Web Content Mining

What is text mining?

- Data mining in text: find something useful and surprising from a text collection;
- text mining vs. information retrieval;
- data mining vs. database queries.

Types of text mining

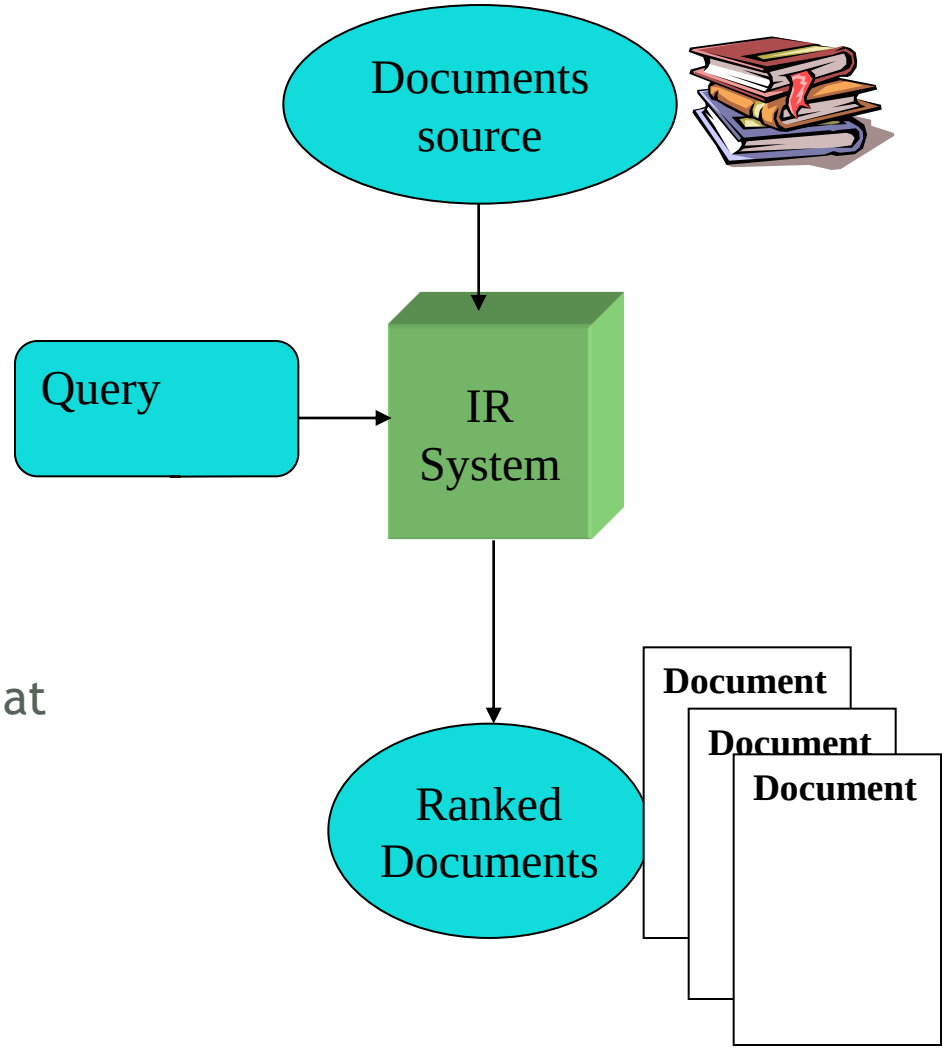
- Keyword (or term) based association analysis
- automatic document (topic) classification
- similarity detection
  - cluster documents by a common author
  - cluster documents containing information from a common source
- sequence analysis: predicting a recurring event, discovering trends
- anomaly detection: find information that violates usual patterns

Types of text mining (cont.)

