Unit 7 : Mining Complex Types of Data

Introduction

Mining complex types of data include:

- Multimedia data mining
- Text data mining
- Web data mining
  - Web content mining
  - Web usage mining
  - Web structure mining

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

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

Mining Multimedia Databases

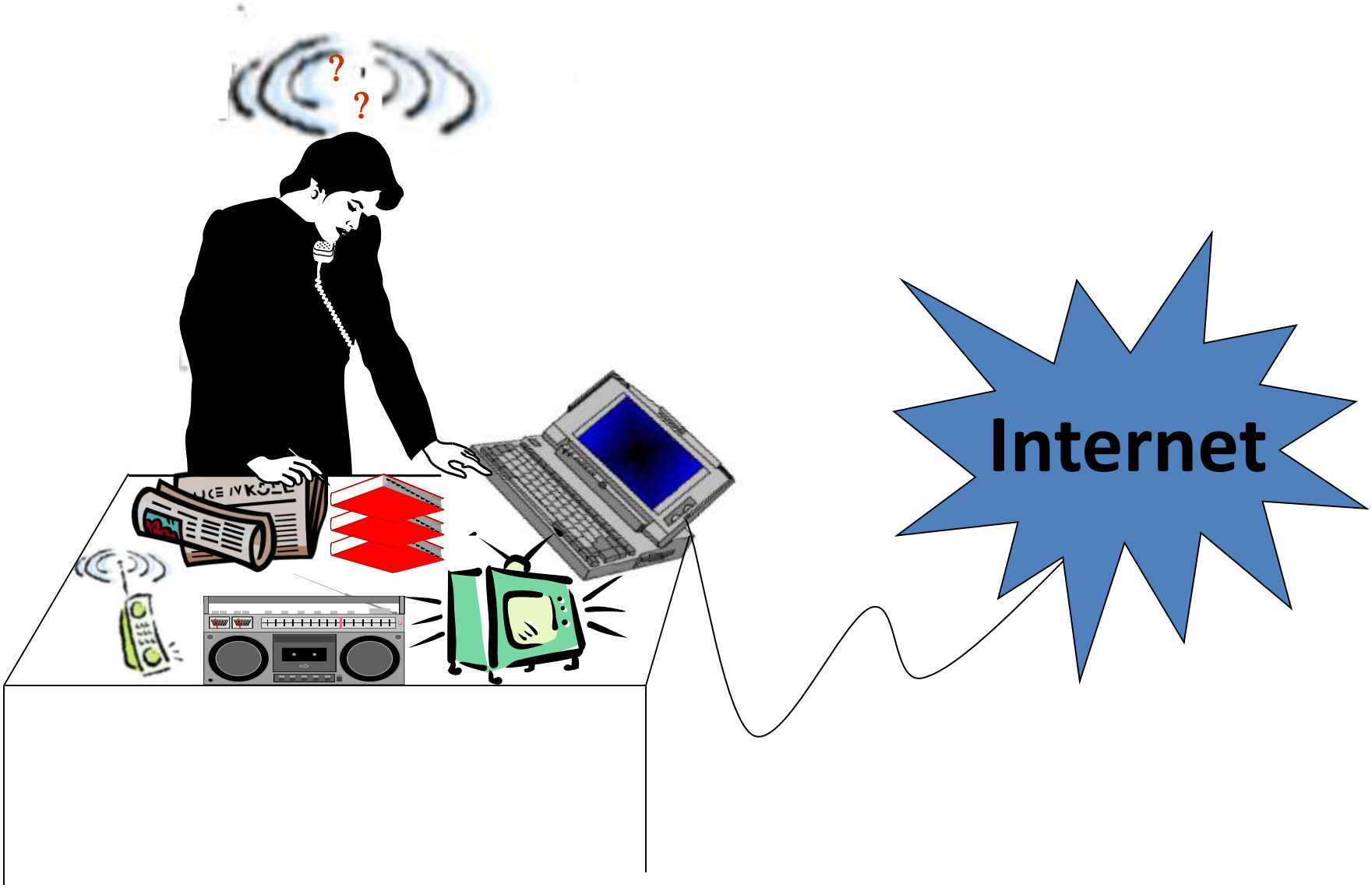
Refining or combining searches

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)

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.

