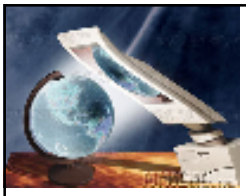
Web Specifics

- Web: A huge, widely-distributed highly heterogeneous, semi-structured hypertext / hypermedia, interconnected information repository.
- Web is a huge collection of documents plus
  - Hyper-link information
  - Access and usage information

1



Web Mining

It is the use of data mining techniques to automatically discover and extract information from Web documents/services

Etzioni CACM

2

A Few Themes in Web Mining

Some interesting problems on Web mining

- Mining what Web search engine finds
- Identification of authoritative Web pages
- Identification of Web communities
- Web document classification
- Warehousing a Meta-Web: Web yellow page service
- Weblog mining (usage, access, and evolution)
- Intelligent query answering in Web search

3

WEB DATA

- Web pages
  - Content of actual web pages.
- Intra-page structures
  - HTML or XML code for web pages. - Anchor
- Inter-page structures
  - Linkage structure between web pages

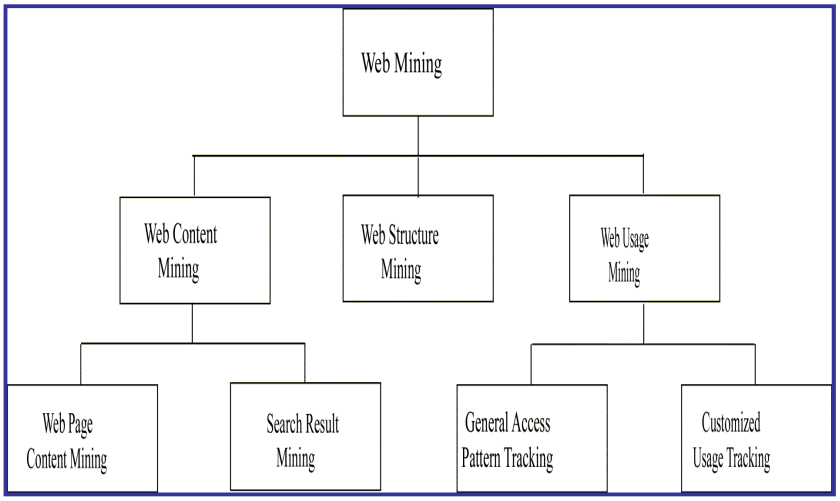
4

WEB DATA

- Usage data
  - How web pages are accessed by end user.
- User Profile
  - demographic and registration information obtained about user.
  - Information found in cookies

5

WEB MINING TAXONOMY



6

Over view of Web Mining

The taxonomy of web mining

- Web page content mining
  Summarization of web page contents
  (Web ML queries)
- Web structure mining
  Summarization of search engine result
  (Page Rank)
- Web usage mining
  mining for user browsing and access pattern, for understanding user's behaviors on the Web.
  (Web log Mining)

7

Web Content Mining

"Web Content Mining is the process of extracting useful information from the contents of Web documents.

It may consist of text, images, audio, video, or structured records such as lists and tables."

8

Web Content Mining

"Web Content Mining examines the content of web pages as well as result of web search"

- The first traditional searching of web pages via content , while second is a further search of web pages found from previous search.

9

Web Content Data Structure

- Web content consists of several types of data : Text, image, audio, video, hyperlinks.
- Unstructured – free text
- Semi-structured – HTML
- More structured – Data in the tables or database generated HTML pages
- Note: much of the Web content data is unstructured text data.

