

Introduction to Web Mining

- ◆ **Mining text databases**
- ◆ Mining the Web
- ◆ Summary

Text Databases and Information Retrieval

◆ 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

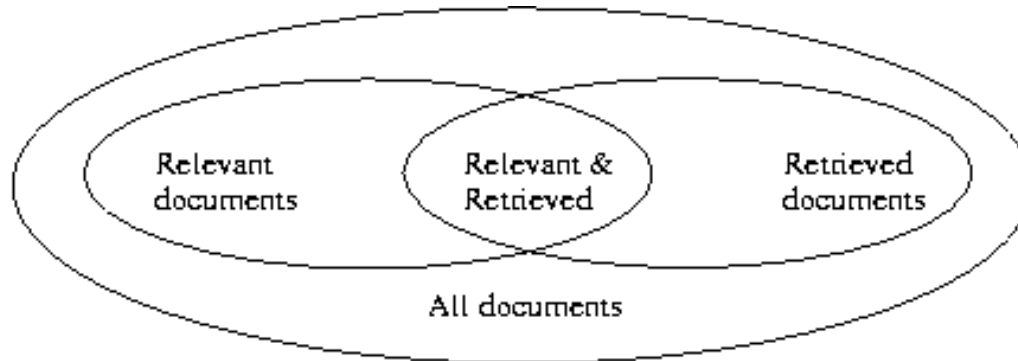
◆ Typical IR systems

- Online library catalogs
- Online document management systems

◆ Information retrieval vs. database systems

- Some DB problems are not present in IR, e.g., update, transaction management, complex objects
- Some IR problems are not addressed well in DBMS, e.g., unstructured documents, **approximate** (in contrast to exact) search using keywords and relevance

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

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

Keyword-Based Retrieval

- ◆ A document is represented by a string, which can be identified by a set of keywords
- ◆ Queries may use **expressions** of keywords
 - E.g., car *and* repair shop, tea *or* coffee, DBMS *but not* Oracle
 - Queries and retrieval should consider **synonyms**, e.g., repair and maintenance
- ◆ Major difficulties of the model
 - **Synonymy**: A keyword T does not appear anywhere in the document, even though the document is closely related to T , e.g., data mining
 - **Polysemy**: The same keyword may mean different things in different contexts, e.g., mining in data mining vs. mining in archeology

Document Data

- ◆ Each document becomes a “term” vector,
 - each term is a component (attribute) of the vector,
 - the value of each component is the number of times the corresponding term occurs in the document.

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Book Number	Word Frequency									
	The	Big-Data	Analytics	Tree	newbie	book	for	Girl	honest	
1	120	80	60	20	1	5	120	0	0	
2	110	0	0	100	10	20	100	40	10	
3	130	0	0	10	11	30	110	20	10	
4	100	0	0	2	20	40	100	10	100	
5	90	0	0	10	30	20	100	100	40	

Similarity-Based Retrieval in Text Databases

- ◆ Finds similar documents based on a set of common keywords
- ◆ Answer should be based on the degree of relevance based on the *nearness of the keywords, relative frequency of the keywords, etc.*
- ◆ Basic techniques
 - Stop list
 - ◆ Set of words that are deemed “irrelevant”, even though they may appear frequently
 - ◆ E.g., *a, the, of, for, with*, etc.
 - ◆ Stop lists may vary when document set varies

Similarity-Based Retrieval in Text Databases (cont.)

◆ Basic techniques (cont.)

- Word stem
 - ◆ Several words are small syntactic variants of each other since they share a common word stem
 - ◆ E.g., *drug, drugs, drugged*
- A term frequency table
 - ◆ Each entry $frequent_table(i, j) = \#$ of occurrences of the word t_j in document d_i
 - ◆ Usually, the *ratio* instead of the absolute number of occurrences is used
- Similarity metrics: measure the closeness of a document to a query (a set of keywords)
 - ◆ Relative term occurrences
 - ◆ Cosine distance:

$$S(v_1, v_2) = \frac{v_1 \cdot v_2}{\|v_1\| \times \|v_2\|}$$

Cosine Similarity

◆ If d_1 and d_2 are two document vectors, then

$$\cos(d_1, d_2) = \frac{d_1 \cdot d_2}{\|d_1\| \times \|d_2\|}$$

where \bullet indicates vector dot product and $\|d\|$ denotes the length of vector d .

