Unit 7 : Mining Complex Types of Data

Lecturer : Bijay Mishra

Introduction

Mining complex types of data include:

- Object data
- Spatial data
- Multimedia data
- Time-series data
- Text data
- Web data

Spatial Data Mining

- **Spatial data mining** is the process of discovering interesting, useful, non-trivial patterns from large spatial datasets.
- Spatial Data Mining = Mining Spatial Data Sets (i.e. Data Mining + Geographic Information Systems)
- Spatial data refer to any data about objects that occupy real physical space.
- Attributes for spatial data usually will include spatial information. Spatial information (metadata) is used to describe objects in space.
- Spatial information includes geometric metadata (e.g., location, shape, size, distance, area, perimeter) and topological metadata (e.g., “neighbor of”, “adjacent to”, “included in”, “includes”).
- Spatial data can contain both spatial and non-spatial features. Spatial data has location or geo-referenced features like:
  - Address, latitude/longitude (explicit)
  - Location-based partitions in databases (implicit)

Spatial Data Mining

Spatial Data Warehouse is an integrated, subject-oriented, time-variant, and nonvolatile spatial data repository for data analysis and decision making.

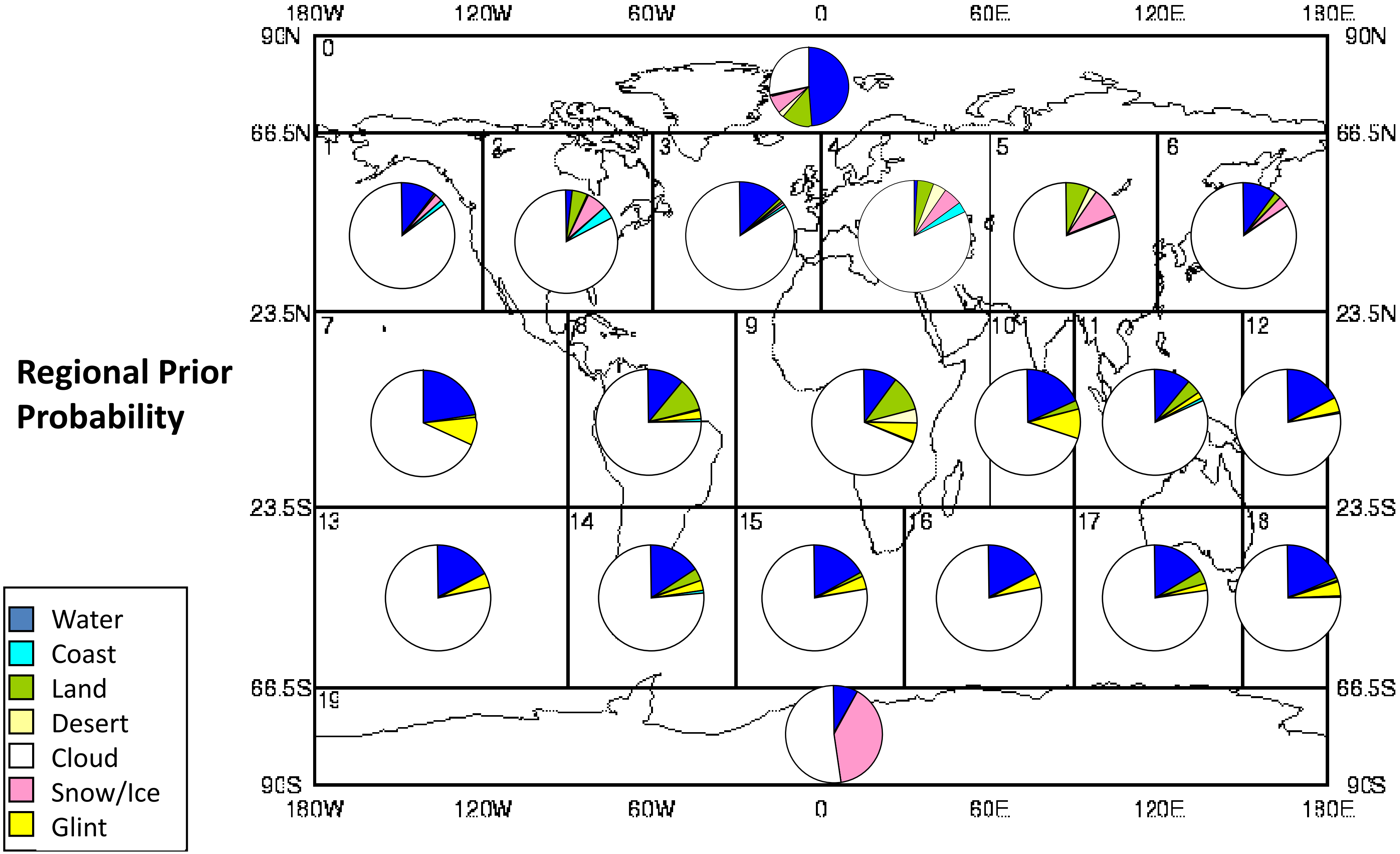
Spatial Data Integration is a big issue. It deals with:

- Structure-specific formats (raster vs. vector-based, Object-Oriented vs. relational models, different storage and indexing, etc.)
- Vendor-specific formats (ESRI, MapInfo, Integraph, etc.)

Spatial Data Cube is a multidimensional spatial database where both dimensions and measures may contain spatial components

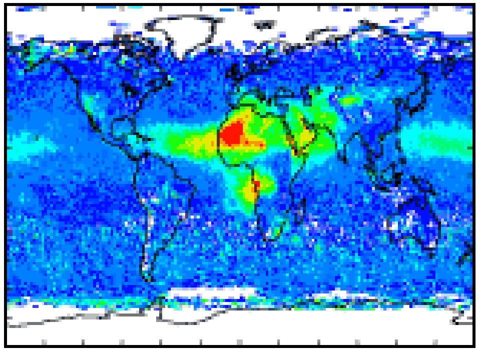
Spatial Data Mining

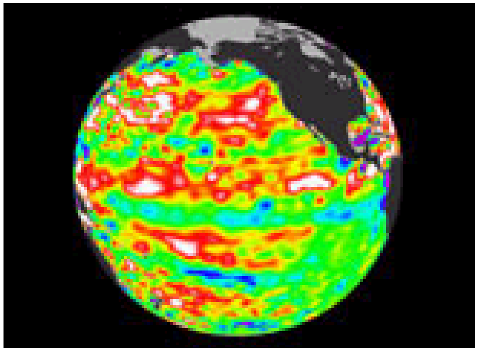
Regional Prior Probability

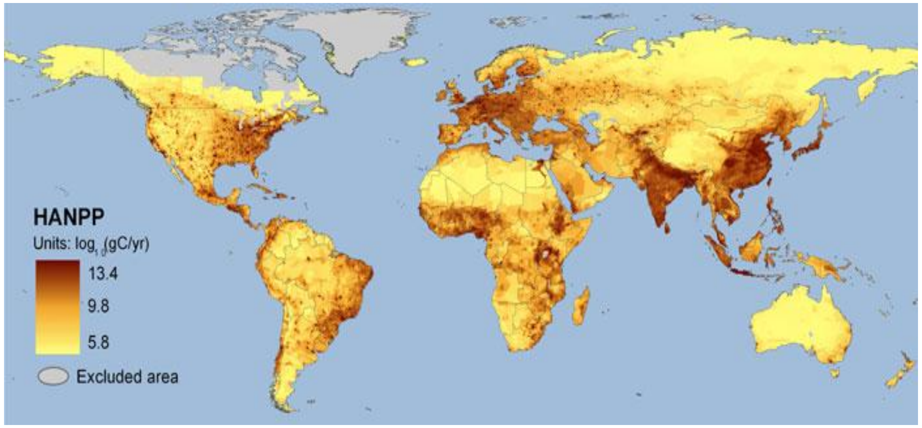


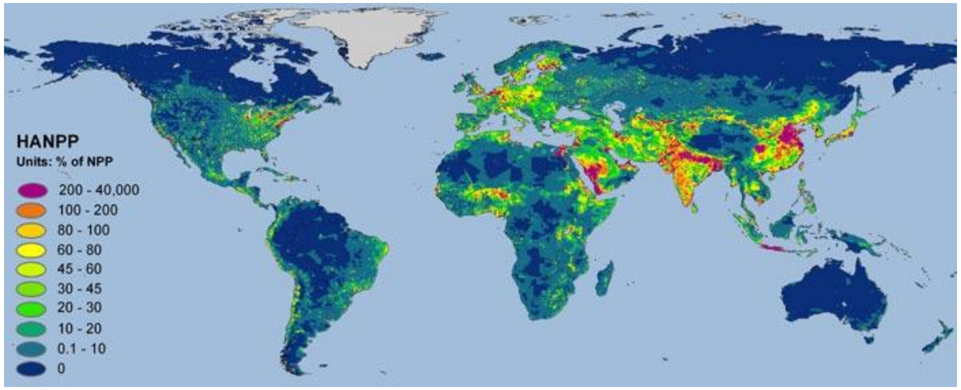
Special Cases

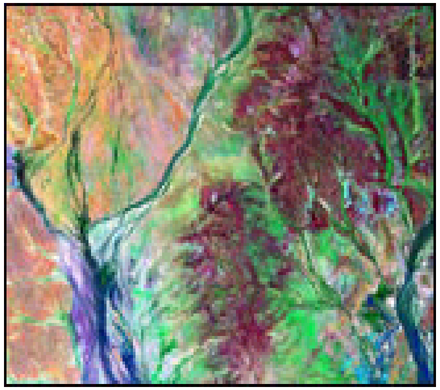
Image databases (Earth or the Sky)

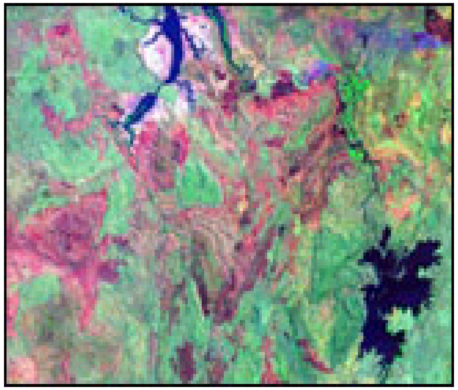












Thematic maps (values of attributes or "themes" are displayed in a spatial distribution = a map!)

Spatial Classification and Spatial Trend Analysis

❖ Spatial Classification

- Analyze spatial objects to derive classification schemes, such as decision trees in relevance to certain spatial properties (district, highway, river, etc.)
- *Example:* Classify regions in a province into *rich* vs. *poor* according to the average family income

❖ Spatial Trend Analysis

- Detect changes and trends along a spatial dimension
- Study the trend of non-spatial or spatial data changing with space
- *Example:* Observe the trend of changes of the climate or vegetation with the increasing distance from an ocean

Common Tasks dealing with Spatial Data

- **Data focusing**
  - Spatial queries
  - Identifying interesting parts in spatial data
  - Progressive refinement can be applied in a tree structure
- **Feature extraction**
  - Extracting important/relevant features for an application
- **Classification or others**
  - Using training data to create classifiers
  - Many mining algorithms can be used
    - Classification, clustering, associations

Spatial Mining Tasks

- Spatial classification
- Spatial clustering
- Spatial association rules

Spatial Classification

- ❖Use spatial information at different (coarse/fine) levels (different indexing trees) for data focusing
- ❖Determine relevant spatial or non-spatial features
- ❖Perform conventional supervised learning algorithms
  - – e.g., Decision trees,

Spatial Clustering

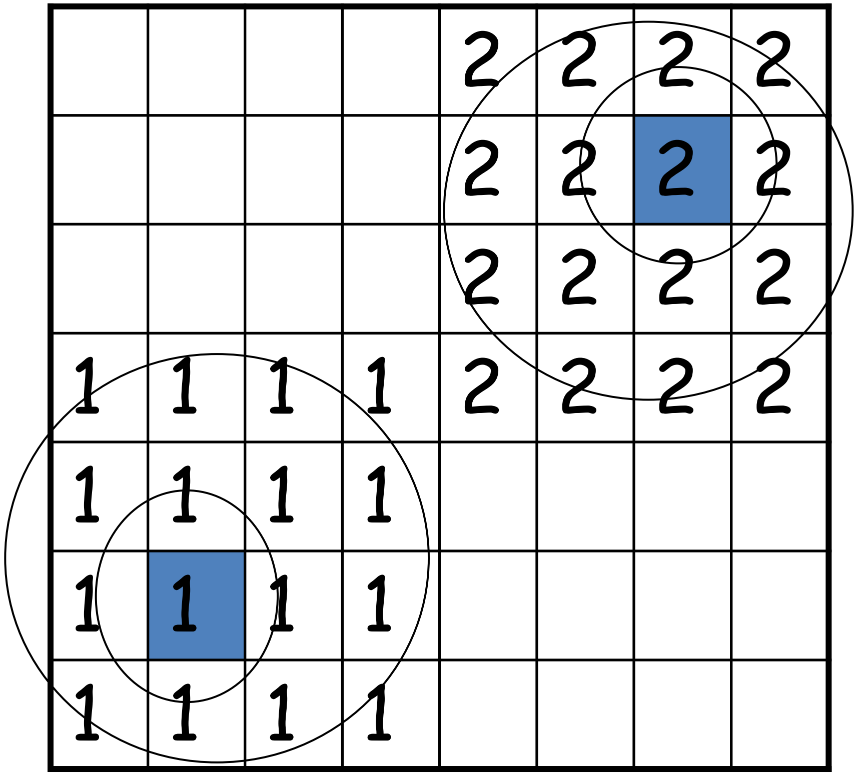
- ❖ Also called **spatial segmentation**
- ❖ Use tree structures to index spatial data
- ❖ Examples: DBSCAN: R-tree, CLIQUE: Grid or Quad tree, etc.

Input

- – a table of area names and their corresponding attributes such as population density, number of adult illiterates etc.
- – Information about the neighbourhood relationships among the areas
- – A list of categories/classes of the attributes

Output

- – Grouped (segmented) areas where each group has areas with similar attribute values

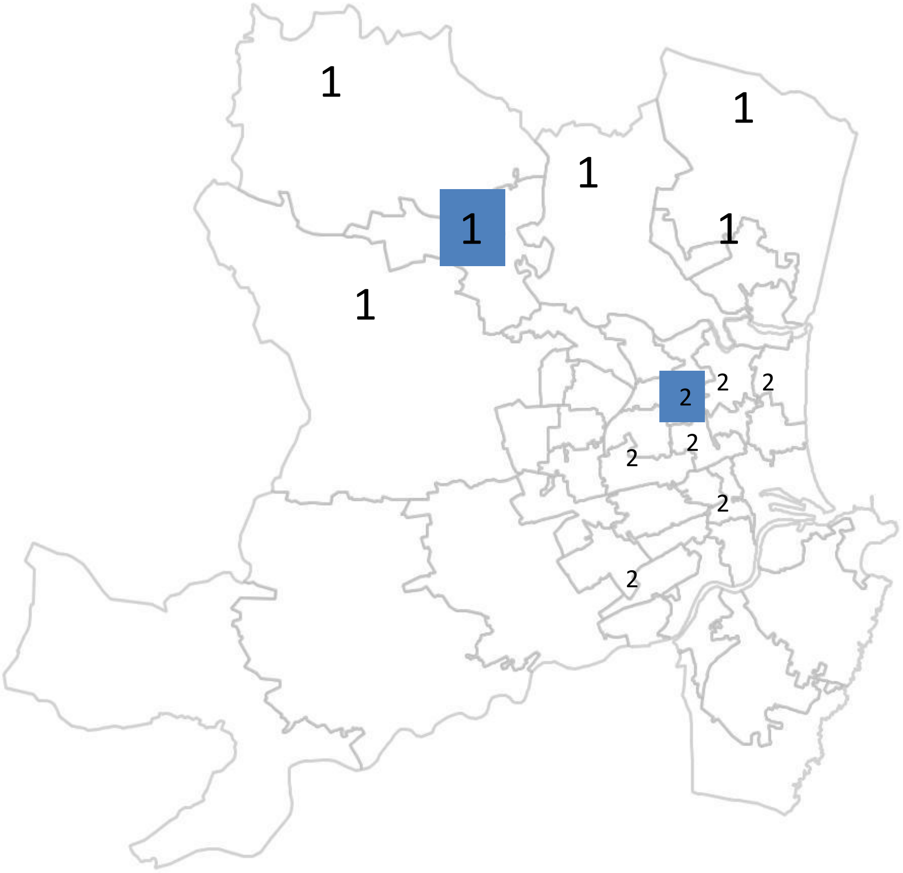


				2	2	2	2
				2	2	2	2
				2	2	2	2
1	1	1	1	2	2	2	2
1	1	1	1				
1	1	1	1				
1	1	1	1				

- Spatial segmentation is performed in image processing
  - Identify regions (areas) of an image that have similar colour (or other image attributes).
  - Many image segmentation techniques are available
    - E.g. region-growing technique

Region Growing Technique

- There are many flavours of this technique
- One of them is described below:
  - Assign seed areas to each of the segments (classes of the attribute)
  - Add neighbouring areas to these segments if the incoming areas have similar values of attributes
  - Repeat the above step until all the regions are allocated to one of the segments
- Functionality to compute spatial relations i.e. neighbours are assumed.