10

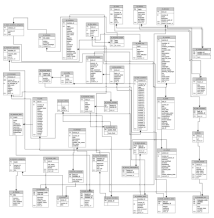
Web Content Mining : IR View

- Unstructured Documents
- Bag of words to represent unstructured documents
  - Takes single word as feature
  - Ignores the sequence in which words occur
- Features could be
  - Boolean
  - Word either occurs or does not occur in a document
  - Frequency based
  - Frequency of the word in a document

11

Web Content Mining : DB View

- The database techniques on the Web are related to the problems of managing and querying the information on the Web.
- DB view tries to infer the structure of a Web site or transform a Web site to become a database
  - Better information management
  - Better querying on the Web
- Can be achieved by:
  - Finding the schema of Web documents
  - Building a Web warehouse
  - Building a virtual database



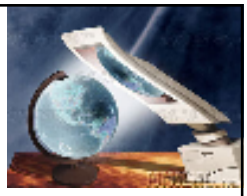
12

Intelligent Search Agents

- Locating documents and services on the Web:
  - **WebCrawler, Alta Vista (http://www.altavista.com):** scan millions of Web documents and create index of words
  - **Web robots**：are software programs that automatically traverse the hyperlink structure of the World Wide Web in order to locate and retrieve information.
- **ShopBot** -Retrieve product information from a variety of vendor sites using only general information about the product domain.

13

Web Outlier Mining



- **Web Outliers are data objects that show significantly different characteristics than other web data.**
- **Web Outlier mining is the discovery and analysis of rare and interesting pattern from the web.**

14

Web Content Outlier Mining

The web consists of interrelated web pages grouped into different categories depending on their contents. A web content outlier is described as page (s) with completely different contents from similar pages within the same category.

Definition: *Given a set of web documents di (i=1, 2,..., n) each with relative weight wi from category C, the document dj constitutes a web content outlier if wj < W_{min}.*

W_{min} is a threshold assigned by the miner based on previous experience with similar data.

15

Web Content Outlier Mining

- N-Gram technique which analyze the contents of the Meta data of the web pages of related category and identify the web pages which are having significantly different contents as compared to other pages.
- takes help of a Domain Dictionary which having the important words belonging to the category of interest.

16

Basic Architecture

17

Web Content Outlier Mining

Extraction of Resources

- Retrieve the desired web pages belonging to the category of interest.
- Use of web search engines or web crawlers.
- Extract web page information from web which contains Meta data field

Win Web Crawler 2.0

18

Extraction of Resources

The Win Web Crawler is a powerful web crawler utility to Extract

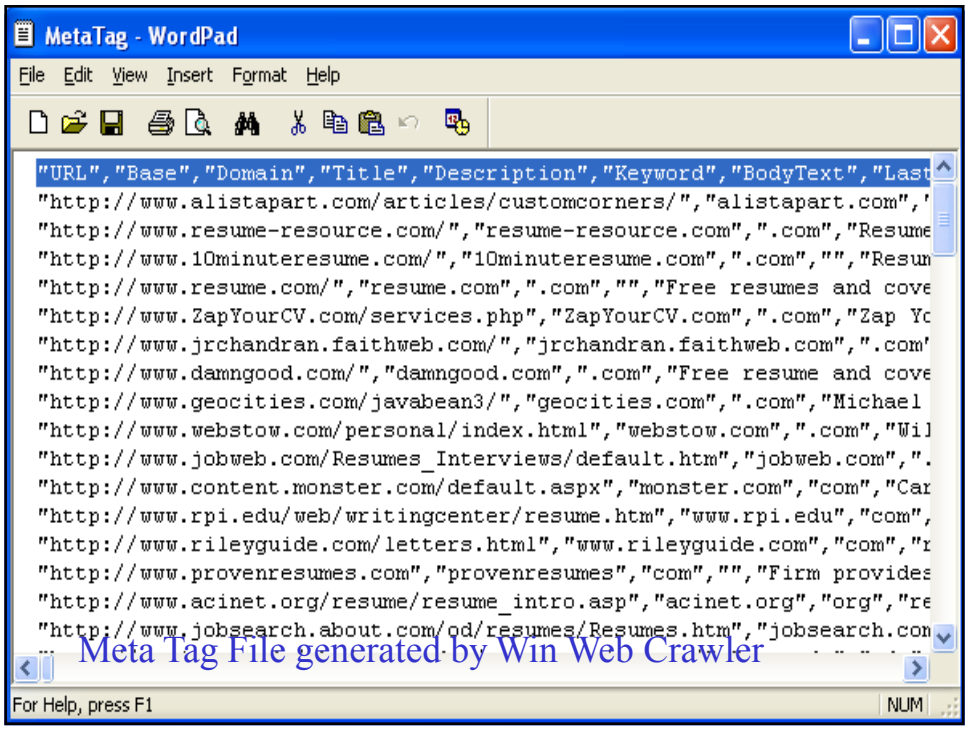
URL

- **meta tag (title, description, keyword),**
- **plain text between <body> to </body> tag**
- **page size**
- **last modified date value**

from

- **Web Site**
- **Web Directories**
- **Search Results**
- **List of URLs from file**

