

Web Technologies and Applications

Winter 2001

CMPUT 499: Web Mining

Dr. Osmar R. Zaiane



University of Alberta

Course Content

- | | |
|--|---|
| <ul style="list-style-type: none">• Introduction• Internet and WWW• Protocols• HTML and beyond• Animation & WWW• Java Script• Dynamic Pages• Perl Intro.• Java Applets | <ul style="list-style-type: none">• Databases & WWW• SGML / XML• Managing servers• Search Engines• Web Mining• CORBA• Security Issues• Selected Topics• Projects |
|--|---|



Objectives of Lecture 14

Web Mining

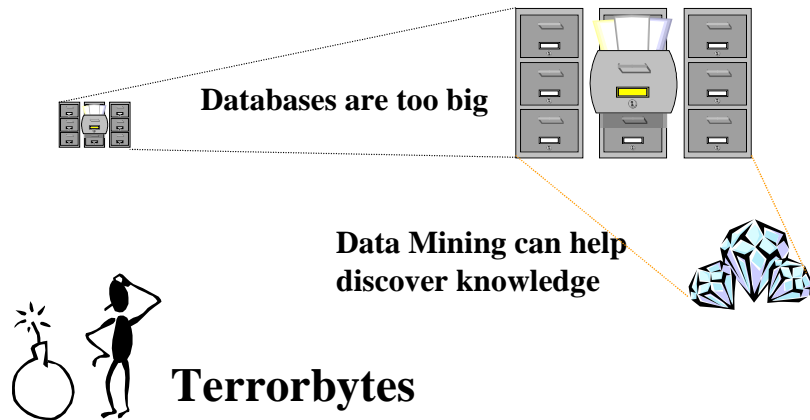
- Get an overview about the functionalities and the issues in data mining.
- Understand the different knowledge discovery issues in data mining from the World Wide Web.
- Distinguish between resource discovery and Knowledge discovery from the Internet.
- Present some problems and explore cutting-edge solutions

Outline of Lecture 14



- Introduction to Data Mining
- Introduction to Web Mining
 - What are the incentives of web mining?
 - What is the taxonomy of web mining?
- Web Content Mining: Getting the Essence From Within Web Pages.
- Web Structure Mining: Are Hyperlinks Information?
- Web Usage Mining: Exploiting Web Access Logs.

We Are Data Rich but Information Poor



What Should We Do?



We are not trying to find the needle in the haystack because DBMSs know how to do that.



We are merely trying to understand the consequences of the presence of the needle, if it exists.

What Led Us To This?

Necessity is the Mother of Invention

- Technology is available to help us collect data
 - Bar code, scanners, satellites, cameras, etc.
- Technology is available to help us store data
 - Databases, data warehouses, variety of repositories...
- We are starving for knowledge (competitive edge, research, etc.)

We are swamped by data that continuously pours on us.

1. We do not know what to do with this data
2. We need to interpret this data in search for new knowledge

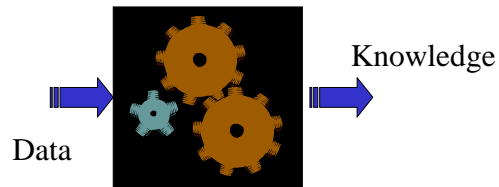
Evolution of Database Technology

- **1950s:** First computers, use of computers for census
- **1960s:** Data collection, database creation (hierarchical and network models)
- **1970s:** Relational data model, relational DBMS implementation.
- **1980s:** Ubiquitous RDBMS, advanced data models (extended-relational, OO, deductive, etc.) and application-oriented DBMS (spatial, scientific, engineering, etc.).
- **1990s:** Data mining and data warehousing, massive media digitization, multimedia databases, and Web technology.

Notice that storage prices have consistently decreased in the last decades

What Is Our Need?

Extract interesting knowledge
(rules, regularities, patterns, constraints)
from data in large collections.



A Brief History of Data Mining Research

- [1989 IJCAI Workshop on Knowledge Discovery in Databases](#) (Piatetsky-Shapiro)



Knowledge Discovery in Databases

(G. Piatetsky-Shapiro and W. Frawley, 1991)

- [1991-1994 Workshops on Knowledge Discovery in Databases](#)



Advances in Knowledge Discovery and Data Mining

(U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1996)

- [1995-1998 International Conferences on Knowledge Discovery in Databases and Data Mining \(KDD'95-98\)](#)



Journal of Data Mining and Knowledge Discovery (1997)

- [1998-2000 ACM SIGKDD'98-2000 conferences](#)

What kind of information are we collecting?



- Business transactions
- Scientific data
- Medical and personal data
- Surveillance video and pictures
- Satellite sensing
- Games



Data Collected (Con't)



- Digital media
- CAD and Software engineering
- Virtual worlds
- Text reports and memos
- The World Wide Web



What are Data Mining and Knowledge Discovery?




Knowledge Discovery:


Process of non trivial extraction of implicit, previously unknown and potentially useful information from large collections of data





Many Steps in KD Process

- Gathering the data together 

- Cleanse the data and fit it in together 

- Select the necessary data 

- Crunch and squeeze the data to extract the *essence* of it 

- Evaluate the output and use it 

So What Is Data Mining?

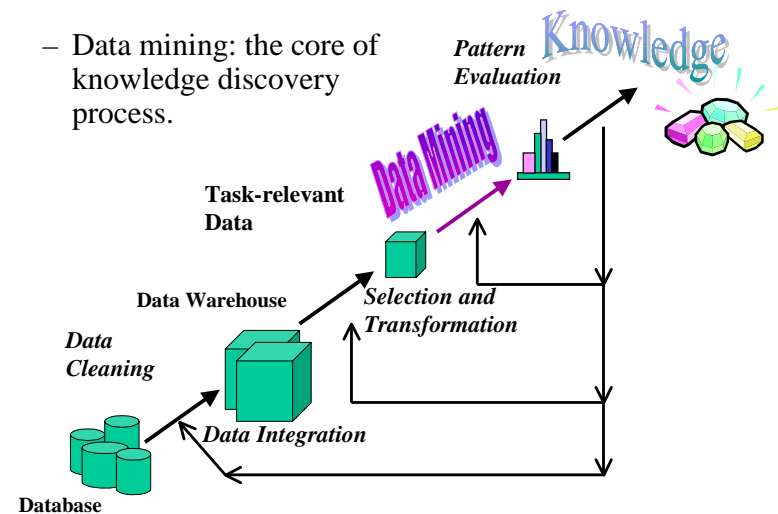


- In theory, *Data Mining* is a step in the knowledge discovery process. It is the extraction of **implicit information from a large dataset**.
- In practice, data mining and knowledge discovery are becoming synonyms.
- There are other equivalent terms: KDD, knowledge extraction, discovery of regularities, patterns discovery, data archeology, data dredging, business intelligence, information harvesting...
- Notice the misnomer for data mining. Shouldn't it be knowledge mining?