- 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.

Web Images Groups[New!] News Desktop **more »**

information retrieval Search Advanced Search Preferences

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

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

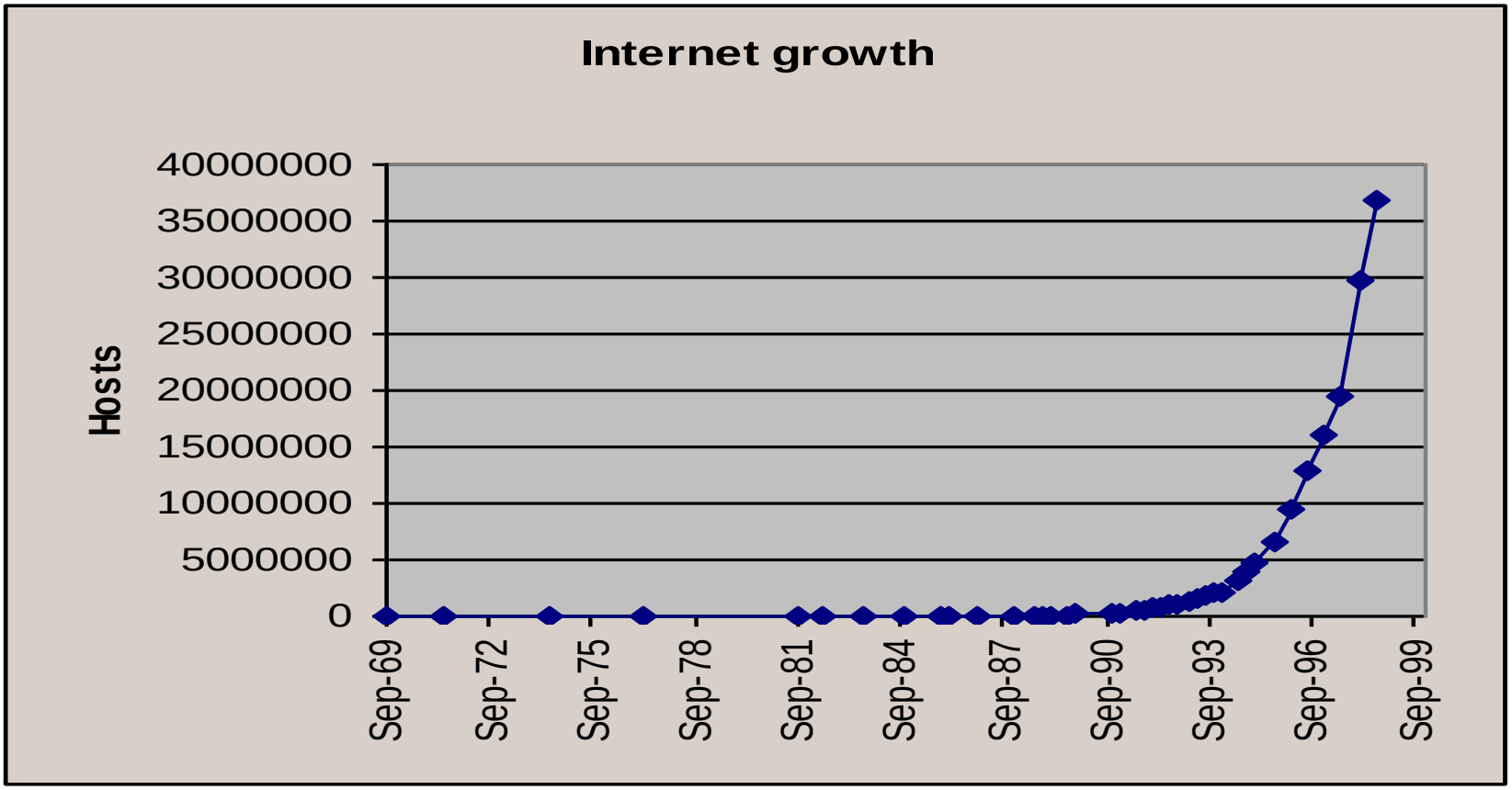
Mining World Wide Web (WWW)

- 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.