19



Meta Tag File generated by Win Web Crawler

Web Content Outlier Mining

Preprocessing

- Transforms the extracted data into a structured form
- Remove stop words (words with frequency greater than some user specified frequency)
- Removal of duplicate words.

21

Web Content Outlier Mining

Web Content Outlier Detection

- **Inputs are**
  - **Preprocessed data**
  - **Domain dictionary.**
- **The algorithm assigns weights to words, of the web pages based on whether words or their N-Gram frequency present in the domain dictionary.**
- **The weight of words on a page are computed and compared with a user defined weight for every page.**

22

N-GRAM TECHNIQUE

- N-grams of a string of length k is an n contiguous slice of the string into substrings each of size n.
- E.g. The string 'computer' has

  'comp', 'ompu','mput', 'pute', 'uter' 4-grams and

  'compu', 'omput', 'mpute', , 'puter' 5- grams.

String length – k → (k-n+1) possible N-grams

n is the size of N-gram

23

Web Content Outlier Mining

- Domain Dictionary

  Domain dictionary containing the important words of the category of interest.

- Weight Assignment

  Assign weights to the words depending upon the presence of the word in the dictionary

- Generate N-Grams for each word which does not match with dictionary contains N-Grams of higher lengths are used because higher order n-grams are capable of capturing similarities between different but related words compared to n-grams of shorter length

24

Web Content Outlier Mining

Analysis of Outliers

- Shows the resultant Outliers Pages in the given domain.
- Provides visualization of the outliers pages.
- Provide comparison of the algorithm using N-Gram technique and without using N-Gram technique .

25

- ALGORITHM
- Input: Dictionary, documents Di
- Outputs: Outlier pages
- Other variable: Total weight of document WDi, Threshold weight WDmin
- 1. Read the contents of the documents (Di) and dictionary
- 2. For (int i =0; i< NoOfDoc i ++) { //Beginning of the first outer loop
- 3. For (int j =0; j< NoOfWords j++) { //Beginning of the first inner loop
- 4. IF (word exist in the dictionary){ //Beginning of the outer IF-ELSE
- 5. Increase Weight of the doc.
- 6. }
- 7. Else {
- 8. Generate n-grams for the word

26

- 9. For (int n =0; n< NoOfNgrams; n++) { //Beginning of the second inner loop
- 10. IF (N-gram exists in dictionary) { //Beginning of the inner IF-ELSE
- 11. Increase Weight of the doc.
- 12 .} Else
- 13. {Weight retained as it is
- 14. } End IF //Ending of inner IF-ELSE
- 15. } // Ending of second inner for loop
- 16. } End IF //Ending of outer IF-ELSE
- 17. } // End of first inner for loop
- 18. WDi =Total weight of the doc.
- 19. Pages with WDi < WDmin are outliers
- 20. } //End of outer for loop
- 21. End of Algorithm.