Spatial Association Rules

- Spatial objects are of major interest, not transactions

$$A \Rightarrow B$$

A, B can be either spatial or non-spatial

Multimedia Data Mining

Multimedia Data Mining is a subfield of data mining that deals with an extraction of implicit knowledge, multimedia data relationships, or other patterns not explicitly stored in multimedia databases

Multimedia Data Types

- any type of information medium that can be represented, processed, stored and transmitted over network in digital form
- Multi-lingual text, numeric, images, video, audio, graphical, temporal, relational, and categorical data.
- Relation with conventional data mining term

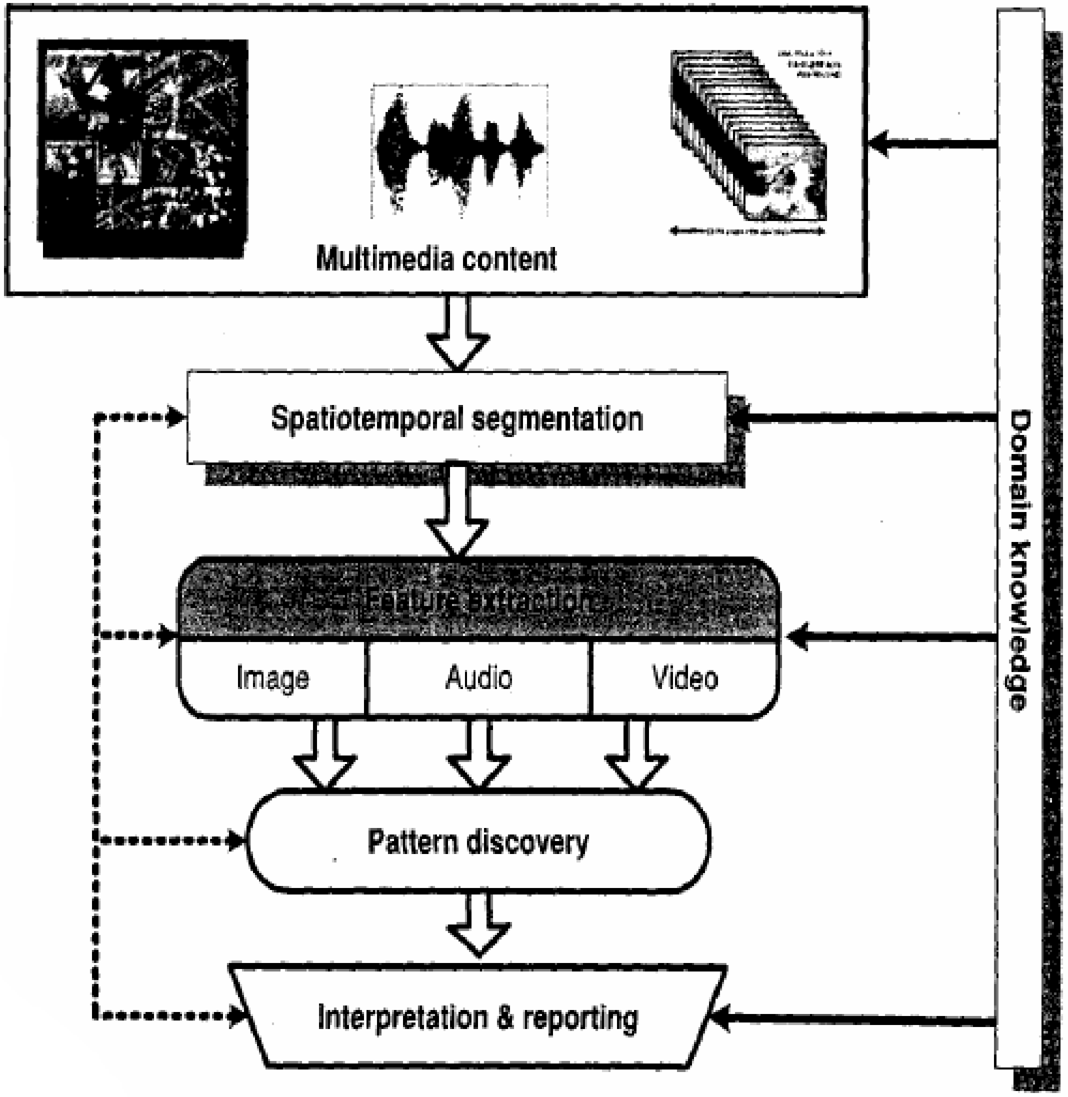


Figure: Multimedia Data Mining Architecture

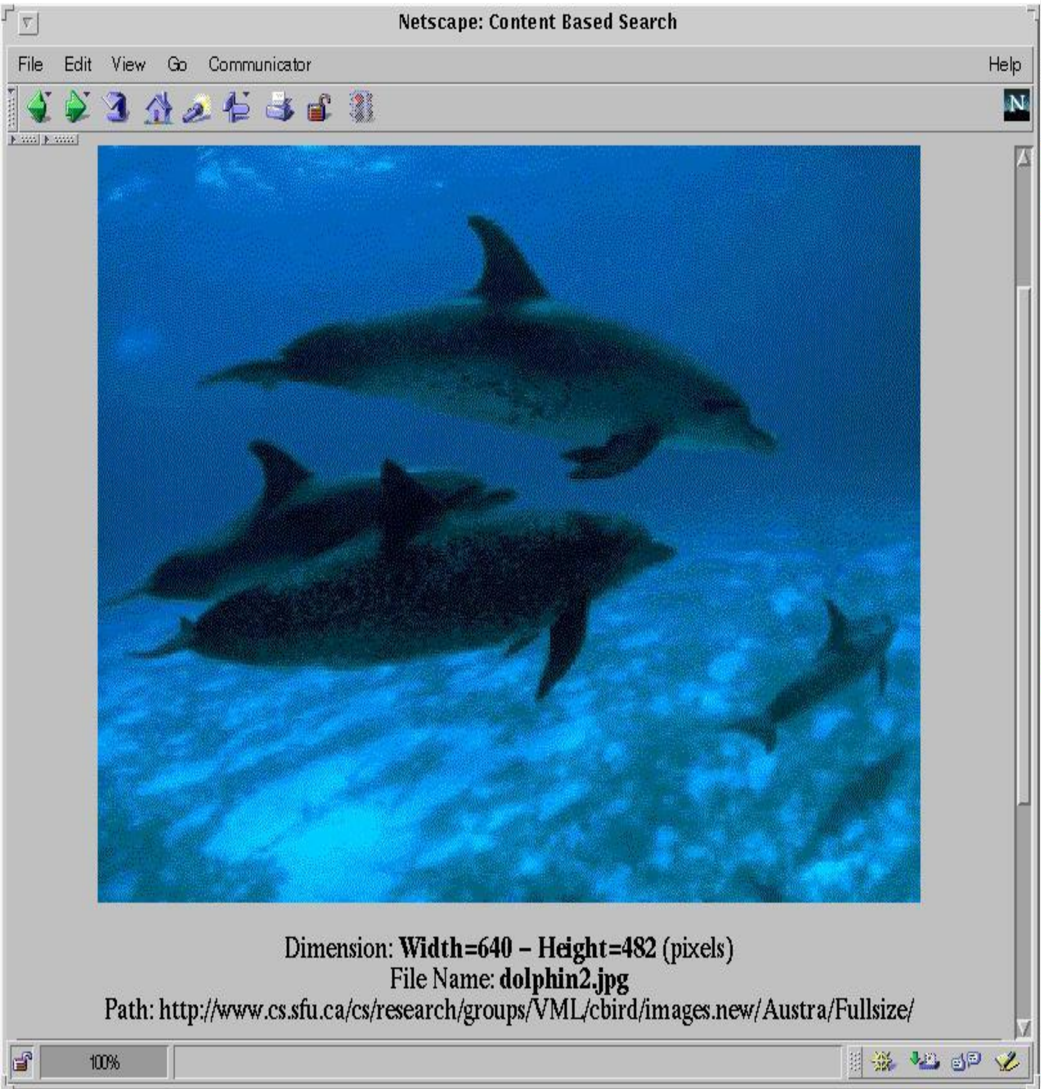
Generalizing Multimedia Data

- **Image data:**
  - Extracted by aggregation and/or approximation
  - Size, color, shape, texture, orientation, and relative positions and structures of the contained objects or regions in the image
- **Music data:**
  - **Summarize its melody**: based on the approximate patterns that repeatedly occur in the segment
  - **Summarized its style**: based on its tone, tempo, or the major musical instruments played
- **Video:**
  - provide news video annotation and indexing
  - traffic monitoring system

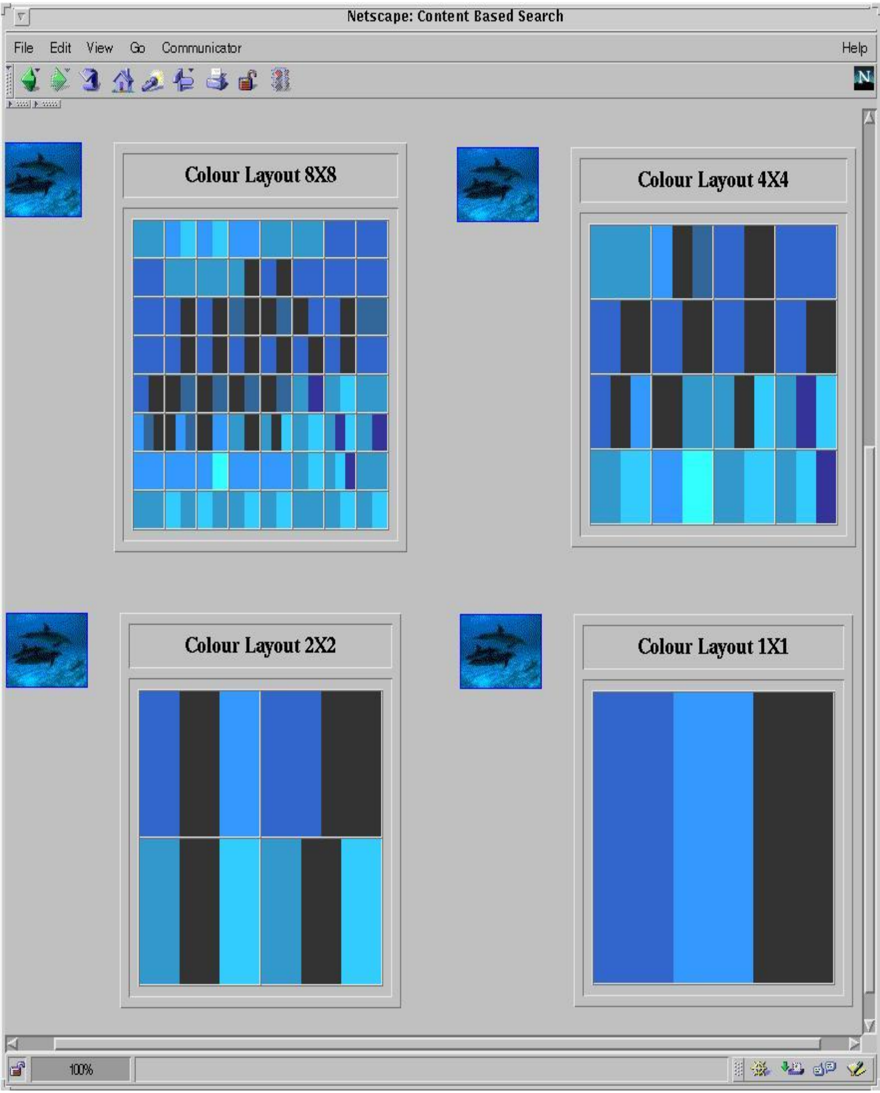
Multidimensional Analysis of Multimedia Data

- **Multimedia Data Cube**
  - Design and construction similar to that of traditional data cubes from relational data
  - Contain additional dimensions and measures for multimedia information, such as color, texture, and shape
- The database does not store images but their descriptors.
  - **Feature descriptor**: a set of vectors for each visual characteristic
    - Color vector: contains the color histogram
    - MFC (Most Frequent Color) vector: five color centroids
    - MFO (Most Frequent Orientation) vector: five edge orientation centroids
  - **Layout descriptor**: contains a color layout vector and an edge layout vector

Multi-Dimensional Search in Multimedia Databases

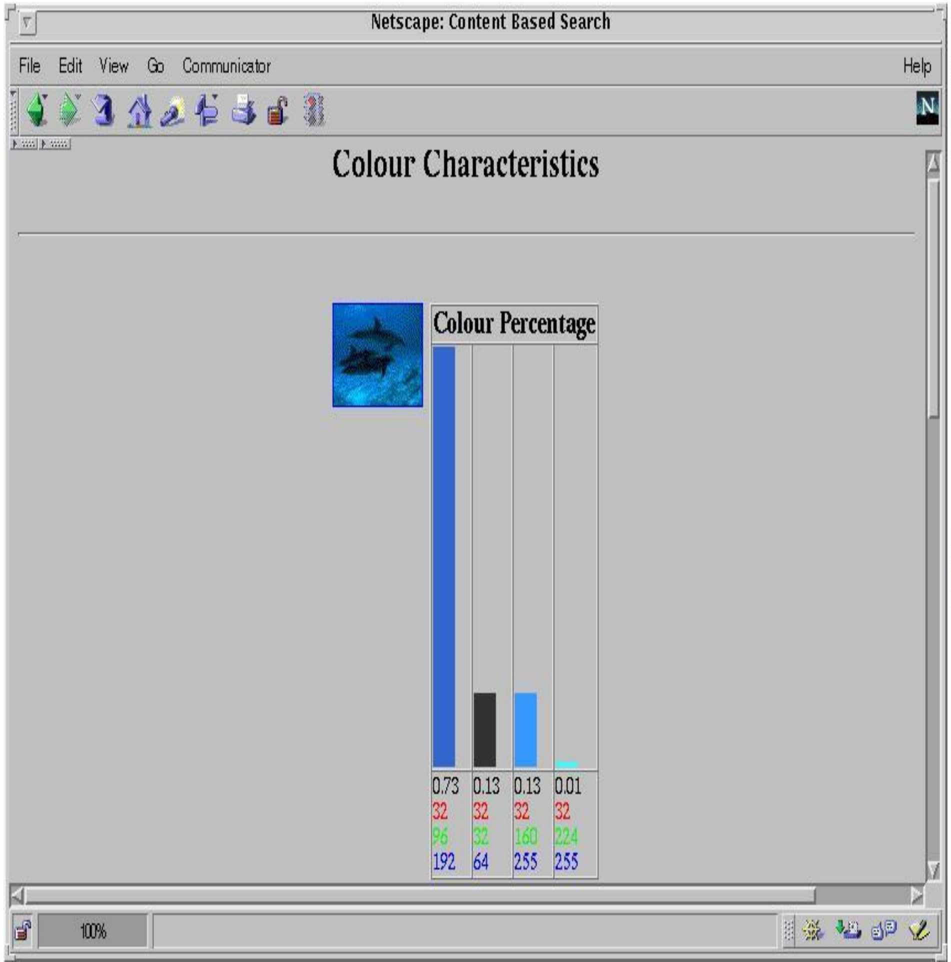

Netscape: Content Based Search
File Edit View Go Communicator Help
100%

Dimension: **Width=640 – Height=482** (pixels)
File Name: **dolphin2.jpg**
Path: http://www.cs.sfu.ca/cs/research/groups/VML/cbird/images.new/Austra/Fullsize/

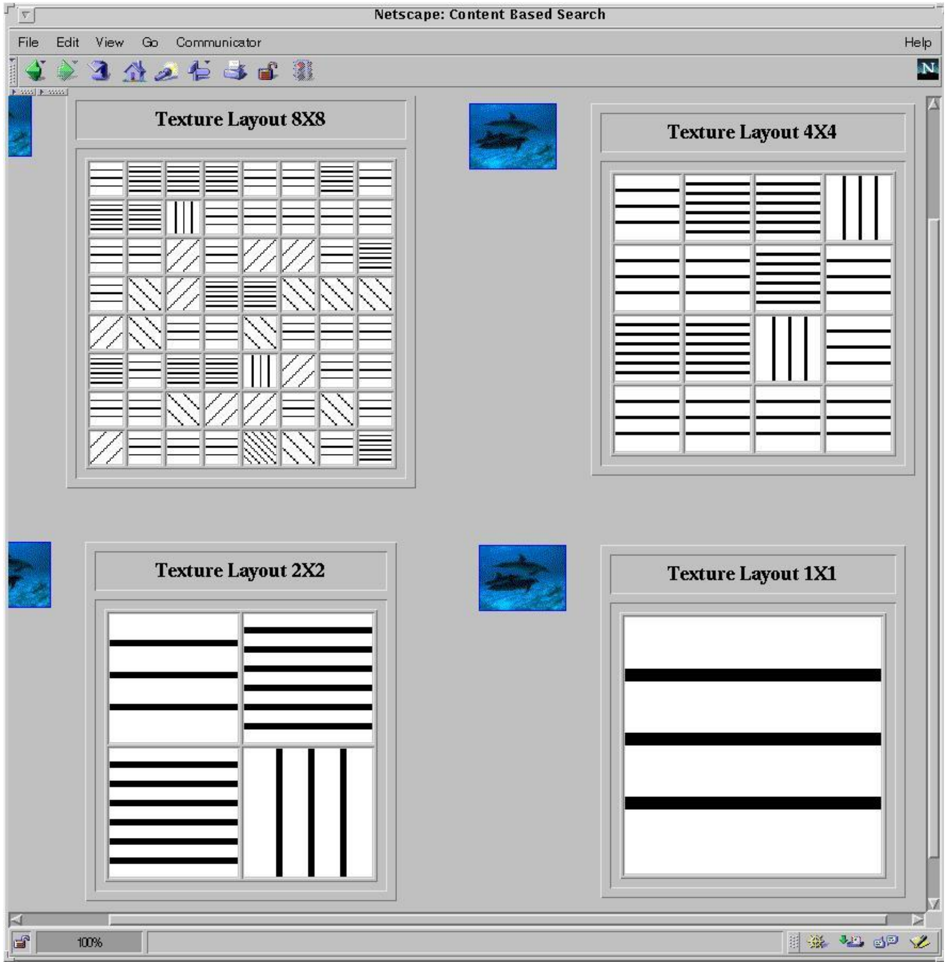


Multi-Dimensional Analysis in Multimedia Databases

Color histogram

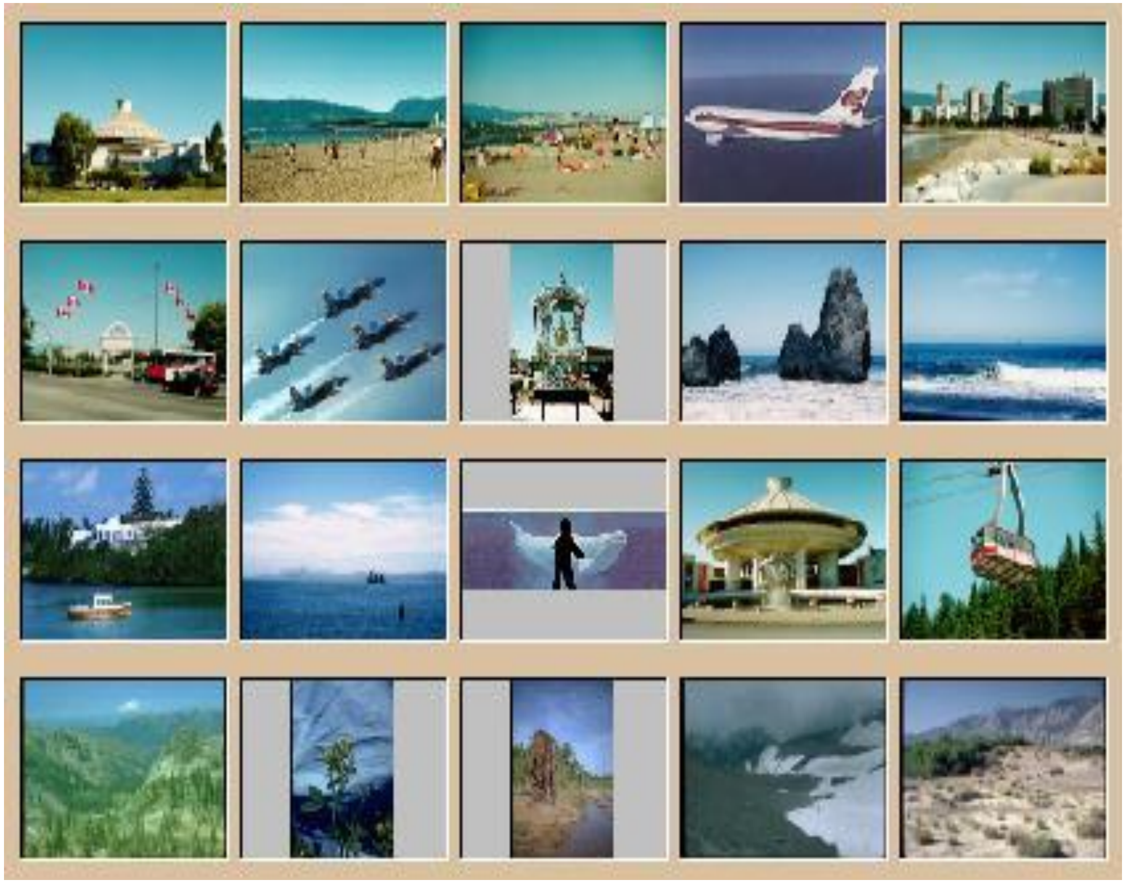


Texture layout



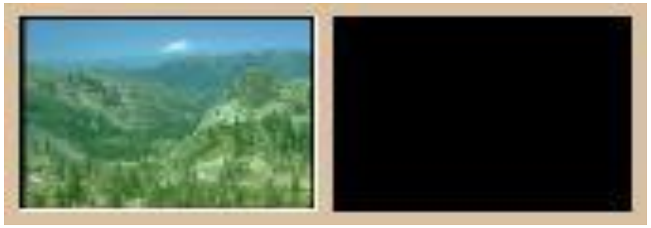
Mining Multimedia Databases

Refining or combining searches



Search for “blue sky” (top layout grid is blue)

