Text Mining

Motivation for Text Mining

- Approximately 90% of the World's data is held in unstructured formats
  - Web pages
  - Emails
  - Technical documents
  - Corporate documents
  - Books
  - Digital libraries
  - Customer complaint letters
- Growing rapidly in size and importance

Text Mining Applications

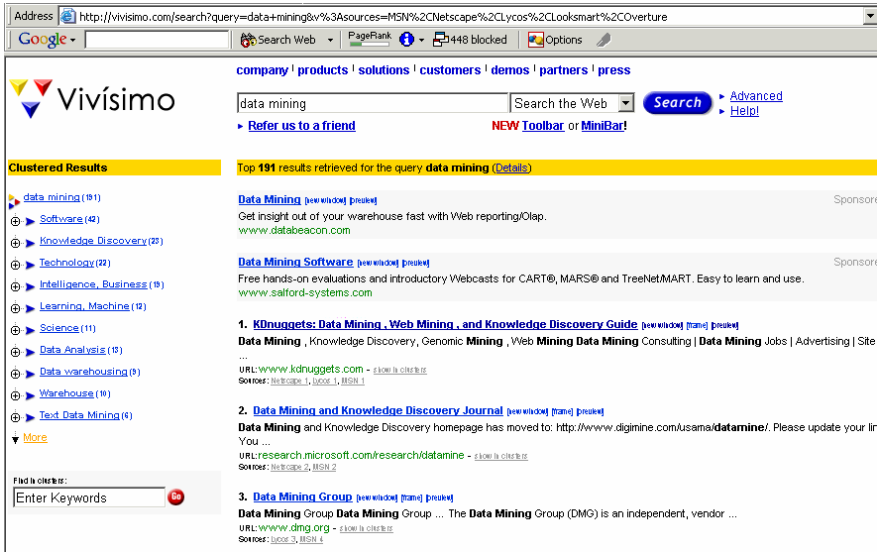
- Classification of news stories, web pages, … , according to their content
- Email and news filtering
- Organize repositories of document-related meta-information for search and retrieval (search engines)
- Clustering documents or web pages
- Gain insights about trends, relations between people, places and/or organizations
- Find associations among entities such as:

Author = Wilson $\Rightarrow$ Author = Holmes
Supervisor = William $\Rightarrow$ Examiner = Ferdinand

Personalizing an Online Newspaper

- Politics
- ~~Economic~~
- UK
- ~~World~~
- Sport
- Entertainment

Clustering Results Of Search Engine Queries



Challenges

- Information is in unstructured textual form
- Large textual data base
  - almost all publications are also in electronic form
- Very high number of possible "dimensions" (but sparse):
  - all possible word and phrase types in the language!!
- Complex and subtle relationships between concepts in text
  - "AOL merges with Time-Warner" "Time-Warner is bought by AOL"
- Word ambiguity and context sensitivity
  - automobile = car = vehicle = Toyota
  - Apple (the company) or apple (the fruit)
- Noisy data
  - Example: Spelling mistakes

Semi-Structured Data

- Text databases are, in general, *semi-structured*
- Example:
  - Title
  - Author
  - Publication_Date
  - Length
  - Category