Mining World Wide Web (WWW)

- 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 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)

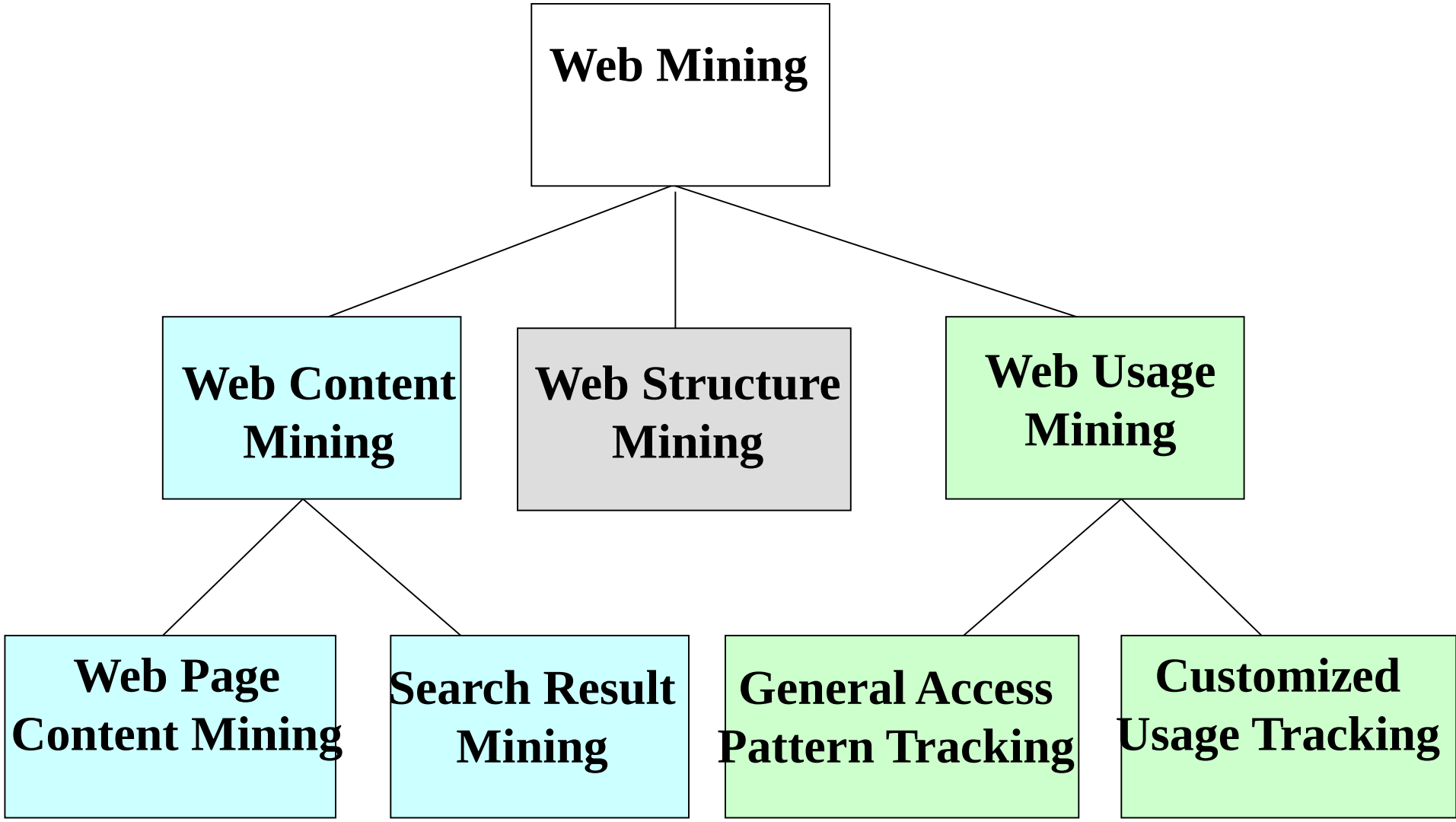
Web Mining: A More Challenging Task

- 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

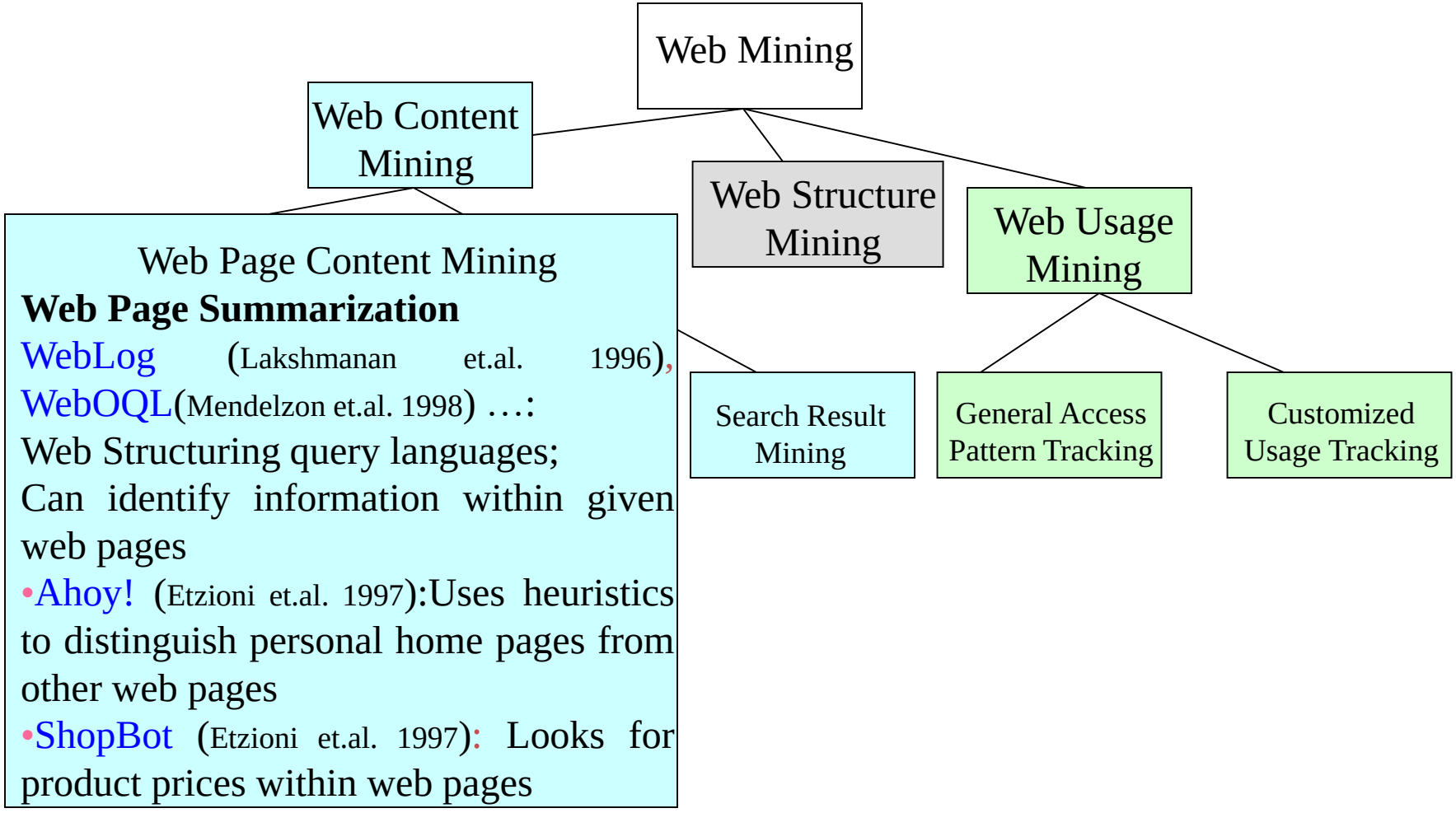
