Principles of Knowledge Discovery in Databases

Fall 1999

Chapter 9: Web Mining

Dr. Osmar R. Zaïane

University of Alberta

Course Content

- Introduction to Data Mining
- Data warehousing and OLAP
- Data cleaning
- Data mining operations
- Data summarization
- Association analysis
- Classification and prediction
- Clustering
- **Web Mining**
- Similarity Search
- *Other topics if time permits*

Chapter 9 Objectives

Understand the different knowledge discovery issues in data mining from the World Wide Web.

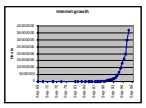
Distinguish between resource discovery and Knowledge discovery from the Internet.

Web Mining Outline

- What are the incentives of web mining?
- What is the taxonomy of web mining?
- What is web content mining?
- What is web structure mining?
- What is web usage mining?
- What is a Virtual Web View?
- Is there a query and discovery language for VWV?

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
- 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.

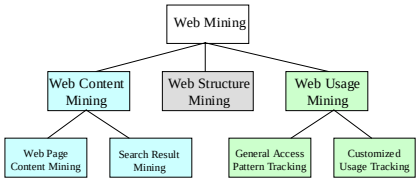
WWW and Web Mining

- 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

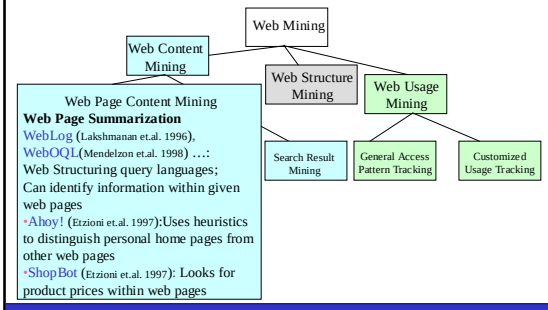
Web Mining Outline

- What are the incentives of web mining?
- What is the taxonomy of web mining?
- What is web content mining?
- What is web structure mining?
- What is web usage mining?
- What is a Virtual Web View?
- Is there a query and discovery language for VWV?

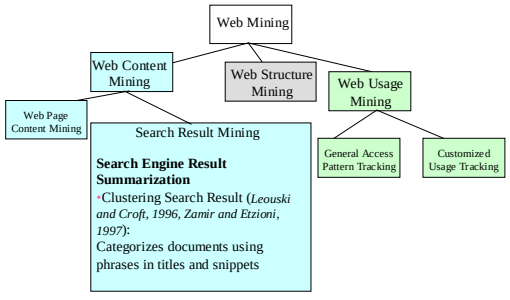
Web Mining Taxonomy



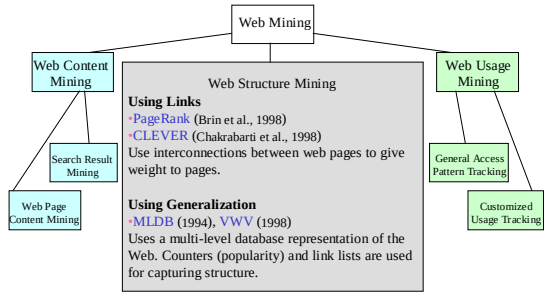
Web Mining Taxonomy



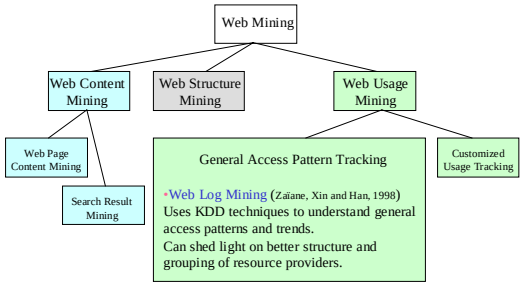
Web Mining Taxonomy



Web Mining Taxonomy

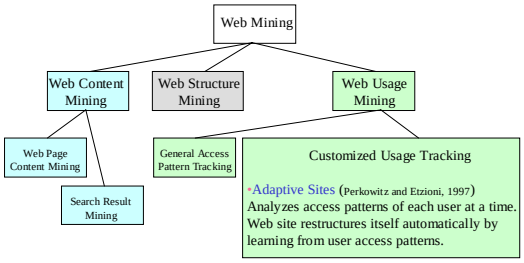


Web Mining Taxonomy



- 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.