- discovery of frequent phrases
- text segmentation (into logical chunks)
- event detection and tracking

Information Retrieval

- **Given:**
  - A source of textual documents
  - A user query (text based)
- **Find:**
  - A set (ranked) of documents that are relevant to the query



Intelligent Information Retrieval

- ***meaning*** of words
  - Synonyms "buy" / "purchase"
  - Ambiguity "bat" (baseball vs. mammal)
- ***order*** of words in the query
  - hot dog stand in the amusement park
  - hot amusement stand in the dog park
- ***user dependency*** for the data
  - direct feedback
  - indirect feedback
- ***authority*** of the source
  - IBM is more likely to be an authorized source then my second far cousin

Intelligent Web Search

- Combine the intelligent IR tools
  - **meaning** of words
  - **order** of words in the query
  - **user dependency** for the data
  - **authority** of the source
- With the unique web features
  - retrieve Hyper-link information
  - utilize Hyper-link as input

What is Information Extraction?

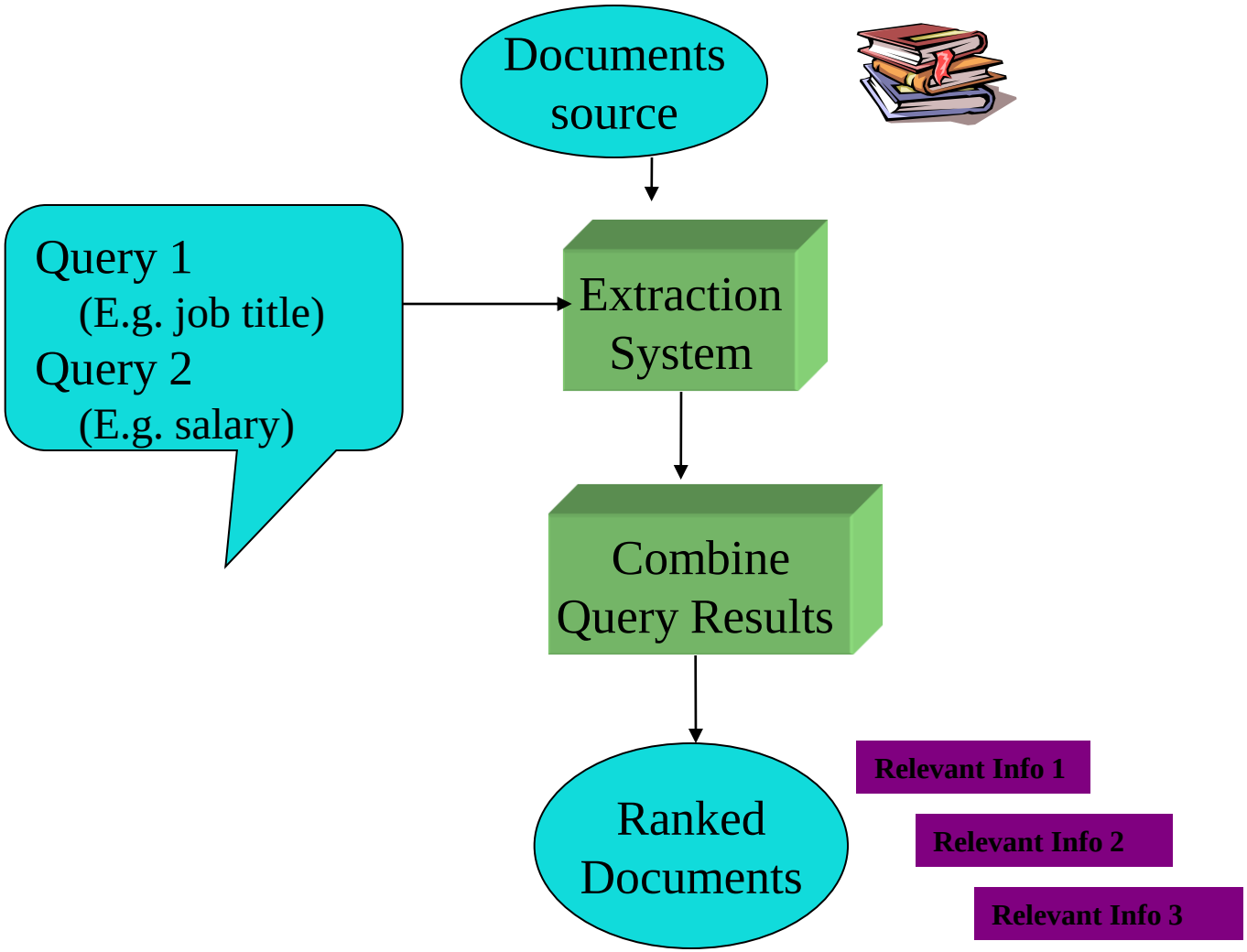
- **Given:**
  - A source of textual documents
  - A well defined limited query (text based)
- **Find:**
  - Sentences with relevant information
  - Extract the relevant information and ignore non-relevant information (important!)
  - Link related information and output in a predetermined format

