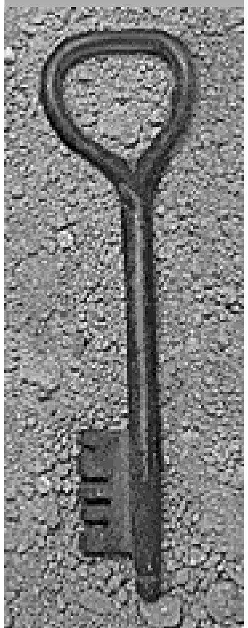
Text Mining / Web Mining

CSE 6331 / CSE 6362

Data Mining

Fall 1999

Diane J. Cook

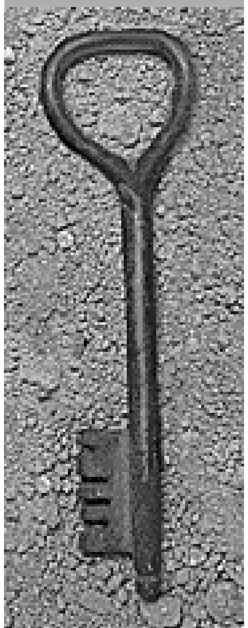


Text Mining as a Data Mining Task

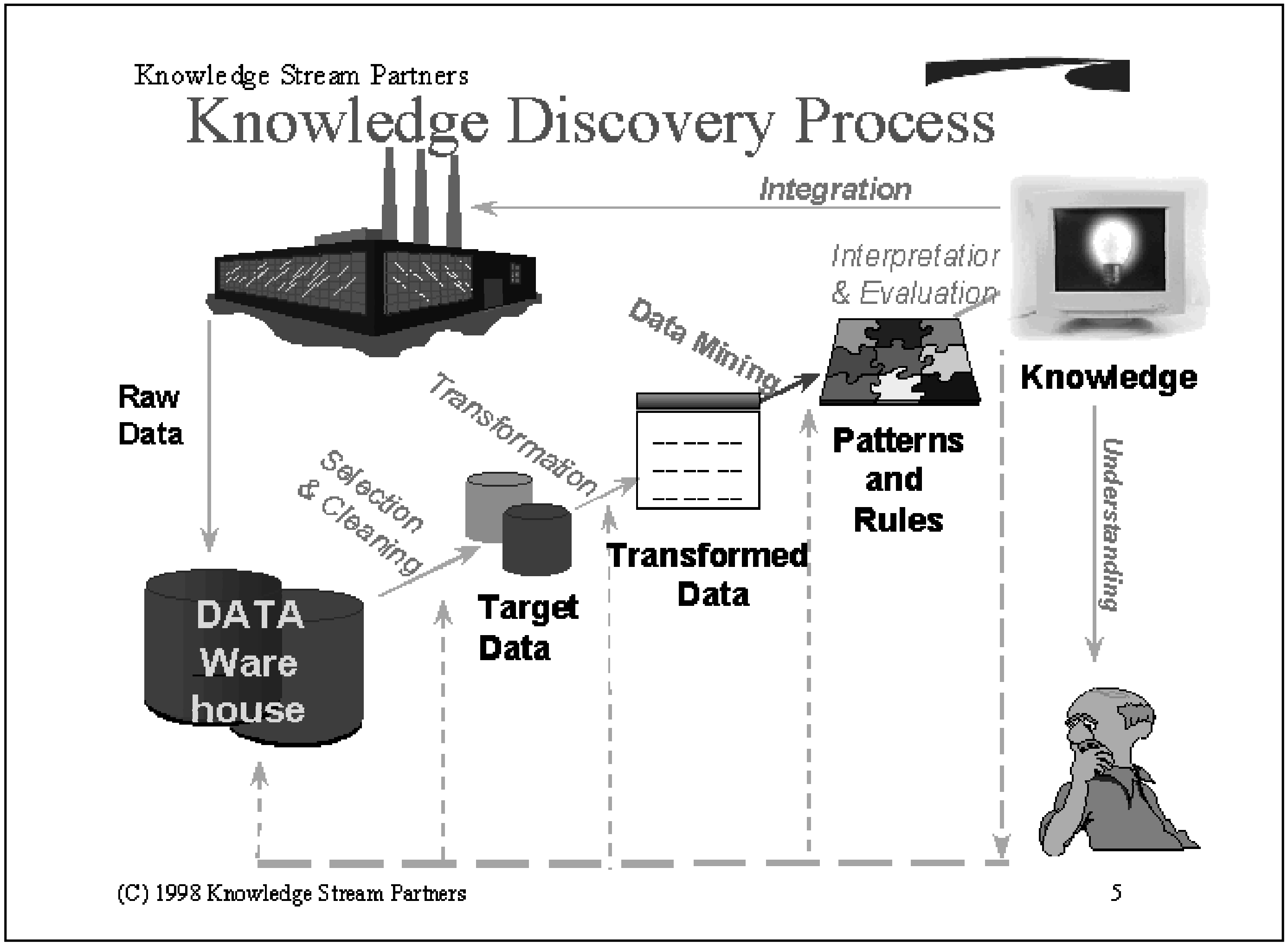
- Speech vs. text mining vs. natural language processing
- Tasks
  - Summarization
  - Cluster documents
  - Classify documents
  - Find similar documents
  - Find commonly occurring terms
  - Find associations between terms
  - Answer queries about documents

Applications

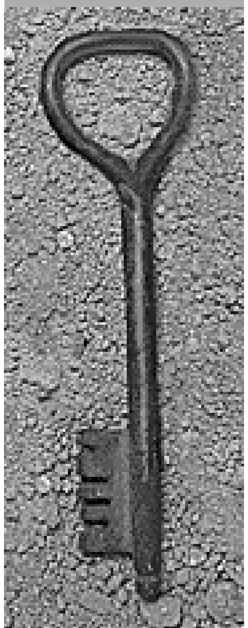
- Web search
- Scan documents for content
- Email processing
- NLP is hard! (AI-Complete?)