Web Mining Taxonomy



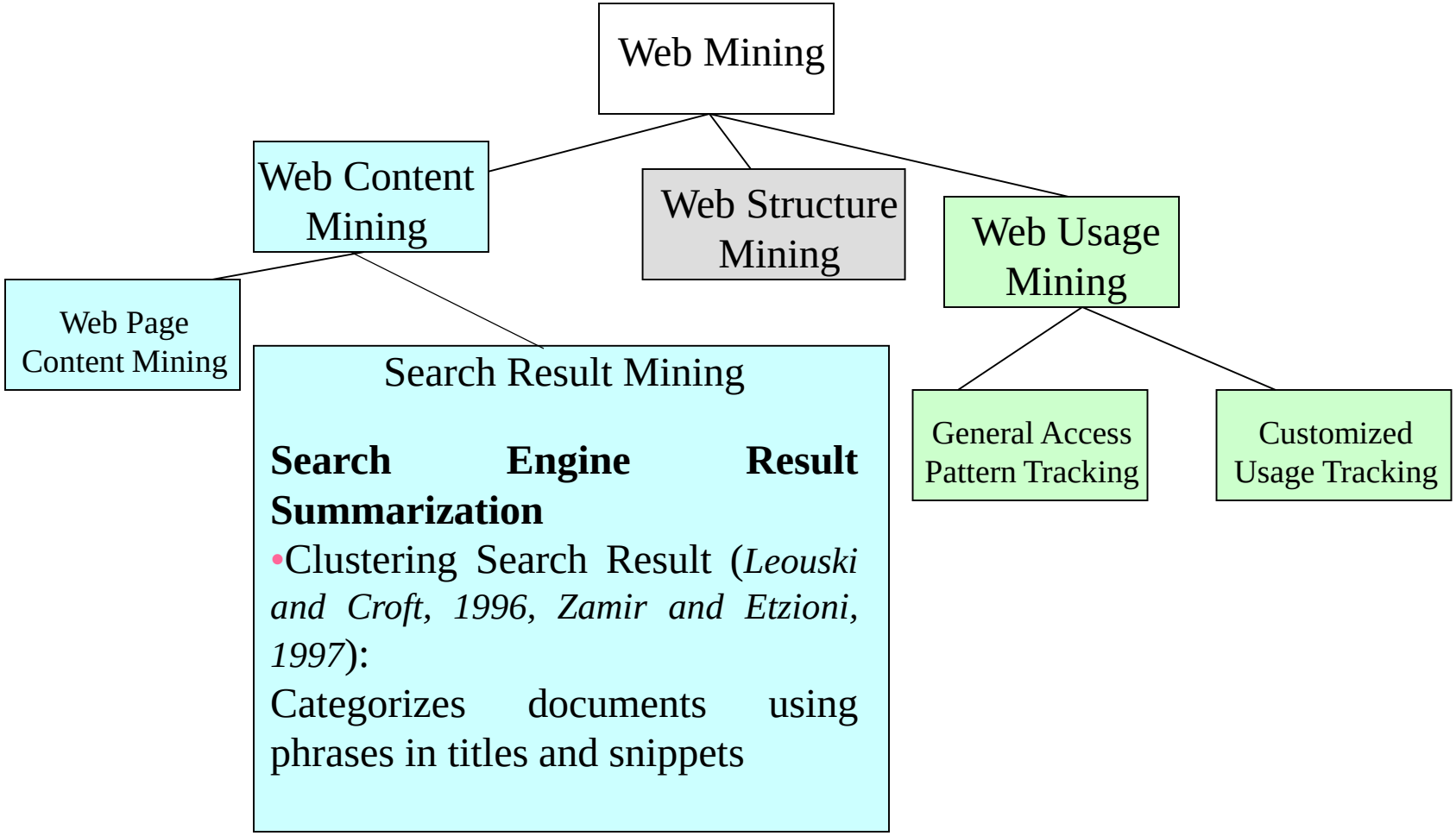
Web Mining Taxonomy

- 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.