Web Mining Taxonomy



- 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 Mining Outline

- What are the incentives of web mining?
- What is the taxonomy of web mining?
- What is web content mining?
- What is web structure mining?
- What is web usage mining?
- What is a Virtual Web View?
- Is there a query and discovery language for VWV?

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

Warehousing a Meta-Web: An MLDB Approach

- *Meta-Web:* A structure which summarizes the contents, structure, linkage, and access of the Web and which evolves with the Web
- $Layer_0$: the Web itself
- $Layer_1$: the lowest layer of the Meta-Web
  - an entry: a Web page summary, including class, time, URL, contents, keywords, popularity, weight, links, etc.
- $Layer_2$ and up: summary/classification/clustering in various ways and distributed for various applications
- Meta-Web can be warehoused and incrementally updated
- Querying and mining can be performed on or assisted by meta-Web (a multi-layer digital library catalogue, yellow page).

Construction of Multi-Layer Meta-Web

- XML: facilitates structured and meta-information extraction
- Hidden Web: DB schema "extraction" + other meta info
- Automatic classification of Web documents:
  - based on Yahoo!, etc. as training set + keyword-based correlation/classification analysis (IR/AI assistance)
- Automatic ranking of important Web pages
  - authoritative site recognition and clustering Web pages
- Generalization-based multi-layer meta-Web construction
  - With the assistance of clustering and classification analysis

Use of Multi-Layer Meta Web

- Benefits of Multi-Layer Meta-Web:
  - Multi-dimensional Web info summary analysis
  - Approximate and intelligent query answering
  - Web high-level query answering (WebSQL, WebML)
  - Web content and structure mining
  - Observing the dynamics/evolution of the Web
- Is it realistic to construct such a meta-Web?
  - Benefits even if it is partially constructed
  - Benefits may justify the cost of tool development, standardization and partial restructuring

Web Mining Outline

- What are the incentives of web mining?
- What is the taxonomy of web mining?
- What is web content mining?
- What is web structure mining?
- What is web usage mining?
- What is a Virtual Web View?
- Is there a query and discovery language for VWV?

Web Structure Mining

- Discovery of influential and authoritative pages in WWW
- Meta-web view can also be viewed as Web structure mining

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.
- Pinski and Narin (1976) proposed a significant variation on this notion, 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*.

Discovery of Authoritative Pages in WWW

- Page-rank method (Brin and Page, 1998):
  - Rank the "importance" of Web pages, based on a model of a "random browser."
- Hub/authority method (Kleinberg, 1998):
  - Prominent authorities often do not endorse one another directly on the Web.
  - Hub pages have a large number of links to many relevant authorities.
  - Thus hubs and authorities exhibit a mutually reinforcing relationship:
- Both the page-rank and hub/authority methodologies have been shown to provide qualitatively good search results for broad query topics on the WWW.

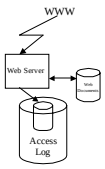
Further Enhancement for Finding Authoritative Pages in WWW

- The CLEVER system (Chakrabarti, et al. 1998)
  - 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) inWeb pages, often referred to as *anchor* text
  - boost the weights of links which occur near instances of query terms.

Web Mining Outline

- What are the incentives of web mining?
- What is the taxonomy of web mining?
- What is web content mining?
- What is web structure mining?
- What is web usage mining?
- What is a Virtual Web View?
- Is there a query and discovery language for VWV?