◆ Example:

$$d_1 = \mathbf{3\ 2\ 0\ 5\ 0\ 0\ 0\ 2\ 0\ 0}$$

$$d_2 = \mathbf{1\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 2}$$

$$d_1 \bullet d_2 = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$$

$$\|d_1\| = (3*3 + 2*2 + 0*0 + 5*5 + 0*0 + 0*0 + 0*0 + 2*2 + 0*0 + 0*0)^{0.5} = (42)^{0.5} = 6.481$$

$$\|d_2\| = (1*1 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0 + 1*1 + 0*0 + 2*2)^{0.5} = (6)^{0.5} = 2.245$$

$$\cos(d_1, d_2) = 0.3150$$

Types of Text Data Mining

- ◆ Keyword-based association analysis
- ◆ Document Clustering (text categorization)
- ◆ Similarity detection
 - Cluster documents by a common author
 - Cluster documents containing information from a common source
- ◆ Link analysis: unusual correlation between entities
- ◆ Sequence analysis: predicting a recurring event
- ◆ Anomaly detection: find information that violates usual patterns
- ◆ Hypertext analysis
 - Patterns in anchors/links
 - ◆ Anchor text correlations with linked objects

Keyword-based association analysis

- ◆ Collect sets of keywords or terms that occur frequently together and then find the **association** or **correlation** relationships among them
- ◆ First preprocess the text data by parsing, stemming, removing stop words, etc.
- ◆ Then evoke association mining algorithms
 - Consider each document as a transaction
 - View a set of keywords in the document as a set of items in the transaction
- ◆ Term level association mining
 - No need for human effort in tagging documents
 - The number of meaningless results and the execution time is greatly reduced

Document Clustering

- ◆ Automatically group related documents based on their contents
- ◆ Require no training sets or predetermined taxonomies, generate a taxonomy at runtime
- ◆ Major steps
 - Preprocessing
 - ◆ Remove stop words, stem, feature extraction, lexical analysis, ...
 - Hierarchical clustering
 - ◆ Compute similarities applying clustering algorithms, ...
 - Slicing
 - ◆ Fan out controls, flatten the tree to configurable number of levels, ...

Web Mining

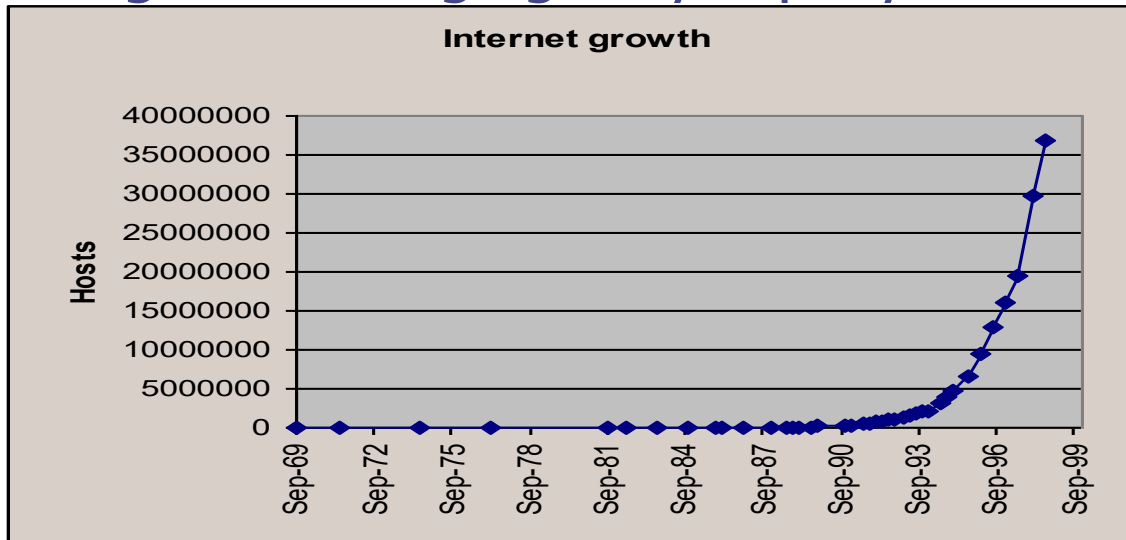
- ◆ Mining text databases
- ◆ **Mining the Web**
- ◆ Summary

Mining the World-Wide Web

- ◆ The WWW is huge, widely distributed, global information service center for
 - Information services: news, advertisements, consumer information, financial management, education, government, e-commerce, etc.
 - Hyper-link information
 - Access and usage information
- ◆ WWW provides rich sources for data mining
- ◆ Challenges
 - Too huge for effective data warehousing and data mining
 - Too complex and heterogeneous: no standards and structure

Mining the World-Wide Web (cont.)