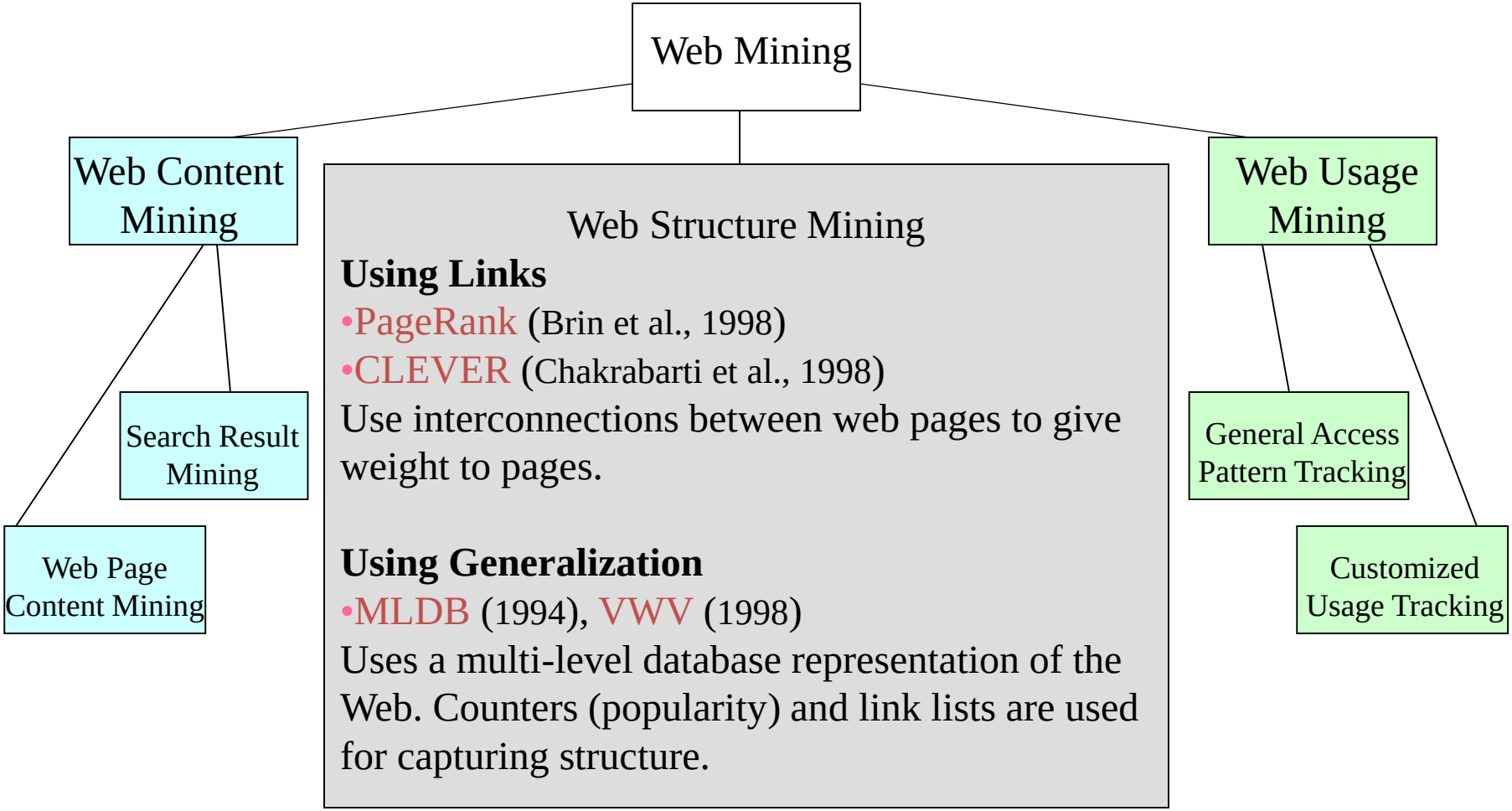
Mining the World-Wide Web



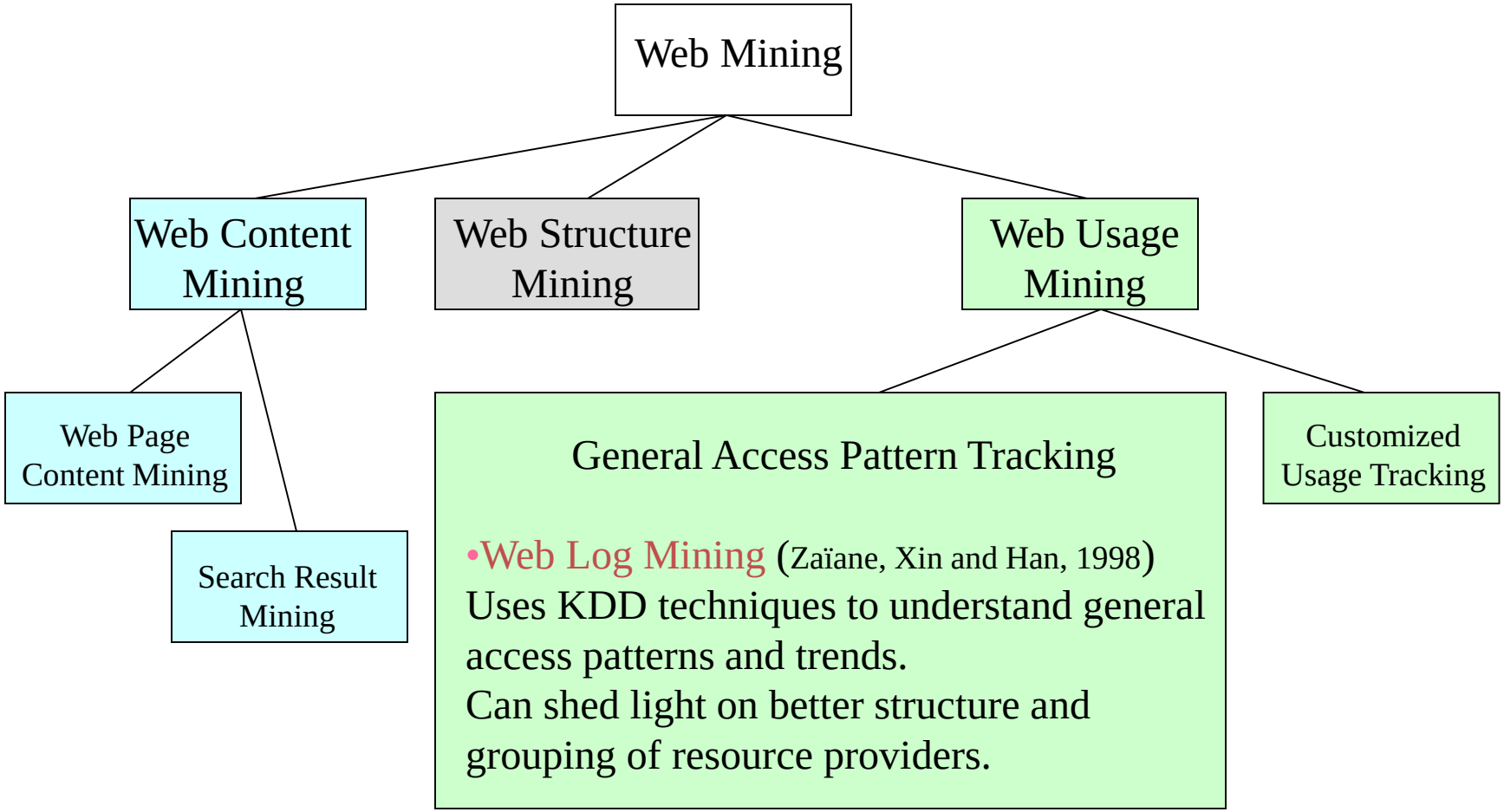
Mining the World-Wide Web



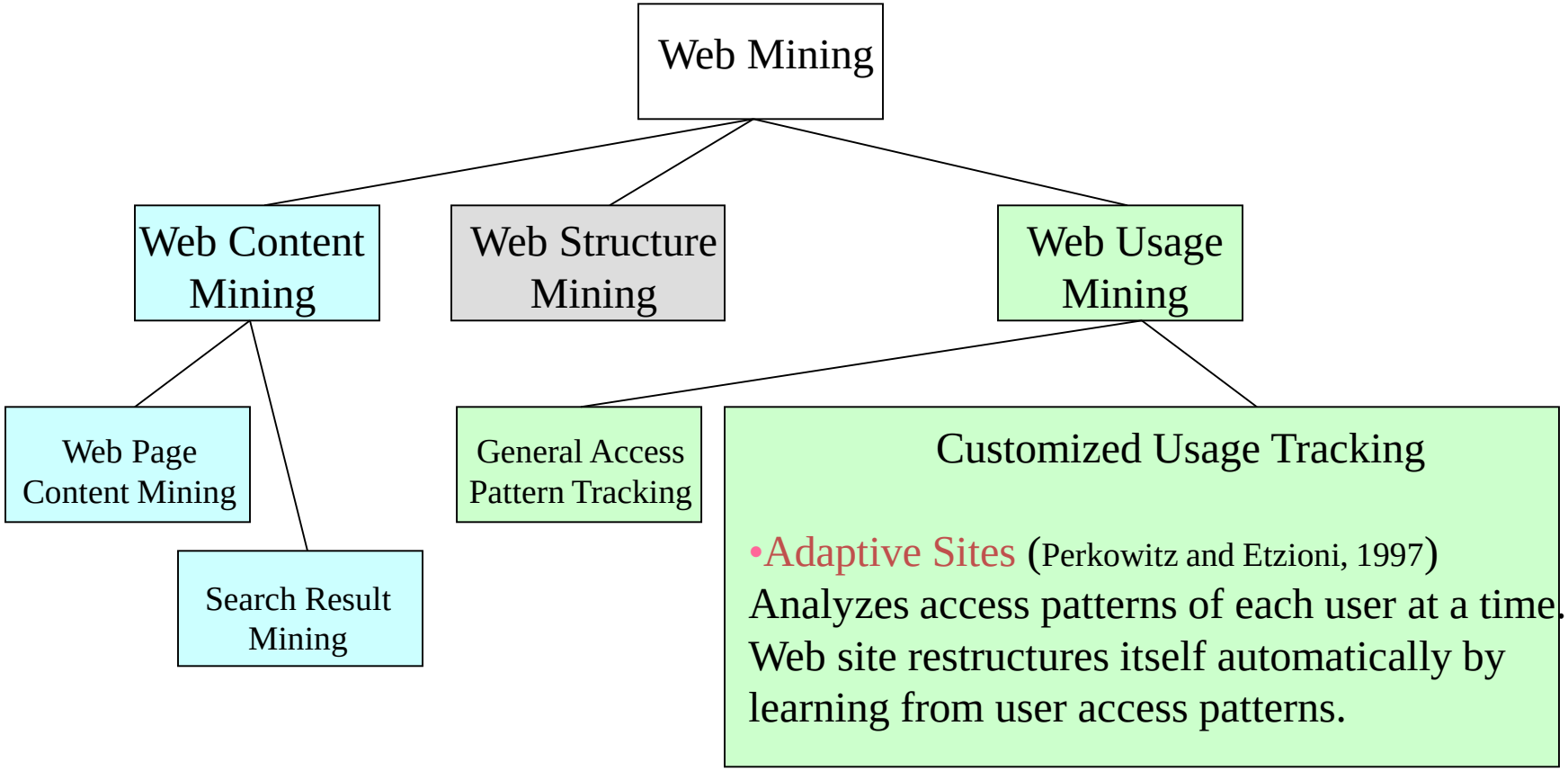
Mining the World-Wide Web



Mining the World-Wide Web



Mining the World-Wide Web



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

- 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.
- 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.

- Many Web applications aim to provide **personalized** information and services to users. **Web usage data** provide an excellent way to learn about users' interest.
  - 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

Web Content Mining

- Web Content Mining is the process of extracting useful information from the contents of Web documents.
- Content data corresponds to the collection of facts a Web page was designed to convey to the users.
- May consist of text, images, audio, video, or structured records such as lists and tables.
- Web content has been the most widely researched.
- Issues addressed in text mining are, topic discovery, extracting association patterns, clustering of web documents and classification of Web Pages.

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
  - **Text clustering algorithms** also have been applied to Web applications.

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.
    - The **qualities** of these referring pages also are determined by PageRank.
  - *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.

Conclusion

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.

Thank you !!!