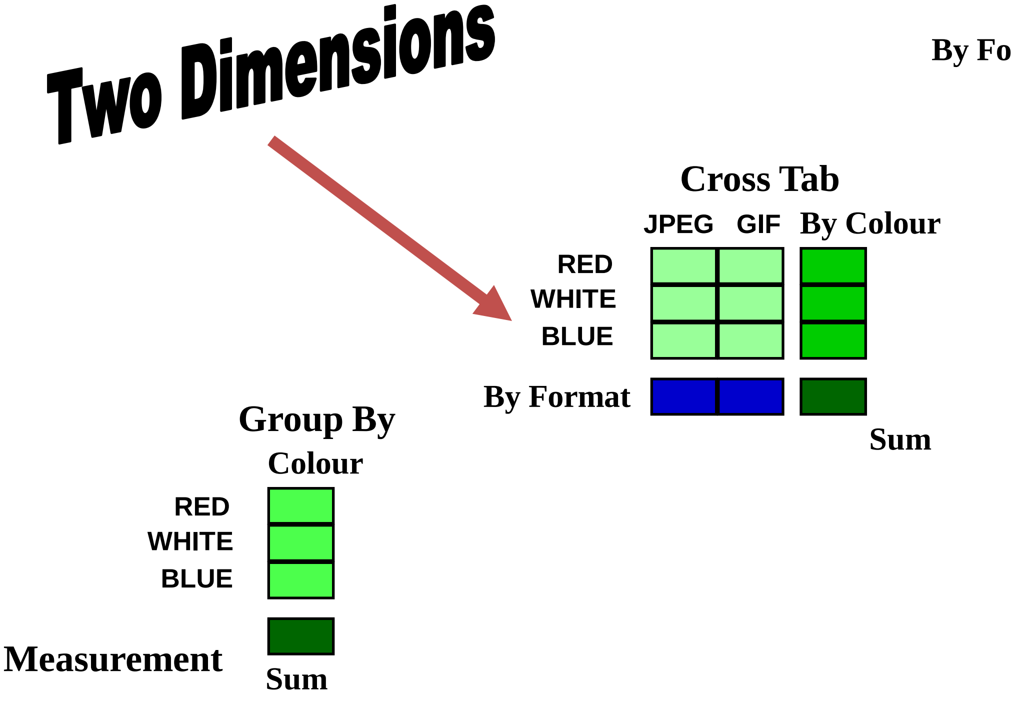
Search for “airplane in blue sky” (top layout grid is blue and keyword = “airplane”)



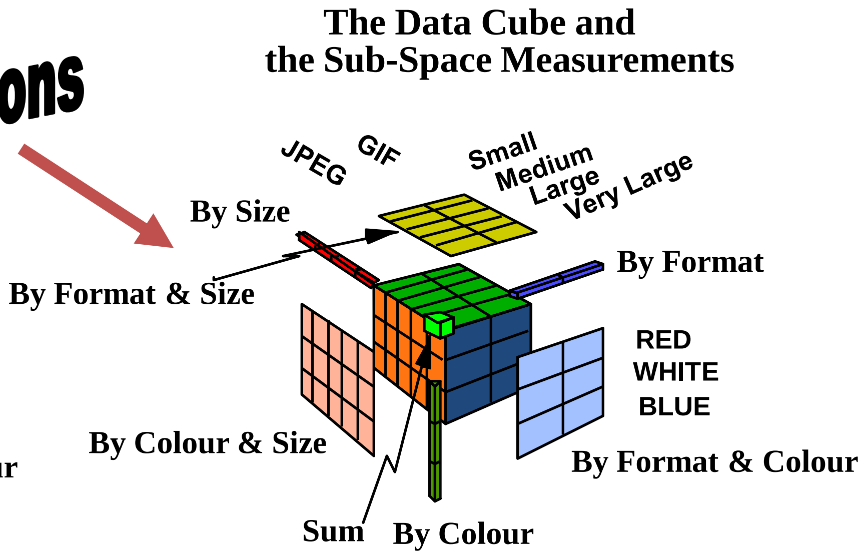
Search for “blue sky and green meadows” (top layout grid is blue and bottom is green)

Mining Multimedia Databases

Two Dimensions

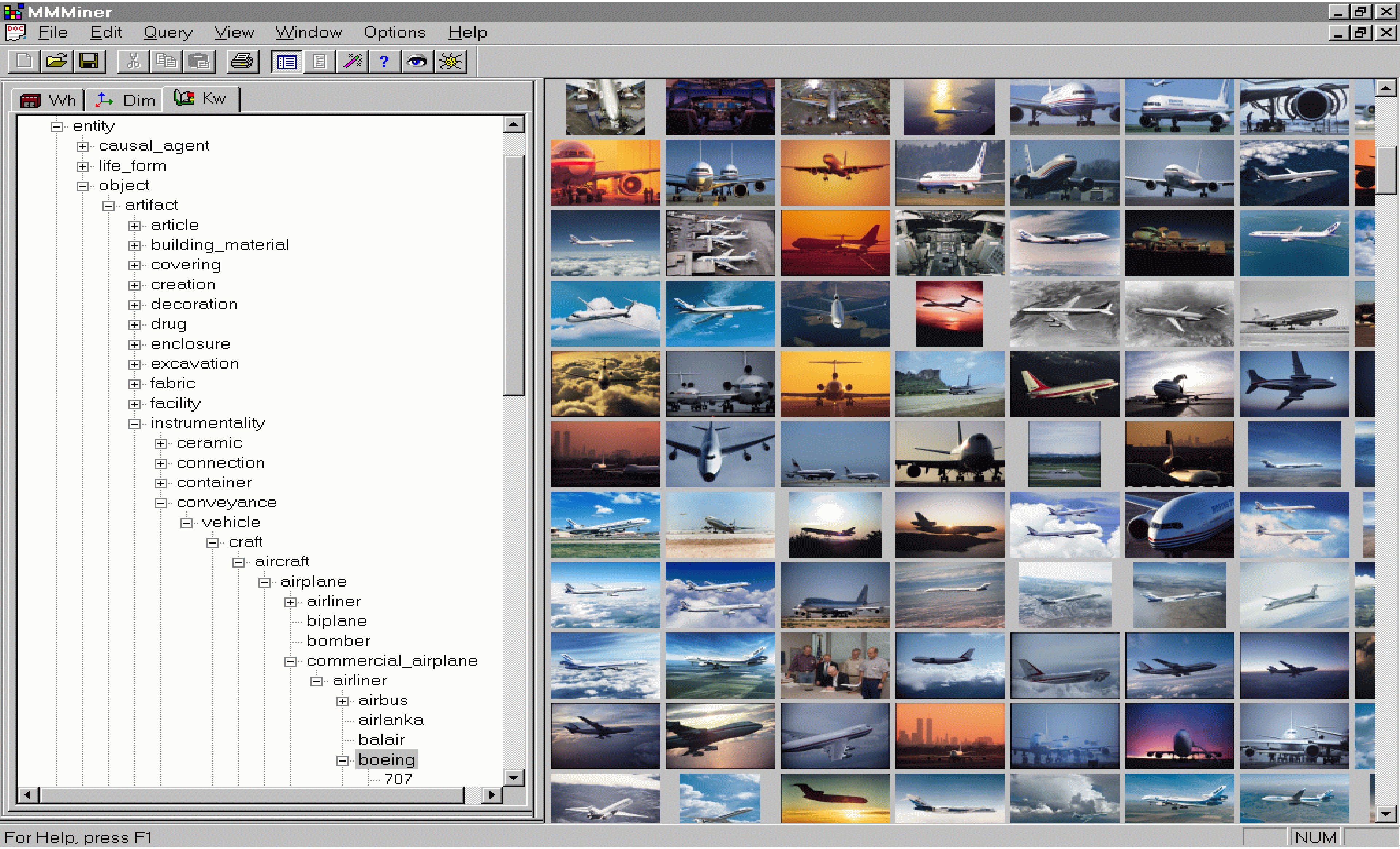


Three Dimensions



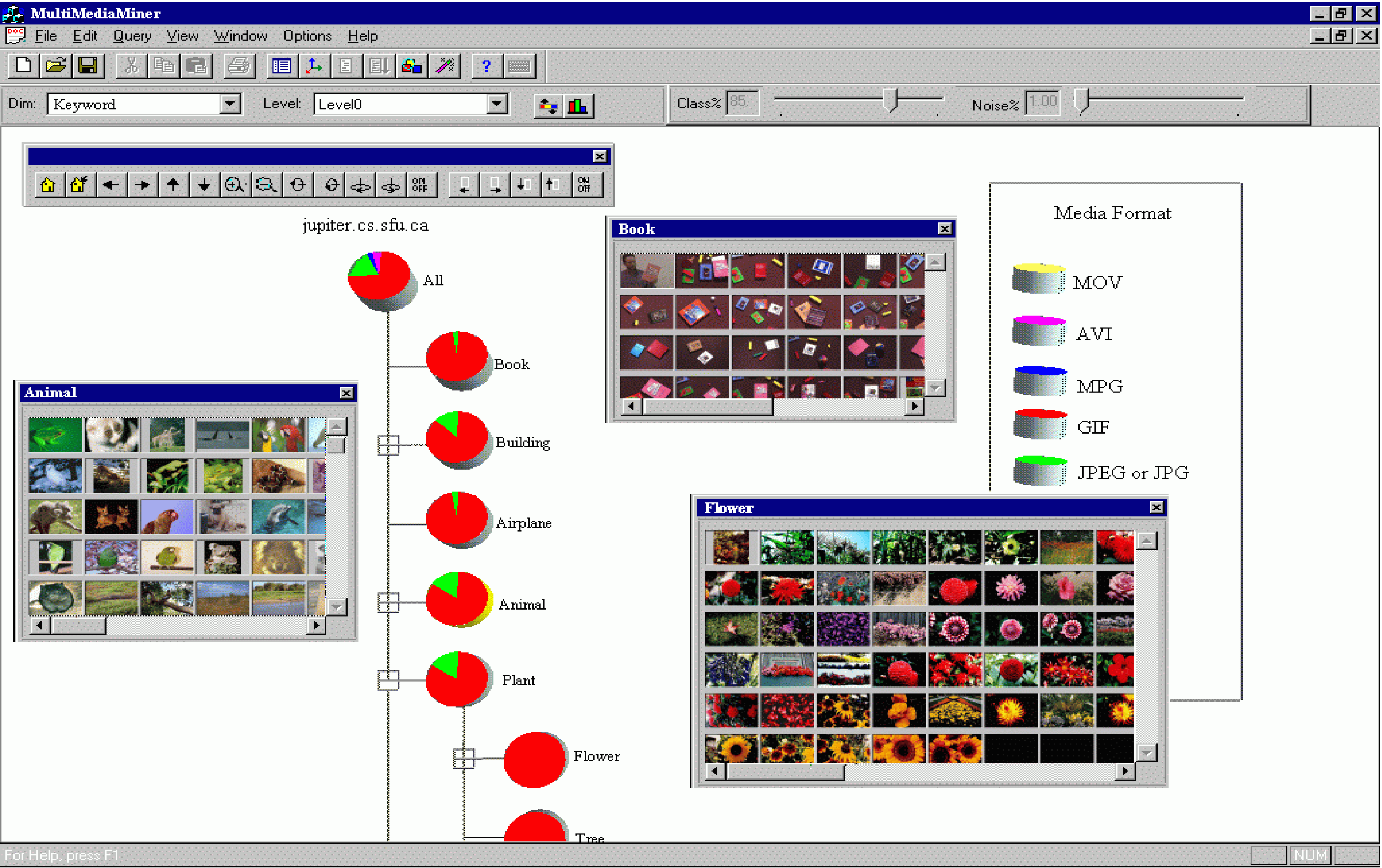
- ❖ Format of image
- ❖ Duration
- ❖ Colors
- ❖ Textures
- ❖ Keywords
- ❖ Size
- ❖ Width
- ❖ Height
- ❖ Internet domain of image
- ❖ Internet domain of parent pages
- ❖ Image popularity

Mining Multimedia Databases in MultimediaMiner



For Help, press F1

Classification in MultimediaMiner



Dim: Keyword

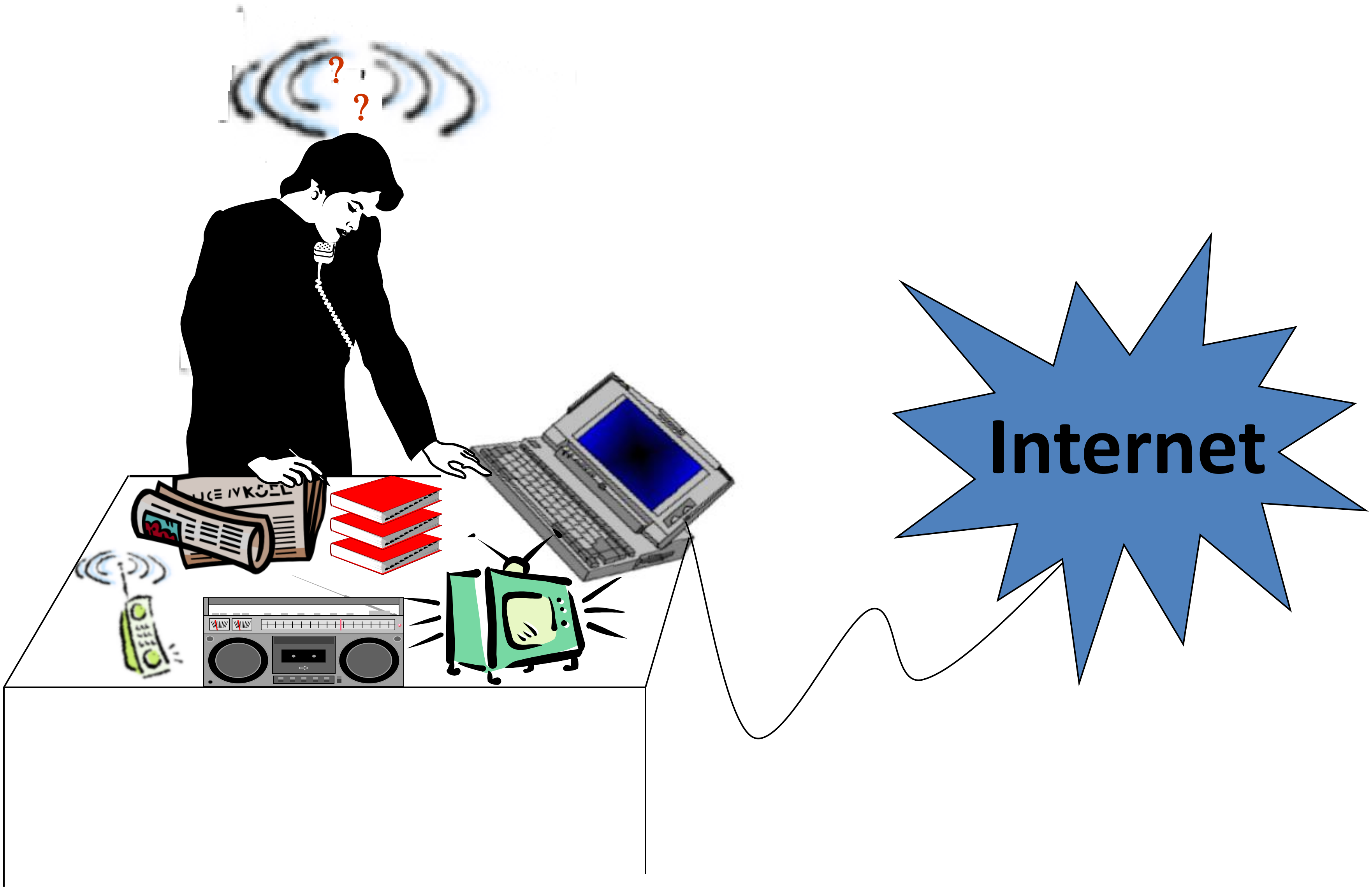
Level: Level0

Class% 85

Noise% 1.00

For Help, press F1

Text Mining



Text mining is the procedure of synthesizing information, by analyzing relations, patterns, and rules among textual data. These procedures contains text summarization, text categorization, and text clustering.

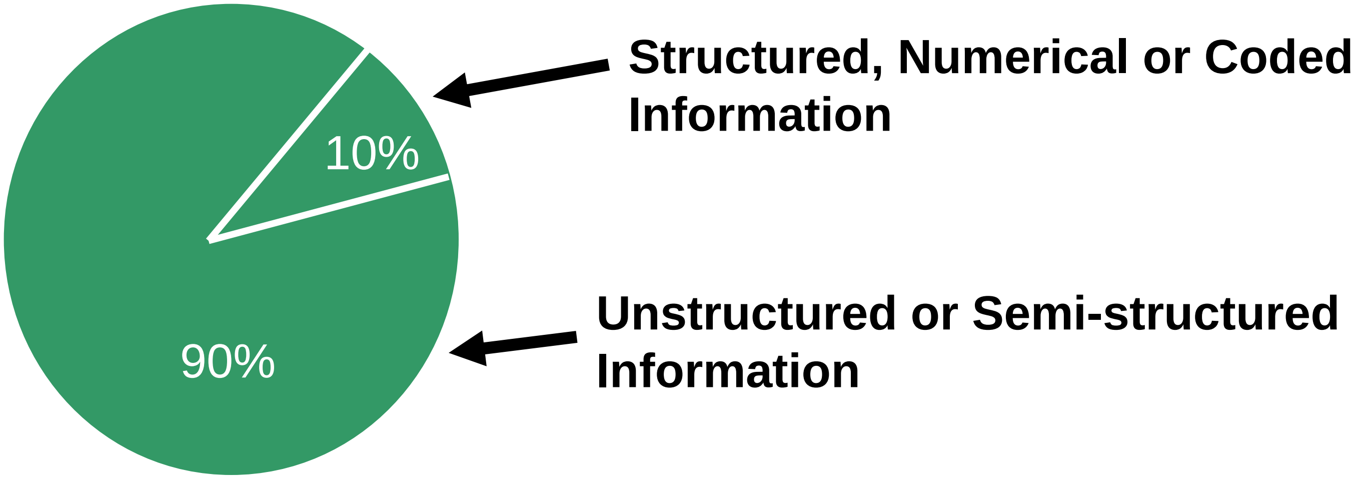
1. *Text summarization* is the procedure to extract its partial content reflecting its whole contents automatically.