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

Diversity of Weblog Mining

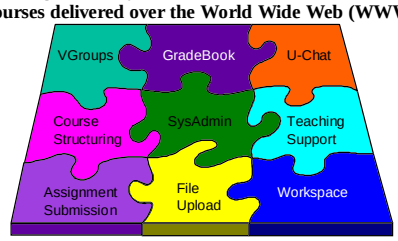
- Weblog provides rich information about Web dynamics
- Multidimensional Weblog analysis:
  - disclose potential customers, users, markets, etc.
- Plan mining (mining general Web accessing regularities):
  - Web linkage adjustment, performance improvements
- Web accessing association/sequential pattern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

Existing Web Log Analysis Tools

- There are more than 30 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.

Virtual-U and Weblog Minir

Virtual-U is a server-based software system that enables customized design, delivery, and enhancement of education and training courses delivered over the World Wide Web (WWW).



Virtual-U Log File Entries

- 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
- Information contained in the log file entries:
  - *dd23-125.compuserve.com* - domain name/IP address of the request
  - *rhuia* - user ID
  - *[01/Apr/1997:00:03:25 -0800]* - timestamp
  - *GET* - method of the request
  - */SFU/* - path root = field site
  - */cgi-bin/VG/VG_dspmsg.cgi?ci=40154&mi=49* - script requested with parameters
  - *200* - server status code
  - *417* - size of the data sent back
- Another log file contains the browser type and the referring page.

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
- Special problems with Virtual-U log files:
  - 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

Use of Log Files

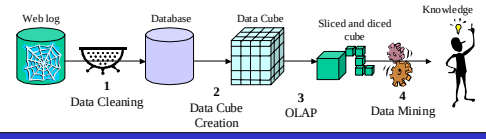
- Basic summarization:
  - Get frequency of individual actions by user, domain and session.
  - Group actions into activities, e.g. reading messages in a conference
  - Get frequency of different errors.
- Questions answerable by such summary:
  - Which components or features are the most/least used?
  - 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

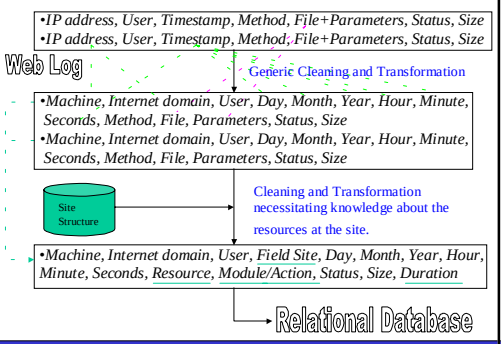
- In-depth analyses:
  - pattern analysis, e.g. between users, over different courses, instructional designs and materials, as Virtual-U features are added or modified
  - trend analysis, e.g. user behaviour change over time, network traffic change over time
- Questions can be answered by in-depth analyses:
  - 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?

Design of a Web Log Miner

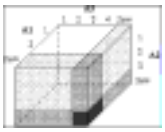
- Web log is filtered to generate a relational database
- A data cube is generated form database
- OLAP is used to drill-down and roll-up in the cube
- OLAM is used for mining interesting knowledge