KDD Process

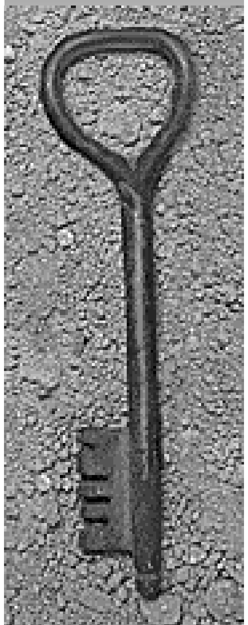


Knowledge Stream Partners
Knowledge Discovery Process
Integration
Interpretation
& Evaluation
Data Mining
Transformation
Selection
& Cleaning
Raw
Data
DATA
Ware
house
Target
Data
Transformed
Data
Patterns
and
Rules
Knowledge
Understanding
(C) 1998 Knowledge Stream Partners
5



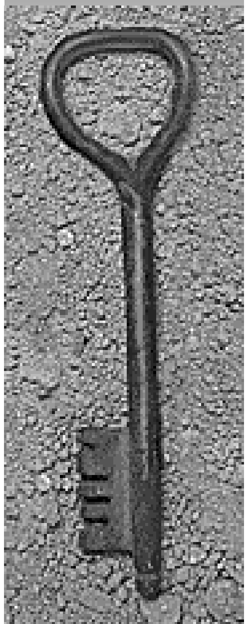
Data Preparation

- Delete supplementary and common words
  - “a”, “an”, “the”, “is”
- Identify stems of words
  - Separate / remove prefixes, suffixes, endings
- Identify senses
  - Which meaning of word is used?
- Normalize
  - Replace similar words with common token



Feature Selection

- Identify important vocabulary terms
  - Names
  - Multiword terms, phrases
  - Abbreviations
  - Numbers, dates, currency

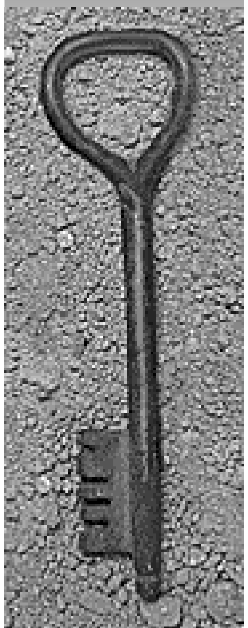


Missing Data

- Almost all data is missing if we handle words individually
- Which meaning is being used?
  - Replace by all, handle separately
- What information is omitted?
- What information is assumed?

Why Text is Rich to Mine

- Organizations have large amounts of online documents with important information
- Examples
  - Electronic mail from customers, containing feedback about products and services
  - Intranet documents such as memos and presentations embodying corporate expertise
  - Technical reports describing new technology
  - News wires with information about business environments and the activities of competitors
- Forrester Research predicted that unstructured data (e.g., text) will dominate online data



Why Text is Tough to Mine

- Little or no structure
- Abstract concepts difficult to represent
- Variations of words
  - – soda, pop, soft drink, cola, carbonated beverage
- Context is important
  - lead time
  - lead the way
  - in the lead
  - heavy as lead
  - you lead the card game
- Other contextual cues give even more info