◆ 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

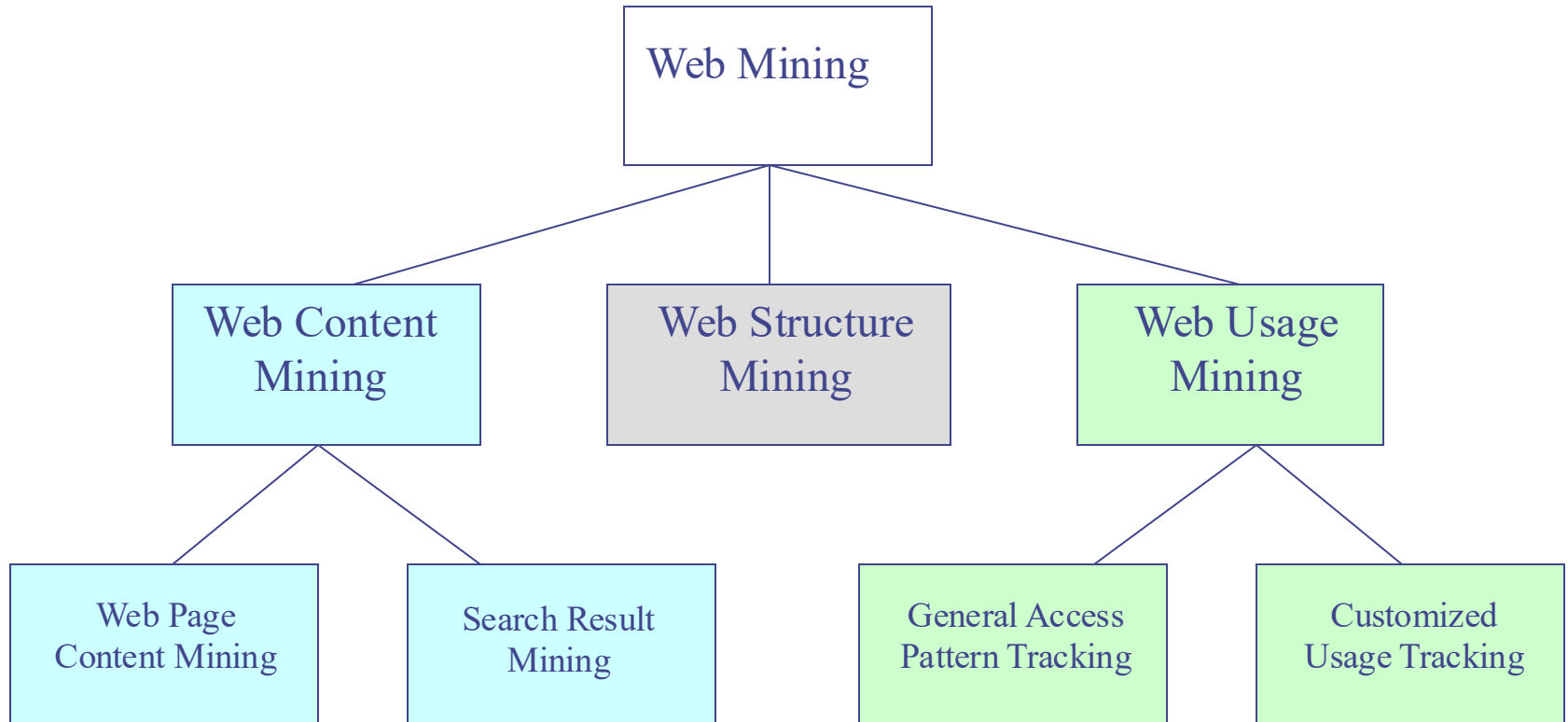
◆ Searches for

- Web access patterns
- Web structures