Data Cleaning and Transformation

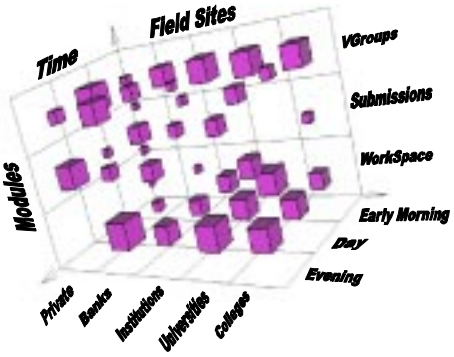


Data Cube Building



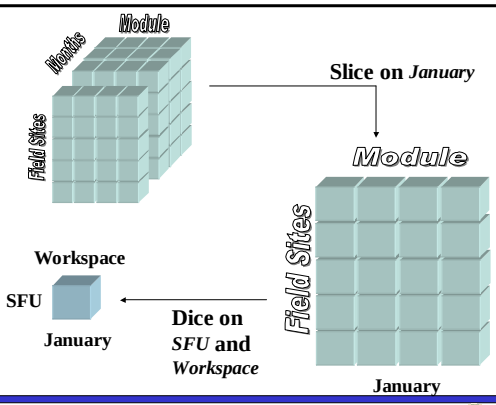
Web Log Data Cube

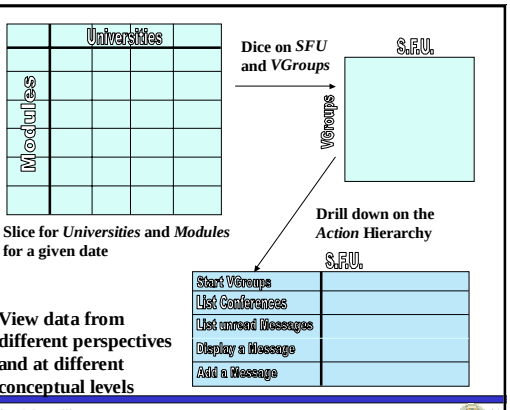
- URL of the Resource
- Action
- Type of the Resource
- Size of the Resource
- Time of the Request
- Time Spent with Resource
- Internet Domain of the Requestor
- Requestor Agent
- User
- Server Status



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.

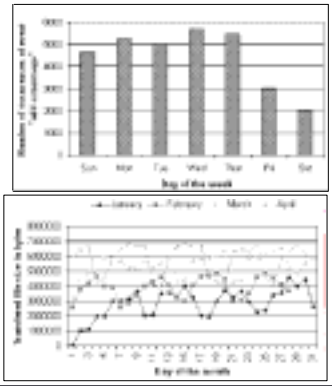




Slice for *Universities* and *Modules* for a given date

View data from different perspectives and at different conceptual levels

OLAP Analysis of Web Log Database



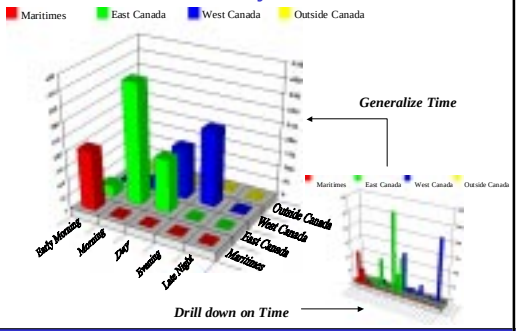
From OLAP to Mining

- 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?

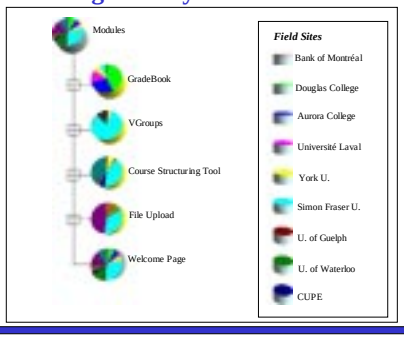
Web Log Data Mining

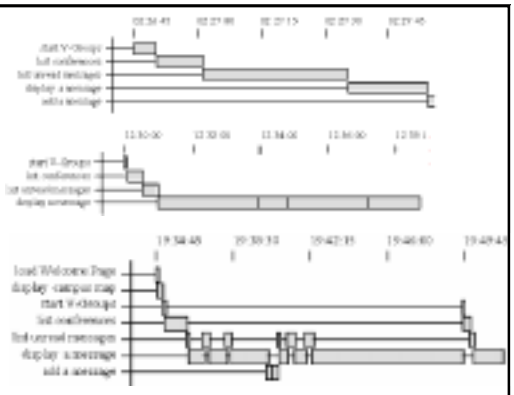
- Data Characterization
- Class Comparison
- Association
- Prediction
- Classification
- Time-Series Analysis
- Web Traffic Analysis
  - Typical Event Sequence and User Behavior Pattern Analysis
  - Transition Analysis
  - Trend Analysis

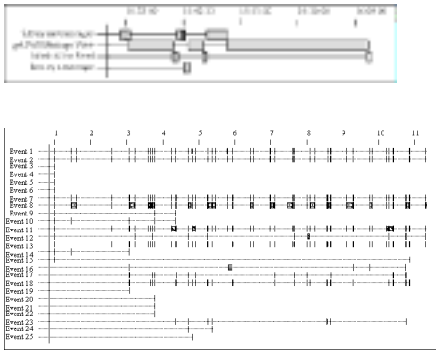
Number of actions registered in Virtual-U server on a day