2. *Text categorization* is the procedure of assigning a category to the text among categories predefined by users

3. *Text clustering* is the procedure of segmenting texts into several clusters, depending on the substantial relevance.

Motivation for Text Mining

- Approximately **90%** of the world's data is held in unstructured formats (Source: Oracle Corporation)
- Information intensive business processes demand that we transcend from simple document retrieval to "knowledge" discovery.



Structured, Numerical or Coded Information

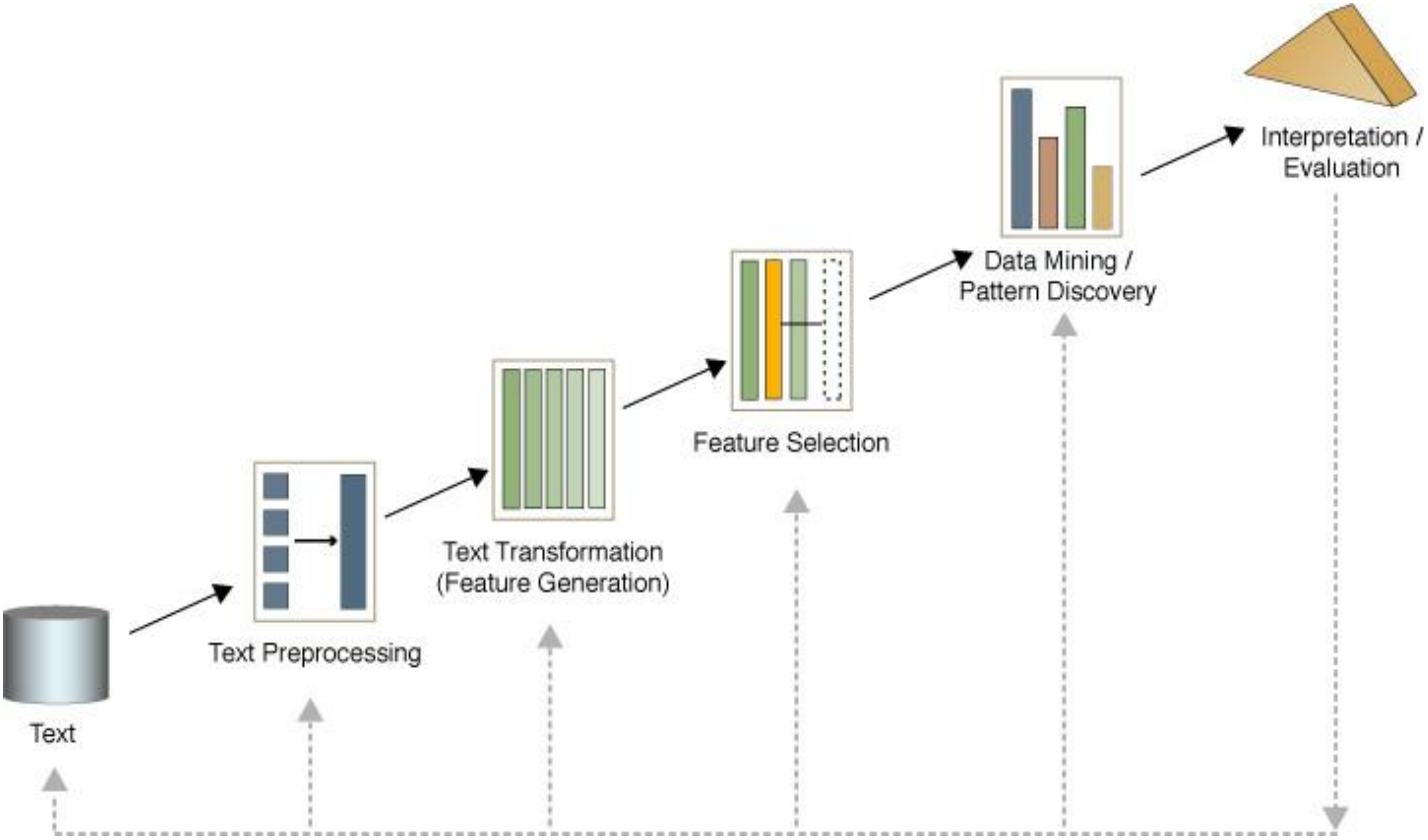
Unstructured or Semi-structured Information

Text mining is well motivated, due to the fact that much of the world’s data can be found in free text form (newspaper articles, emails, literature, etc.). There is a lot of information available to mine.

While mining free text has the same goals as data mining, in general, extracting useful knowledge/stats/trends), text mining must overcome a major difficulty – there is no explicit structure.

Machines can reason will relational data well since schemas are explicitly available. Free text, however, encodes all semantic information within natural language. Our text mining algorithms, then, must make some sense out of this natural language representation. Humans are great at doing this, but this has proved to be a problem for machines.

Text Mining Process



Text

Text Preprocessing

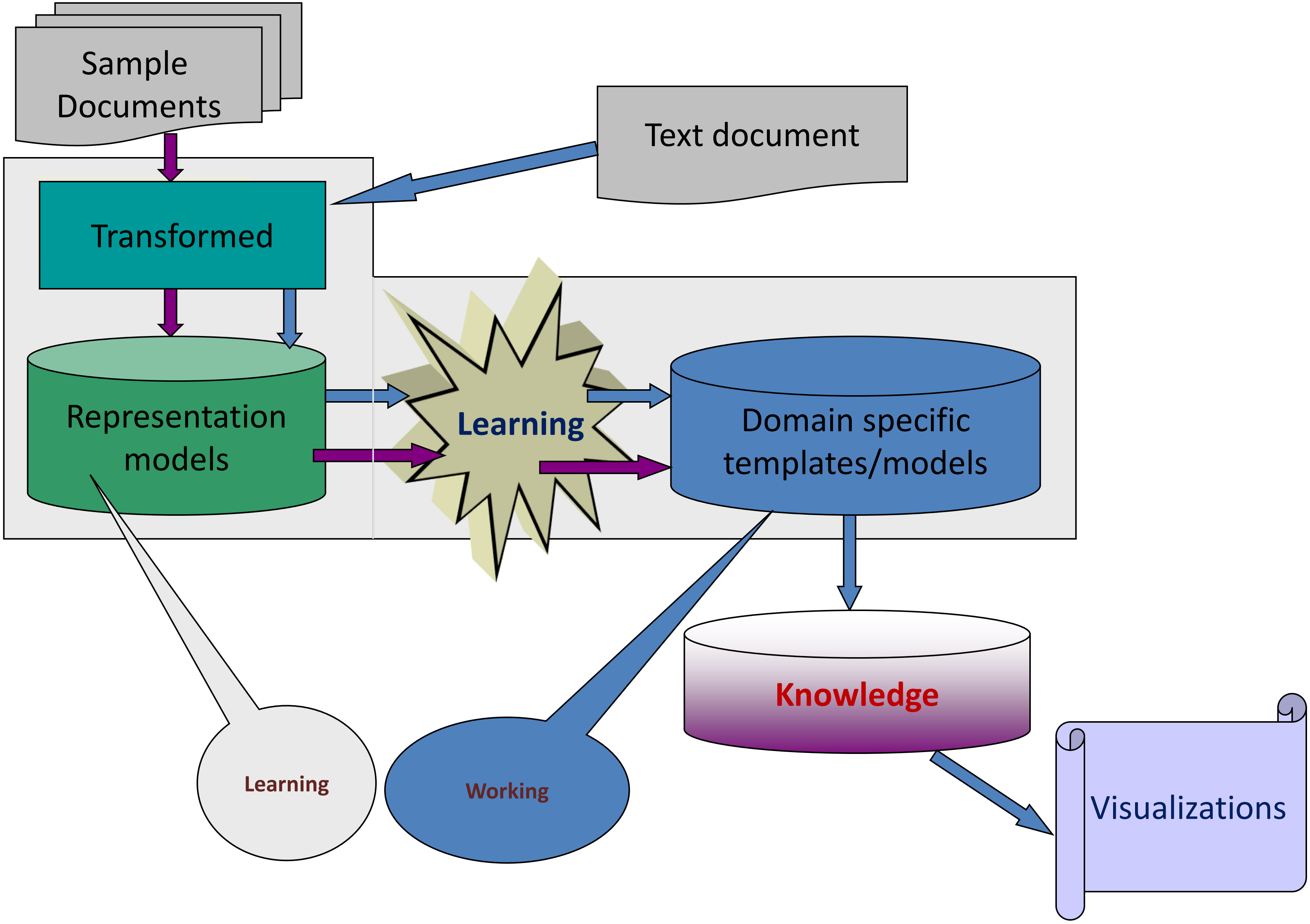
Text Transformation (Feature Generation)

Feature Selection

Data Mining / Pattern Discovery

Interpretation / Evaluation

What's Text Mining



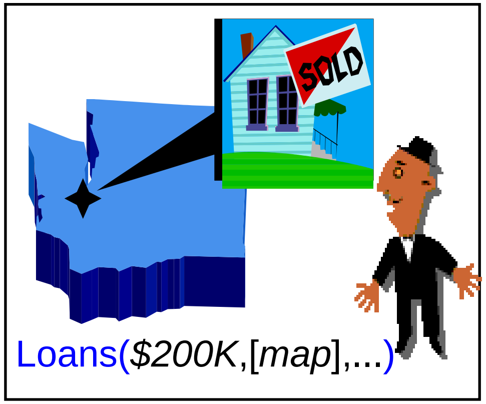
Mining Text Data: An Introduction

Data Mining / Knowledge Discovery

Structured Data

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

Multimedia



Loans(*$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* </a> *Bought* <a hef>*this home*</a> *from* <a href>*Lake View Real Estate*</a> *In* <b>*1992*</b>*.* <p>...

Text Representation Issues

- Each word has a dictionary meaning, or meanings
  - **Run** – (1) the verb. (2) the noun, in **cricket**
  - **Cricket –** (1) The game. (2) The insect.
  - Apple (the company) or apple (the fruit)
- Ambiguity and context sensitivity - Each word is used in various “senses”
  - Tendulkar made 100 runs
  - Because of an injury, Tendulkar can not run and will need a runner between the wickets
- Capturing the “meaning” of sentences is an important issue as well. Grammar, parts of speech, time sense could be easy!
- ***Order*** of words in the query
  - hot dog stand in the amusement park
  - hot amusement stand in the dog park

Text Databases and IR

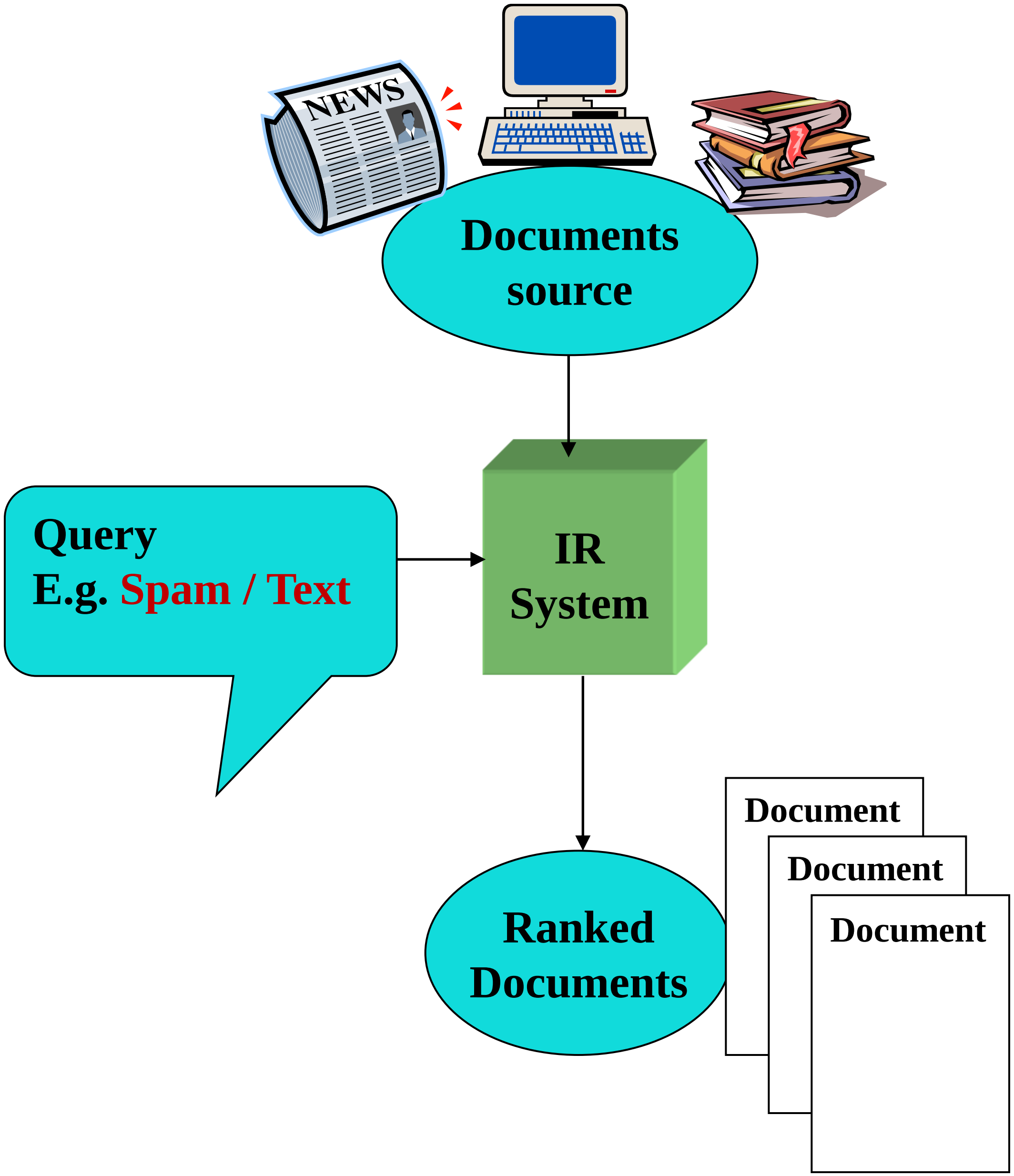
Text databases (document databases)

- 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*
- Traditional information retrieval techniques become inadequate for the increasingly vast amounts of text data

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



Documents source

Query
E.g. Spam / Text

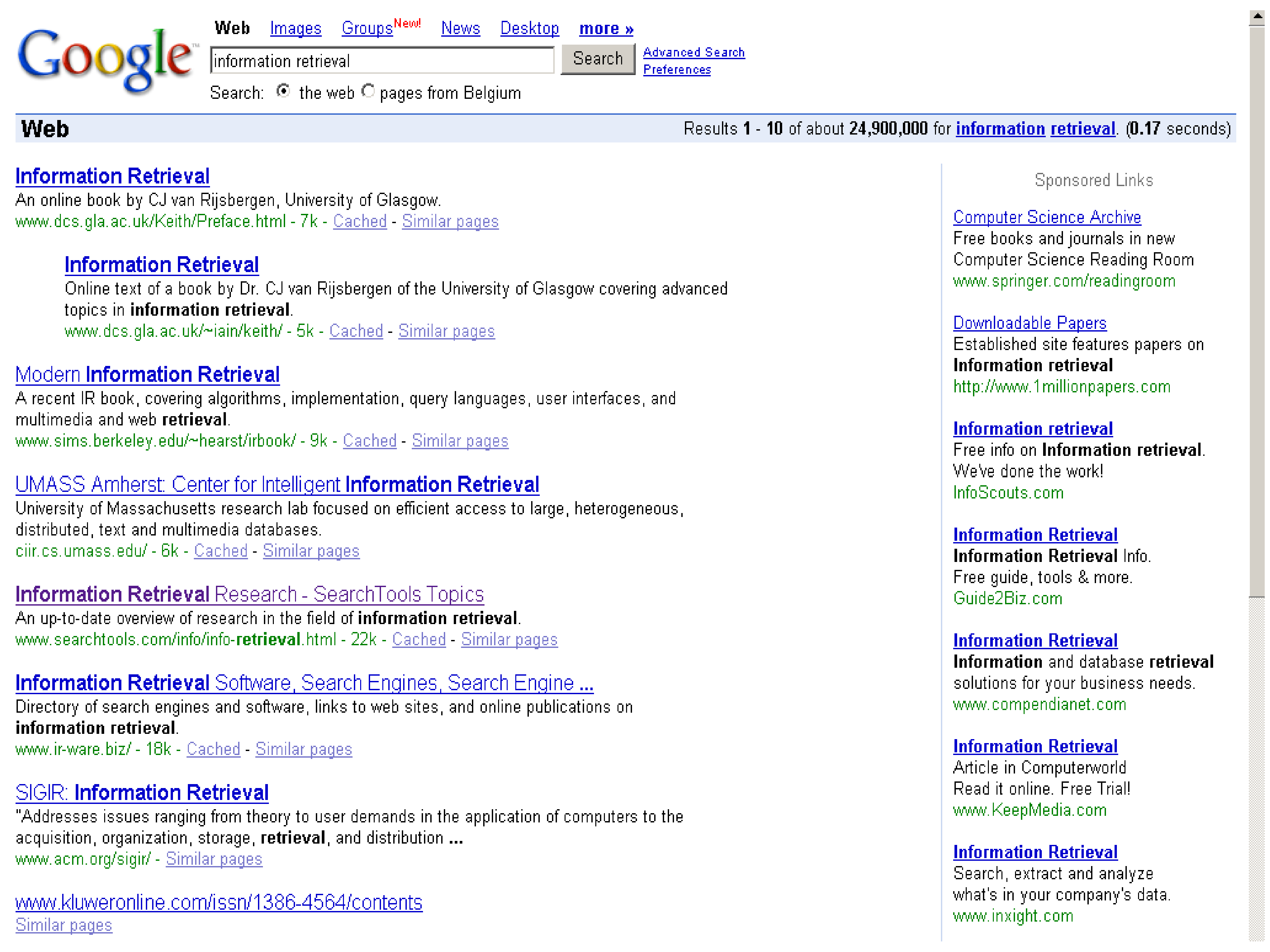
IR System

Ranked Documents

Document

Document

Document



information retrieval

Search: ● the web ○ pages from Belgium

Web Results **1** - **10** of about **24,900,000** for **information retrieval**. (**0.17** seconds)

Information Retrieval
An online book by CJ van Rijsbergen, University of Glasgow.
www.dcs.gla.ac.uk/Keith/Preface.html - 7k - Cached - Similar pages

Information Retrieval
Online text of a book by Dr. CJ van Rijsbergen of the University of Glasgow covering advanced topics in **information retrieval**.
www.dcs.gla.ac.uk/~iain/keith/ - 5k - Cached - Similar pages

Modern **Information Retrieval**
A recent IR book, covering algorithms, implementation, query languages, user interfaces, and multimedia and web **retrieval**.
www.sims.berkeley.edu/~hearst/irbook/ - 9k - Cached - Similar pages

UMASS Amherst: Center for Intelligent **Information Retrieval**
University of Massachusetts research lab focused on efficient access to large, heterogeneous, distributed, text and multimedia databases.
ciir.cs.umass.edu/ - 6k - Cached - Similar pages

Information Retrieval Research - SearchTools Topics
An up-to-date overview of research in the field of **information retrieval**.
www.searchtools.com/info/info-**retrieval**.html - 22k - Cached - Similar pages