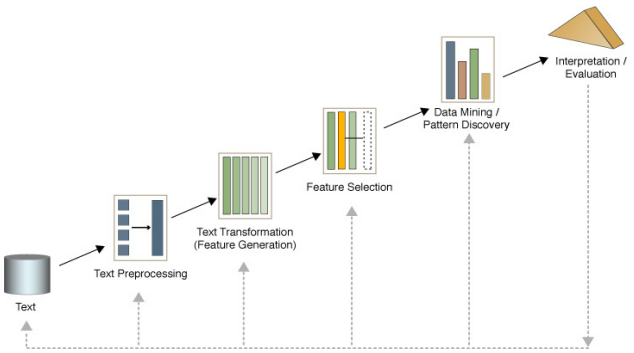
  } Structured attribute/value pairs

  - Abstract
  - Content

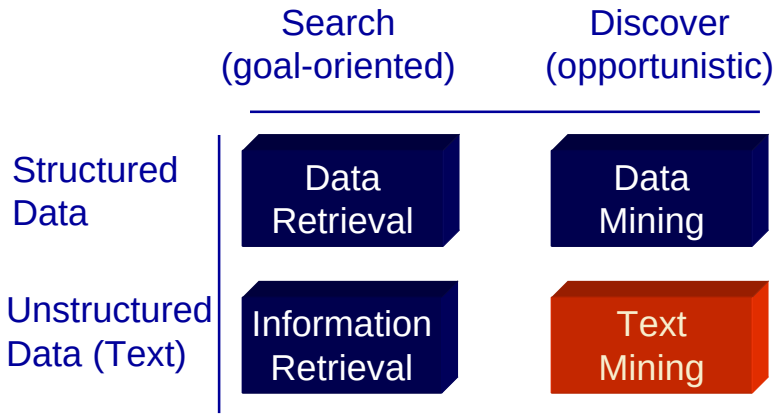
  } Unstructured

Text Mining Process

- Text preprocessing
  - Syntactic/Semantic text analysis
- Features Generation
  - Bag of words
- Features Selection
  - Simple counting
  - Statistics
- Text/Data Mining
  - Classification
  - Clustering
  - Associations
- Analyzing results



"Search" versus "Discover"



Handling Text Data

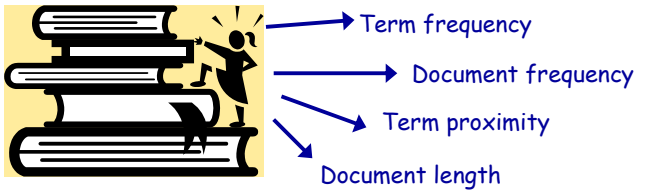
- Modeling semi-structured data
- Information Retrieval (IR) from unstructured documents
  - Locates relevant documents and Ranks documents
    - Keyword based (Boolean matching)
    - Similarity based
- Text mining
  - Classify documents
  - Cluster documents
  - Find patterns or trends across documents

Information Retrieval (IR)

- Information retrieval problem: locating relevant documents (e.g., given a set of keywords) in a corpus of documents
  - Major application: Web search engines

Structuring Textual Information

- Many methods designed to analyze structured data
- If we can represent documents by a set of attributes we will be able to use existing data mining methods
- How to represent a document?
  - Vector based representation
    - (referred to as "bag of words" as it is invariant to permutations)
- Use statistics to add a numerical dimension to unstructured text



Document Representation

- A document representation aims to capture what the document is about
- One possible approach:
  - Each entry describes a document
  - Attribute describe whether or not a term appears in the document

Example

	Terms				
	Camera	Digital	Memory	Pixel	…
Document 1	**1**	**1**	**0**	**1**	
Document 2	**1**	**1**	**0**	**0**	
…	…	…	…	…	

Document Representation

- Another approach:
  - Each entry describes a document
  - Attributes represent the frequency in which a term appears in the document

Example: Term frequency table

	Terms				
	Camera	Digital	Memory	Print	…
Document 1	**3**	**2**	**0**	**1**	
Document 2	**0**	**4**	**0**	**3**	
…	…	…	…	…	

Document Representation

- But a term is mentioned more times in longer documents
- Therefore, use relative frequency (% of document):
  - No. of occurrences/No. of words in document

	Terms				
	Camera	Digital	Memory	Print	…
Document 1	**0.03**	**0.02**	**0**	**0.01**	
Document 2	**0**	**0.004**	**0**	**0.003**	
…	…	…	…	…	

More on Document Representation

- **Stop Word removal:** Many words are not informative and thus irrelevant for document representation
  - the, and, a, an, is, of, that, …
- **Stemming:** reducing words to their root form
  - A document may contain several occurrences of words like
    - fish, fishes, fisher, and fishers
  - But would not be retrieved by a query with the keyword
    - fishing
  - Different words share the same word stem and should be represented with its stem, instead of the actual word
    - fish

Weighting Scheme for Term Frequencies

- **TF-IDF weighting:** give higher weight to terms that are rare
  - TF: term frequency (increases weight of frequent terms)
    - If a term is frequent in lots of documents it does not have discriminative power
  - IDF: inverse term frequency

For a given term w_j and document d_i

n_{ij} is the number of occurrences of w_j in document d_i

$|d_i|$ is the number of words in document d_i

n is the number of documents

n_j is the number of documents that contain w_j

$$TF_{ij} = \frac{n_{ij}}{|d_i|}$$

$$IDF_j = \log \frac{n_j}{n}$$

$$x_{ij} = TF_{ij} \cdot IDF_j$$

There is no compelling motivation for this method but it has been shown to be superior to other methods

Locating Relevant Documents

- Given a set of keywords
- Use similarity/distance measure to find similar/relevant documents
- Rank documents by their relevance/similarity

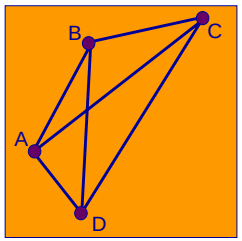
How to determine if two documents are similar?