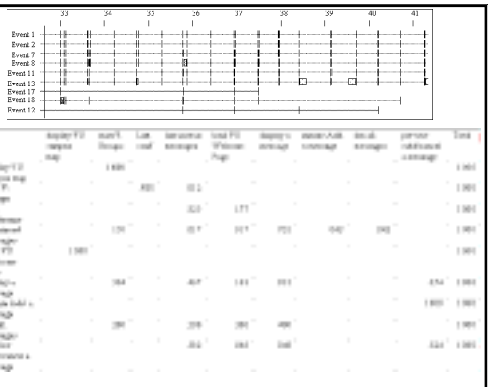


Classification of Modules/Actions by Field Site on a given day









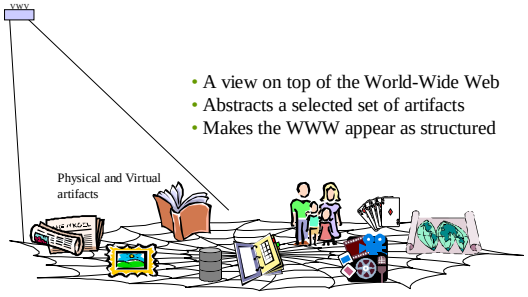
Discussion

- 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.

Web Mining Outline

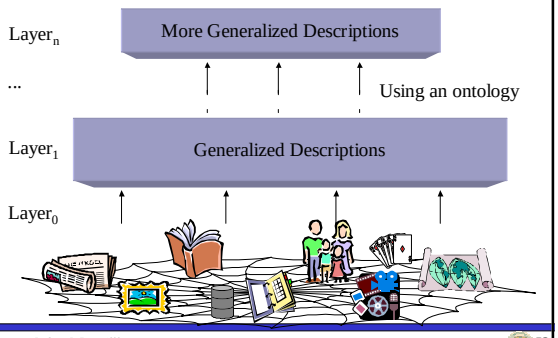
- What are the incentives of web mining?
- What is the taxonomy of web mining?
- What is web content mining?
- What is web structure mining?
- What is web usage mining?
- What is a Virtual Web View?
- Is there a query and discovery language for VWV?

Virtual Web View

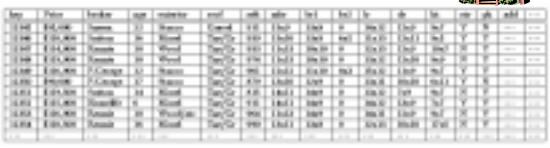


- A view on top of the World-Wide Web
- Abstracts a selected set of artifacts
- Makes the WWW appear as structured

Multiple Layered Database Architecture



Observation



Area	Class	Type	Price	Size	Age	Count
Richmond	Aprt	1 bdr	$75,000-$85,000	500-700	10-12	23
Richmond	Aprt	1 bdr	$85,000-$95,000	701-899	5-10	18
Richmond	Aprt	2 bdr	$95,000-$110,000	900-955	10-12	12
...	...	...	...	...	...	...

Transformed and generalized database

•User may be satisfied with the abstract data associated with statistics
•Higher layers are smaller. Retrieval is faster
•Higher layers may assist the user to browse the database content progressively

Multiple Layered Database Strength

- Distinguishes and separates meta-data from data
- Semantically indexes objects served on the Internet
- Discovers resources without overloading servers and flooding the network
- Facilitates progressive information browsing
- Discovers implicit knowledge (data mining)

Multiple Layered Database First Layers

Layer-0: Primitive data

Layer-1: dozen database relations representing types of objects (metadata)

document, organization, person, software, game, map, image,...

• **document**(file_addr, authors, title, publication, publication_date, abstract, language, table_of_contents, category_description, keywords, index, multimedia_attached, num_pages, format, first_paragraphs, size_doc, timestamp, access_frequency, links_in, links_out,...)