- Regularity and dynamics of Web contents

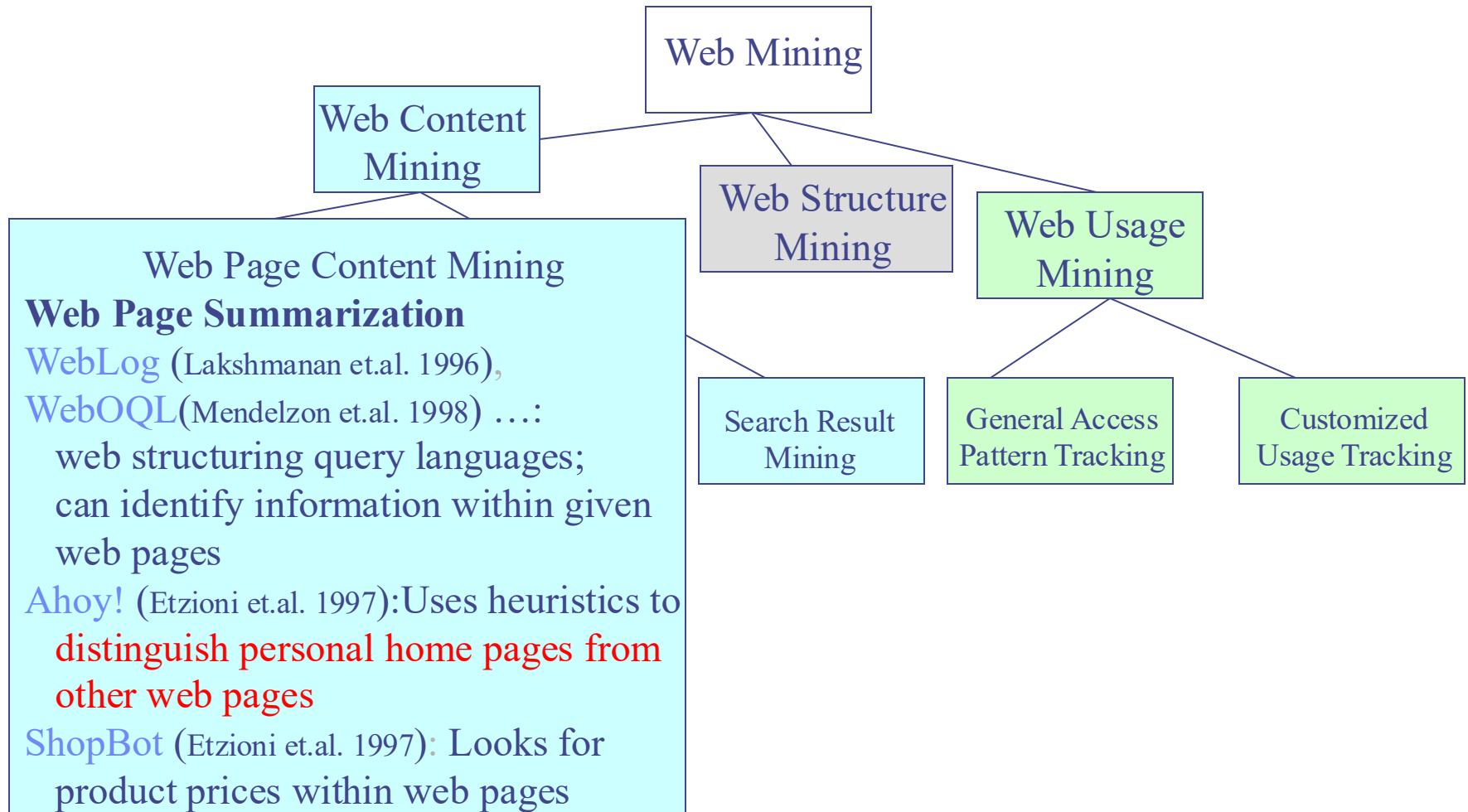
◆ Problems

- 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⇒Personalization Issue