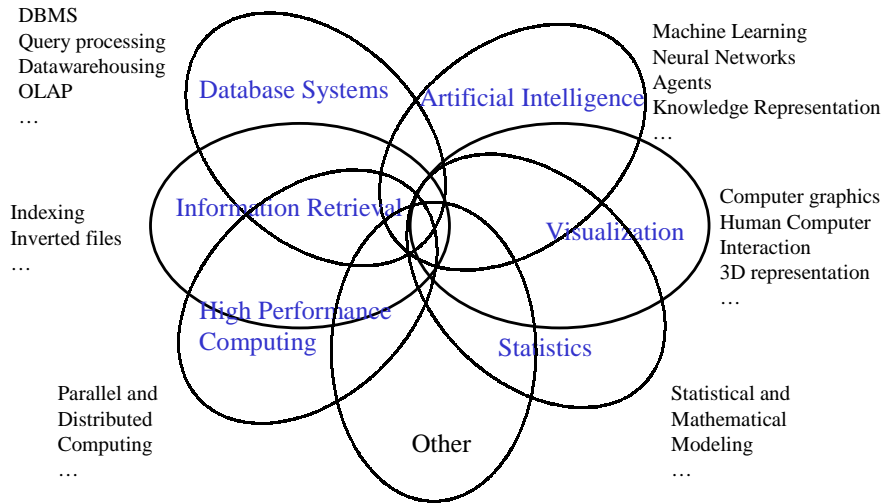


Data Mining: A KDD Process

- Data mining: the core of knowledge discovery process.



KDD at the Confluence of Many Disciplines

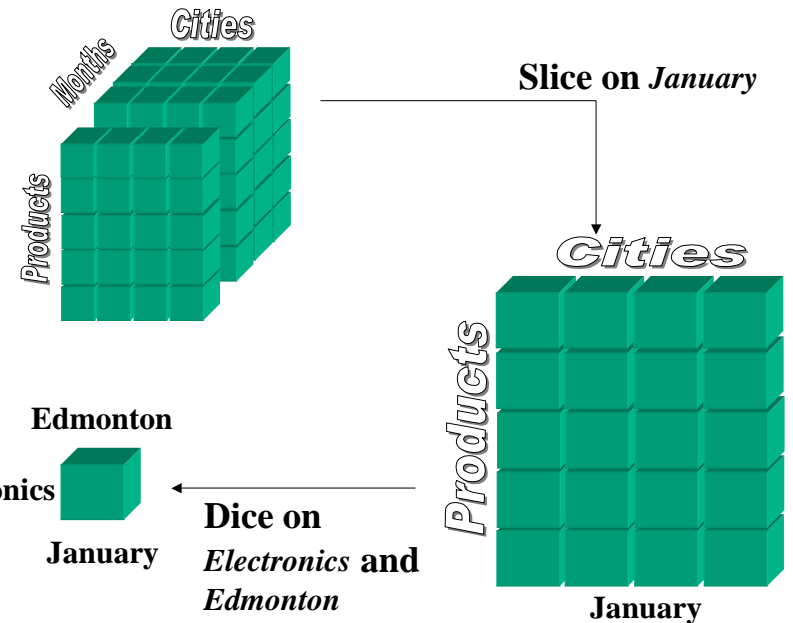
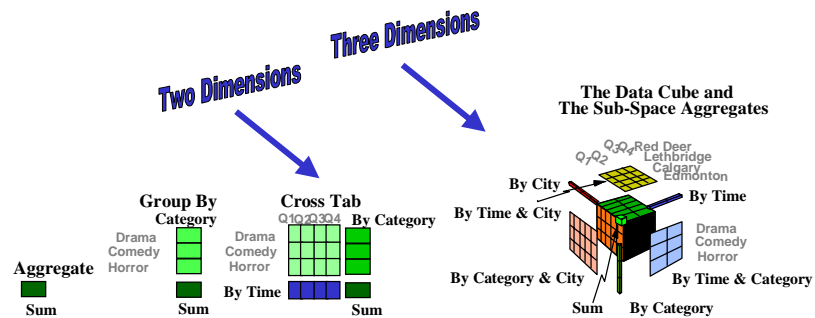


Data Mining: On What Kind of Data?

- Flat Files
- Heterogeneous and legacy databases
- Relational databases
and other DB: Object-oriented and object-relational databases
- Transactional databases
Transaction(TID, Timestamp, UID, {item1, item2,...})

Data Mining: On What Kind of Data?

- Data warehouses



Data Mining: On What Kind of Data?

- Multimedia databases



- Spatial Databases



- Time Series Data and Temporal Data



Data Mining: On What Kind of Data?

- Text Documents



- The World Wide Web

➤ The content of the Web

➤ The structure of the Web

➤ The usage of the Web



What Can Be Discovered?

What can be discovered depends upon the data mining task employed.

- Descriptive DM tasks
Describe general properties
- Predictive DM tasks
Infer on available data



Data Mining Functionality

- **Characterization:**
Summarization of general features of objects in a target class.
(Concept description)
Ex: Characterize grad students in Science
- **Discrimination:**
Comparison of general features of objects between a target class and a contrasting class. (Concept comparison)
Ex: Compare students in Science and students in Arts
- **Association:**
Studies the frequency of items occurring together in transactional databases.
Ex: buys(x, bread) → buys(x, milk).

Data Mining Functionality (Con't)

- **Prediction:**
Predicts some unknown or missing attribute values based on other information.
Ex: Forecast the sale value for next week based on available data.
- **Classification:**
Organizes data in given classes based on attribute values. (supervised classification)
Ex: classify students based on final result.
- **Clustering:**
Organizes data in classes based on attribute values. (unsupervised classification)
Ex: group crime locations to find distribution patterns.
Minimize inter-class similarity and maximize intra-class similarity

Data Mining Functionality (Con't)

- **Outlier analysis:**
Identifies and explains exceptions (surprises)
- **Time-series analysis:**
Analyzes trends and deviations; regression, sequential pattern, similar sequences...