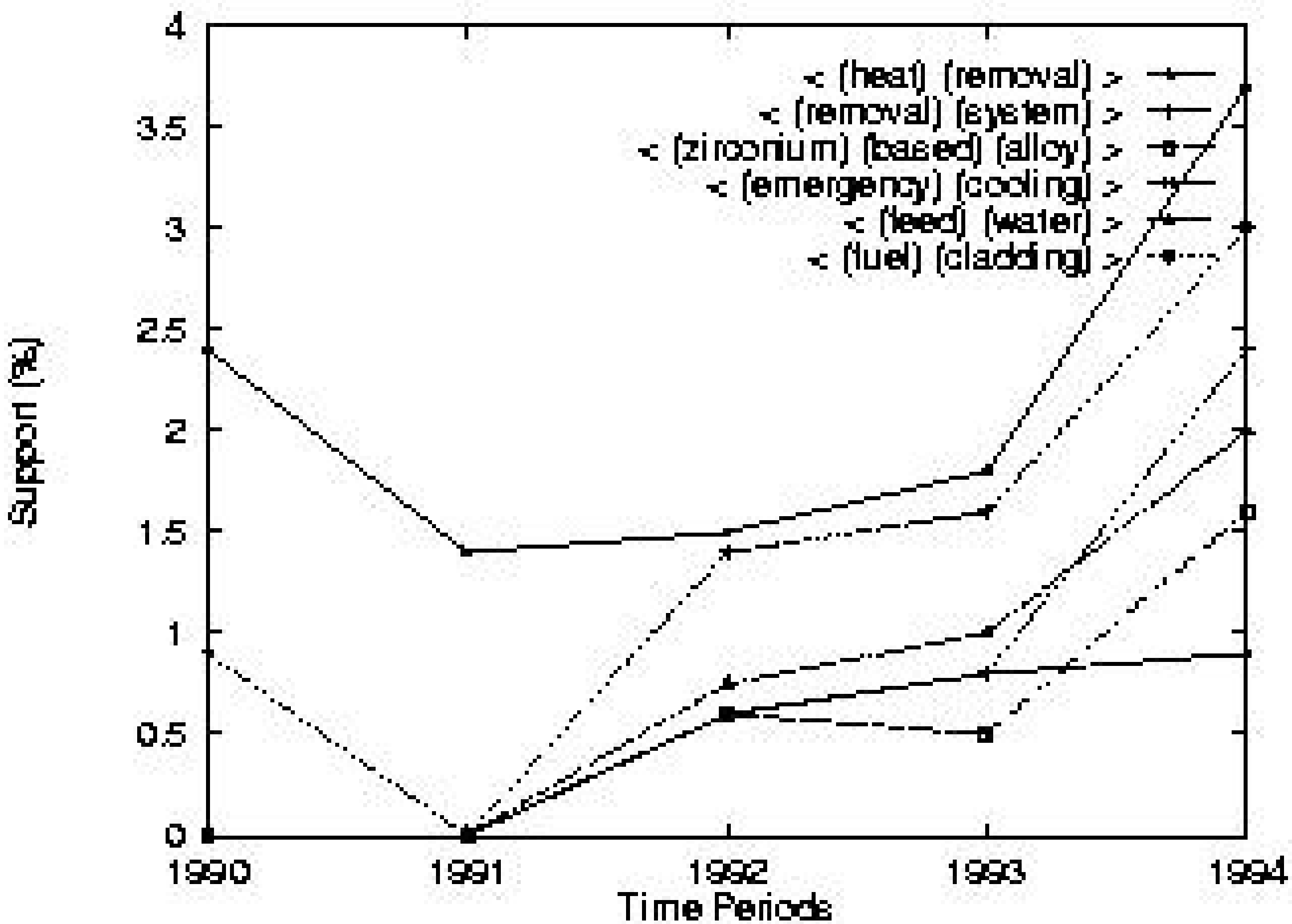
Classify documents

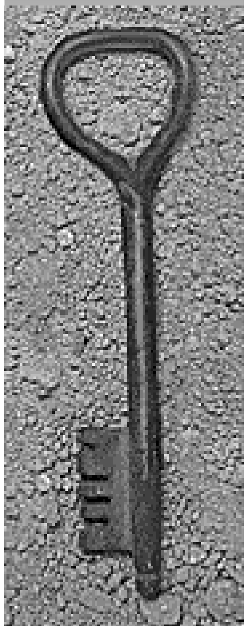
- Naïve Bayesian classification
- Each document represented as vector of words (attributes)
- Remember:
  $v(new) = \text{argmax}_{v_j} P(v_j) \Pi_{a_i} P(a_i|v_j)$
- 1000 documents, 20 categories
- 89% classification accuracy

Discovering Trends in Text

Upward trends in patent databases

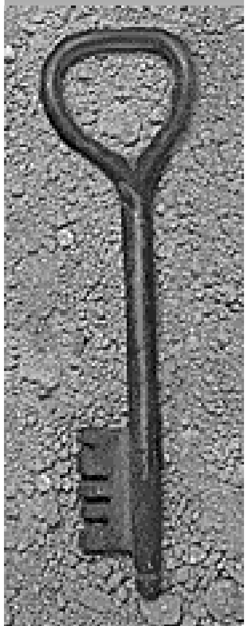
Induced Nuclear Reactions





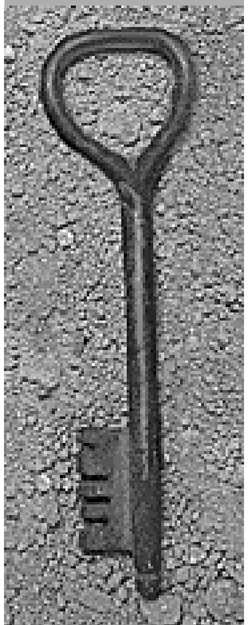
Hierarchy of Phrases

- Basic unit is word
- Phrase (0-phrase) is sequence of words
- 1-phrase is sequence of 0-phrases
- k-phrase list of phrases, nesting level = k
- Results are 2-phrases
- Allow gap between items in sequence
  - User defined
  - [MinGap .. MaxGap]
  - Guess gap for sentences, paragraphs, sections



Grow Sequences

- Generalized Sequential Patterns (GSP)
- Similar to mining association rules
- Find n-element sequences
- Calculate support
- Grow sequence by adding new element to beginning, end, or middle
- Compress database using discoveries
- Run again to generate hierarchy

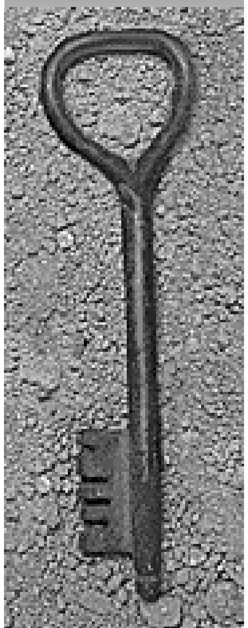


Shape Matching

- Specific shapes only
  - upward / downward trend
  - spike
  - resurgence
- Fuzzy match

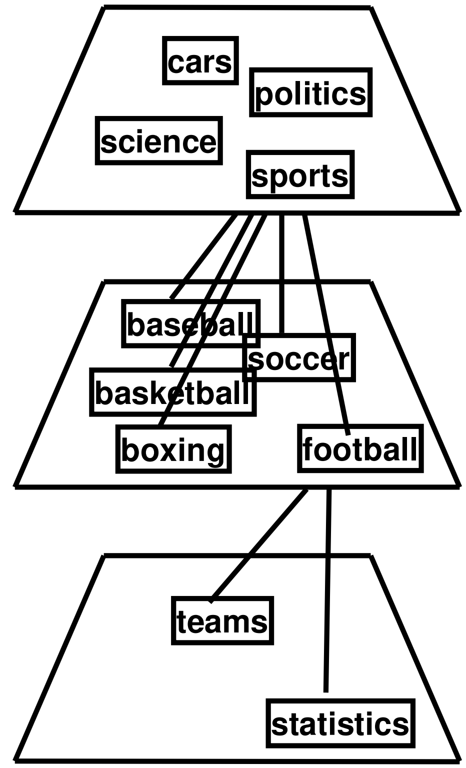
Data Mining to Create Taxonomy

- Chakrabarti et al
- Organize documents by topics
- Use multi-level classifier
- Enable search and navigation in text documents
- Apply to newswire stories, patent applications, and Yahoo! documents



Organization of Web Pages Using WordNet and Self-Organizing Maps

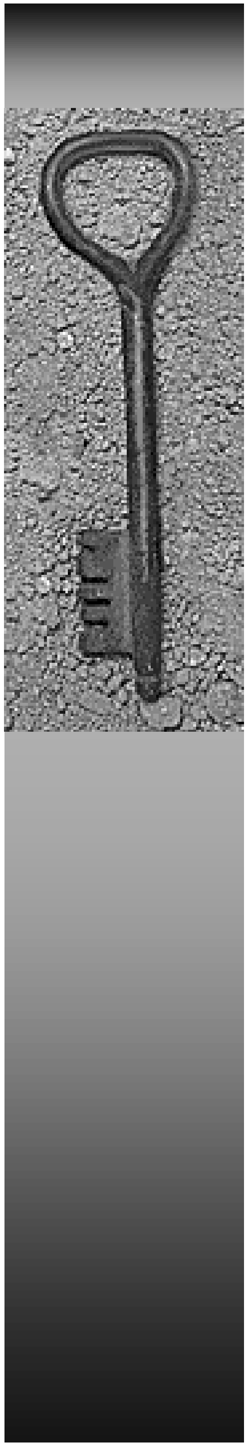
- Darin Brezeale
- Wordnet: Semantic network organizing English words
- Self-organizing map: neural network that creates 2D/3D shape or organization
  - – unsupervised