• **person**(last_name, first_name, home_page_addr, position, picture_attached, phone, e-mail, office_address, education, research_interests, publications, size_of_home_page, timestamp, access_frequency, ...)

• **image**(image_addr, author, title, publication_date, category_description, keywords, size, width, height, duration, format, parent_pages, colour_histogram, Colour_layout, Texture_layout, Movement_vector, localisation_vector, timestamp, access_frequency, ...)

Examples

URL	title	set of authors	pub_data	format	language	size	set of keywords	set of media	set of links-out	set of links-in	access-freq	timestamp

Documents

URL	format	size	height	width	Start_frame	duration	set of keywords	set of parent pages	visual feature vectors	access-freq	timestamp

Images and Videos

Multiple Layered Database Higher Layers

Layer-2: simplification of layer-1

•**doc_brief**(file_addr, authors, title, publication, publication_date, abstract, language, category_description, key_words, major_index, num_pages, format, size_doc, access_frequency, links_in, links_out)

•**person_brief** (last_name, first_name, publications,affiliation, e-mail, research_interests, size_home_page, access_frequency)

Layer-3: generalization of layer-2

•**cs_doc**(file_addr, authors, title, publication, publication_date, abstract, language, category_description, keywords, num_pages, form, size_doc, links_in, links_out)

•**doc_summary**(affiliation, field, publication_year, count, first_author_list, file_addr_list)

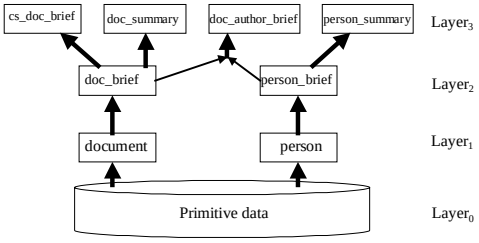
•**doc_author_brief**(file_addr, authors, affiliation, title, publication, pub_date, category_description, keywords, num_pages, format, size_doc, links_in, links_out)

•**person_summary**(affiliation, research_interest, year, num_publications, count)

Multiple Layered Database doc_summary example

affiliation	field	pub_year	count	first_author_list	file_addr_list	…
Simon Fraser Univ.	Database Systems	1994	15	Han, Kameda, Luk, ...	…	…
Univ. of Colorado	Global Network Systems	1993	10	Danzig, Hall, ...	…	…
MIT	Electromagnetic Field	1993	53	Bernstein, Phillips, ...	…	…
…	…	…	…	…	…	…

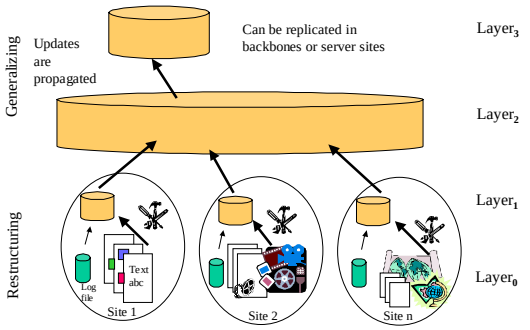
Construction of the Stratum



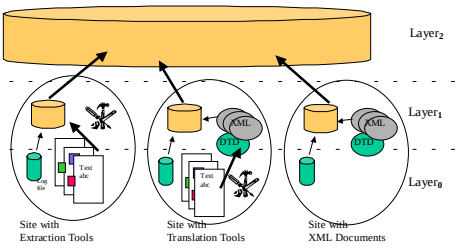
•The multi-layer structure should be constructed based on the study of frequent accessing patterns

•It is possible to construct high layered databases for special interested users ex: *computer science documents, ACM papers, etc.*

Construction and Maintenance of Layer-1



Options for the Layer-1 Construction



The Need for Metadata

Can XML help to extract the right needed descriptors?