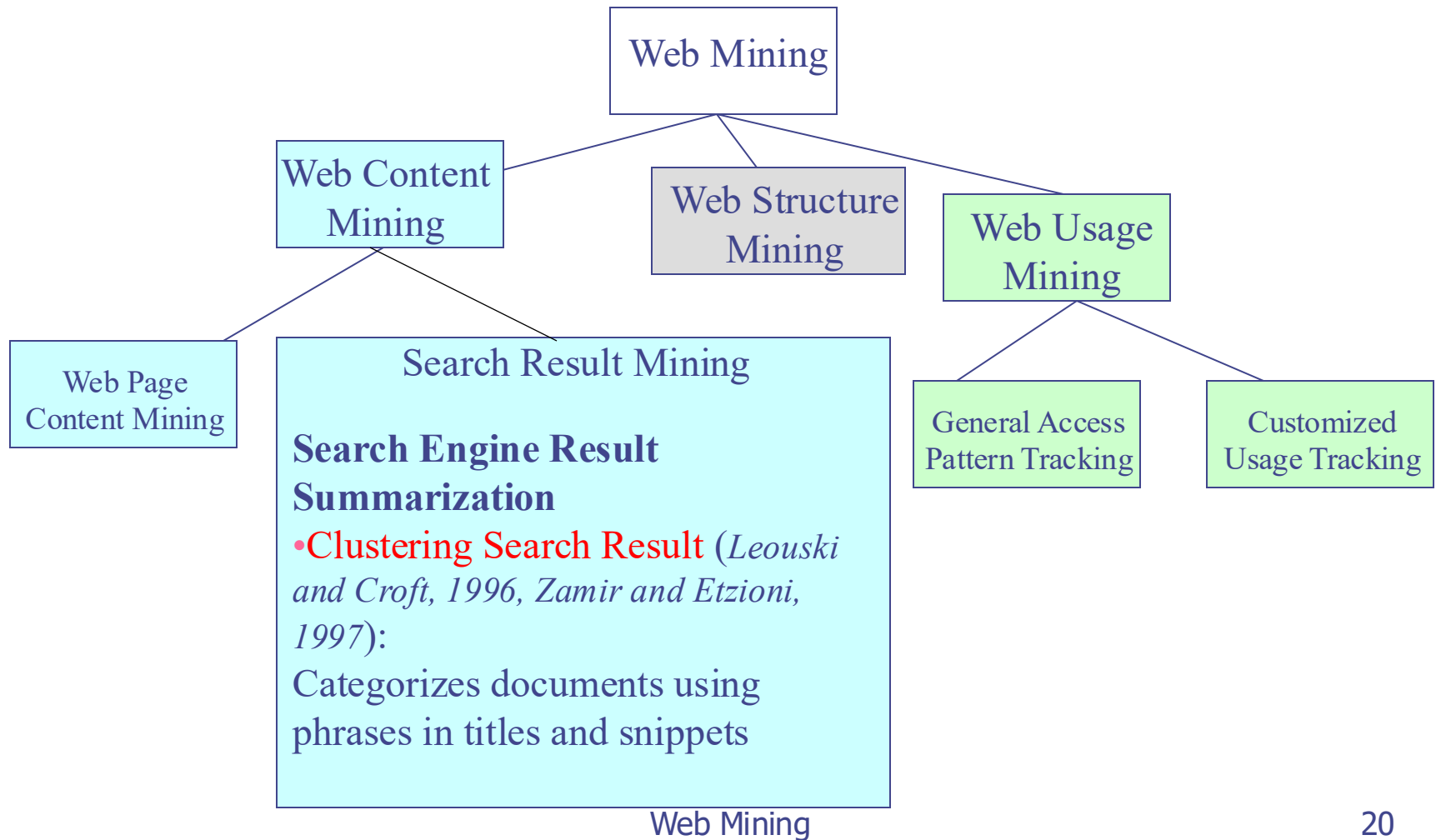
Web Mining Taxonomy



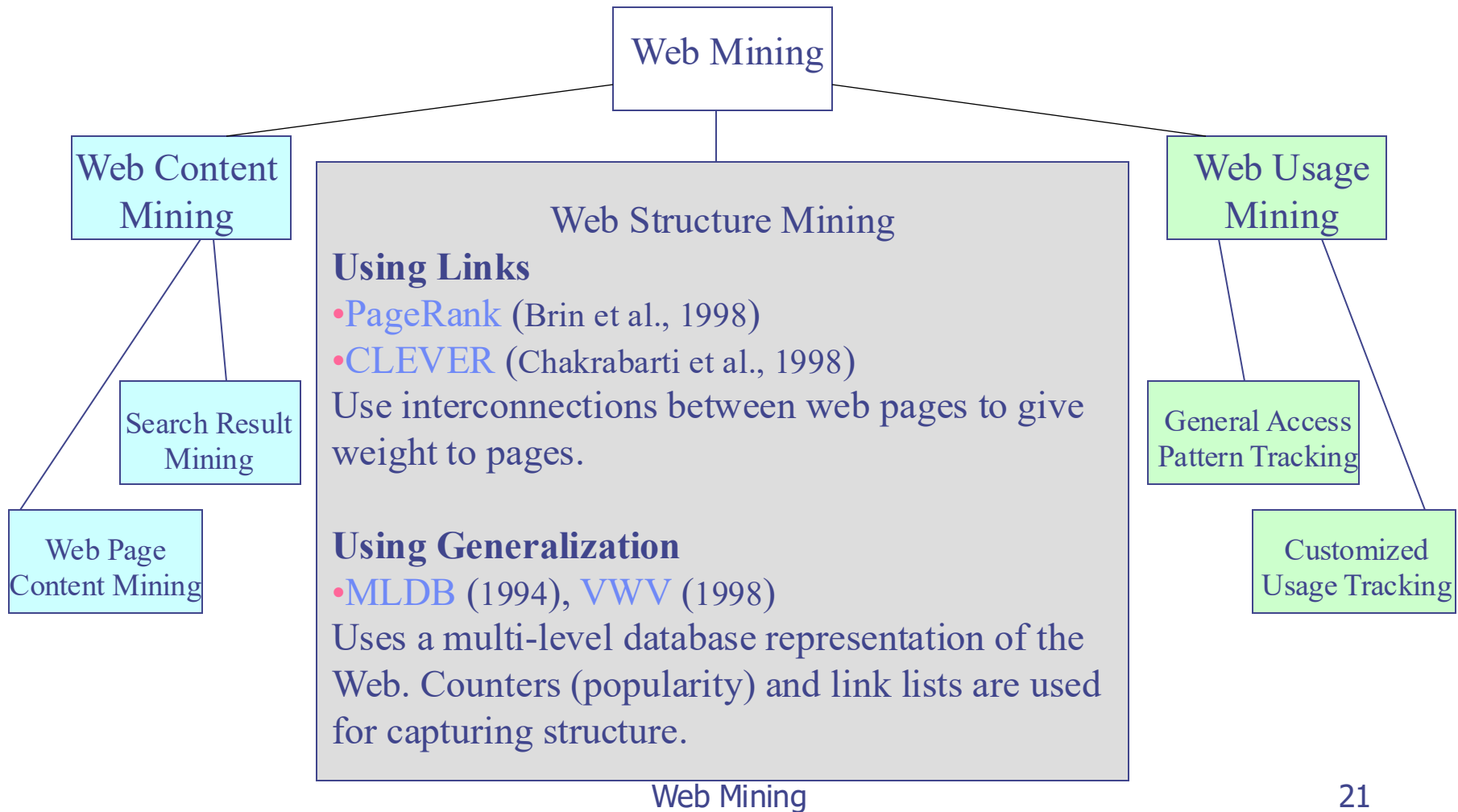
Web Mining Taxonomy (cont.)