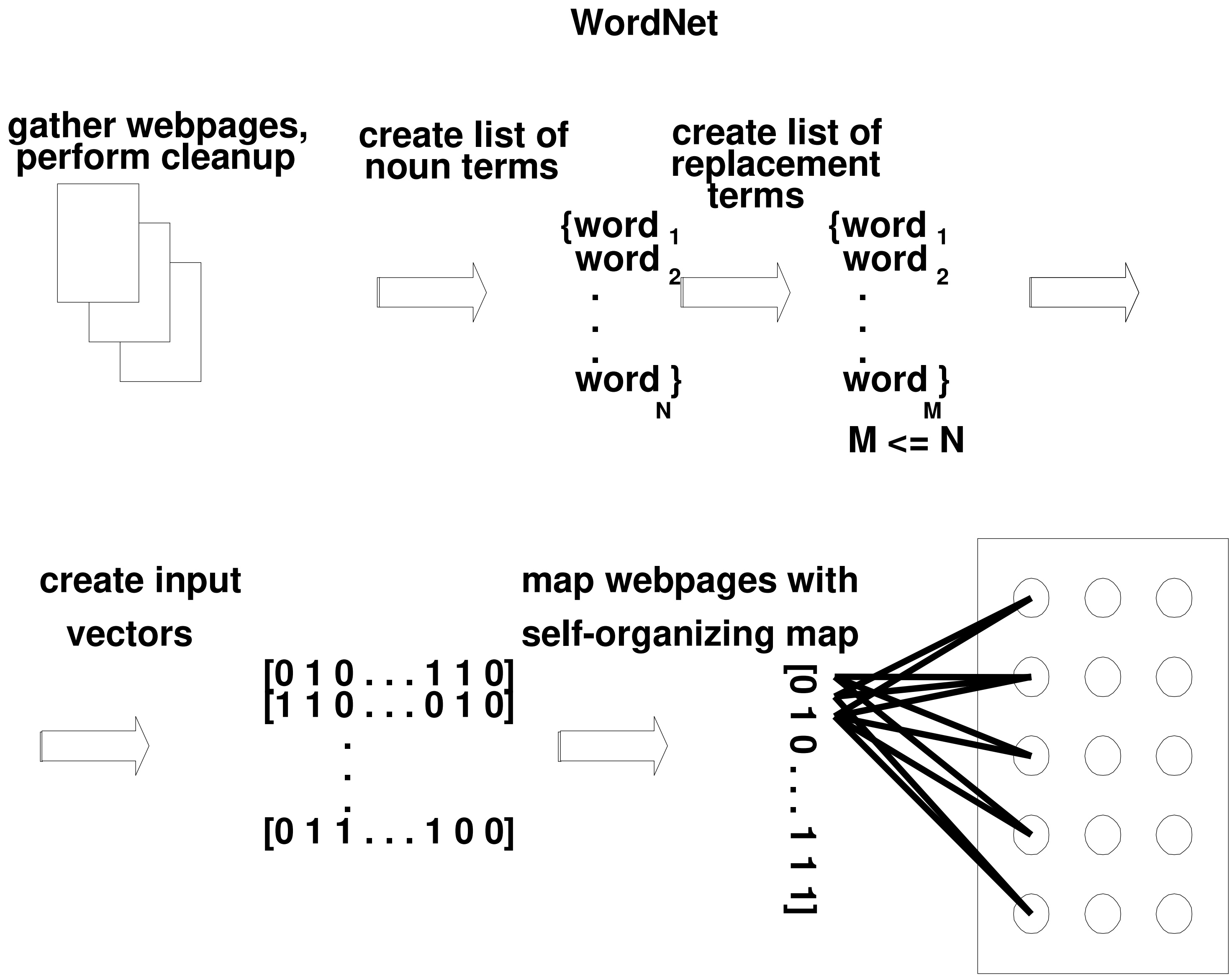
Test Pages

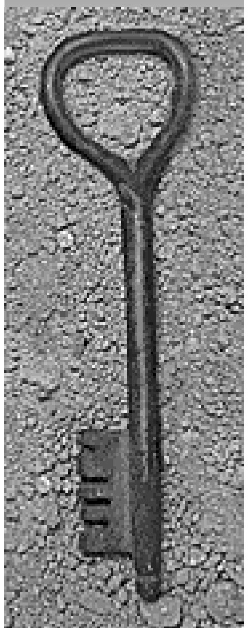
10 topics, 10 pages per topic, collected using AltaVista

- apache helicopter military – Apache helicopter used by U.S. military
- census bureau statistics – census statistics
- cortisol memory loss – hormone cortisol and relationship to memory loss
- fantasy football rules – the rules for playing fantasy football
- fermats last theorem – information about Fermat’s last theorem
- growing fruit trees – the raising of fruit trees
- nearest neighbor classifier – classifier used in machine learning
- performing clustering analysis – how to analyze clustering methods
- solar system planets – a collection of planets that orbit a star
- wolfgang amadeus mozart biography – biography of the composer Mozart