Information Retrieval Software, Search Engines, Search Engine ...
Directory of search engines and software, links to web sites, and online publications on **information retrieval**.
www.ir-ware.biz/ - 18k - Cached - Similar pages

SIGIR: **Information Retrieval**
"Addresses issues ranging from theory to user demands in the application of computers to the acquisition, organization, storage, **retrieval**, and distribution **...**
www.acm.org/sigir/ - Similar pages

www.kluweronline.com/issn/1386-4564/contents
Similar pages

Sponsored Links

Computer Science Archive
Free books and journals in new Computer Science Reading Room
www.springer.com/readingroom

Downloadable Papers
Established site features papers on **Information retrieval**
http://www.1millionpapers.com

Information retrieval
Free info on **Information retrieval**. We've done the work!
InfoScouts.com

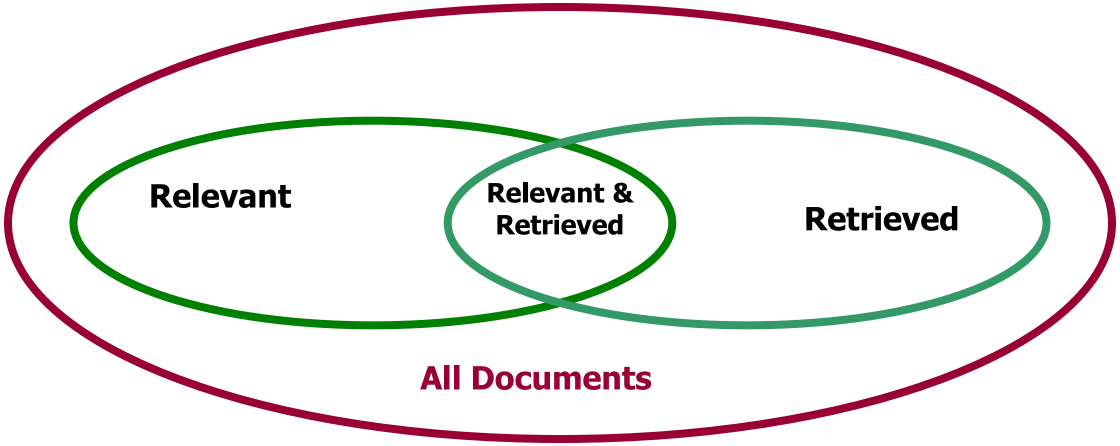
Information Retrieval
Information Retrieval Info. Free guide, tools & more.
Guide2Biz.com

Information Retrieval
Information and database **retrieval** solutions for your business needs.
www.compendianet.com

Information Retrieval
Article in Computerworld
Read it online. Free Trial!
www.KeepMedia.com

Information Retrieval
Search, extract and analyze what's in your company's data.
www.inxight.com

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

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

Application of Text Mining

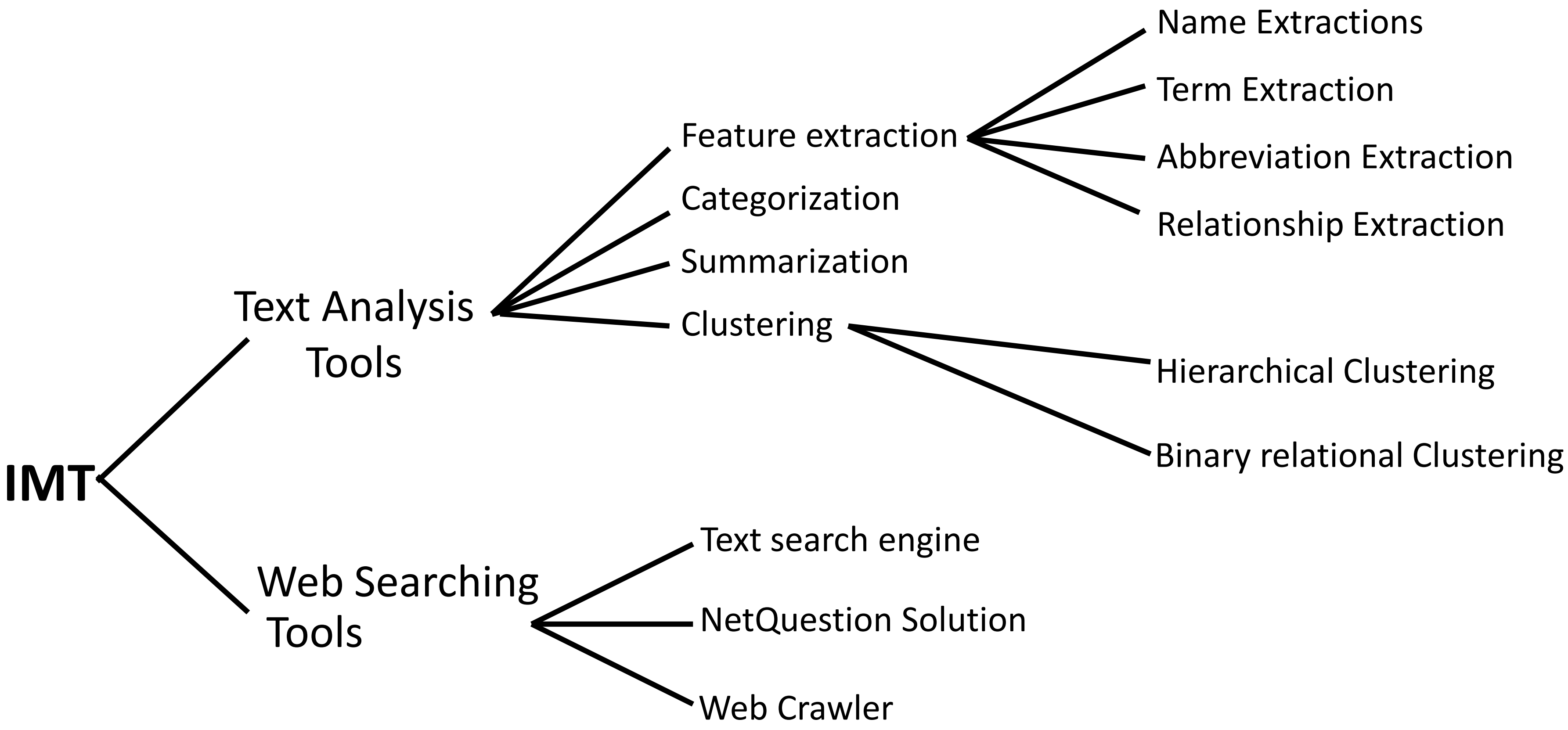
Text mining system provides a competitive edge for a company to process and take advantage of a large quantity of textual information. The potential applications are countless. We highlight a few below.

- Customer profile analysis, e.g., mining incoming emails for customers' complaint and feedback.
- Patent analysis, e.g., analyzing patent databases for major technology players, trends, and opportunities.
- Information dissemination, e.g., organizing and summarizing trade news and reports for personalized information services.
- Company resource planning, e.g., mining a company's reports and correspondences for activities, status, and problems reported.

Text Mining vs. Data Mining

	Data Mining	Text Mining
Data Object	Numerical & categorical data	Textual data
Data structure	Structured	Unstructured &semi-structured
Data representation	Straightforward	Complex
Space dimension	< tens of thousands	> tens of thousands
Methods	Data analysis, machine learning, Statistic, neural networks	Data mining, information retrieval, NLP, ...
Maturity	Broad implementation since1994	Broad implementation starting 2000
Market	10^5 analysts at large and mid size companies	10^8 analysts corporate workers and individual users

Product : Intelligent Miner for Text(IMT)



- IMT
  - Text Analysis Tools
    - Feature extraction
      - Name Extractions
      - Term Extraction
      - Abbreviation Extraction
      - Relationship Extraction
    - Categorization
    - Summarization
    - Clustering
      - Hierarchical Clustering
      - Binary relational Clustering
  - Web Searching Tools
    - Text search engine
    - NetQuestion Solution
    - Web Crawler

1. Feature extraction tools

It recognizes significant **vocabulary** items in documents, and measures their importance to the document content.

2. Clustering tools

Clustering is used to segment a document collection into subsets, called **clusters**.

3. Summarization tools

Summarization is the process of **condensing a source text** into a shorter version **preserving its information content**.

4. Categorization tool

Categorization is used to assign objects to **predefined categories**, or **classes** from a taxonomy.

1. Feature Extraction Tools

1.1 Information extraction

- ❖ Extract linguistic items that represent document contents

1.2 Feature extraction

- ❖ Assign of different categories to vocabulary in documents,
- ❖ Measure their importance to the document content.

1.3 Name extraction

- ❖ Locate names in text,
- ❖ Determine what type of entity the name refers to

1.4 Term extraction

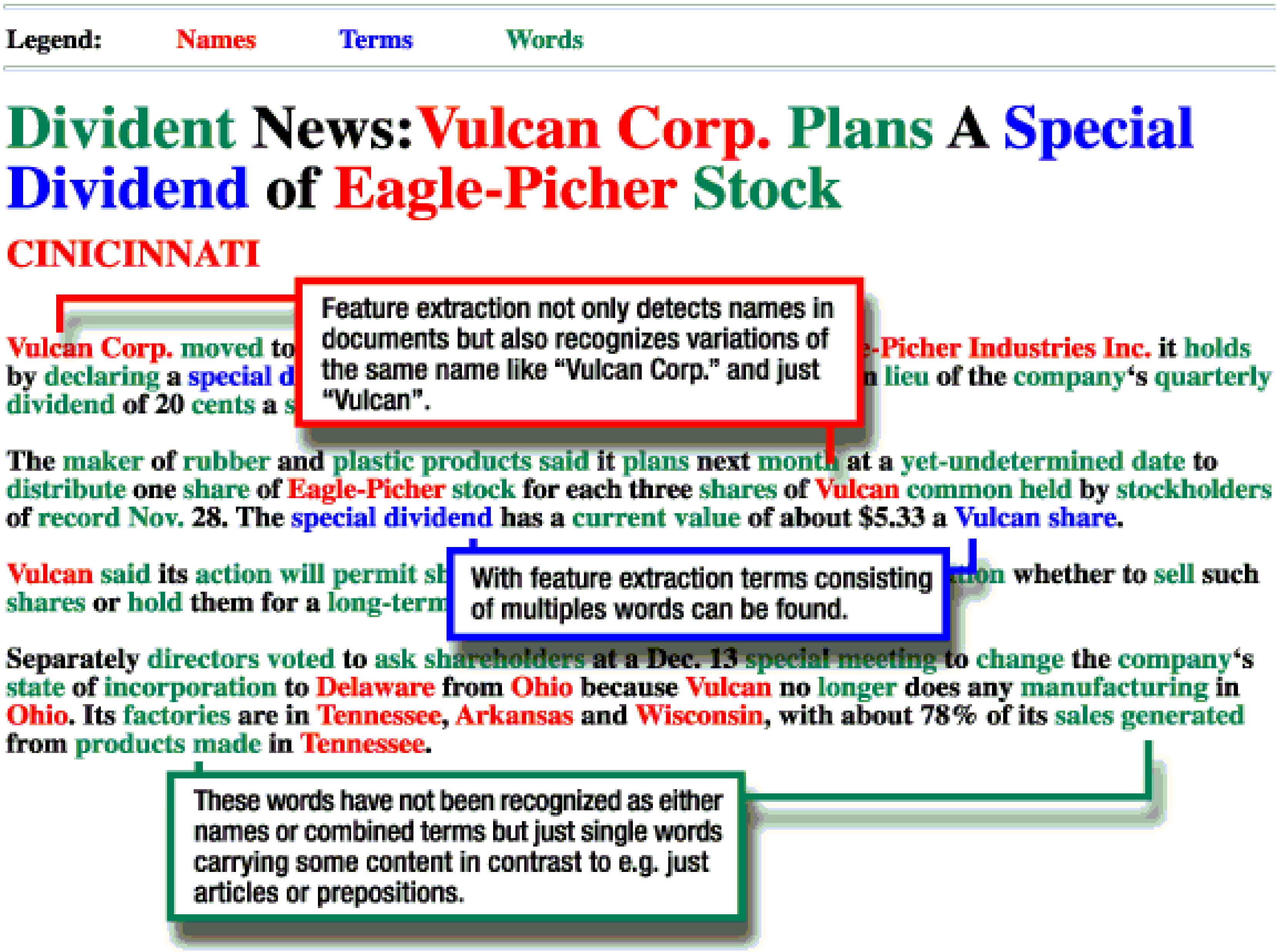
- ❖ Discover terms in text. Multiword technical terms
- ❖ Recognize variants of the same concept

1.5 Abbreviation recognition

- ❖ Find abbreviation and math them with their full forms.

1.6 Relation extraction

Feature Extraction Demo



Legend: Names Terms Words

Divident News: Vulcan Corp. Plans A Special Dividend of Eagle-Picher Stock

CINICINNATI

Vulcan Corp. moved to -Picher Industries Inc. it holds
by declaring a special d n lieu of the company's quarterly
dividend of 20 cents a s

Feature extraction not only detects names in documents but also recognizes variations of the same name like "Vulcan Corp." and just "Vulcan".

The maker of rubber and plastic products said it plans next month at a yet-undetermined date to distribute one share of Eagle-Picher stock for each three shares of Vulcan common held by stockholders of record Nov. 28. The special dividend has a current value of about $5.33 a Vulcan share.

Vulcan said its action will permit sh ion whether to sell such
shares or hold them for a long-term

With feature extraction terms consisting of multiples words can be found.

Separately directors voted to ask shareholders at a Dec. 13 special meeting to change the company's state of incorporation to Delaware from Ohio because Vulcan no longer does any manufacturing in Ohio. Its factories are in Tennessee, Arkansas and Wisconsin, with about 78% of its sales generated from products made in Tennessee.

These words have not been recognized as either names or combined terms but just single words carrying some content in contrast to e.g. just articles or prepositions.

2. Clustering Tools

2.1 Application

- Provide a overview of content in a large document collection
- Identify hidden structures between groups of objects
- Improve the browsing process to find similar or related information
- Find outstanding documents within a collection

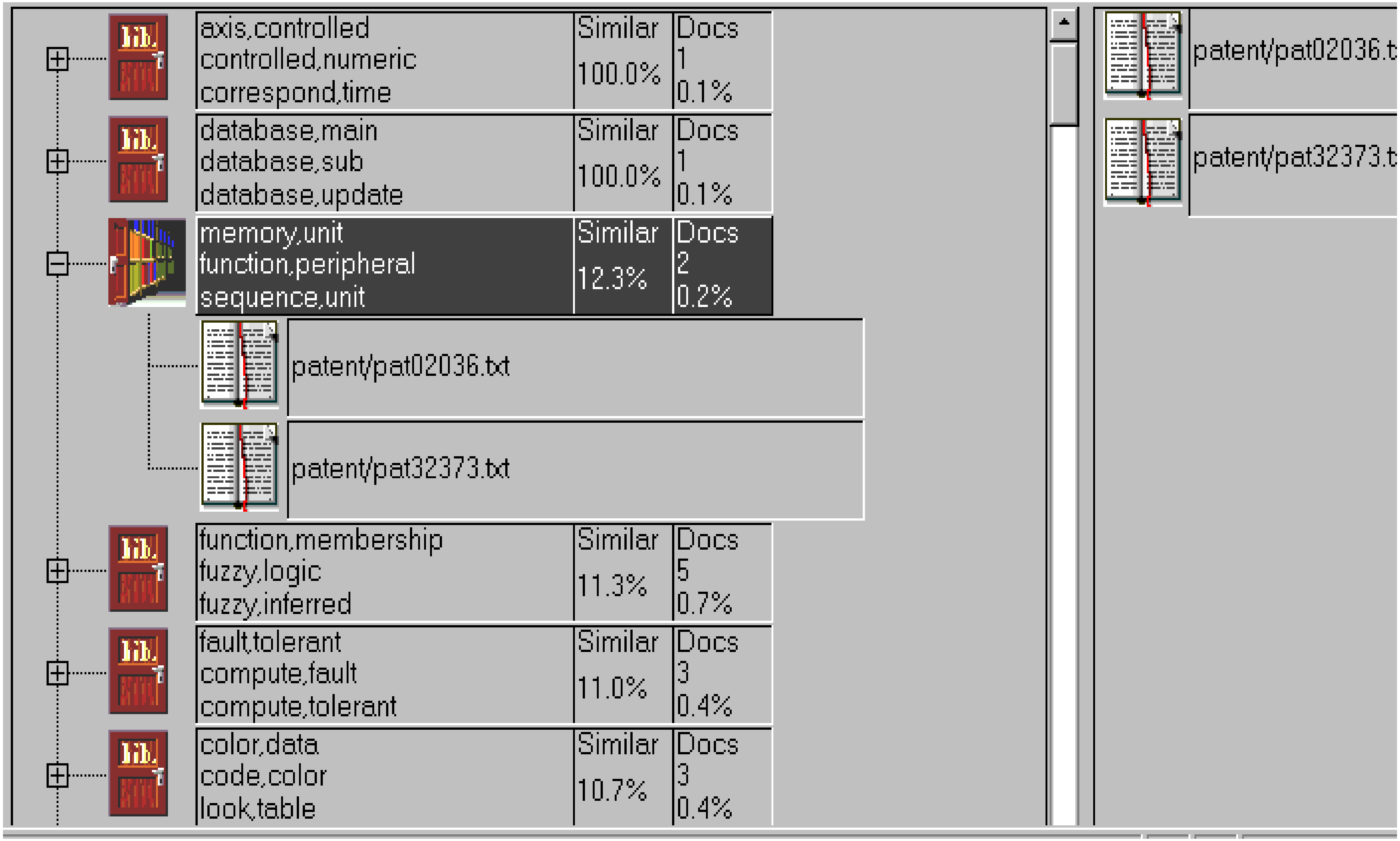
2.2 Hierarchical clustering

Clusters are organized in a clustering tree and related clusters occurs in the same branch of tree.

2.3 Binary relational clustering

With Binary Relational Clustering, the tool finds topics hidden in a document collection and establishes *links* or *relations* between these topics.

Clustering Demo: Navigation of document collection



axis,controlled controlled,numeric correspond,time	Similar 100.0%	Docs 1 0.1%
database,main database,sub database,update	Similar 100.0%	Docs 1 0.1%
memory,unit function,peripheral sequence,unit	Similar 12.3%	Docs 2 0.2%
patent/pat02036.txt		
patent/pat32373.txt		
function,membership fuzzy,logic fuzzy,inferred	Similar 11.3%	Docs 5 0.7%
fault,tolerant compute,fault compute,tolerant	Similar 11.0%	Docs 3 0.4%
color,data code,color look,table	Similar 10.7%	Docs 3 0.4%

patent/pat02036.t

patent/pat32373.t

3. Summarization Tools

3.1 Steps

- ❖ the most relevant sentences → the relevancy of a sentence to a document

  → a summary of the document with length set by user

3.2 Applications

- ❖ Judge the relevancy of a full text

  Easily determine whether the document is relevant to read.

- ❖ Enrich search results

  The results of a query to a search engine can be enriched with a short summary of each document.