Information Extraction: Example

- Salvadoran President-elect Alfredo Cristiania condemned the terrorist killing of Attorney General Roberto Garcia Alvarado and accused the Farabundo Marti Natinal Liberation Front (FMLN) of the crime. … Garcia Alvarado, 56, was killed when a bomb placed by urban guerillas on his vehicle exploded as it came to a halt at an intersection in downtown San Salvador. … According to the police and Garcia Alvarado's driver, who escaped unscathed, the attorney general was traveling with two bodyguards. One of them was injured.

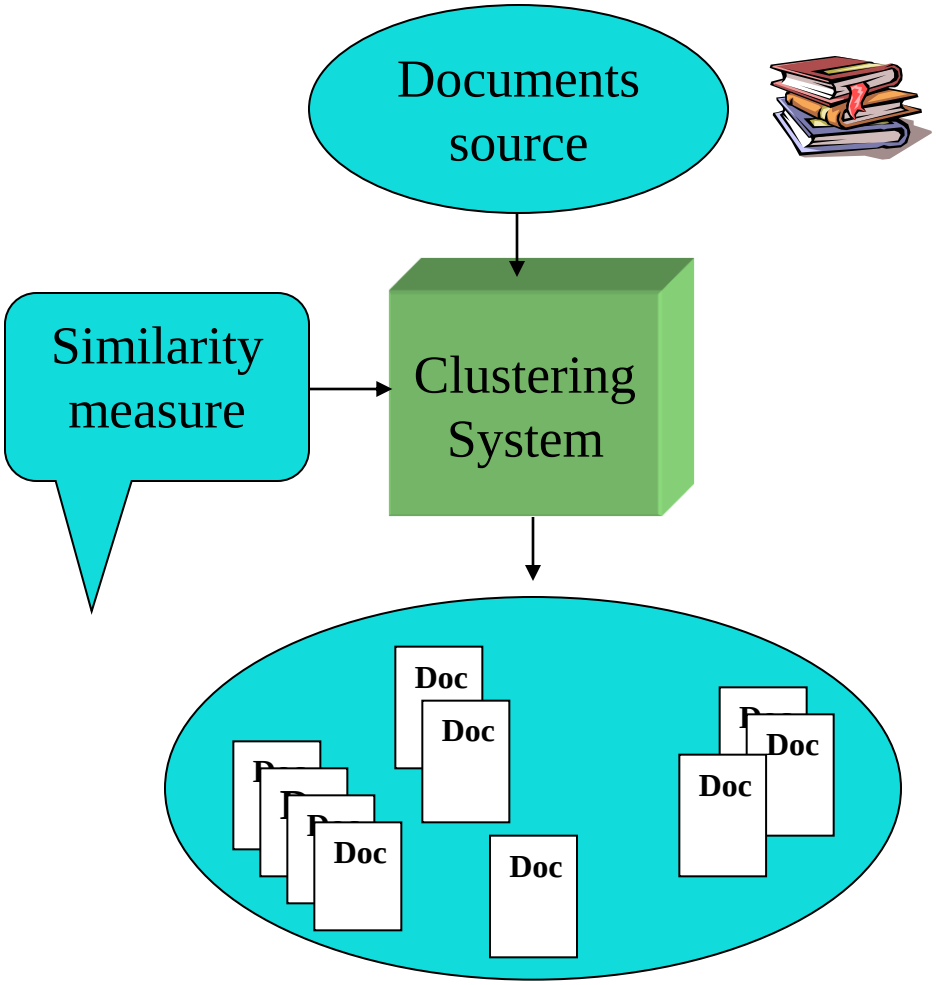
- **Incident Date:** 19 Apr 89
- **Incident Type:** Bombing
- **Perpetrator Individual ID:** "urban guerillas"
- **Human Target Name:** "Roberto Garcia Alvarado"
- **…**

Querying Extracted Information



What is Clustering ?