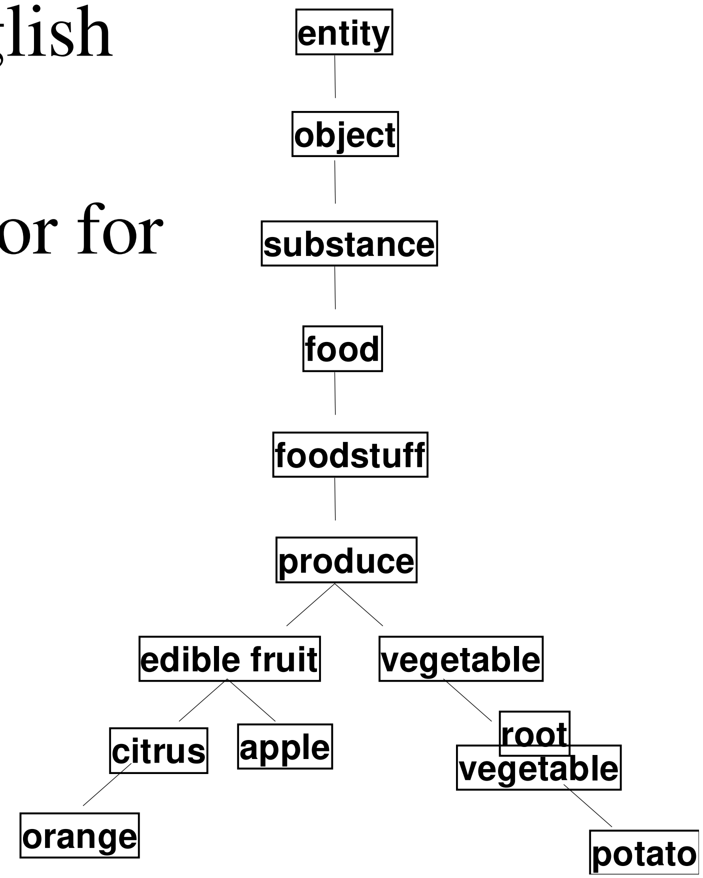
Overall Process

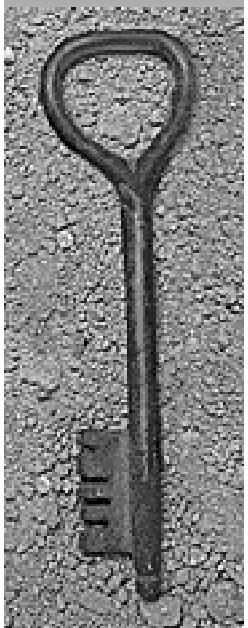




WordNet

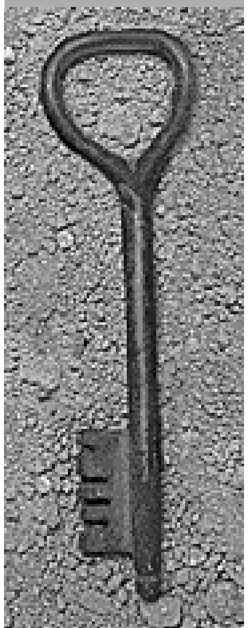
- Words grouped in hierarchy
- Intended to encode all of English language
- Used to generate feature vector for documents
  - Use only nouns
  - Use only first sense of words
  - No context





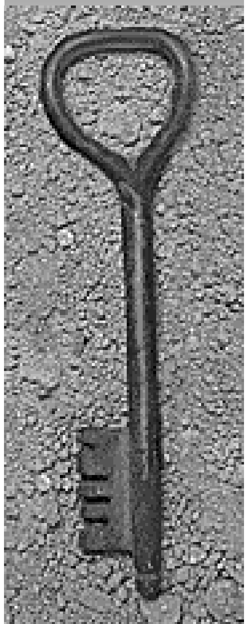
Creating Feature Vectors

- Replace original words with hypernyms
- Example
  - Page 1: (apple, tree, box)
  - Page 2: (orange, tree, man)
  - Master list: (apple, box, man, orange, tree), size=5
  - Vector 1: (apple, box, *, *, tree)
  - Vector 2: (*, *, man, orange, tree)
  - Replacement list: (box, edible_fruit, man, tree), size=4
  - Vector 1: (box, edible_fruit, *, tree)
  - Vector 2: (*, edible_fruit, man, tree)