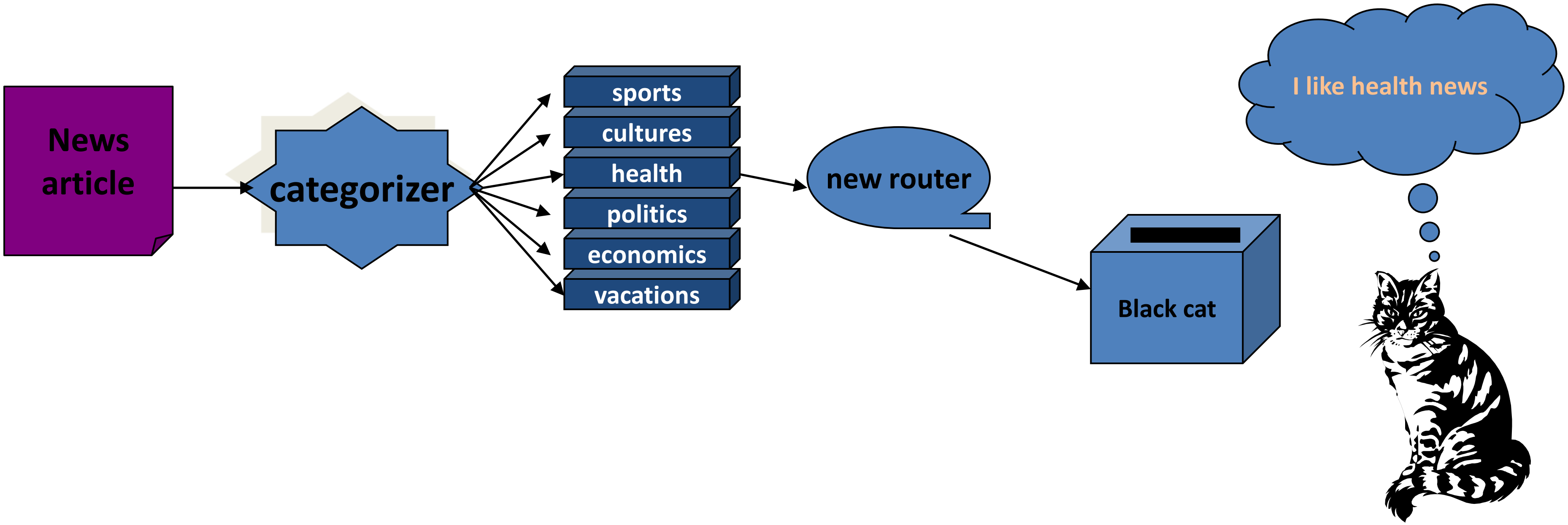
- ❖ Get a fast overview over document collections

  summary → full document

4. Categorization Tool

Applications

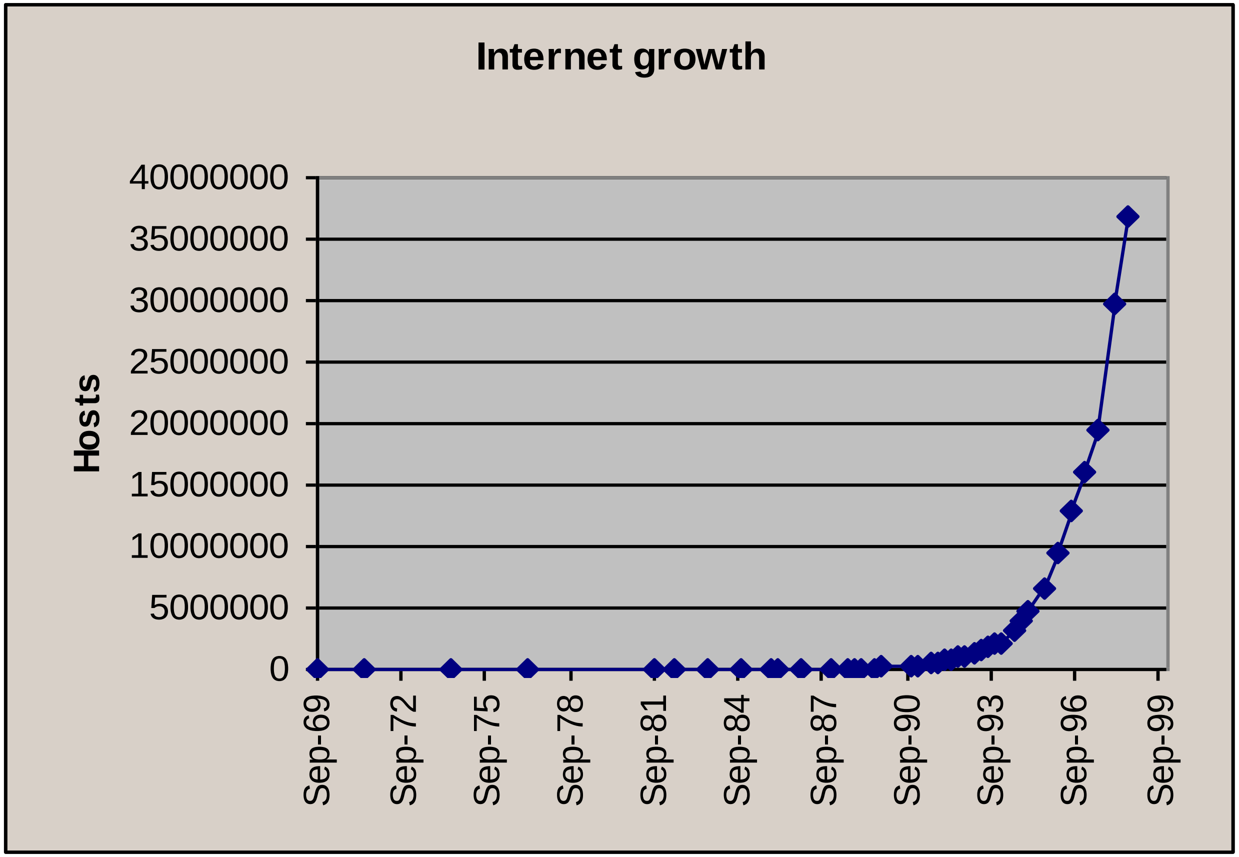
- ❖ Organize intranet documents
- ❖ Assign documents to folders
- ❖ Dispatch requests
- ❖ Forward news to subscribers



Mining World Wide Web (WWW)

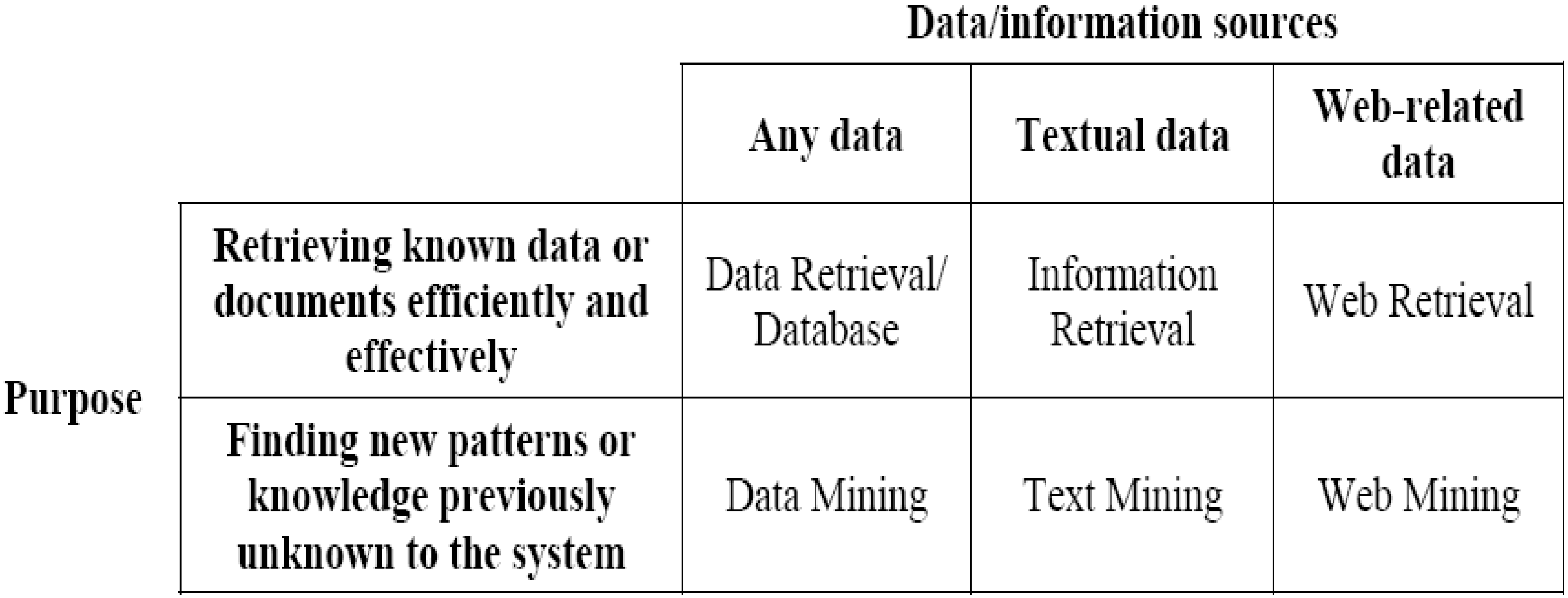
- The term **Web Mining** was coined by Orem Etzioni (1996) to denote the use of data mining techniques to automatically discover Web documents and services, extract information from Web resources, and uncover general patterns on the Web.
- The World Wide Web is a rich, enormous knowledge base that can be useful to many applications. The WWW is huge, widely distributed, global information service centre for news, advertisements, consumer information, financial management, education, government, e-commerce, hyperlink information, access and usage information.
- The Web's large size and its unstructured and dynamic content, as well as its multilingual nature make extracting useful knowledge from it a challenging research problem.

Why Mining the World-Wide Web



- Growing and changing very rapidly
- Broad diversity of user communities
- Only a small portion of the information on the Web is truly relevant or useful
  - 99% of the Web information is useless to 99% of Web users
  - How can we find high-quality Web pages on a specified topic?

Web mining research overlaps substantially with other areas, including data mining, text mining, information retrieval, and web retrieval.



		Data/information sources		
		Any data	Textual data	Web-related data
Purpose	Retrieving known data or documents efficiently and effectively	Data Retrieval/ Database	Information Retrieval	Web Retrieval
	Finding new patterns or knowledge previously unknown to the system	Data Mining	Text Mining	Web Mining

Table 1. A classification of retrieval and mining techniques and applications.

Web Search Engines

- **Index-based:** search the Web, index Web pages, and build and store huge keyword-based indices
- Help locate sets of Web pages containing certain keywords

Deficiencies

- A topic of any breadth may easily contain hundreds of thousands of documents
- Many documents that are highly relevant to a topic may not contain keywords defining them (polysemy)

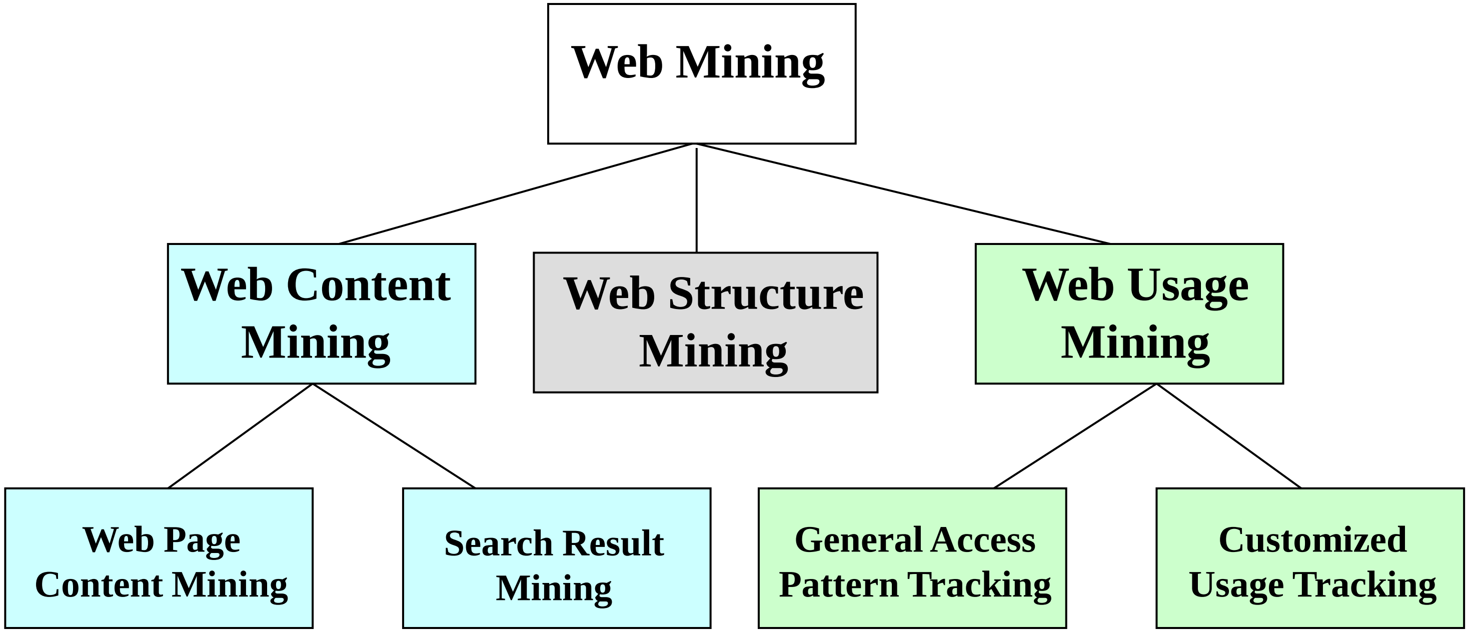
Web Mining: A More Challenging Task

- ❖ Searches for
  - – Web access patterns
  - – Web structures
  - – Regularity and dynamics of Web contents

Problems

- – The "abundance" problem
- – Limited coverage of the Web: hidden Web sources, majority of data in DBMS
- – Limited query interface based on keyword-oriented search
- – Limited customization to individual users

Web Mining Taxonomy



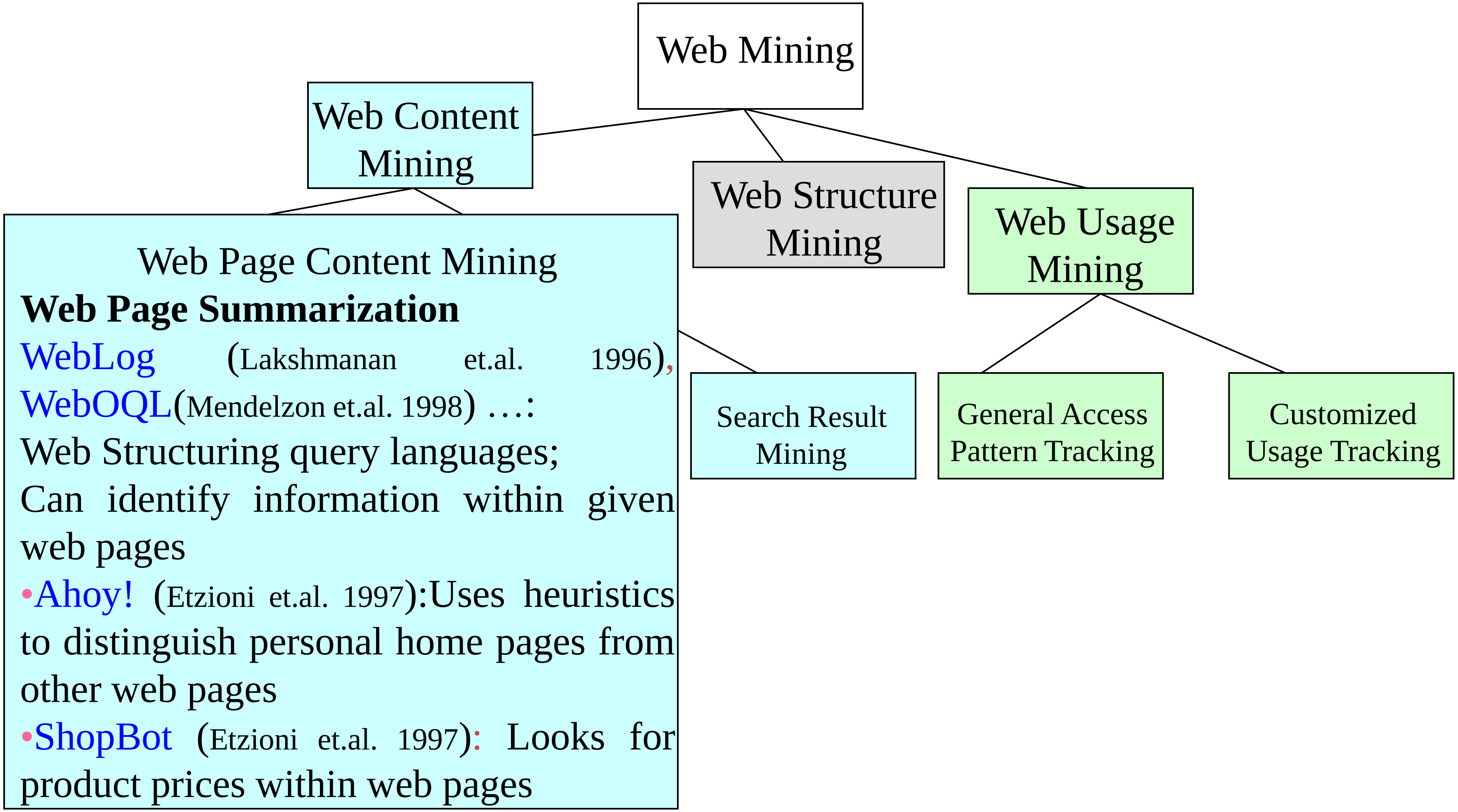
Web Mining research can be classified into three categories:

Web content mining refers to the discovery of useful information from Web contents, including text, images, audio, video, etc.

Web structure mining studies the model underlying the link structures of the Web. It has been used for search engine result ranking and other Web applications.

Web usage mining focuses on using data mining techniques to analyze search logs to find interesting patterns. One of the main applications of Web usage mining is its use to learn user profiles.