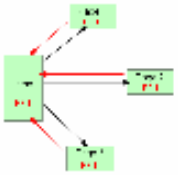
27

WEB CONTENT OUTLIER MINING APPLICATIONS

- Identification of junk materials.
- Topic identification and structured search
- Information retrieval communities.
- Determine pages with entirely different contents from their parent web sites.
- Discovery of competitors in electronic business and commerce

28

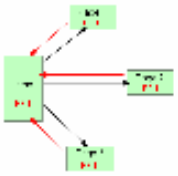
Web Structure Mining



- Interested in the structure of the hyperlinks within the Web
- Inspired by the study of social networks and citation analysis
- Can discover specific types of pages(such as hubs, authorities, etc.) based on the incoming and outgoing links.
- Application:
  - Discovering micro-communities in the Web ,
  - measuring the "completeness" of a Web site

29

Web Structure Mining



- There are two approaches:
- page rank: for discovering the most important pages on the Web (as used in Google)
- hubs and authorities: a more detailed evaluation of the importance of Web pages
- Basic definition of importance:
  - A page is important if important pages link to it

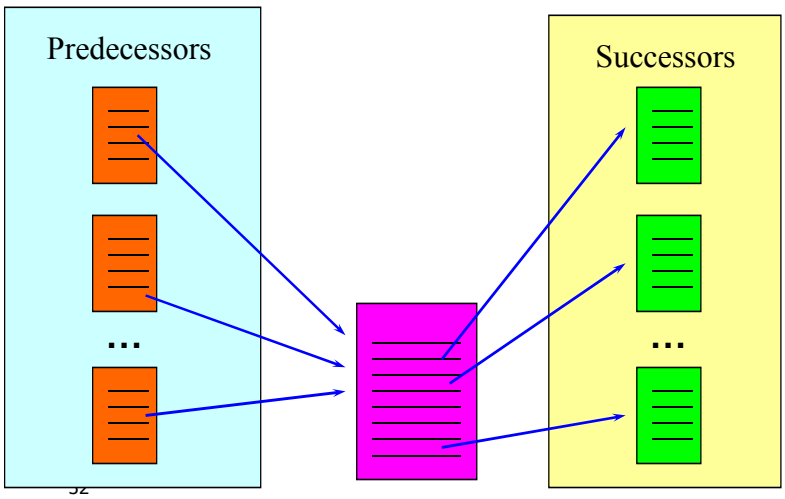
30

Web Structure Mining

- (1970) Researchers proposed methods of using citations among journal articles to evaluate the quality of reserach papers.
- Unlike journal citations, the Web linkage has some unique features:
  - one authority page will seldom have its Web page point to its competitive authorities (CocaCola --→ Pepsi)
  - authoritative pages are seldom descriptive (Yahoo! may not contain the description „Web search engine")

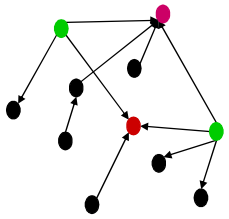
31

Predecessors and Successors of a Web Page



Link Structure Analysis

- **Improve information retrieval by scoring Web pages according to their importance in the Web or a thematic sub-domain of it.**
- **Nodes with large fan-in (authorities) provide high quality information.**
- **Nodes with large fan-out (hubs) are good starting points.**



33

Page Rank

Simple solution: create a stochastic matrix of the Web:

- Each page i corresponds to row i and column i of the matrix
- If page j has n successors (links) then the ijth cell of the matrix is equal to 1/n if page i is one of these n succesors of page j, and 0 otherwise.

34

Page Rank

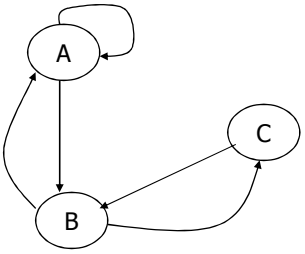
The intuition behind this matrix:

- initially each page has 1 unit of importance.
- At each round, each page shares importance it has among its successors, and receives new importance from its predecessors.
- The importance of each page reaches a limit after some steps

Page Rank – Example

- Assume that the Web consists of only three pages - A, B, and C. The links among these pages are shown below.