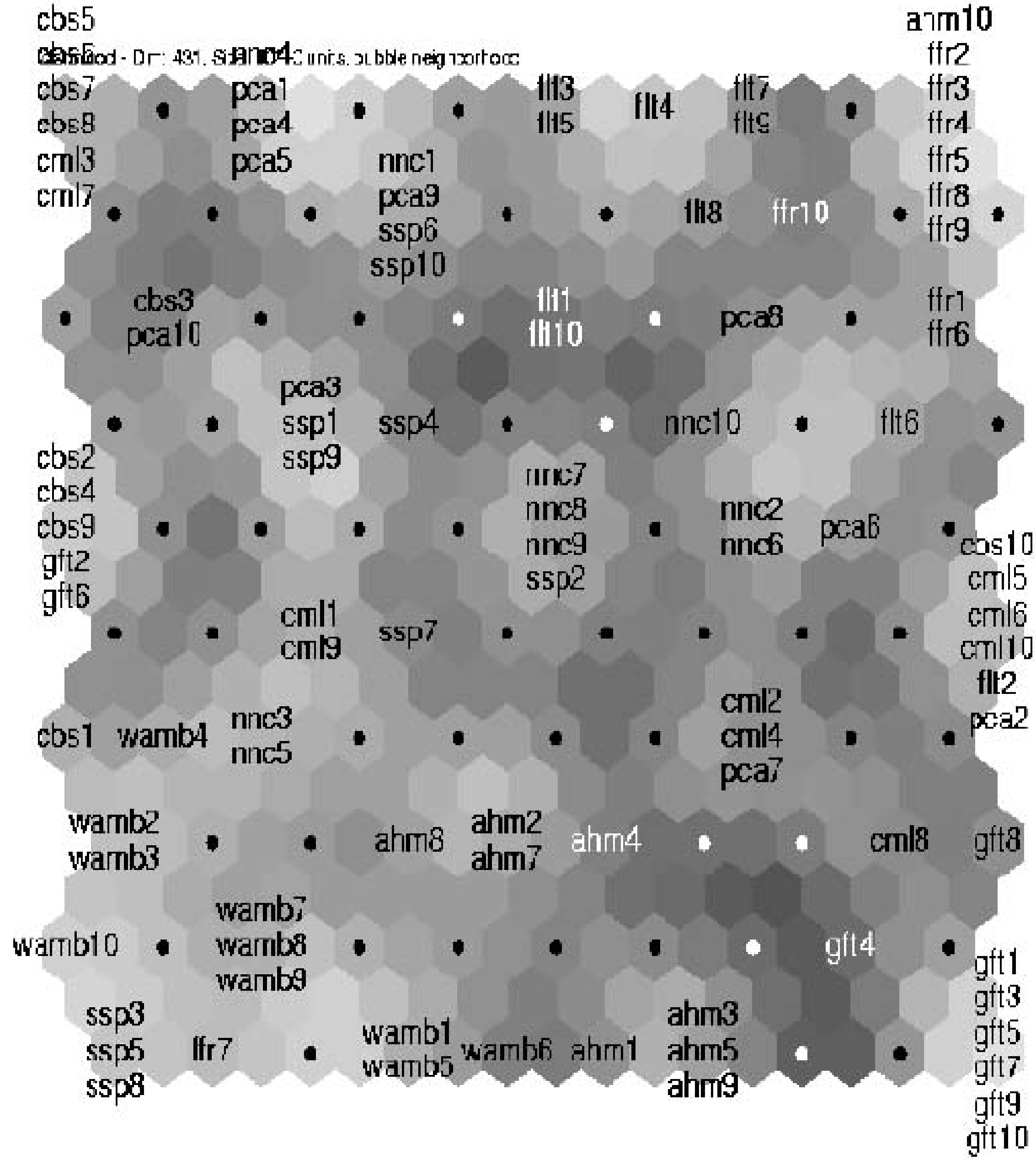


Self-Organizing Map

- Feature vectors are input to SOM
- Value (position) of each SOM node randomly initialized
- For each input vector, find closest node (Euclidean distance)
- Adjust weight of winning node so will be even closer to vector (more likely to be picked next time)

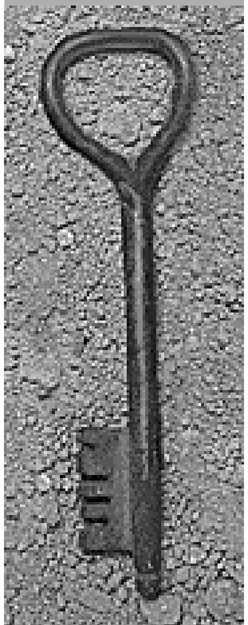


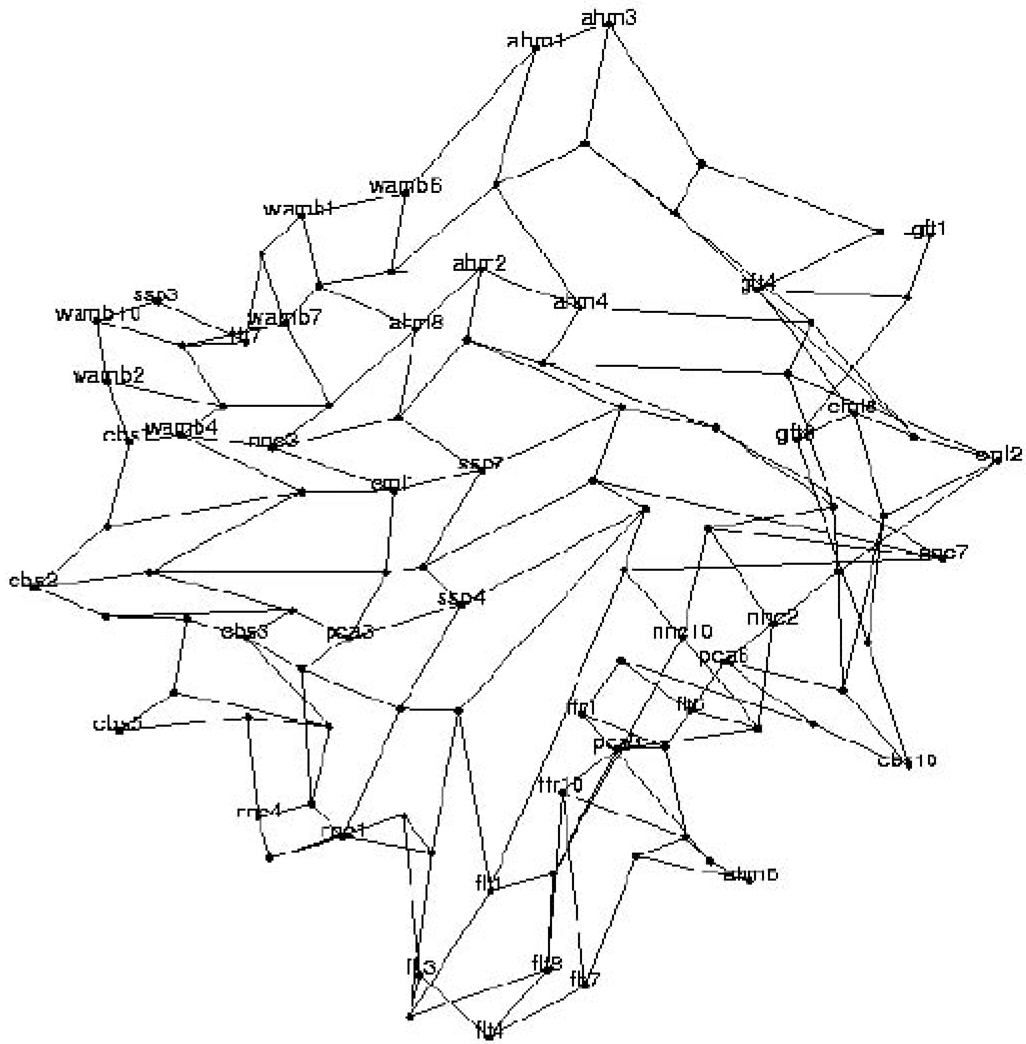
Maps



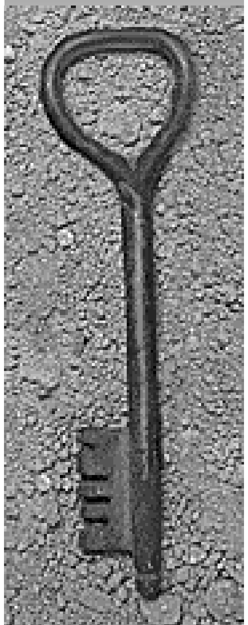
cbs5
ahm10
ffr2
cbs7
cbs8
cml3
cml7
pca1
pca4
pca5
flt3
flt5
flt4
flt7
flt9
ffr3
ffr4
ffr5
ffr8
ffr9
nnc1
pca9
ssp6
ssp10
flt8
ffr10
cbs3
pca10
flt1
flt10
pca8
ffr1
ffr6
pca3
ssp1
ssp9
ssp4
nnc10
flt6
cbs2
cbs4
cbs9
gft2
gft6
nnc7
nnc8
nnc9
ssp2
nnc2
nnc6
pca6
cbs10
cml5
cml6
cml10
cml1
cml9
ssp7
flt2
pca2
cml2
cml4
pca7
cbs1
wamb4
nnc3
nnc5
wamb2
wamb3
ahm8
ahm2
ahm7
ahm4
cml8
gft8
wamb7
wamb8
wamb9
wamb10
gft4
gft1
gft3
gft5
gft7
gft9
gft10
ssp3
ssp5
ssp8
ffr7
wamb1
wamb5
wamb6
ahm1
ahm3
ahm5
ahm9