Mining the World-Wide Web



Web Mining

Web Content Mining

Web Structure Mining

Web Usage Mining

Web Page Content Mining

Web Page Summarization

WebLog (Lakshmanan et.al. 1996), WebOQL(Mendelzon et.al. 1998) …: Web Structuring query languages; Can identify information within given web pages

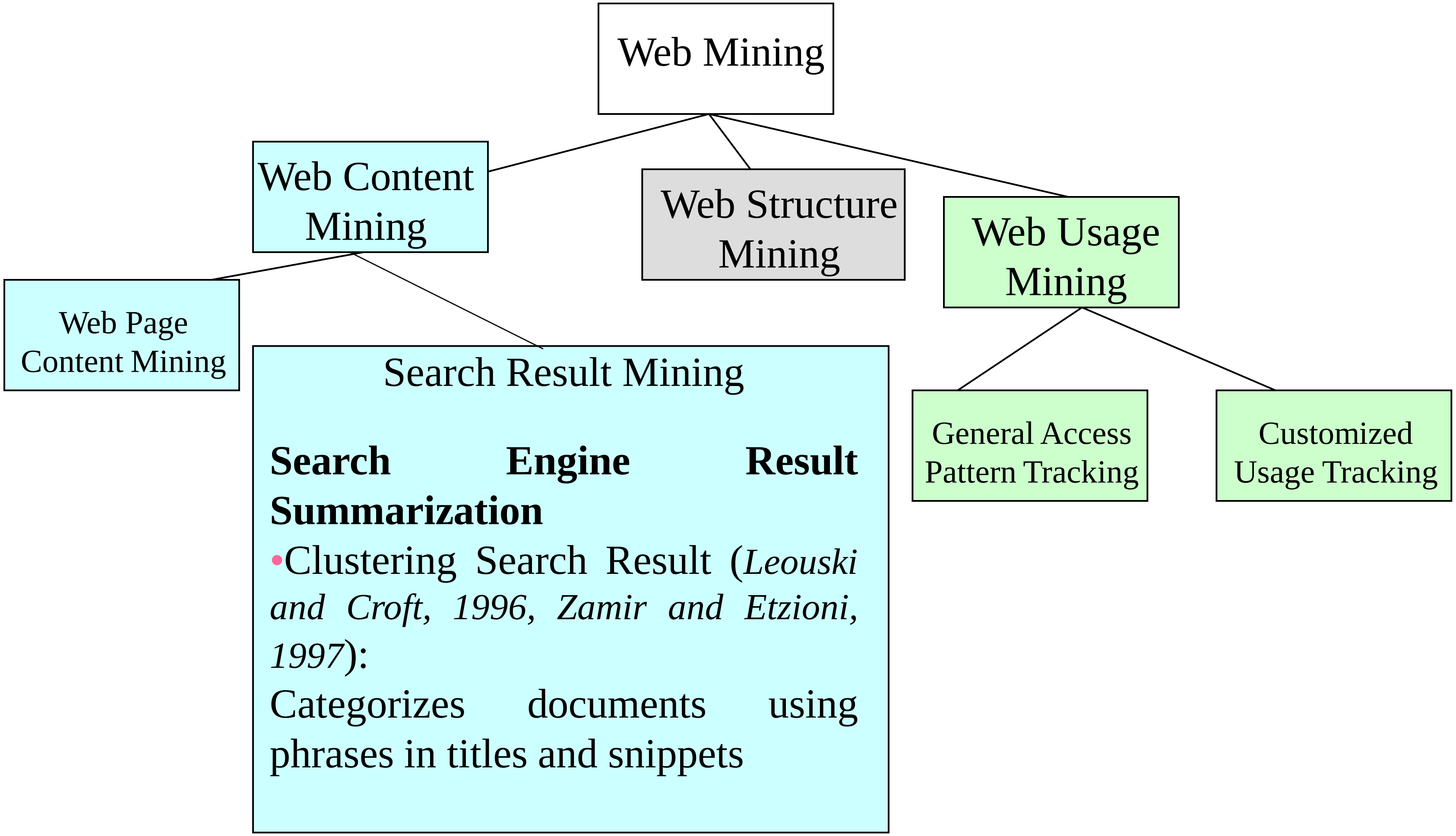
- Ahoy! (Etzioni et.al. 1997):Uses heuristics to distinguish personal home pages from other web pages
- ShopBot (Etzioni et.al. 1997): Looks for product prices within web pages

Search Result Mining

General Access Pattern Tracking

Customized Usage Tracking

Mining the World-Wide Web



Web Mining

Web Content Mining

Web Structure Mining

Web Usage Mining

Web Page Content Mining

Search Result Mining

Search Engine Result Summarization

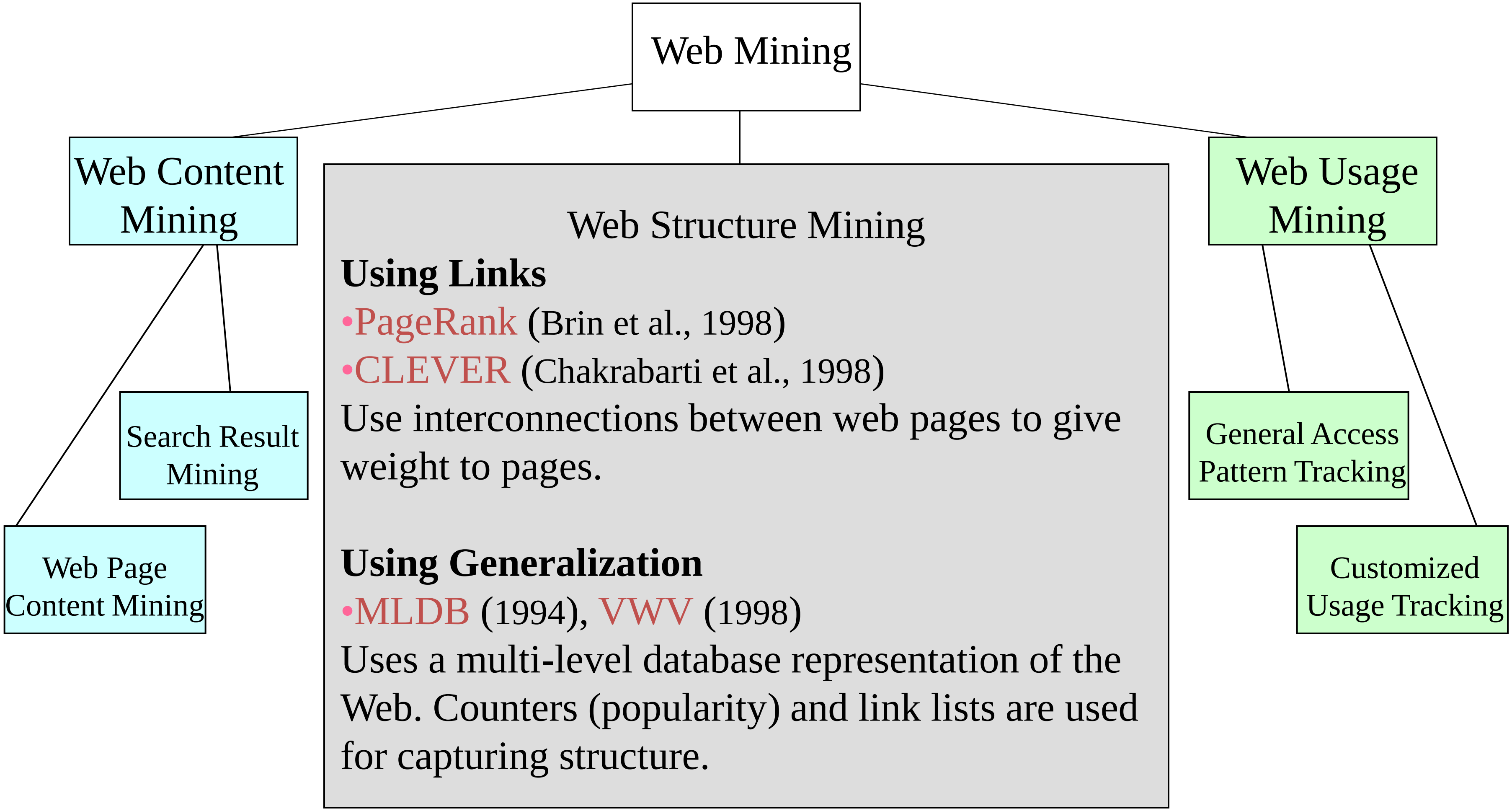
•Clustering Search Result (*Leouski and Croft, 1996, Zamir and Etzioni, 1997*):

Categorizes documents using phrases in titles and snippets

General Access Pattern Tracking

Customized Usage Tracking

Mining the World-Wide Web



Web Mining

Web Content Mining

Search Result Mining

Web Page Content Mining

Web Structure Mining

Using Links

- PageRank (Brin et al., 1998)
- CLEVER (Chakrabarti et al., 1998)

Use interconnections between web pages to give weight to pages.

Using Generalization

- MLDB (1994), VWV (1998)

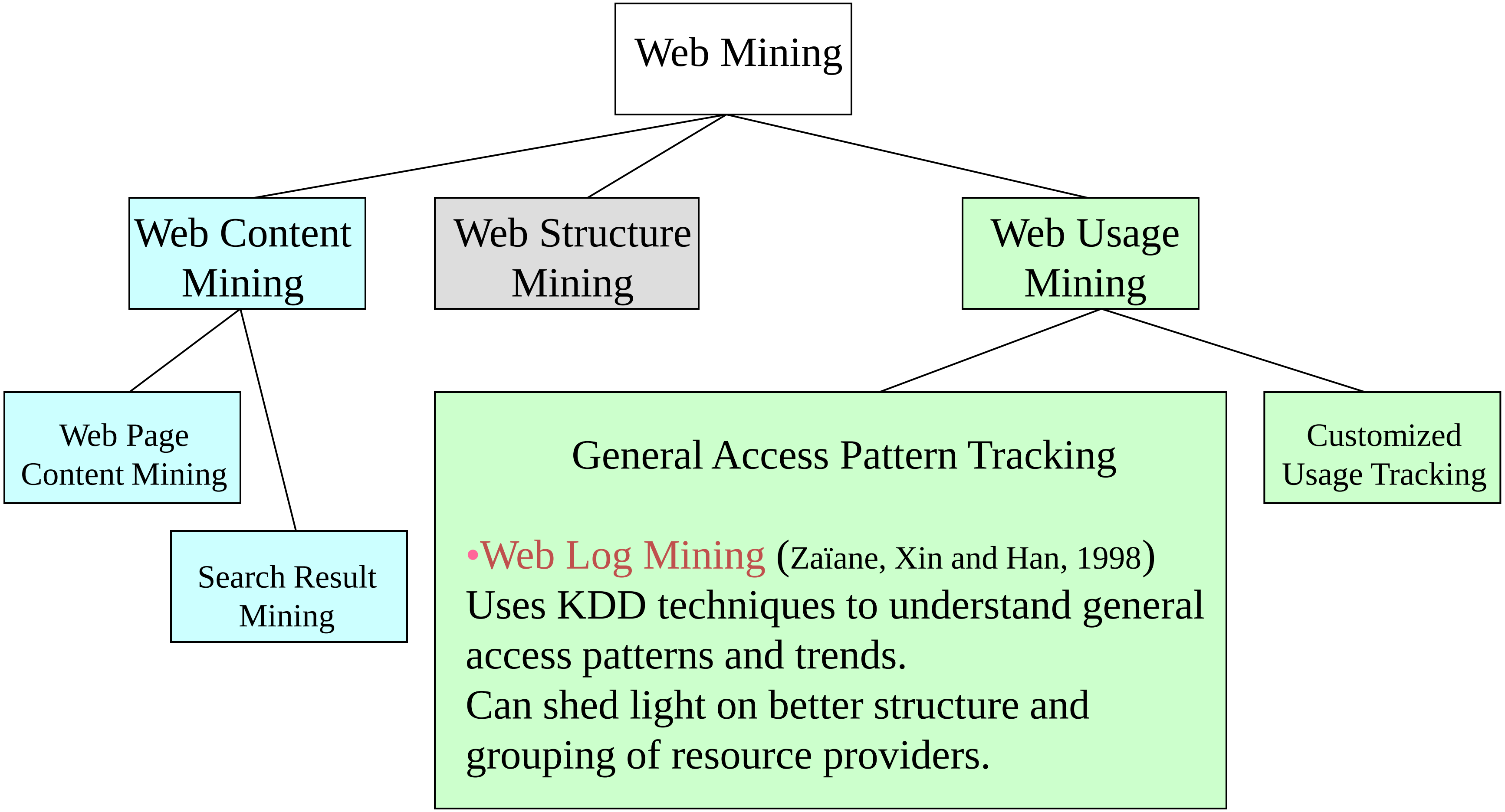
Uses a multi-level database representation of the Web. Counters (popularity) and link lists are used for capturing structure.

Web Usage Mining

General Access Pattern Tracking

Customized Usage Tracking

Mining the World-Wide Web



Web Mining

Web Content Mining

Web Structure Mining

Web Usage Mining

Web Page Content Mining

Search Result Mining

General Access Pattern Tracking

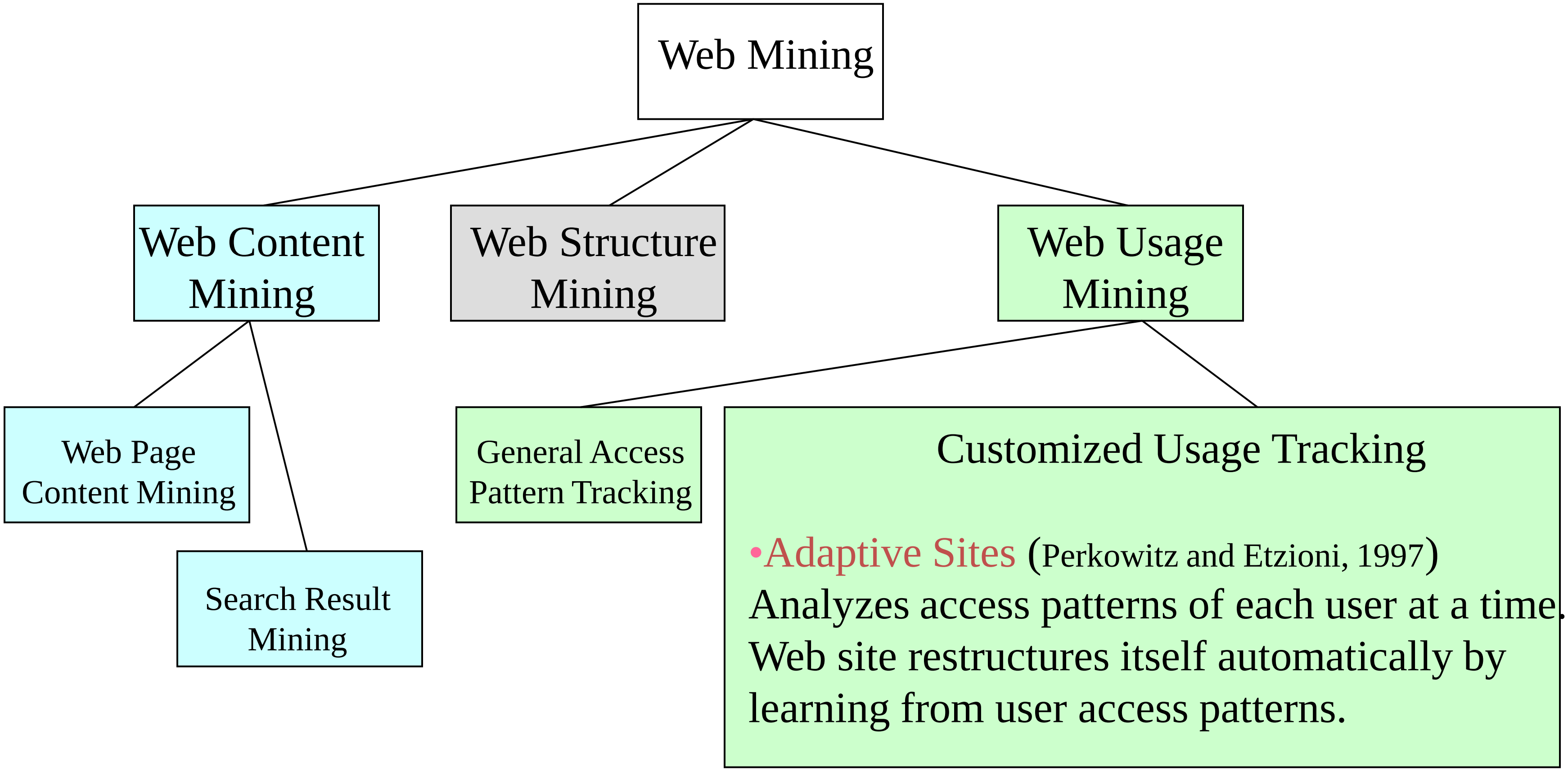
- Web Log Mining (Zaïane, Xin and Han, 1998)

Uses KDD techniques to understand general access patterns and trends.

Can shed light on better structure and grouping of resource providers.

Customized Usage Tracking

Mining the World-Wide Web



Web Mining

Web Content Mining

Web Structure Mining

Web Usage Mining

Web Page Content Mining

Search Result Mining

General Access Pattern Tracking

Customized Usage Tracking

- Adaptive Sites (Perkowitz and Etzioni, 1997)

Analyzes access patterns of each user at a time.

Web site restructures itself automatically by learning from user access patterns.

Web Usage Mining

- ❖ Web servers, Web proxies, and client applications can quite easily capture **Web Usage data**.
  - – Web server log: Every visit to the pages, what and when files have been requested, the IP address of the request, the error code, the number of bytes sent to user, and the type of browser used…
- ❖ By analyzing the Web usage data, web mining systems can discover useful knowledge about a **system's usage characteristics** and the **users' interests** which has various applications:
  - – Personalization and Collaboration in Web-based systems
  - – Marketing
  - – Web site design and evaluation
  - – Decision support (*e.g., Chen & Cooper, 2001; Marchionini, 2002*).

❖ Mining Web log records to discover user access patterns of Web pages

Applications

- Target potential customers for electronic commerce
- Enhance the quality and delivery of Internet information services to the end user
- Improve Web server system performance
- Identify potential prime advertisement locations

❖ Web logs provide rich information about Web dynamics

- Typical Web log entry includes the URL requested, the IP address from which the request originated, and a timestamp

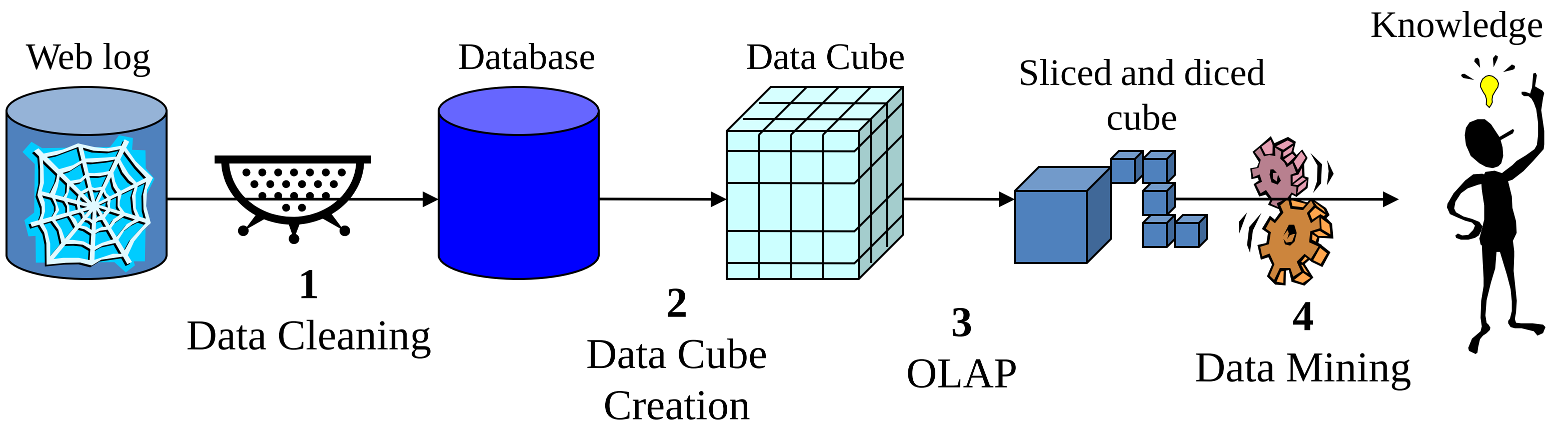
Why Web Usage Mining?

