Knowledge Discovery and Data Mining

Unit # 17

KNIME DEMO: ASSOCIATION RULES

Mining Text Data

Data Mining / Knowledge Discovery

Structured Data

HomeLoan (
Loanee: *Frank Rizzo*
Lender: *MWF*
Agency: *Lake View*
Amount: *$200,000*
Term: *15 years*
)

Multimedia

Loan($200K,[map],...)

Free Text

Frank Rizzo bought his home from Lake View Real Estate in 1992.
He paid $200,000 under a15-year loan from MW Financial.

Hypertext

<a href>Frank Rizzo Bought <a hef>this home from <a href>Lake View Real Estate In 1992. <p>...

Bag-of-Tokens Approaches

Documents

Four score and seven years ago our fathers brought forth on this continent, a new nation, conceived in Liberty, and dedicated to the proposition that all men are created equal.
Now we are engaged in a great civil war, testing whether that nation, or …

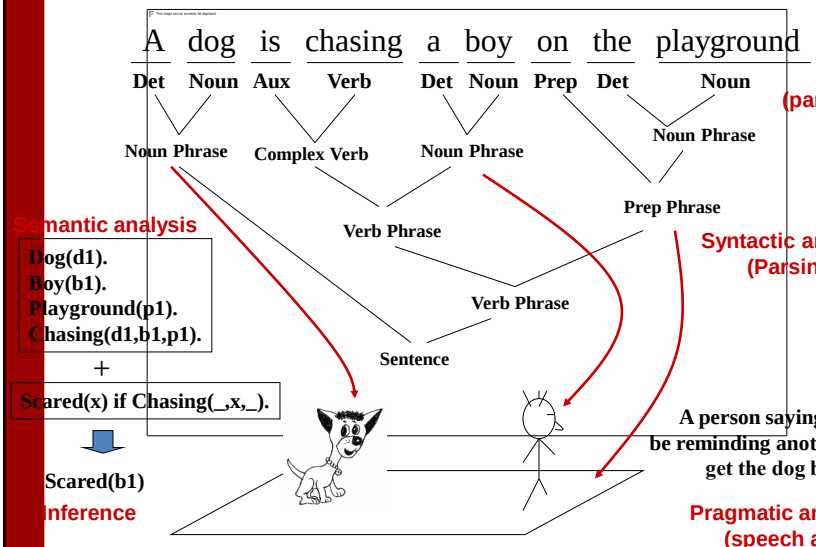
Feature Extraction

Token Sets

nation – 5
civil - 1
war – 2
men – 2
died – 4
people – 5
Liberty – 1
God – 1
…

Loses all order-specific information!
Severely limits context!

Natural Language Processing



Lexical analysis (part-of-speech tagging)

Syntactic analysis (Parsing)

Semantic analysis

Dog(d1).
Boy(b1).
Playground(p1).
Chasing(d1,b1,p1).
+
Scared(x) if Chasing(_,x,_).

Scared(b1)

Inference

A person saying this may be reminding another person to get the dog back…

Pragmatic analysis (speech act)

General NLP—Too Difficult!

- Word-level ambiguity
  - **"design" can be a noun or a verb** (Ambiguous POS)
  - **"root" has multiple meanings** (Ambiguous sense)
- Syntactic ambiguity
  - **"natural language processing"** (Modification)
  - **"A man saw a boy *with a telescope*."** (PP Attachment)
- Anaphora resolution
  - **"John persuaded Bill to buy a TV for *himself*."**
    (*himself* = John or Bill?)
- Presupposition
  - **"He has quit smoking." implies that he smoked before.**

Humans rely on context to interpret (when possible). This context may extend beyond a given document!

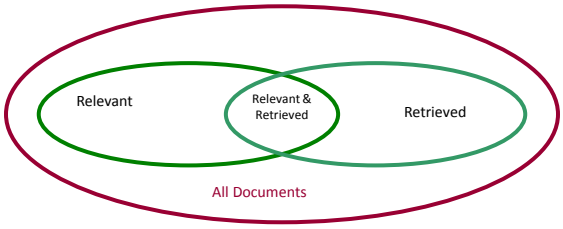
Text Databases and IR

- Text databases (document databases)
  - Large collections of documents from various sources: news articles, research papers, books, digital libraries, e-mail messages, and Web pages, library database, etc.
  - Data stored is usually *semi-structured*
- Information retrieval
  - A field developed in parallel with database systems
  - Information is organized into (a large number of) documents
  - Information retrieval problem: locating relevant documents based on user input, such as keywords or example documents