<NAME> eXtensible Markup Language</NAME>
<RECOM>World-Wide Web Consortium</RECOM>
<SINCE>1998</SINCE>
<VERSION>1.0</VERSION>
<DESC>Meta language that facilitates more meaningful and precise declarations of document content</DESC>
<HOW>Definition of new tags and DTDs</HOW>

Dublin Core Element Set

TITLE
CREATOR
SUBJECT
DESCRIPTION
PUBLISHER
CONTRIBUTOR
DATE
TYPE
FORMAT
IDENTIFIER
SOURCE
LANGUAGE
RELATION
COVERAGE
RIGHTS

XML can help solve heterogeneity for vertical applications, but the freedom to define tags can make horizontal applications on the Web more heterogeneous.

Concept Hierarchy

All	**contains**:	Science, Art, …
Science	**contains**:	Computing Science, Physics,Mathematics,…
Computing Science	**contains**:	Theory, Database Systems, Programming Languages,…
Computing Science	**alias**:	*Information Science, Computer Science, Computer Technologies,* …
Theory	**contains**:	Parallel Computing, Complexity, Computational Geometry, …
Parallel Computing	**contains**:	Processors Organization, Interconnection Networks, RAM, …
Processor Organization	**contains**:	Hypercube, Pyramid, Grid, Spanner, X-tree,…
Interconnection Networks	**contains**:	Gossiping, Broadcasting, …
Interconnection Networks	**alias**:	*Intercommunication Networks,* …
Gossiping	**alias**:	*Gossip Problem, Telephone Problem, Rumour,* …
Database Systems	**contains**:	Data Mining, Transaction Management, Query Processing, …
Database Systems	**alias**:	*Database Technologies, Data Management,* …
Data Mining	**alias**:	*Knowledge Discovery, Data Dredging, Data Archaeology,* …
Transaction Management	**contains**:	Concurrency Control, Recovery, …
Computational Geometry	**contains**:	Geometry Searching, Convex Hull, Geometry of Rectangles, Visibility, …

Web Mining Outline

- What are the incentives of web mining?
- What is the taxonomy of web mining?
- What is web content mining?
- What is web structure mining?
- What is web usage mining?
- What is a Virtual Web View?
- Is there a query and discovery language for VWV?

WebML

Since concepts in a MLDB are generalized at different layers, search conditions may not exactly match the concept level of the inquired layers. Can be too general or too specific.

Introduction of new operators

Primitives for additional relational operations

WebML primitive	Operation	Name of the operation
covers	⊃	Coverage
covered-by	⊂	Subsumption
like	≈	Synonymy
close-to	~	Approximation

User-defined primitives can also be added

Top Level Syntax

<*WebML*> ::= <*Mine Header*> **from** relation_list
[**related-to** name_list] [**in** location_list]
where where_clause
[**order by** attributes_name_list]
[**rank by** {inward | outward | access}]

<*Mine Header*> ::= {{**select** | **list**} {attribute_name_list | *}
| <*Describe Header*> | <*Classify Header*>}

<*Describe Header*> ::= **mine description**
in-relevance-to {attribute_name_list | *}

<*Classify Header*> ::= **mine classification**
according-to attribute_name_list
in-relevance-to {attribute_name_list | *}

WebML Example: Resource Discovery

Locate the documents related to "computer science" written by "Ted Thomas" and about "data mining".

```
select    *
from      document
related-to "computer science"
where     "Ted Thomas" in authors and one of keywords like "data mining"
```

Discovering Resources

Returns a list of URL addresses together with important attributes of the documents.

WebML Example: Resource Discovery

Locate the documents about "data mining" linked from Osmar's web page and rank them by importance.

```
select    *
from      document
where     exact "http://www.cs.sfu.ca/~zaiane" in links_in
          and one of keywords like "data mining"
rank by inward, access
```

Discovering Resources

Returns a list of URL addresses together with important attributes of the documents.

WebML Example: Resource Discovery

Locate the documents about "Intelligent Agents" published at SFU and that link to Osmar's web pages.

```
select    *
from      document
in "http://www.sfu.ca"
related-to "computer science"
where     "http://www.cs.sfu.ca/~zaiane" in links_out
          and one of keywords like "Agents"
```

No "**exact**" ⇒ prefix substring

Discovering Resources

Returns a list of URL addresses together with important attributes of the documents.