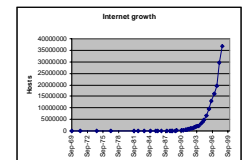
Outline of Lecture 14



- Introduction to Data Mining
- Introduction to Web Mining
 - What are the incentives of web mining?
 - What is the taxonomy of web mining?
- Web Content Mining: Getting the Essence From Within Web Pages.
- Web Structure Mining: Are Hyperlinks Information?
- Web Usage Mining: Exploiting Web Access Logs.

WWW: Facts

- No standards, unstructured and heterogeneous
- Growing and changing very rapidly
 - One new WWW server every 2 hours
 - 5 million documents in 1995
 - 320 million documents in 1998
 - More than 1 billion in 2000
- Indices get stale very quickly



Need for better resource discovery and knowledge extraction.



The Asilomar Report urges the database research community to contribute in deploying new technologies for resource and information retrieval from the World-Wide Web.

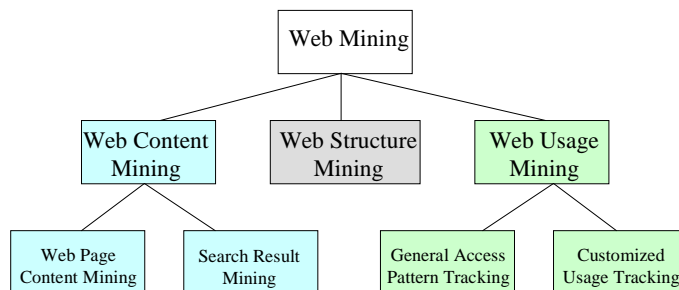
WWW: Incentives

- Enormous wealth of information on web
- The web is a huge collection of:
 - Documents of all sorts
 - Hyper-link information
 - Access and usage information
- Mine interesting nuggets of information leads to wealth of information and knowledge
- Challenge: Unstructured, huge, dynamic.

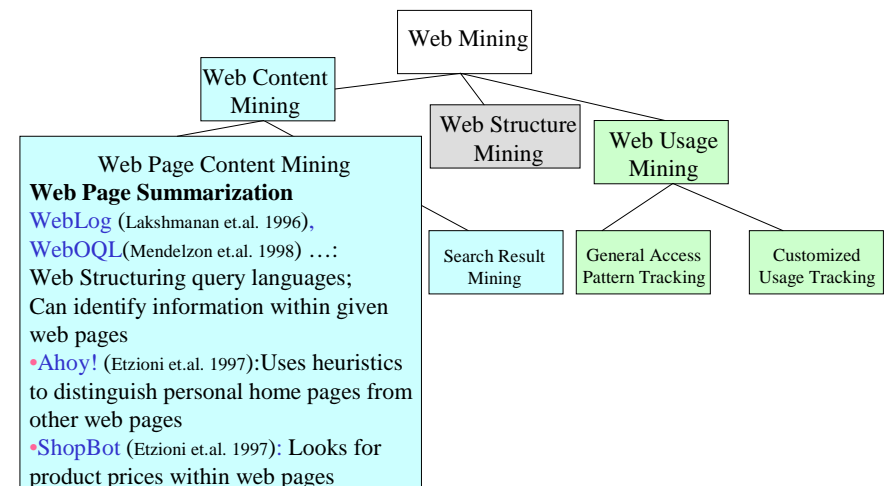
WWW and Web Mining

- Web: A huge, widely-distributed, highly heterogeneous, semi-structured, interconnected, evolving, hypertext/hypermedia information repository.
- Problems:
 - the “*abundance*” problem:
 - 99% of info of no interest to 99% of people
 - *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