Sammon Map





ahm3
wamb6
wamb1
gtt1
gtt4
wamb10
wamb7
ahm4
wamb2
wamb4
cbs2
ssp4
cbs3
pca3
pca6
cbs10



Results

			Smaller Avg Distance per Category		Total of Avg Distances		
Experimen	Threshol	# of categorie	W/O replacement	With replacements	W/O replacements	With replacement	% Diff.
1	4	37	7	7	18.30	18.40	-0.55%
2	2	37	7	6	18.30	15.43	15.70%
3	4	10	8	2	16.80	20.35	-21.16%
4	2	10	6	4	16.80	19.07	-13.52%

			Average Intercluster Distance		Average Intracluster Distance	
Experiment	Threshold	# of categories	W/O replacements	With replacements	W/O replacements	With replacements
1	4	37	2.10	4.59	1.22	1.23
2	2	37	2.10	2.71	1.22	1.03
3	4	10	1.74	5.43	1.68	2.04
4	2	10	1.74	3.32	1.68	1.91

In the Spotlight

Daily stock forecast from textual Web data

http://www.cs.ust.hk/~beat/Predict

DAILY STOCK MARKET FORECAST

This Web site provides Daily Stock Market predictions for five major indices.

Forecasts are available daily at 7.45 am HK time.

(At 8 am Tokyo starts as the first major stock market).

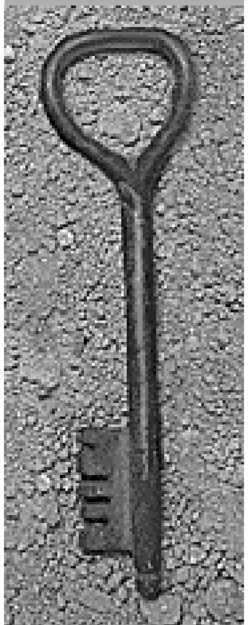
Disclaimer : We are not liable for any losses incurred when following these predictions.

PERFORMANCE COMING SOON !

Dow Jones Indus Avg	Forecasted index closing value for 10/08/99	
	Less than 0.5% change	**10708.25**

EXPLAIN DOW PERFORMANCE

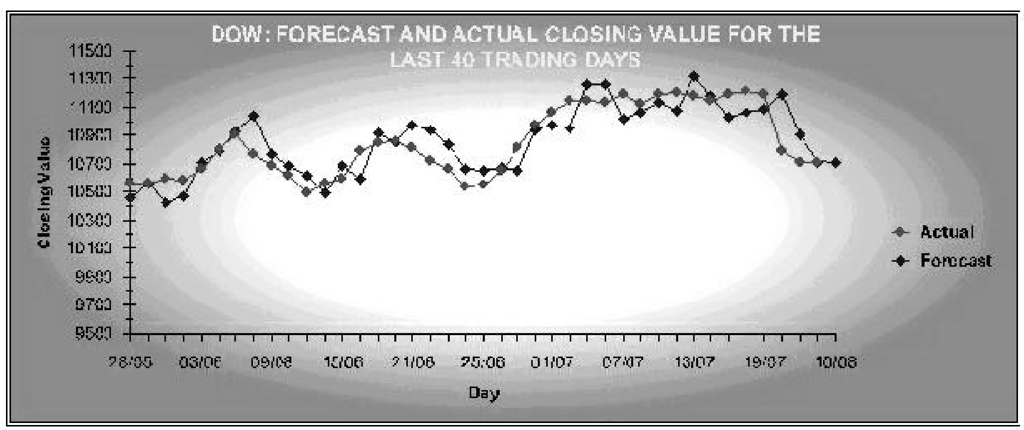
The Latest Update for Dow Jones Indus Avg				
Current Value	***Local Time***	***Status***	***Net Change***	***% Change***
10714.03	21:06 PD	Closed	-79.79	-0.74%



In the Spotlight

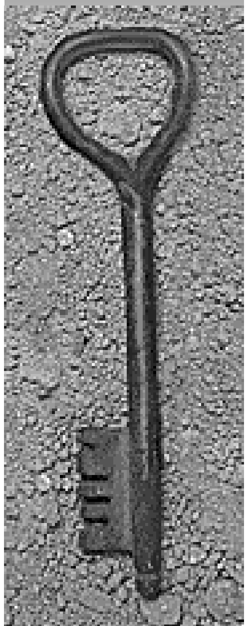
- Wuthrich et al.
  - "Daily stock market forecast from textual data", 1998
  - "Discovering probabilistic decision rules", 1997

FORECASTED AND ACTUAL CLOSING VALUES FOR THE LAST 40 TRADING DAYS



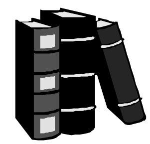
WWW Data Mining

- Web mining tasks
  - Resource discovery: locating documents and services
    - WebCrawler, AltaVista: too many irrelevant, outdated responses
    - Future: automatic text categorization, construction of directories
    - Information extraction: Automatic information extraction from newly discovered Web sources.
    - Harvest: uses a model of semi-structured documents.
    - Internet Learning Agent and Shopbot : learn about Web services.
  - Generalization: Uncover general patterns at individual and multiple sites
    - Relying on feedback from users to solve the labeling problem



Challenges to Web Mining

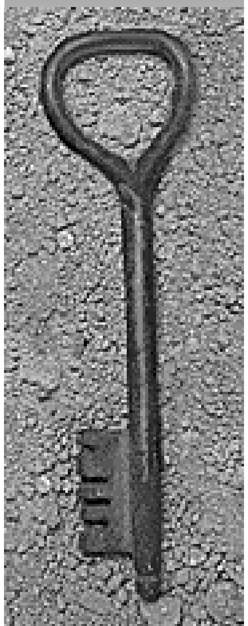
- Web: A huge, widely-distributed, highly heterogeneous, semi-structured, interconnected, evolving, hypertext/hypermedia information repository.
- Problems:
  - the "*abundance*" problem



  - *limited* coverage of the Web (hidden Web sources)
  - *limited* query interface: keyword-oriented search
  - *limited* customization to individual users
  - difficult to enforce *standards*
- DBMS, DBers, and data miners will play an increasingly important role in the new generation of Internet

Web Mining: Much Can Be Done!

