Introduction to Data Retrieval

Virendra Singh

Professor, Indian Institute of Technology Bombay
And
Adjunct Professor, Indian Institute of Technology Jammu
http://www.ee.iitb.ac.in/~viren/
E-mail: viren@ee.iitb.ac.in, virendra.singh@iitjammu.ac.in

CSPL201: Data Organization & Retrieval

Lecture 3 (22 September 2020)

Related Areas

- Database Management
- Library and Information Science
- Artificial Intelligence
- Natural Language Processing
- Machine Learning

Database Management

- Focused on *structured* data stored in relational tables rather than free-form text
- Focused on efficient processing of well-defined queries in a formal language (SQL)
- Clearer semantics for both data and queries
- Recent move towards *semi-structured* data (XML) brings it closer to IR

Library and Information Science

- Focused on the human user aspects of information retrieval (human-computer interaction, user interface, visualization)
- Concerned with effective categorization of human knowledge
- Concerned with citation analysis and *bibliometrics* (structure of information)
- Recent work on *digital libraries* brings it closer to CS & IR

Artificial Intelligence

- Focused on the representation of knowledge, reasoning, and intelligent action
- Formalisms for representing knowledge and queries
  - First-order Predicate Logic
  - Bayesian Networks
- Recent work on web ontologies and intelligent information agents brings it closer to IR

Natural Language Processing

- Focused on the syntactic, semantic, and pragmatic analysis of natural language text and discourse
- Ability to analyze syntax (phrase structure) and semantics could allow retrieval based on *meaning* rather than keywords

Natural Language Processing: IR Directions

- Methods for determining the sense of an ambiguous word based on context (*word sense disambiguation*)
- Methods for identifying specific pieces of information in a document (*information extraction*)
- Methods for answering specific NL questions from document corpora or structured data like FreeBase or Google's Knowledge Graph.

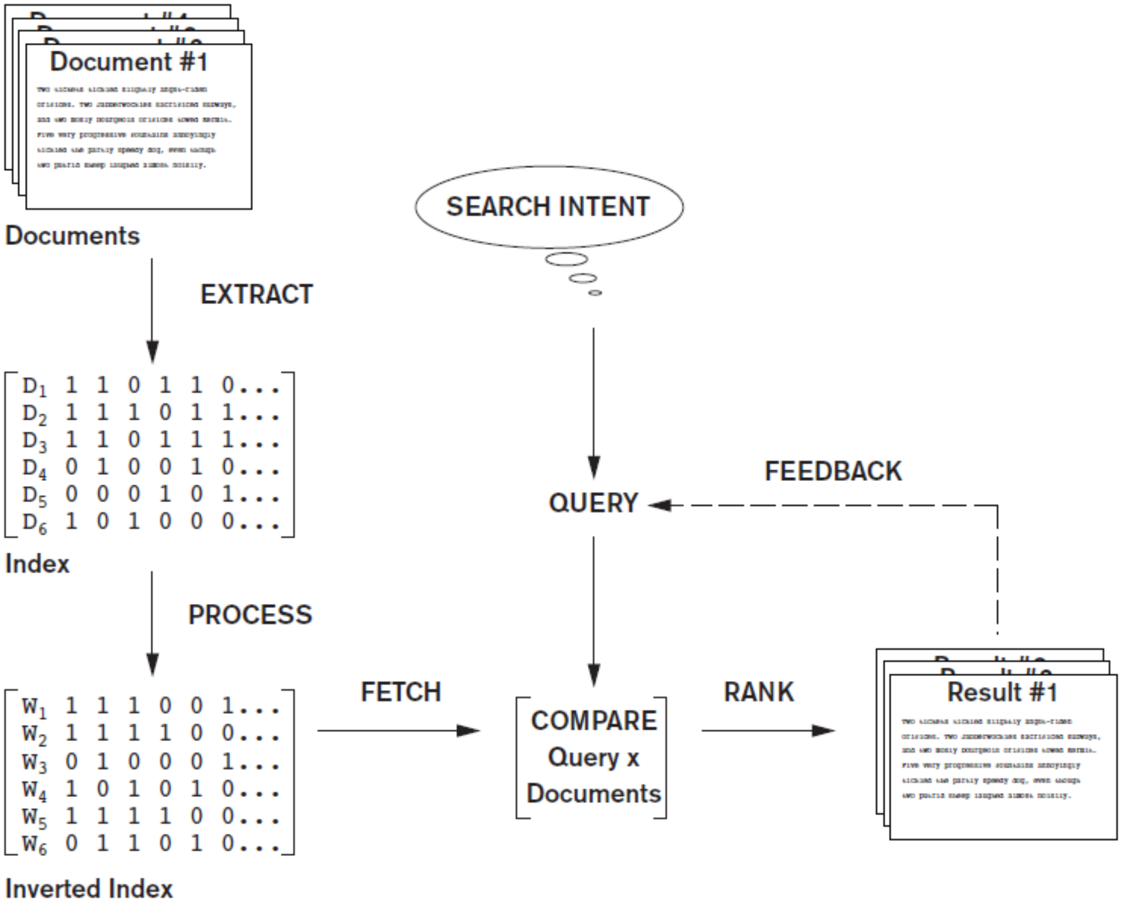
Machine Learning

- Focused on the development of computational systems that improve their performance with experience.
- Automated classification of examples based on learning concepts from labeled training examples (*supervised learning*).
- Automated methods for clustering unlabeled examples into meaningful groups (*unsupervised learning*).

Machine Learning: IR Directions

- Text Categorization
  - Automatic hierarchical classification (Yahoo).
  - Adaptive filtering/routing/recommending.
  - Automated spam filtering.
- Text Clustering
  - Clustering of IR query results.
  - Automatic formation of hierarchies (Yahoo).
- Learning for Information Extraction
- Text Mining
- Learning to Rank

Generic IR Pipeline



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...

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.

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

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
  - 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

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

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**

More Information

http://www.ee.iitb.ac.in/~viren/Courses/2020/DOR.htm

Thank You