- **Given:**
  - A source of textual documents
  - Similarity measure
    - e.g., how many words are common in these documents
- **Find:**
  - Several clusters of documents that are relevant to each other



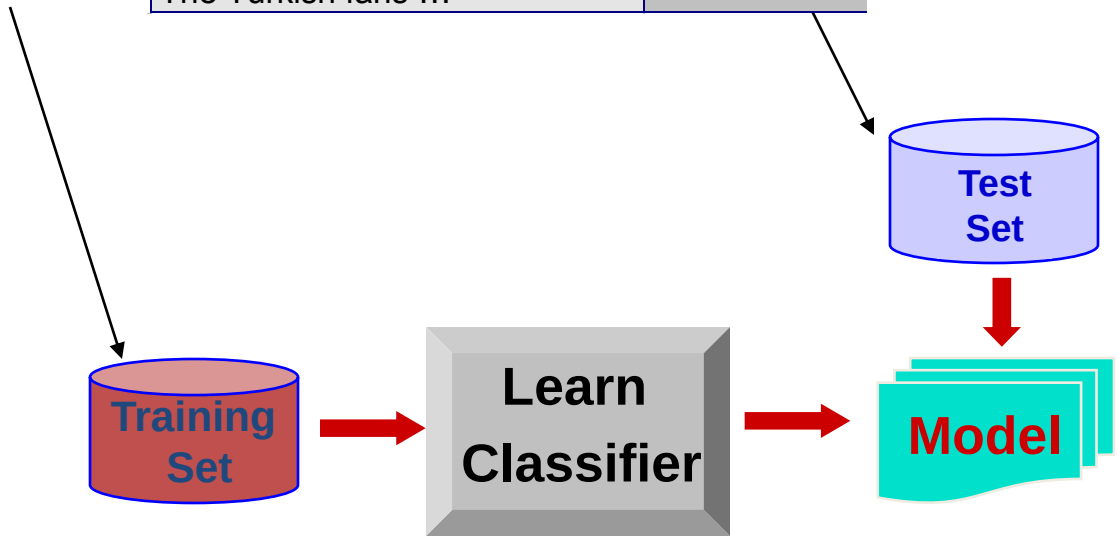
Text Classification definition

- **Given:** a collection of labeled records *(training set)*
  - Each record contains a set of features *(attributes)*, and the true class *(label)*
- **Find:** a model for the class as a function of the values of the features
- **Goal:** previously unseen records should be assigned a class as accurately as possible
  - A test set is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it

Text Classification: An Example

Ex#	text	class: Hooligan
1	An English football fan …	Yes
2	During a game in Italy …	Yes
3	England has been beating France …	Yes
4	Italian football fans were cheering …	No
5	An average USA salesman earns 75K	No
6	The game in London was horrific	Yes
7	Manchester city is likely to win the championship	Yes
8	Rome is taking the lead in the football league	Yes

text	Hooligan
A Danish football fan	?
Turkey is playing vs. France. The Turkish fans …	?



Discovery of frequent sequences (1)

- Find all frequent maximal sequences of words (=phrases) from a collection of documents
  - frequent: frequency threshold is given; e.g. a phrase has to occur in at least 15 documents
  - maximal: a phrase is not included in another longer frequent phrase
  - other words are allowed between the words of a sequence in text

Discovery of frequent sequences (2)

- Frequency of a sequence cannot be decided locally: all the instances in the collection has to be counted
- however: already a document of length 20 contains over million sequences
- only small fraction of sequences are frequent

Basic idea: bottom-up

- 1. Collect all pairs from the documents, count them, and select the frequent ones
- 2. Build sequences of length p + 1 from frequent sequences of length p
- 3. Select sequences that are frequent
- 4. Select maximal sequences

Summary

- There are many **scientific and statistical text mining methods** developed, see e.g.:
  - http://www.cs.utexas.edu/users/pebronia/text-mining/
  - http://filebox.vt.edu/users/wfan/text_mining.html
- Also, it is important to study **theoretical foundations** of data mining.
  - Data Mining Concepts and Techniques / J.Han & M.Kamber
  - Machine Learning, / T.Mitchell