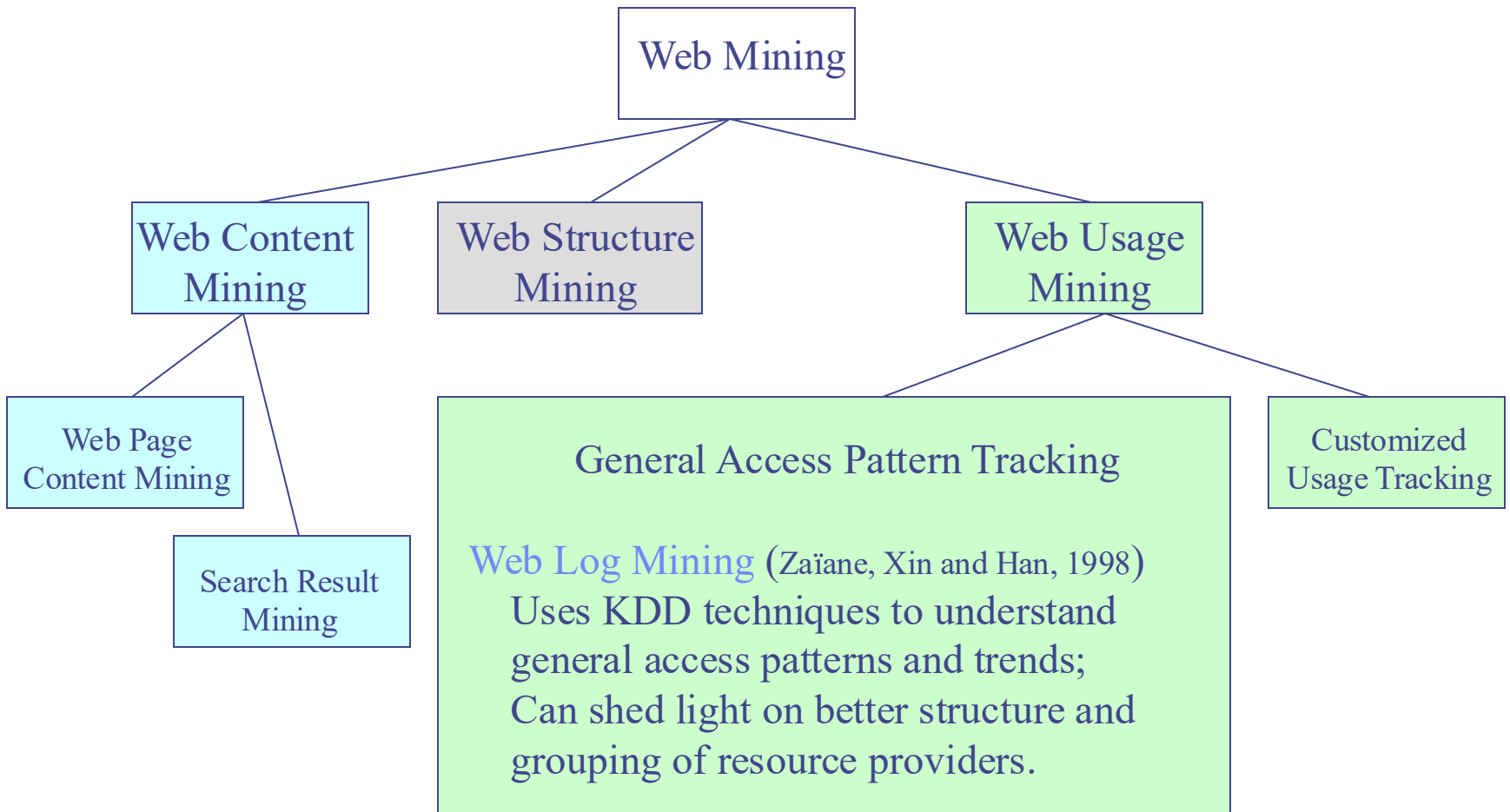
Web Mining Taxonomy (cont.)