Let [a, b, c] be the vector of importances for these three pages



	A	B	C
A	1/2	1/2	0
B	1/2	0	1
C	0	1/2	0

Page Rank

- The equation describing the asymptotic values of these three variables is:

$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1 \\ 0 & 1/2 & 0 \end{bmatrix} \times \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

We can solve the equations like this one by starting with the assumption a = b = c = 1, and applying the matrix to the current estimate of these values repeatedly. The first four iterations give the following estimates:

a =	1	1	5/4	9/8	5/4	...	**6/5**
b =	1	3/2	1	11/8	17/16	...	**6/5**
c =	1	1/2	3/4	1/2	11/16	...	**3/5**

37

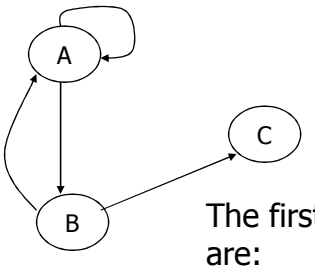
Problems with Real Web Graphs

- In the limit, the solution is a=b=6/5, c=3/5. That is, *a* and *b* each have the same importance, and twice of *c*.
- Problems with Real Web Graphs
  - dead ends: a page that has no succesors has nowhere to send its importance.
  - Pages with no outlinks are "dead ends" for the random surfer
  - Nowhere to go on next step

38

Dead End– Page Rank

- Assume now that the structure of the Web has changed. The new matrix describing transitions is:



$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} ½ & ½ & 0 \\ ½ & 0 & 0 \\ 0 & ½ & 0 \end{bmatrix} \times \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

The first four steps of the iterative solution are:

a =	1	1	3/4	5/8	1/2
b =	1	1/2	1/2	3/8	5/16
c =	1	1/2	1/4	1/4	3/16

Eventually, each of a, b, and c become 0.

39

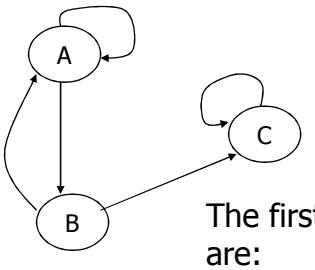
Spider traps

- A group of pages is a spider trap if there are no links from within the group to outside the group
  - Random surfer gets trapped
- spider traps: a group of one or more pages that have no links out.

40

Spider Traps

- Assume now once more that the structure of the Web has changed. The new matrix describing transitions is:



$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} ½ & ½ & 0 \\ ½ & 0 & 0 \\ 0 & ½ & 1 \end{bmatrix} \times \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

The first four steps of the iterative solution are:

a =	1	1	3/4	5/8	1/2
b =	1	1/2	1/2	3/8	5/16
c =	1	3/2	7/4	2	35/16

c converges to 3, and a=b=0.

41

Google Solution

- Instead of applying the matrix directly, „tax" each page some fraction of its current importance, and distribute the taxed importance equally among all pages.
- Example: if we use 20% tax, the equation of the previous example becomes:

  a = 0.8 * (½*a + ½ *b +0*c)

  b = 0.8 * (½*a + 0*b + 0*c)

  c = 0.8 * (0*a + ½*b + 1*c)

The solution to this equation is a=7/11, b=5/11, and c=21/11

42

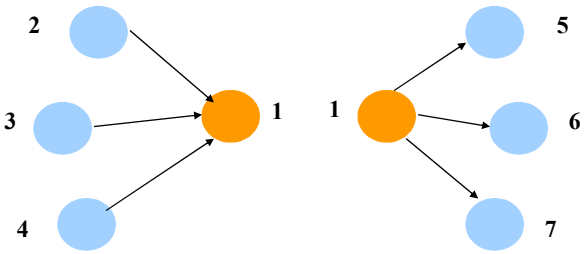
HITS (hypertext-induced topic selection)

- Kleinberg's hypertext-induced topic selection (HITS) algorithm is also developed for ranking documents based on the link information among a set of documents.

Authorities and hubs

- The algorithm produces two types of pages:
  - Authority: pages that provide an important, trustworthy information on a given topic.
  - Hub: pages that contain links to authorities.
- Authorities and hubs exhibit a mutually reinforcing relationship: a better hub points to many good authorities, and a better authority is pointed to by many good hubs.