Distance Based Matching

- In order retrieve documents similar to a given document we need a measure of similarity
  - Euclidean distance (example of a metric distance):
    - The Euclidean distance between

      $X=(x_1, x_2, x_3,\ldots x_n)$ and $Y=(y_1, y_2, y_3,\ldots y_n)$
    - is defined as:

$$D(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$



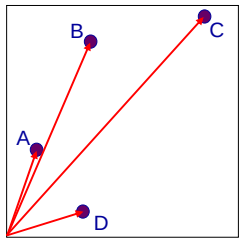
Properties of a metric distance:

- D(X,X)=0
- D(X,Y)=D(Y,X)
- D(X,Z)+D(Z,Y) ≥ D(X,Y)

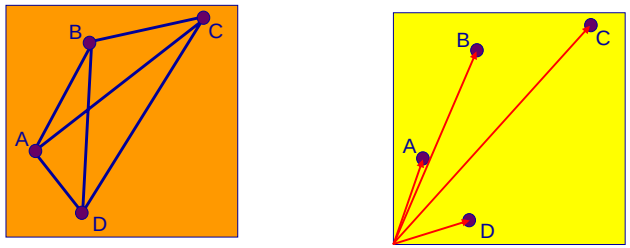
Angle Based Matching

- Cosine of the **angle** between the vectors representing the document and the query
- Documents "in the same direction" are closely related.
- Transforms the angular measure into a measure ranging from 1 for the highest similarity to 0 for the lowest

$$D(X,Y) = \cos(X,Y) = \frac{X^T Y}{\|X\| \cdot \|Y\|} = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2} \cdot \sqrt{\sum y_i^2}}$$



Distance vs. Angle



Performance Measure

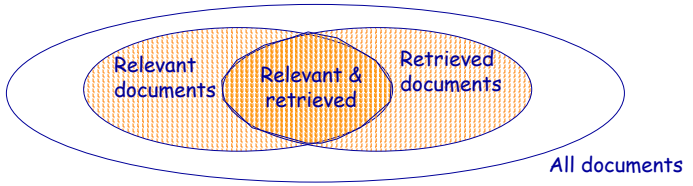
- The set of retrieved documents can be formed by collecting the top-ranking documents according to a similarity measure
- The quality of a collection can be compared by the two following measures

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

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

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

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



Text Mining

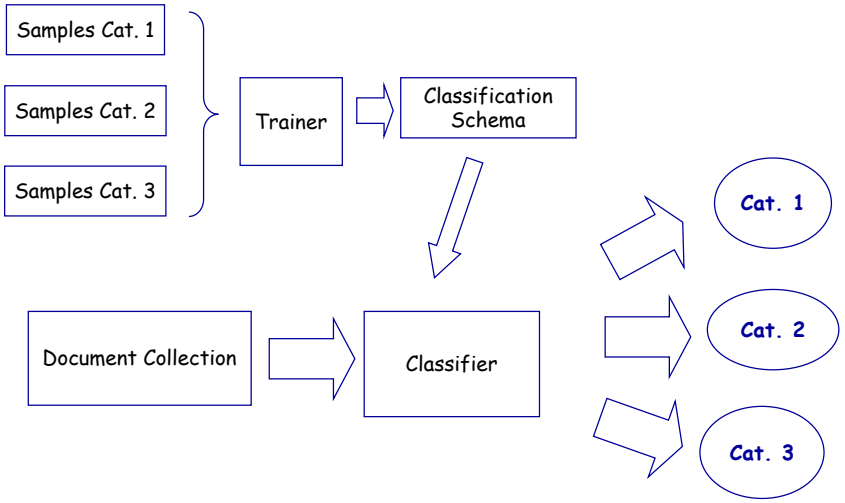
- Document classification
- Document clustering
- Key-word based association rules