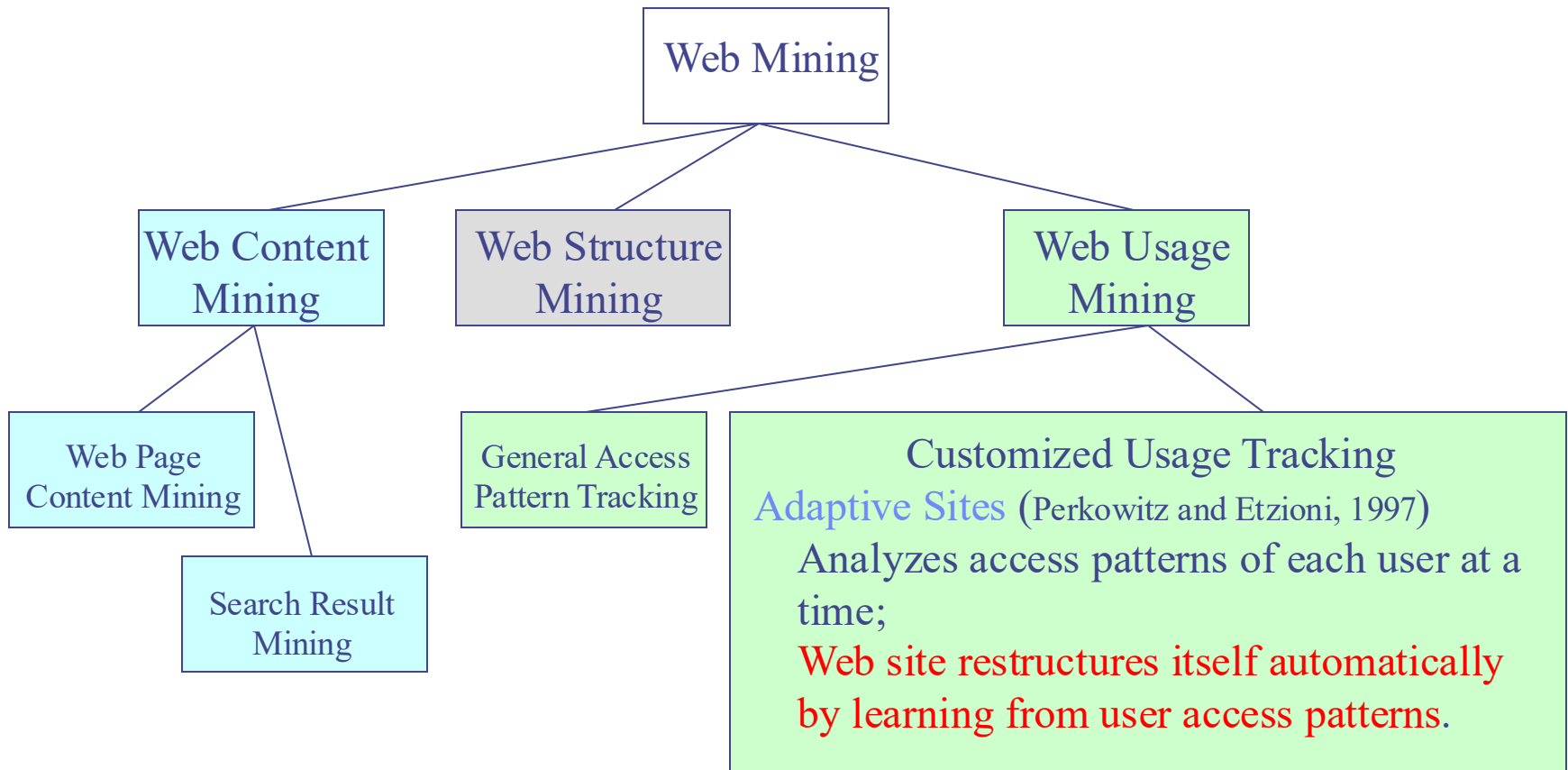
Web Mining Taxonomy (cont.)



Web Mining Taxonomy (cont.)



Web Mining Taxonomy (cont.)



Web Content Mining: Automatic Classification of Web Documents

- ◆ Assign a class label to each document from a set of predefined topic categories
- ◆ Based on a set of examples of preclassified documents
- ◆ Example
 - Use Yahoo!'s taxonomy and its associated documents as training and test sets
 - Derive a Web document classification scheme
 - Use the scheme classify new Web documents by assigning categories from the same taxonomy
- ◆ Keyword-based document classification methods
- ◆ Statistical models, Classification models, etc.