WebML Example: Resource Discovery

List the documents published in North America and related to "data mining".

```
list      *
from      document
in        "North_America"
related-to "computer science"
where     one of keywords covered_by "data mining"
```

Discovering Resources

Returns a list of documents at a high conceptual level and allows browsing of the list with slicing and drilling through to the appropriate physical documents.

WebML Example: Knowledge Discovery

Inquire about European universities *productive* in publishing on-line *popular* documents related to database systems since 1990.

```
select    affiliation
from      document
in        "Europe"
where     affiliation belong_to "university" and
          one of keywords covered-by "database systems"
          and publication_year > 1990 and count = "high"
          and f(links_in) = "high"
```

Weight (heuristic formula)

Discovering Knowledge

Does not return a list of document references, but rather a list of universities.

WebML Example: Knowledge Discovery

Describe the general characteristics in relevance to authors' affiliations, publications, etc. for those documents which are popular on the Internet (in terms of access) and are about "data mining".

```
mine description
in-relevance-to author.affiliation, publication, pub_date
from document related-to Computing Science
where one of keywords like "database systems"
      and access_frequency = "high"
```

Discovering Knowledge

Retrieves information according to the 'where clause', then generalizes and collects it in a data cube for interactive OLAP-like operations.

WebML Example: Knowledge Discovery

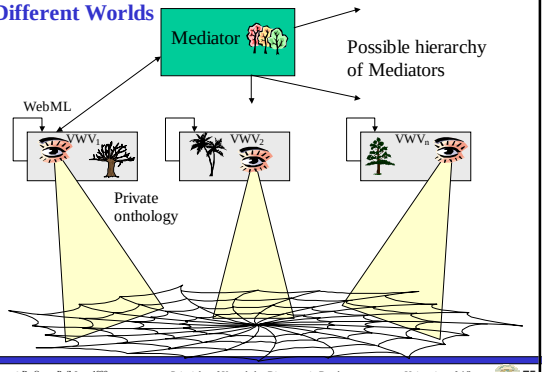
Classify, according to update time and access popularity, the documents published on-line in sites in the Canadian and commercial Internet domain after 1993 and about IR from the Internet.

```
mine classification
according-to timestamp, access_frequency
in-relevance-to *
from document in Canada, Commercial
where one of keywords covered-by "Information Retrieval"
      and one of keywords like "Internet"
      and publication_year > 1993
```

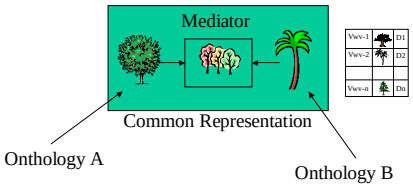
Discovering Knowledge

Generates a classification tree where documents are classified by access frequency and modification date.

Different Worlds



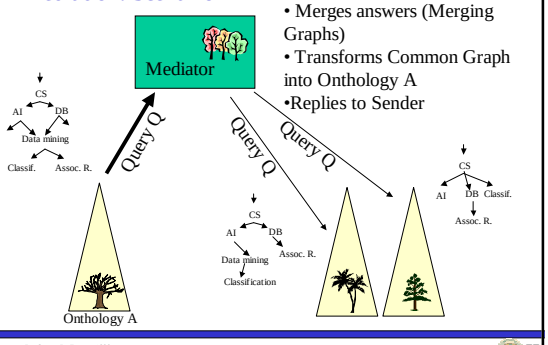
Standard Onthology Representation



Mapping between concept hierarchies (one-to-one or one-to-many)
Reduction of semantic ambiguities

Mediation: Scenario 1

- Broadcasts Query Q
- Merges answers (Merging Graphs)
- Transforms Common Graph into Onthology A
- Replies to Sender



Mediation: Scenario 2

- Re-expresses Query Q into Q' (or set of queries)
- Submits queries
- Merges result with other answers using onthology of Q
- Replies to Sender