Document Classification

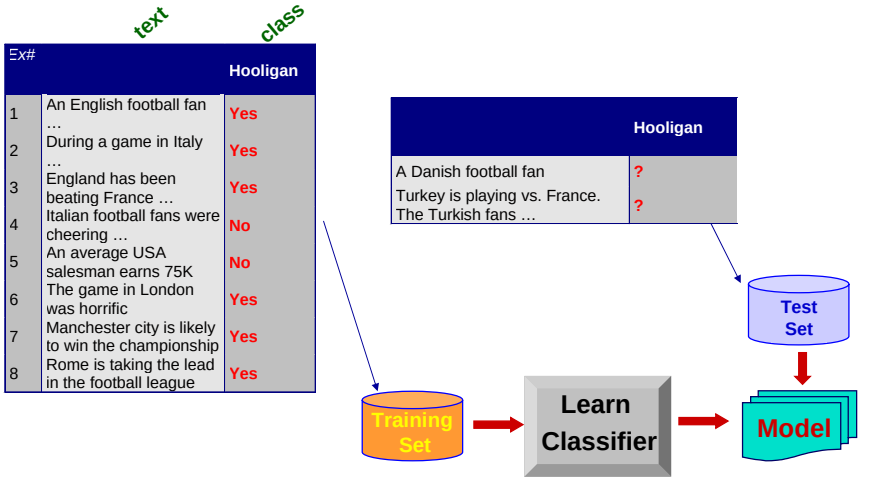
- Human experts classify a set of documents
  - training data set
- Induce a classification model

	Terms					Class
	Oil	Iraq	build	France	…	Interesting/Not interesting
Document 1	**0.01**	**0.05**	**0.03**	**0**		Interesting
Document 2	**0**	**0.05**	**0**	**0.01**		Not interesting
…	…	…	…	…		…

Classification Schema



Text Classification: An Example



Classification Techniques

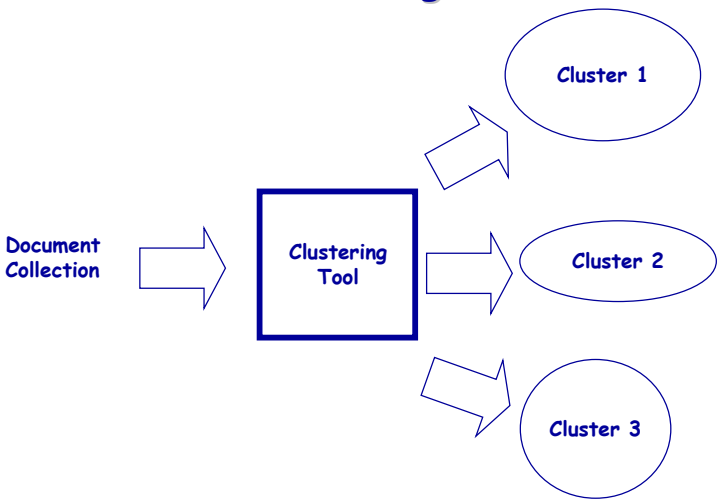
- Decision Trees
- K-nearest neighbors
  - Training examples are points in a vector space
  - Compute distance between new instance and all training instances and the k-closest vote for the class
- Naïve Bayes Classifier
  - Classify using probabilities and assuming independence among terms
    - $P(x_i/C)$ is estimated as the relative frequency of examples having value x_i as feature in class C
    - P(C/ Xi Xj Xk) = P(C) P(Xi/C) P(Xj/C) P(Xk/C)
- Neural networks, support vector machines,…

Document Clustering

- Finding Groups of Similar Documents
  - Partitioning Methods: k-means
  - Hierarchical Methods: Agglomerative or Divisive

	Terms					Class
	Oil	Iraq	build	France	…	?
Document 1	0.01	0.05	0.03	0		?
Document 2	0	0.05	0	0.01		?
…	…	…	…	…		…

Clustering Schema



Associations: Keyword-Based Associations

- Each document is a "basket"/collection of terms
  - Apriori, FP-tree, ...

	Terms				
	HP	Digital	Compaq	IBM	...
Story 1	1	1	0	1	
Story 2	1	1	0	0	
...	...	...	...	...	

If HP → Digital, Compaq

IF Business_Intelligence → ClearForest

Text is tricky to process, but "ok" results are easily achieved

References

- Pierre Baldi, Paolo Frasconi, Padhraic Smyth "Modeling the Internet and the Web, Probabilistic Methods and Algorithms", 2003 (chapter 4) [http://ibook.ics.uci.edu/Chapter4.pdf]
- David J. Hand, Heikki Mannila and Padhraic Smyth, "Principles of Data Mining", 2001
- Yair Even-Zohar, "Introduction to Text Mining" [slides: http://algdocs.ncsa.uiuc.edu/PR-20021116-2.ppt]
- Jochen Dijrre, Peter Gerstl, Roland Seiffert, "Text Mining: Finding Nuggets in Mountains of Textual Data", KDD 1999.

Thank you !!!