Web Usage Mining

- ◆ 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

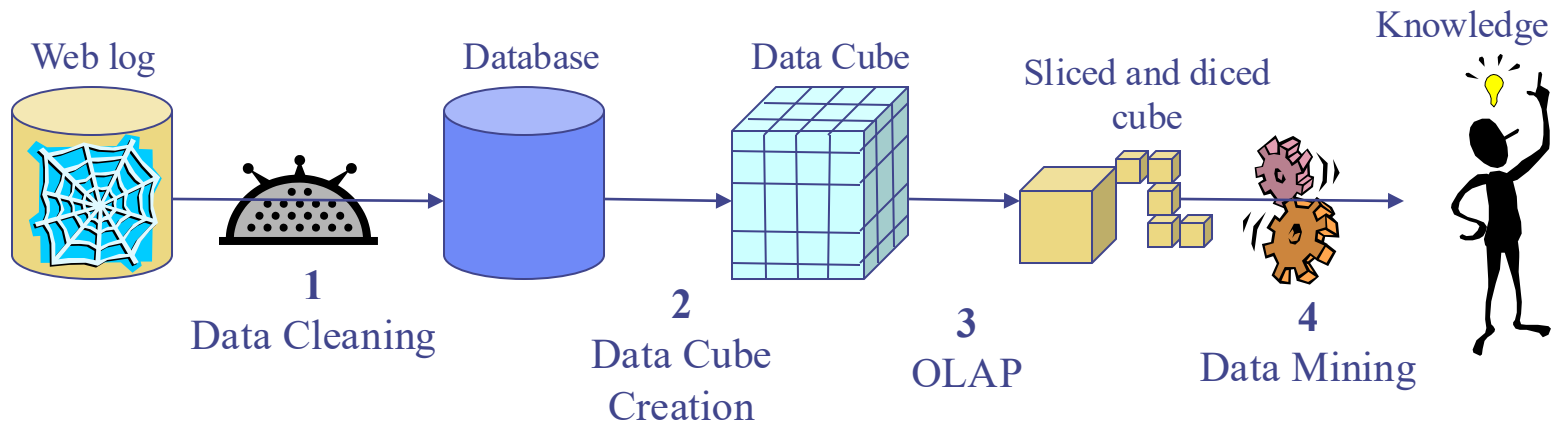
Techniques for Web Usage Mining

- ◆ Perform data mining on Weblog records
 - Find association patterns, sequential patterns, and trends of Web accessing
 - May need additional information, e.g., user browsing sequences of the Web pages in the Web server buffer
 - Treat the URL information as items
 - Treat the IP address as customer
- ◆ Conduct studies to
 - Analyze system performance, improve system design by Web caching, Web page prefetching, and Web page swapping
- ◆ Construct multidimensional view on the Weblog database
 - Perform multidimensional OLAP analysis to find the top N users, top N accessed Web pages, most frequently accessed time periods, etc.

Techniques for Web Usage Mining (Cont.)

◆ Design of a Web Log Miner

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



Web Structure Mining: Mining the Web's Link Structures

- ◆ Finding authoritative Web pages
 - Retrieving pages that are not only relevant, but also of high quality, or **authoritative** on the topic
- ◆ Hyperlinks can infer the notion of authority
 - The Web consists not only of pages, but also of hyperlinks pointing from one page to another
 - These hyperlinks contain an enormous amount of latent human annotation
 - These called upon web link analysis models

Web Link Analysis Model: PageRank

- ◆ The year 1998 was an eventful year for Web link analysis models. Both the PageRank and HITS algorithms were reported in that year.
- ◆ The connections between PageRank and HITS are quite striking.
- ◆ Since that eventful year, PageRank has emerged as the dominant link analysis model,
 - due to its query-independence,
 - its ability to combat spamming, and
 - Google's huge business success.

PageRank: The Intuitive Idea

- ◆ PageRank relies on the democratic nature of the Web by using its vast link structure as an indicator of an individual page's value or quality.
- ◆ PageRank interprets a hyperlink from page x to page y as a vote, by page x , for page y .
- ◆ However, PageRank looks at more than the sheer number of votes; it also analyzes the page that casts the vote.
 - Votes casted by “important” pages weigh more heavily and help to make other pages more “important.”
- ◆ This is exactly the idea of **rank prestige** in social network.

More specifically (PageRank)

- ◆ A hyperlink from a page to another page is an implicit conveyance of authority to the target page.
 - The more in-links that a page i receives, the more prestige the page i has.
- ◆ Pages that point to page i also have their own prestige scores.
 - A page of a higher prestige pointing to i is more important than a page of a lower prestige pointing to i .
 - In other words, a page is important if it is pointed to by other important pages.

Mining the Web's Link Structures

◆ Problems with the Web linkage structure

- Not every hyperlink represents an endorsement
 - ◆ Other purposes are for navigation or for paid advertisements
 - ◆ If the majority of hyperlinks are for endorsement, the collective opinion will still dominate
- One authority will seldom have its Web page point to its rival authorities in the same field
- Authoritative pages are seldom particularly descriptive

◆ Hub

- Set of Web pages that provides collections of links to authorities

Summary

- ◆ 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 (Web structure mining), the automatic classification of Web documents (Web content mining), and Weblog mining (Web usage mining)