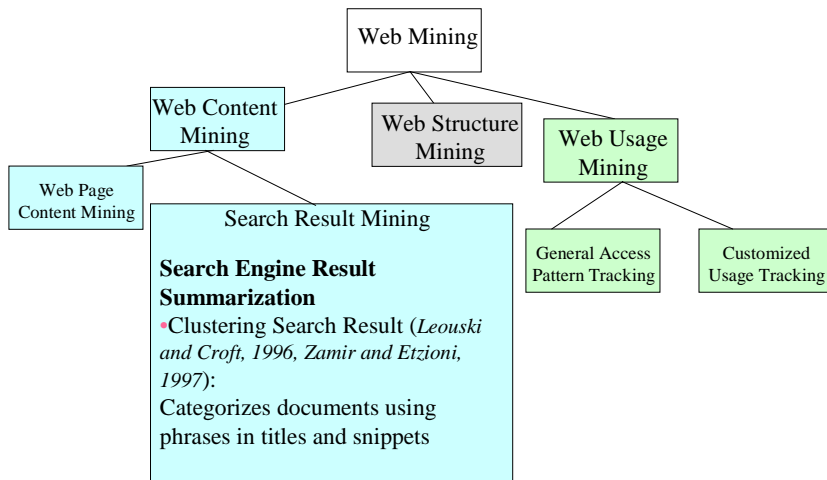
Web Mining Taxonomy



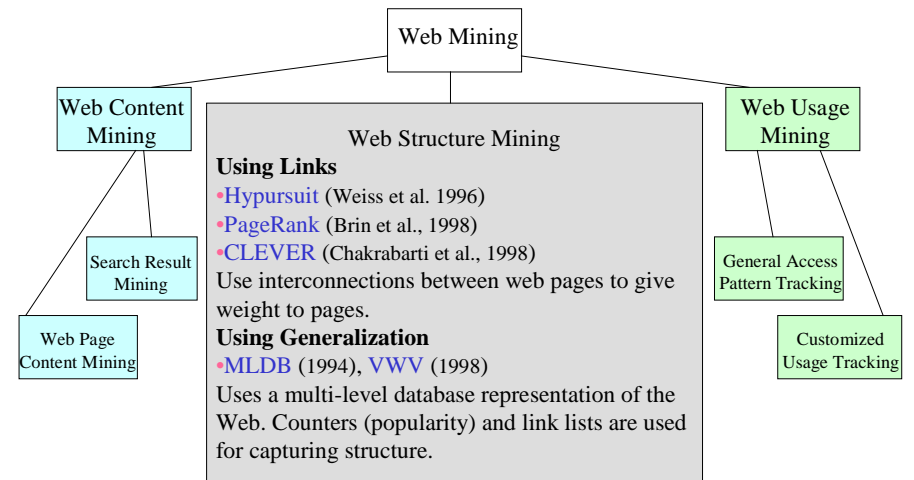
Web Mining Taxonomy



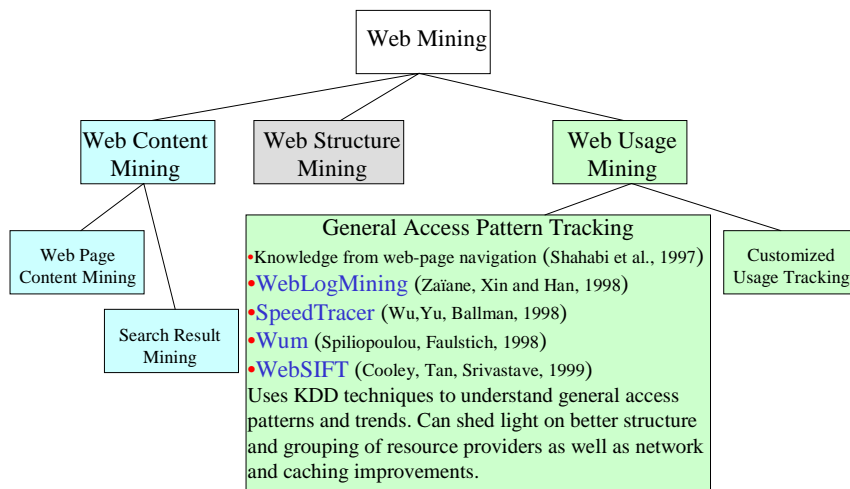
Web Mining Taxonomy



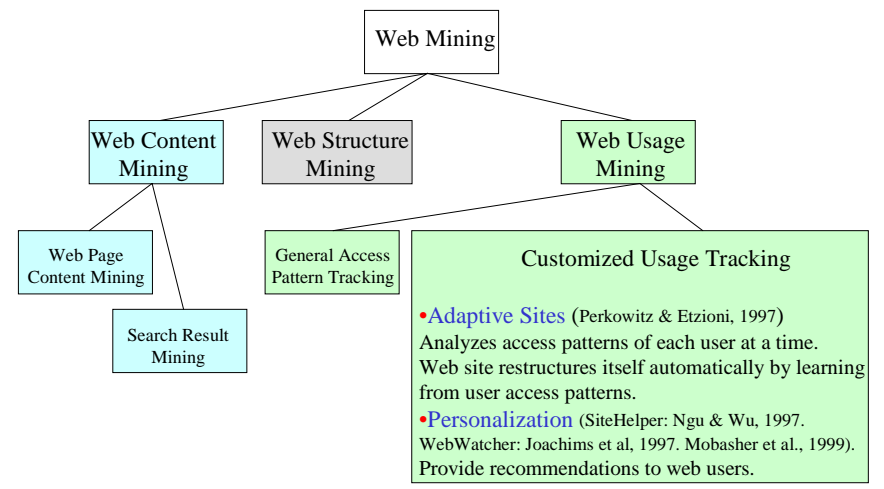
Web Mining Taxonomy



Web Mining Taxonomy



Web Mining Taxonomy

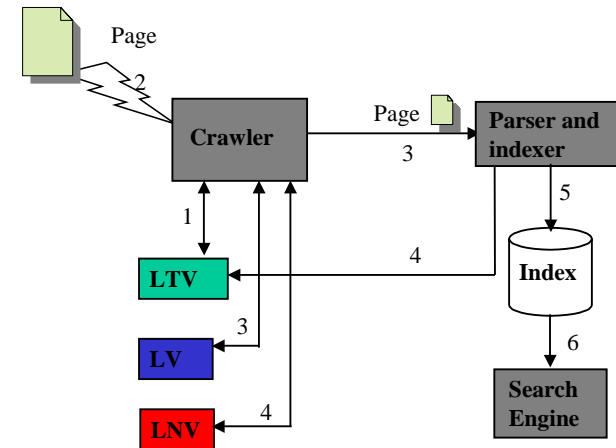


Outline of Lecture 14



- Introduction to Data Mining
- Introduction to Web Mining
 - What are the incentives of web mining?
 - What is the taxonomy of web mining?
- **Web Content Mining: Getting the Essence From Within Web Pages.**
- Web Structure Mining: Are Hyperlinks Information?
- Web Usage Mining: Exploiting Web Access Logs.

Search Engine General Architecture

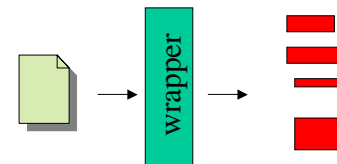


Search Engines are not Enough

- Most of the knowledge in the World-Wide Web is buried inside documents.
- Search engines (and crawlers) barely scratch the surface of this knowledge by extracting keywords from web pages.
- There is text mining, text summarization, natural language statistical analysis, etc., but not the scope of this tutorial.

Web page Summarization or Web Restructuring

- Most of the suggested approaches are limited to known groups of documents, and use custom-made wrappers.



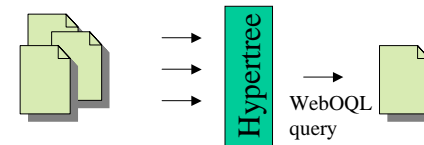
Ahoy!
WebOQL
Shopbot
...

Discovering Personal Homepages

- Ahoy! (shakes et al. 1997) uses Internet services like search engines to retrieve resources a person's data.
- Search results are parsed and using heuristics, typographic and syntactic features are identified inside documents.
- Identified features can betray personal homepages.

Query Language for Web Page Restructuring

- WebOQL (Arocena et al. 1998) is a declarative query language that retrieves information from within Web documents.
- Uses a graph hypertree representation of web documents.



- CNN pages
- Tourist guides
- Etc.

Shopbot

- Shopbot (Doorendos et al. 1997) is shopping agent that analyzes web page content to identify price lists and special offers.
- The system learns to recognize document structures of on-line catalogues and e-commerce sites.
- Has to adjust to the page content changes.

