieval and web search approaches
  - Evaluate retrieval results

What was this all about?

Guidelines for your own review of today's session

- In information retrieval, our aim is…
  - The main difference between unstructured…
- IR systems generally work as follows…
  - The three main retrieval models…
  - TF-IDF is…
- Queries can be specified as…
- Information Extraction is about…
- We evaluate results using…
- Web analysis is related in that it…
- The idea underlying algorithms like PageRank is…