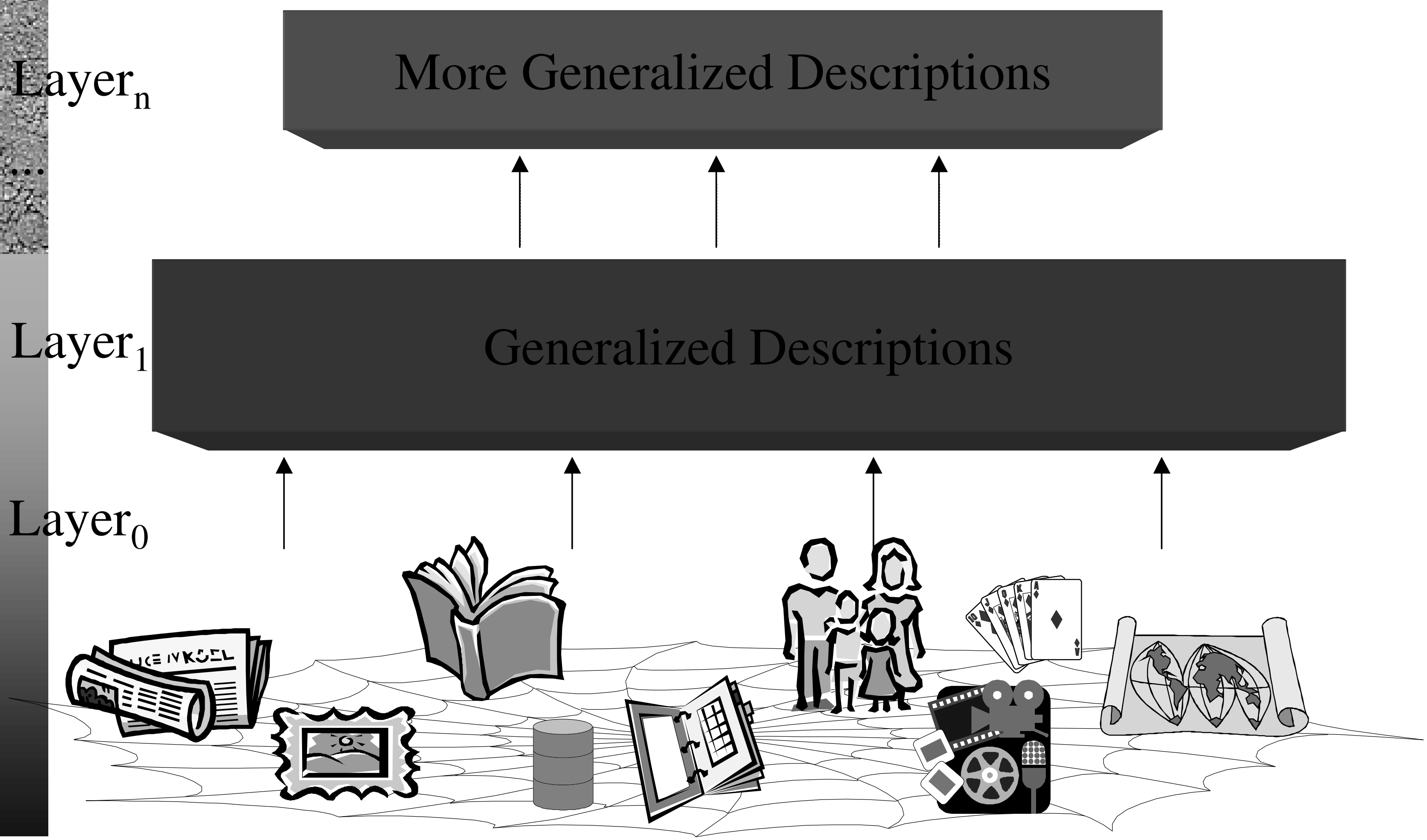
- A taxonomy of Web mining:
  - Web content mining
  - Web structure mining
  - Web usage mining
- Some interesting examples of Web mining
  - Mining integrated with Web search engines
  - Weblog mining (usage, access, and evolution)
  - Warehousing a Meta-Web: An MLDB approach

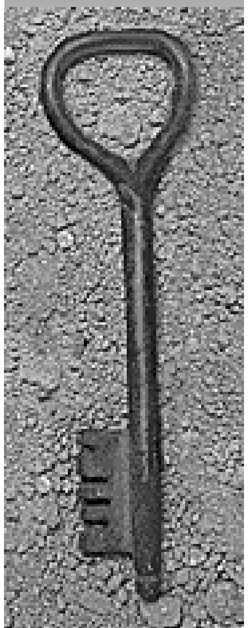


Web Log Mining

- Weblog provides rich information about Web dynamics
- Multidimensional Weblog analysis:
  - disclose potential customers, users, markets, etc.
- Plan mining (mining general Web accessing regularities):
  - Web linkage adjustment, performance improvements
- Web accessing association/sequential pattern analysis:
  - Web cashing, prefetching, swapping
- Trend analysis:
  - Dynamics of the Web: what has been changing?
- Customized to individual users

A Multiple Layered Meta-Web Architecture





Future

- Will traditional mining techniques (e.g., clustering, classification) be able to cope with scale, heterogeneity and dynamic nature of the Web?
  - New technologies:
- What key innovation will be required going forward?
  - Web warehouse
- Security restrictions