Mine What Web Search Engine Finds

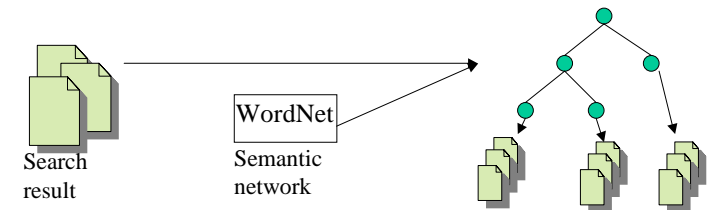
- Current Web search engines: convenient source for mining
 - keyword-based, return too many answers, low quality answers, still missing a lot, not customized, etc.
- Data mining will help:
 - coverage: “Enlarge and then shrink,” using synonyms and conceptual hierarchies
 - better search primitives: user preferences/hints
 - linkage analysis: authoritative pages and clusters
 - Web-based languages: XML + WebSQL + WebML
 - customization: home page + Weblog + user profiles

Refining and Clustering Search Engine Results

- WebSQL (Mendelzon et al. 1996) is an SQL-like declarative language that provides the ability to retrieve pertinent documents.
- Web documents are parsed and represented in tables to allow result refining.
- [Zamir et al. 1998] present a technique using COBWEB that relies on snippets from search engine results to cluster documents in significant clusters.

Ontology for Search Results

- There are still too many results in typical search engine responses.
- Reorganize results using a semantic hierarchy (Zaiane et al. 2001).



Outline of Lecture 14



- Introduction to Data Mining
- Introduction to Web Mining
 - What are the incentives of web mining?
 - What is the taxonomy of web mining?
- Web Content Mining: Getting the Essence From Within Web Pages.
- Web Structure Mining: Are Hyperlinks Information?
- Web Usage Mining: Exploiting Web Access Logs.

Web Structure Mining

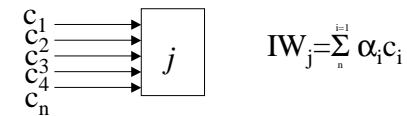
- Hyperlink structure contains an enormous amount of concealed human annotation that can help automatically infer notions of “authority” in a given topic.
- Web structure mining is the process of extracting knowledge from the interconnections of hypertext document in the world wide web.
- Discovery of influential and authoritative pages in WWW.

Citation Analysis in Information Retrieval

- Citation analysis was studied in information retrieval long before WWW came into scene.
- Garfield's *impact factor* (1972): It provides a numerical assessment of journals in the journal citation.
- Kwok (1975) showed that using citation titles leads to good cluster separation.

Citation Analysis in Information Retrieval

- Pinski and Narin (1976) proposed a significant variation on the notion of impact factor, based on the observation that not all citations are equally important.
 - A journal is influential if, recursively, it is heavily cited by other influential journals.
 - *influence weight*: The influence of a journal j is equal to the sum of the influence of all journals citing j , with the sum weighted by the amount that each cites j .



HyPursuit

- Hypursuit (Weiss et al. 1996) groups resources into clusters according to some criteria. Clusters can be clustered again into clusters of upper level, and so on into a hierarchy of clusters.
- Clustering Algorithm
 - Computes clusters: set of related pages based on the semantic info embedded in hyperlink structure and other criteria.
 - abstraction function

Search for Authoritative Pages

A good authority is a page pointed by many good hubs, while a good hub is a page that point to many good authorities.

This mutually enforcing relationship between the hubs and authorities serves as the central theme in our exploration of link based method for search, and the automated compilation of high-quality web resources.

Hyperlink Induced Topic Search (HITS)

- Kleinberg's HITS algorithm (1998) uses a simple approach to finding quality documents and assumes that if document A has a hyperlink to document B, then the author of document A thinks that document B contains valuable information.
- If A is seen to point to a lot of good documents, then A's opinion becomes more valuable and the fact that A points to B would suggest that B is a good document as well.

General HITS Strategy

HITS algorithm applies two main steps.

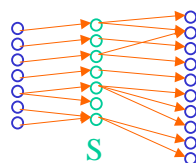
- A sampling component which constructs a focused collection of thousand web pages likely to be rich in authorities.
- A weight-propagation component, which determines the numerical estimates of hub and authority weights by an iterative procedure.

Steps of HITS Algorithm

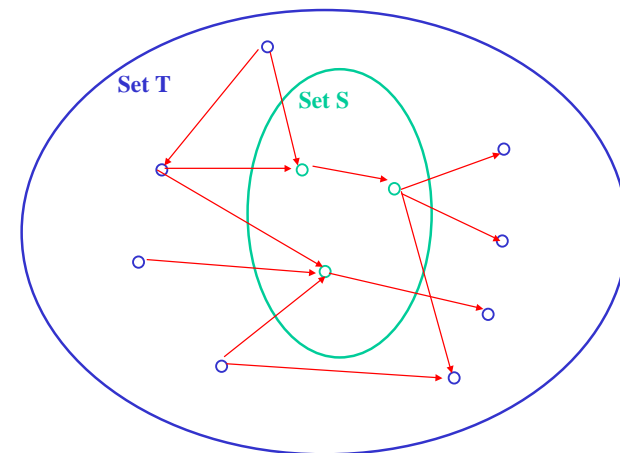
- Starting from a user supplied query, HITS assembles an initial set S of pages:

The initial set of pages is called root set.

These pages are then expanded to a larger root set T by adding any pages that are linked to or from any page in the initial set S.



- HITS then associates with each page p a hub weight $h(p)$ and an authority weight $a(p)$, all initialized to one.



- HITS then iteratively updates the hub and authority weights of each page.

Let $p \rightarrow q$ denote “page p has an hyperlink to page q ”. HITS updates the hubs and authorities as follows:

$$a(p) = \sum_{p \rightarrow q} h(q)$$

$$h(p) = \sum_{q \rightarrow p} a(q)$$

Further Enhancement for Finding Authoritative Pages in WWW

- The CLEVER system (Chakrabarti, et al. 1998-1999)
 - builds on the algorithmic framework of extensions based on both content and link information.
- Extension 1: mini-hub pagelets
 - prevent "*topic drifting*" on large hub pages with many links, based on the fact: Contiguous set of links on a hub page are more focused on a single topic than the entire page.
- Extension 2. Anchor text
 - make use of the text that surrounds hyperlink definitions (href's) in Web pages, often referred to as *anchor* text
 - boost the weights of links which occur near instances of query terms.