- Explosive growth of E-commerce
  - Provides an cost-efficient way doing business
  - Amazon.com: “online Wal-Mart”
- Hidden Useful information
  - Visitors’ profiles can be discovered
  - Measuring online marketing efforts, launching marketing campaigns, etc.

Mining the World-Wide Web

Design of a Web Log Miner

- Web log is filtered to generate a relational database
- A data cube is generated form database
- OLAP is used to drill-down and roll-up in the cube
- OLAM is used for mining interesting knowledge



- Web usage mining has been used for various purposes:
  - **A knowledge discovery process** for mining marketing intelligence information from Web data. *Buchner and Mulvenna (1998)*
  - **Web traffic patterns** also can be extracted from Web usage logs in order to improve the performance of a Web site *(Cohen et al., 1998).*
  - **Commercial products**: *Web Trends developed by NetIQ, WebAnalyst by Megaputer and NetTracker by Sane Solutions.*

- Search engine transaction logs also provide valuable knowledge about **user behavior** on Web searching.

- Such information is very useful for a better understanding of users’ Web searching and information seeking behavior and can improve the design of Web search systems.

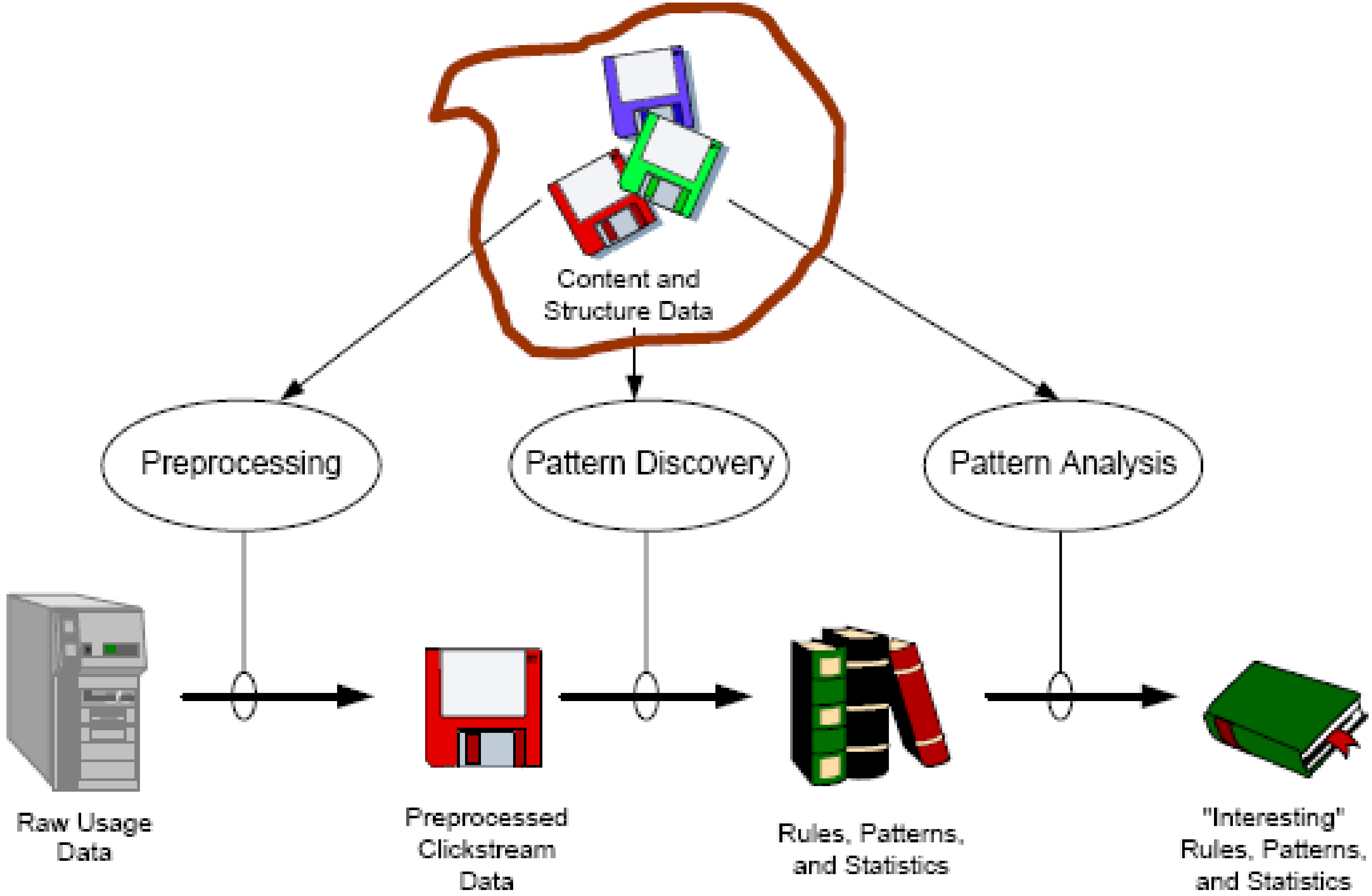
One of the major goals of Web usage mining is to reveal **interesting trends and patterns** which can often provide important knowledge about the users of a system.

The **Framework** for Web usage mining. *Srivastava et al. (2000)*

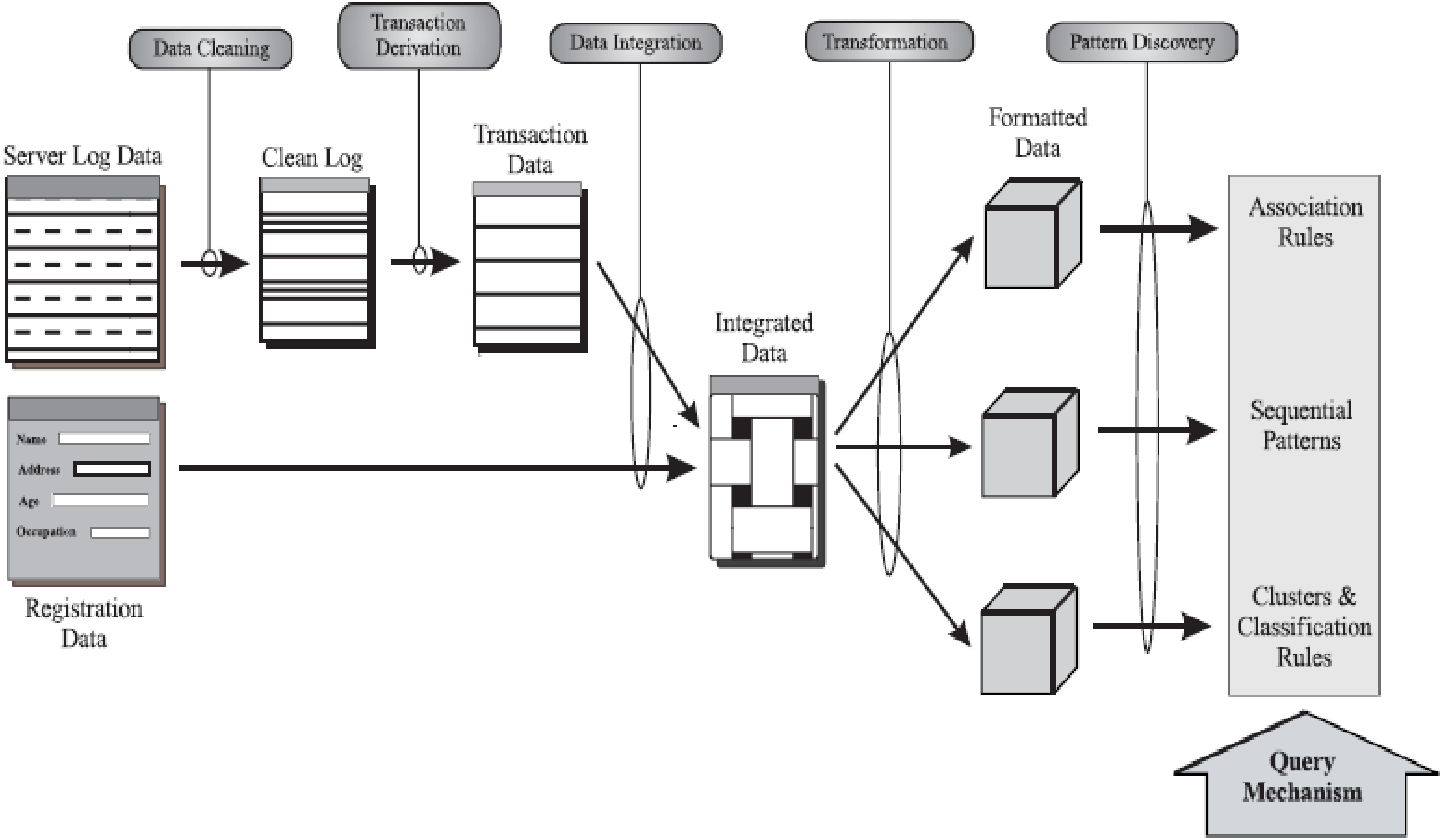
- **Preprocessing:** Data cleansing
- **Pattern discovery:**
- **Pattern analysis:**

Generic machine learning and Data mining techniques, such as association rule mining, classification, and clustering, often can be applied.

Web Usage Mining - Procedure



Web Usage Mining - Model



- Many Web applications aim to provide **personalized** information and services to users. **Web usage data** provide an excellent way to learn about users' interest (*Srivastava et al., 2000*).
  - WebWatcher (*Armstrong et al., 1995*)
  - Letizia (*Lieberman, 1995*)

- Web usage mining on Web logs can help identify users who have accessed similar Web pages. The patterns that emerge can be very useful in **collaborative** Web searching and filtering.

  - *Amazon.com* uses **collaborative filtering** to recommend books to potential customers based on the preferences of other customers having similar interests or purchasing histories.

  - *Huang et al. (2002)* used **Hopfield Net** to model user interests and product profiles in an online bookstore in Taiwan.

How to perform Web Usage Mining

- Obtain web traffic data from
  - Web server log files
  - Corporate relational databases
  - Registration forms
- Apply data mining techniques and other Web mining techniques
- Two categories:
  - Pattern Discovery Tools
  - Pattern Analysis Tools

Pattern Analysis Tools

- Answer Questions like:
  - “How are people using this site?”
  - “which Pages are being accessed most frequently?”
- This requires the analysis of the structure of hyperlinks and the contents of the pages

Pattern Discovery Tools

- Data Pre-processing
  - Filtering/clean Web log files
    - eliminate outliers and irrelevant items
  - Integration of Web Usage data from:
    - Web Server Logs
    - Referral logs
    - Registration file
    - Corporate Database
- Converting IP addresses to Domain Names
  - Domain Name System does the conversion
  - Discover information from visitors' domain names:
    - Ex: .ca(Canada), .cn(China), etc
- Converting URLs to Page Titles
  - Page Title: between <title> and </title>

Pattern Discovery Techniques

❖ Path Analysis

- Uses Graph Model
- Provide insights to navigational problems
- Example of info. Discovered by Path analysis:
    - 78% “company”-> “what’s new”->“sample”-> “order”
    - 60% left sites after 4 or less page references

      => most important info must be within the first 4 pages of site entry points.

❖ Grouping

- Groups similar info. to help draw higher-level conclusions
- Ex: all URLs containing the word “Yahoo”…

❖ Filtering

- Allows to answer specific questions like:
    - how many visitors to the site in ***this week***?

Pattern Discovery Techniques

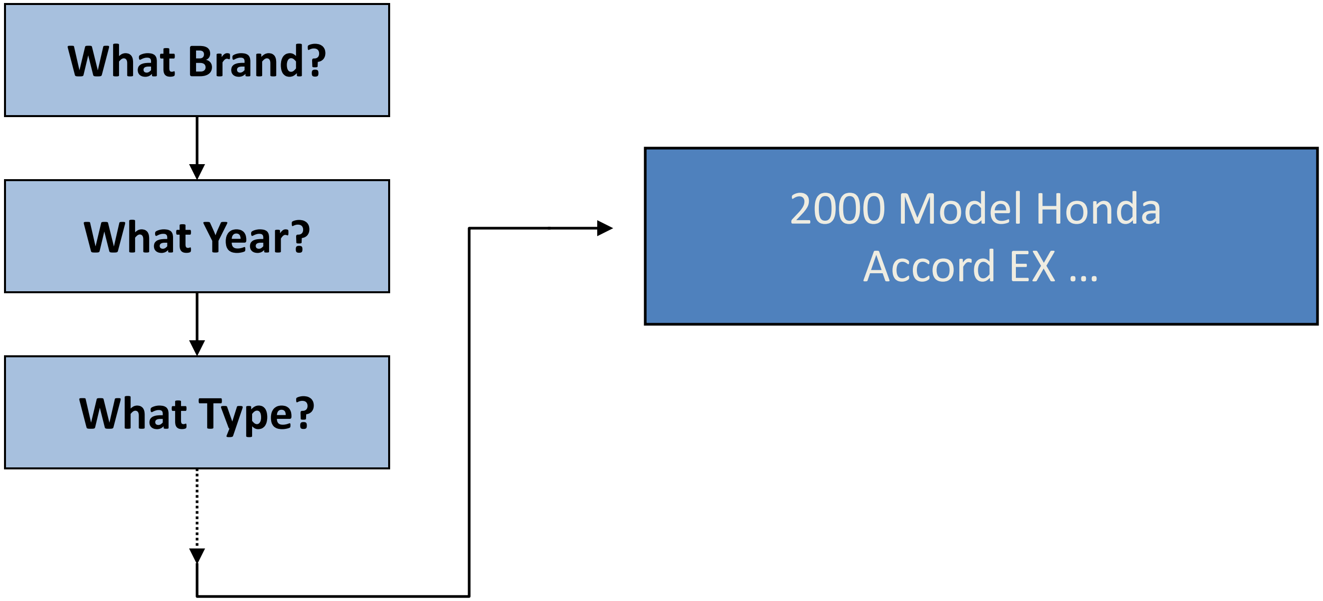
- Dynamic Site Analysis
  - Dynamic html links to the database, and requires parameters appended to URLs
  - http://search.netscape.com/cgi-in/search?search=Federal+Tax+Return+Form&cp=ntserch
  - Knowledge:
    - What the visitors looked for
    - What keywords S/B purchased from Search engineer
- Cookies
  - Randomly assigned ID by web server to browser
  - Cookies are beneficial to both web site developers and visitors
  - Cookie field entry in log file can be used by Web traffic analysis software to track repeat visitors → loyal customers.

Pattern Discovery Techniques

- Association Rules
  - help find spending patterns on related products
  - 30% who accessed/company/products/bread.html, also accessed /company/products/milk.htm.
- Sequential Patterns
  - help find inter-transaction patterns
  - 50% who bought items in /pcworld/computers/, also bought in /pcworld/accessories/ within 15 days
- Clustering
  - Identifies visitors with common characteristics based on visitors' profiles
  - 50% who applied discover platinum card in /discovercard/customerService/newcard, were in the 25-35 age group, with annual income between $40,000 – 50,000.

Pattern Discovery Techniques

- Decision Trees
  - a flow chart of questions leading to a decision
  - Ex: car buying decision tree



Web Content Mining

❖ Text Mining for Web Documents

- Text mining for Web documents can be considered a sub-field of **Web content mining**.

- **Information extraction techniques** have been applied to Web HTML documents
  - ➢ E.g., *Chang and Lui (2001)* used a PAT tree to construct automatically a set of rules for information extraction.

- **Text clustering algorithms** also have been applied to Web applications.
  - ➢ E.g., *Chen et al. (2001; 2002)* used a combination of noun phrasing and SOM to cluster the search results of search agents that collect Web pages by meta-searching popular search engines.

Web Structure Mining

- **Web link structure** has been widely used to infer important web pages information.

- Web structure mining has been largely influenced by research in
  - **Social network analysis**
  - **Citation analysis** (bibliometrics).
    - *in-links:* the hyperlinks pointing to a page
    - *out-links:* the hyperlinks found in a page.
    - Usually, the **larger** the number of in-links, the **better** a page is.

- By analyzing the pages containing a **URL**, we can also obtain
  - **Anchor text**: how other Web page authors annotate a page and can be useful in predicting the content of the target page.

❖**Web structure mining algorithms**:

- **The PageRank algorithm** is computed by weighting each in-link to a page **proportionally** to the quality of the page containing the in-link (*Brin & Page, 1998*).
- The **qualities** of these referring pages also are determined by PageRank. Thus, a page *p* is calculated **recursively** as follows:

$$PageRank(p) = (1-d) + d \times \sum_{\substack{all\ q\ linking \\ to\ p}} \left(\frac{PageRank(q)}{c(q)} \right)$$

where *d is a damping factor between 0 and 1,*
$c(q)$ is the number of out-going links in a page q.

- **Web structure mining algorithms:**
  - *Kleinberg (1998)* proposed the **HITS** (Hyperlink-Induced Topic Search) algorithm, which is similar to PageRank.
    - **Authority pages**: high-quality pages related to a particular search query.
    - **Hub pages:** pages provide pointers to other authority pages.
    - A page to which many others point should be a good authority, and a page that points to many others should be a good hub.

$$AuthorityScore(p) = \sum_{\substack{all\ q\ linking \\ to\ p}} \left(HubScore(q)\right)$$

$$HubScore(p) = \sum_{\substack{all\ r\ linking \\ from\ p}} \left(AuthorityScore(r)\right)$$

- Another application of Web structure mining is to understand the structure of the Web **as a whole**.

- The core of the Web is a **strongly connected** component and that the Web's graph structure is shaped like a bowtie. *Broder et al. (2000)*

  - **Strongly Connected Component** (**SCC**); 28% of the Web.
  - **IN**: every Web page contains a direct path to the SCC; 21% of Web
  - **OUT**: a direct path from SCC linking to it; 21% of Web
  - **TENDRILS**: pages hanging off from IN and OUT but without direct path to SCC; 22% of Web
  - **Isolated, Disconnected Components** that are not connected to the other 4 groups; 8% of Web

Conclusion

Spatial data mining is facilitated by Spatial warehousing, OLAP and mining and finds spatial associations, classifications and trends.

Multimedia data mining needs content-based retrieval and similarity search integrated with mining methods

Text mining goes beyond keyword-based and similarity-based information retrieval and discovers knowledge from semi-structured data using methods like keyword-based association and document classification.

Web mining includes mining Web link structures to identify authoritative Web pages, the automatic classification of Web documents, building a multilayered Web information base, and Weblog mining.

Questions?

References

1. D. Pyle. Data Preparation for Data Mining. Morgan Kaufmann, 1999.
2. Petrushin, Valery A: Introduction into Multimedia Data Mining and Knowledge Discovery. Edited by Valery A Petrushin and Latifur Khan. London: Springer-Verlag, 2007. P. 3-13.J.
3. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2000.
4. *Web Mining Research: A Survey* – Raymond Kosala, Hendrik Blockeel Dept of CS Katholieke Universiteit Leuven
5. W. H. Inmon. Building the Data Warehouse. John Wiley, 1996
6. R. Kimball and M. Ross. The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling. 2ed. John Wiley, 2002
7. S. Chakrabarti, *"Mining the Web: Statistical Analysis of Hypertext and Semi-Structured Data"*, Morgan Kaufmann, 2002.

End of Unit 7

Thank you !!!