Information Retrieval

- Typical IR systems
  - Online library catalogs
  - Online document management systems
- Information retrieval vs. database systems
  - Some DB problems are not present in IR, e.g., update, transaction management, complex objects
  - Some IR problems are not addressed well in DBMS, e.g., unstructured documents, approximate search using keywords and relevance

Basic Measures for Text Retrieval



- Precision: the percentage of retrieved documents that are in fact relevant to the query (i.e., "correct" responses)

$$precision = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Retrieved\}|}$$

- Recall: the percentage of documents that are relevant to the query and were, in fact, retrieved

$$recall = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Relevant\}|}$$

Information Retrieval Techniques

- Basic Concepts
  - A document can be described by a set of representative keywords called index terms.
  - Different index terms have varying relevance when used to describe document contents.
  - This effect is captured through the assignment of numerical weights to each index term of a document. (e.g.: frequency, tf-idf)
- DBMS Analogy
  - Index Terms → Attributes
  - Weights → Attribute Values

Keyword-Based Retrieval

- A document is represented by a string, which can be identified by a set of keywords
- Queries may use expressions of keywords
  - E.g., car *and* repair shop, tea *or* coffee, DBMS *but not* Oracle
  - Queries and retrieval should consider synonyms, e.g., repair and maintenance
- Major difficulties of the model
  - Synonymy: A keyword *T* does not appear anywhere in the document, even though the document is closely related to *T*, e.g., data mining
  - Polysemy: The same keyword may mean different things in different contexts, e.g., mining

Similarity-Based Retrieval in Text Data

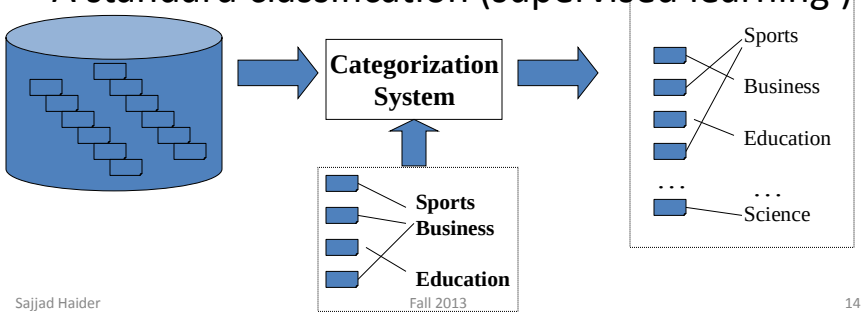
- Finds similar documents based on a set of common keywords
- Answer should be based on the degree of relevance based on the nearness of the keywords, relative frequency of the keywords, etc.
- Basic techniques
- Stop list
  - Set of words that are deemed "irrelevant", even though they may appear frequently
  - E.g., *a, the, of, for, to, with*, etc.
  - Stop lists may vary when document set varies

Document Clustering

- Motivation
  - Automatically group related documents based on their contents
  - No predetermined training sets or taxonomies
  - Generate a taxonomy at runtime
- Clustering Process
  - Data preprocessing: remove stop words, stem, feature extraction, lexical analysis, etc.
  - Hierarchical clustering: compute similarities applying clustering algorithms.
  - Model-Based clustering (Neural Network Approach): clusters are represented by "exemplars". (e.g.: SOM)

Text Categorization

- Pre-given categories and labeled document examples (Categories may form hierarchy)
- Classify new documents
- A standard classification (supervised learning)



Applications

- News article classification
- Automatic email filtering
- Webpage classification
- Word sense disambiguation
- And many more

Tag Clouds

- A **tag cloud** (**word cloud**) is a visual representation for text data, typically used to depict keyword metadata (tags) on websites, or to visualize free form text.
- Tags are usually single words, and the importance of each tag is shown with font size or color.
- This format is useful for quickly perceiving the most prominent terms and for locating a term alphabetically to determine its relative prominence.

TagCrowd.com

action agents algorithm annotation based bayesian best bipedal blend building collaboration comparison coordinated data dynamic evolving extraction files finding fourier frameworks imitation influence information intelligence kick modelling nets networks ontology ontology-based opponent optimization particle portfolio reference-set-based reporting robocup robot rule-based semantic simulation skills soccer sources strategies study testing unstructured via

Wordle.com (1)

Wordle.com (2)

KNIME Demo: Tag cloud