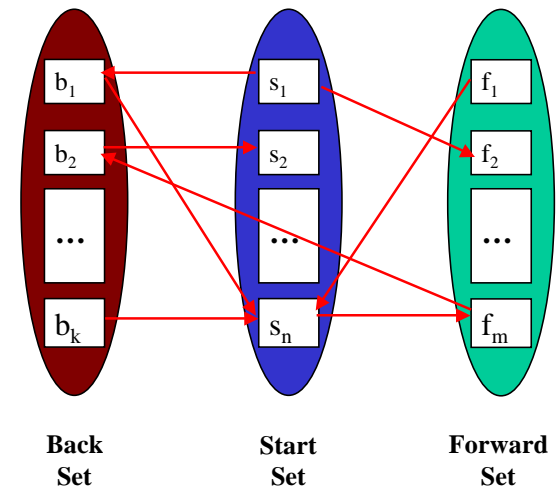
CLEVER System

- The output of the HITS algorithm for the given search topic is a short list consisting of the pages with largest hub weights and the pages with largest authority weights.
- HITS uses a purely link-based computation once the root set has been assembled, with no further regard to the query terms.
- In HITS all the links out of a hub page propagate the same weight, the algorithm does not take care of hubs with multiple topics.

Connectivity Server

- Connectivity server (Bharat et al. 1998) also exploit linkage information to find most relevant pages for a query.
- HITS algorithm and CLEVER uses the 200 pages indexed by the AltaVista search engine as the base set.
- Connectivity Server uses entire set of pages returned by the AltaVista search engines to find result of the query.

- Connectivity server in its base operation, the server accept a query consisting of a set L of one or more URLs and returns a list of all pages that point to pages in L (predecessors) and list of all pages that are pointed to from pages in L (successors).
- Using this information Connectivity Server includes information about all the links that exist among pages in the neighborhood.



- The neighborhood graph is the graph produced by a set L of start pages and the predecessors of L, and all the successors of L and the edges among them.
- Once the neighborhood graph is created, the Connectivity server uses Kleinberg's method to analyze and detect useful pages and to rank computation on it.
- Outlier filtering (Bharat & Henzinger 1998-1999) integrates textual content: nodes in neighborhood graph are term vectors. During graph expansion, prune nodes distant from query term vector. Avoids contamination from irrelevant links.

Ranking Pages Based on Popularity

- Page-rank method (Brin and Page, 1998): Rank the "importance" of Web pages, based on a model of of a "random browser."
 - Initially used to select pages to revisit by crawler.
 - Ranks pages in Google's search results.
- In a simulated web crawl, following a random link of each visited page may lead to the revisit of popular pages (pages often cited).
- Brin and Page view Web searches as random walks to assign a topic independent "rank" to each page on the world wide web, which can be used to reorder the output of a search engine.
- The number of visits to each page is its PageRank. PageRank estimates the visitation rate => popularity score.

Each Page p has a number of links coming out of it $C(p)$ (C for citation), and number of pages pointing at page p_1, p_2, \dots, p_n .

PageRank of P is obtained by

$$PR(p) = (1-d) + \left(\sum_{k=1}^n \frac{PR(p_k)}{C(p_k)} \right)$$

Comparison

- Google assigns initial ranking and retains them independently of any queries. This makes it faster.
- CLEVER and Connectivity server assembles different root set for each search term and prioritizes those pages in the context of the particular query.
- Google works in the forward direction from link to link.
- CLEVER and Connectivity server looks both in the forward and backward direction.
- Both the page-rank and hub/authority methodologies have been shown to provide qualitatively good search results for broad query topics on the WWW.
- Hyperclass (Chakrabarti 1998) uses content and links of exemplary page to focus crawling of relevant web space.

Nepotistic Links

- Nepotistic links are links between pages that are present for reasons other than merit.
- Spamming is used to trick search engines to rank some documents high.
- Some search engines use hyperlinks to rank documents (ex. Google) it is thus necessary to identify and discard nepotistic links.
- Recognizing Nepotistic Links on the Web (Davidson 2000).
- Davidson uses C4.5 classification algorithm on large number of page attributes, trained on manually labeled pages.

Outline of Lecture 14



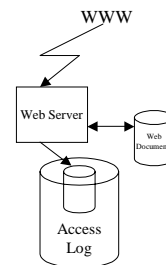
- Introduction to Data Mining
- Introduction to Web Mining
 - What are the incentives of web mining?
 - What is the taxonomy of web mining?
- Web Content Mining: Getting the Essence From Within Web Pages.
- Web Structure Mining: Are Hyperlinks Information?
- Web Usage Mining: Exploiting Web Access Logs.

Existing Web Log Analysis Tools

- There are many commercially available applications.
 - Many of them are slow and make assumptions to reduce the size of the log file to analyse.
- Frequently used, pre-defined reports:
 - Summary report of hits and bytes transferred
 - List of top requested URLs
 - List of top referrers
 - List of most common browsers
 - Hits per hour/day/week/month reports
 - Hits per Internet domain
 - Error report
 - Directory tree report, etc.
- Tools are limited in their performance, comprehensiveness, and depth of analysis.

What Is Weblog Mining?

- Web Servers register a log entry for every single access they get.
- A huge number of accesses (hits) are registered and collected in an ever-growing web log.
- Weblog mining:
 - Enhance server performance
 - Improve web site navigation
 - Improve system design of web applications
 - Target customers for electronic commerce
 - Identify potential prime advertisement locations



Web Server Log File Entries

IP address	User ID	Timestamp	Method	URL/Path	Status	Size	Referrer	Agent	Cookie
------------	---------	-----------	--------	----------	--------	------	----------	-------	--------

dd23-125.compuserve.com - rhuia [01/Apr/1997:00:03:25 -0800] "GET /SFU/cgi-bin/VG/VG_dspmsg.cgi?ci=40154&mi=49 HTTP/1.0" 200 417

129.128.4.241 - [15/Aug/1999:10:45:32 - 0800] "GET /source/pages/chapter1.html" 200 618 /source/pages/index.html Mozilla/3.04(Win95)

Diversity of Weblog 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 caching, prefetching, swapping
- Trend analysis:
 - Dynamics of the Web: what has been changing?
- Customized to individual users

More on Log Files

- Information NOT contained in the log files:
 - use of browser functions, e.g. backtracking within-page navigation, e.g. scrolling up and down
 - requests of pages stored in the cache
 - requests of pages stored in the proxy server
 - Etc.
- Special problems with dynamic pages:
 - different user actions call same cgi script
 - same user action at different times may call different cgi scripts
 - one user using more than one browser at a time
 - Etc.

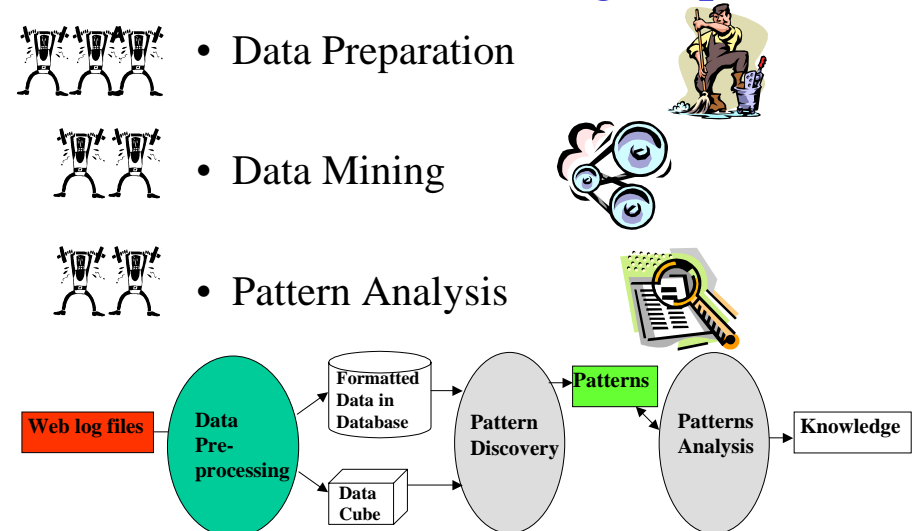
Use of Log Files

- Basic summarization:
 - Get frequency of individual actions by user, domain and session.
 - Group actions into activities, e.g. reading messages in a conference
 - Get frequency of different errors.
- Questions answerable by such summary:
 - Which components or features are the most/least used?
 - Which events are most frequent?
 - What is the user distribution over different domain areas?
 - Are there, and what are the differences in access from different domains areas or geographic areas?

In-Depth Analysis of Log Files

- In-depth analyses:
 - pattern analysis, e.g. between users, over different courses, instructional designs and materials, as Virtual-U features are added or modified
 - trend analysis, e.g. user behaviour change over time, network traffic change over time
- Questions can be answered by in-depth analyses:
 - In what context are the components or features used?
 - What are the typical event sequences?
 - What are the differences in usage and access patterns among users?
 - What are the differences in usage and access patterns over courses?
 - What are the overall patterns of use of a given environment?
 - What user behaviors change over time?
 - How usage patterns change with quality of service (slow/fast)?
 - What is the distribution of network traffic over time?

Main Web Mining steps



Data Pre-Processing

Problems:

- Identify types of pages: content page or navigation page.
- Identify visitor (user)
- Identify session, transaction, sequence, episode, action,...
- Inferring cached pages
- Identifying visitors:
 - Login / Cookies / Combination: IP address, agent, path followed
- Identification of session (division of clickstream)
 - We do not know when a visitor leaves → use a timeout (usually 30 minutes)
- Identification of user actions
 - Parameters and path analysis

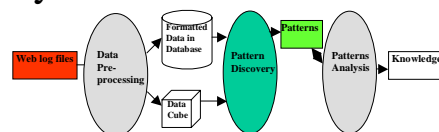
Use of Content and Structure in Data Cleaning

- Structure:
 - The structure of a web site is needed to analyze session and transactions.
 - Hypertree of links between pages.
- Content
 - Content of web pages visited can give hints for data cleaning and selection.
 - Ex: grouping web transactions by terminal page content.
 - Content of web pages gives a clue on type of page: navigation or content.

Data Mining: Pattern Discovery

Kinds of mining activities (drawn upon typical methods)

- Clustering
- Classification
- Association mining
- Sequential pattern analysis
- Prediction



Clustering

• Clustering

Grouping together objects that have “similar” characteristics.

- Clustering of transactions
 - Grouping same behaviours regardless of visitor or content
- Clustering of pages and paths
 - Grouping same pages visited based on content and visits
- Clustering of visitors
 - Grouping of visitors with same behaviour

Classification

- Classification of visitors
- Categorizing or profiling visitors by selecting features that best describe the properties of their behaviour.
- 25% of visitors who buy fiction books come from Ontario, are aged between 18 and 35, and visit after 5:00pm.
- The behaviour (ie. class) of a visitor may change in time.

Association Mining

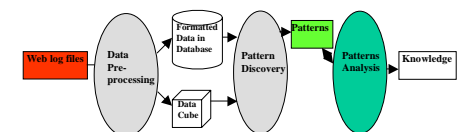
- Association of frequently visited pages
- Pages visited in the same session constitute a transaction. Relating pages that are often referenced together regardless of the order in which they are accessed (may not be hyperlinked).
- Inter-session and intra-session associations.

Sequential Pattern Analysis

- Sequential Patterns are inter-session ordered sequences of page visits. Pages in a session are time-ordered sets of episodes by the same visitor.
- ($\langle A,B,C \rangle$, $\langle A,D,C,E,F \rangle$, B, $\langle A,B,C,E,F \rangle$)
- $\langle A,B,C \rangle$ $\langle E,F \rangle$ $\langle A,*,F \rangle$,...

Pattern Analysis

- Set of rules discovered can be very large
- Pattern analysis reduces the set of rules by filtering out uninteresting rules or directly pinpointing interesting rules.
 - SQL like analysis
 - OLAP from datacube
 - Visualization

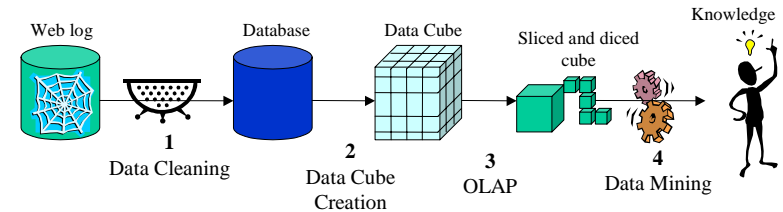


Web Usage Mining Systems

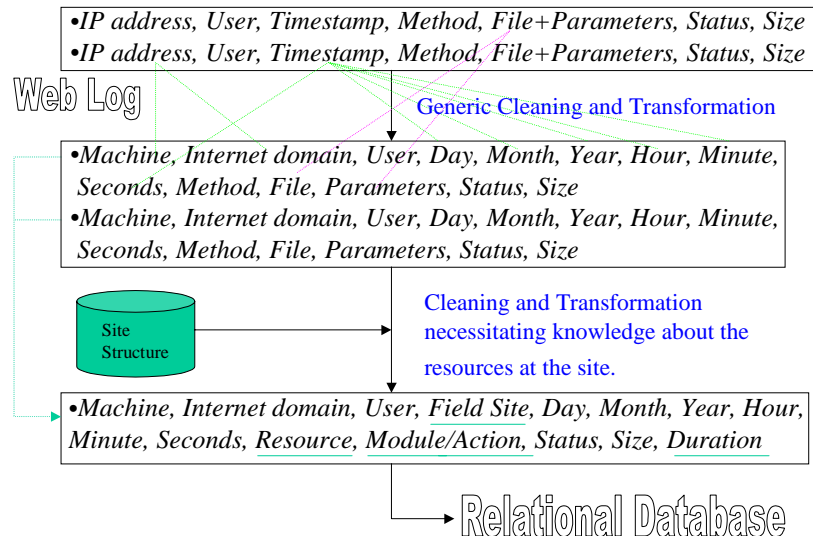
- General web usage mining:
 - WebLogMiner (Zaiane et al. 1998)
 - WUM (Spiliopoulou et al. 1998)
 - WebSIFT (Cooley et al. 1999)
- Adaptive Sites (Perkowitz et al. 1998).
- Personalization and recommendation
 - WebWatcher (Joachims et al. 1997)
 - Clustering of users (Mobasher et al. 1999)
- Traffic and caching improvement
 - (Cohen et al. 1998)

Design of 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

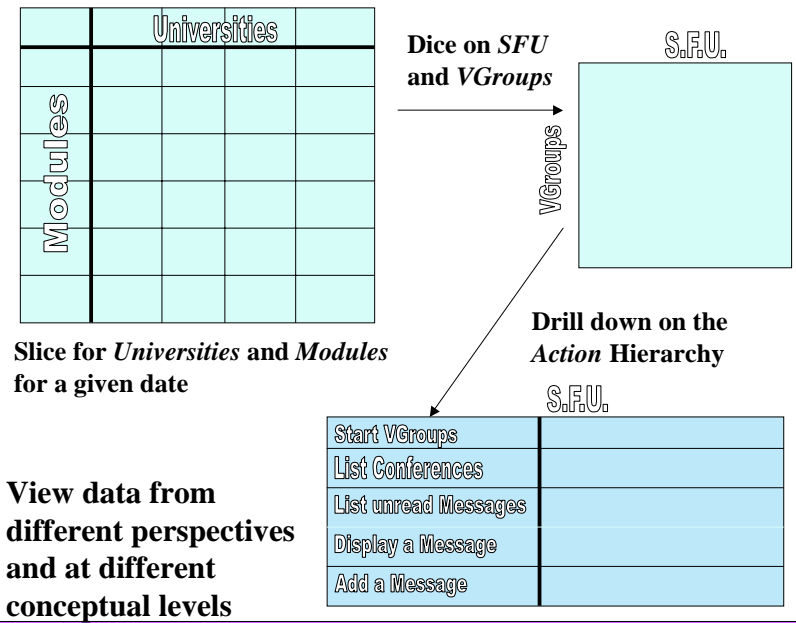
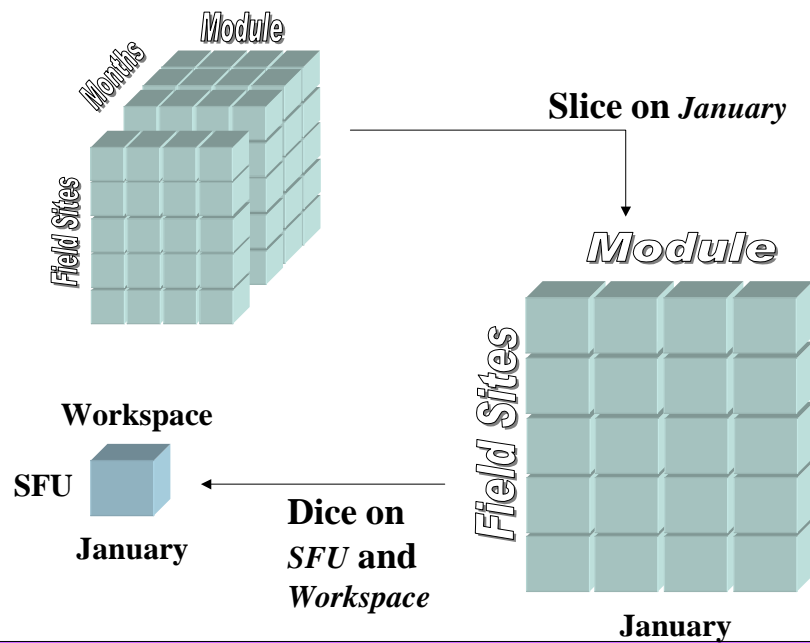


Data Cleaning and Transformation



Typical Summaries

- *Request summary*: request statistics for all modules/pages/files
- *Domain summary*: request statistics from different domains
- *Event summary*: statistics of the occurring of all events/actions
- *Session summary*: statistics of sessions
- *Bandwidth summary*: statistics of generated network traffic
- *Error summary*: statistics of all error messages
- *Referring Organization summary*: statistics of where the users were from
- *Agent summary*: statistics of the use of different browsers, etc.



Slice for Universities and Modules for a given date

View data from different perspectives and at different conceptual levels

From OLAP to Mining

- OLAP can answer questions such as:
 - Which components or features are the most/least used?
 - What is the distribution of network traffic over time (hour of the day, day of the week, month of the year, etc.)?
 - What is the user distribution over different domain areas?
 - Are there and what are the differences in access for users from different geographic areas?
- Some questions need further analysis: mining.
 - In what context are the components or features used?
 - What are the typical event sequences?
 - Are there any general behavior patterns across all users, and what are they?
 - What are the differences in usage and behavior for different user population?
 - Whether user behaviors change over time, and how?

Web Log Data Mining

- Data Characterization
- Class Comparison
- Association
- Prediction
- Classification
- Time-Series Analysis
- Web Traffic Analysis
 - Typical Event Sequence and User Behavior Pattern Analysis
 - Transition Analysis
 - Trend Analysis

Discussion

- Analyzing the web access logs can help understand user behavior and web structure, thereby improving the design of web collections and web applications, targeting e-commerce potential customers, etc.
- Web log entries do not collect enough information.
- Data cleaning and transformation is crucial and often requires site structure knowledge (Metadata).
- OLAP provides data views from different perspectives and at different conceptual levels.
- Web Log Data Mining provides in depth reports like time series analysis, associations, classification, etc.