Authorities and hubs (2)

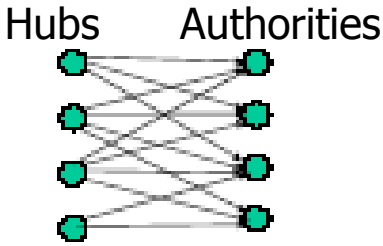


a(1) = h(2) + h(3) + h(4)

h(1) = a(5) + a(6) + a(7)

HITS Algorithm

- Hubs point to lots of authorities.
- Authorities are pointed to by lots of hubs.
- Together they form a bipartite graph:



Web Usage Mining

- A Web is a collection of inter-related files on one or more Web servers.
- Web Usage Mining.
  - Discovery of meaningful patterns from data generated by client-server transactions.
- Typical Sources of Data:
  - automatically generated data stored in server access logs, referrer logs, agent logs, and client-side cookies.
  - user profiles

47

Web Usage Mining

Tries to predict user behavior from interaction with the Web

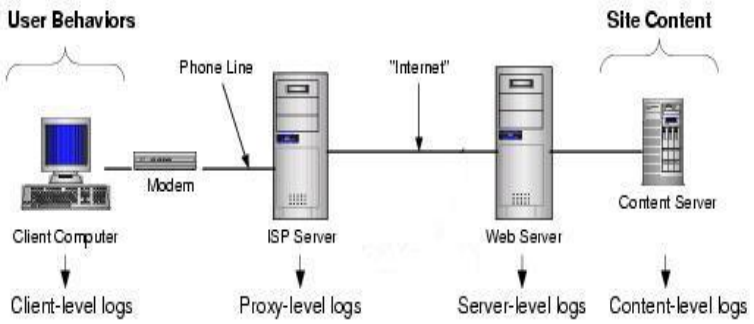
- Wide range of data (logs)
  - Web client data
  - Proxy server data
  - Web server data

- Two common approaches
  - Maps the usage data of Web server into relational tables before an adapted data mining techniques
  - Uses the log data directly by utilizing special pre-processing techniques

48

Data Sources- Logs

Every time an user link up at the web site the server keeps track in the log file of the click flow (click stream).



49

Sample Web log

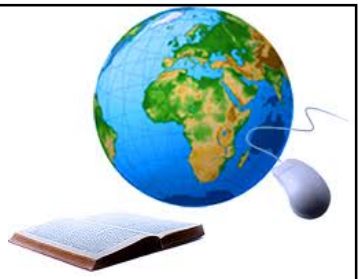
2006-10-04 04:24:39 203.200.95.194 - W3SVC195 NS 69.41.233.13 80 GET /STUDYCENTRE/StudyCentreSearchPage.asp - 200 0 0 568 210 HTTP/1.0 www.mcu.ac.in Mozilla/4.0+(compatible;+MSIE+5.0;+Windows+98;) - http://www.mcu.ac.in/search_a_study_institute.htm

date(2006-10-04) **time**(04:24:39)
Client side IP (203.200.95.194) **Server site name**(W3SVC195)
Server Computer Name (NS), **Server ip**(69.41.233.13)
Server port(80) **method** (GET)
Client url(/STUDYCENTRE/StudyCentreSearchPage.asp)
Server status code(200) **Client side byte received** (568)
Time taken(210),
User agent Mozilla/4.0+(compatible;+MSIE+5.0;+Windows+98;)

CS_referer (http://www.mcu.ac.in/search_a_study_institute.htm)

50

Web Usage Mining



- Typical problems:
  - Distinguishing among unique users, server sessions, episodes, etc. in the presence of caching and proxy servers.
  - Often Usage Mining uses some background or domain knowledge
    - E.g. site topology, Web content, etc.

51

Web Usage Mining

Applications:

- Two main categories:
  - Learning a user profile (personalized)
  - Web users would be interested in techniques that learn their needs and preferences automatically
- Learning user navigation patterns (impersonalized)
  - Information providers would be interested in techniques that improve the effectiveness of their Web site

52

Web Usage Mining

Applications:

- **Personalization**
- **Improve structure of a site's Web pages**
- **Aid in caching and prediction of future page references**
- **Improve design of individual pages**
- **Improve effectiveness of e-commerce (sales and advertising)**

53

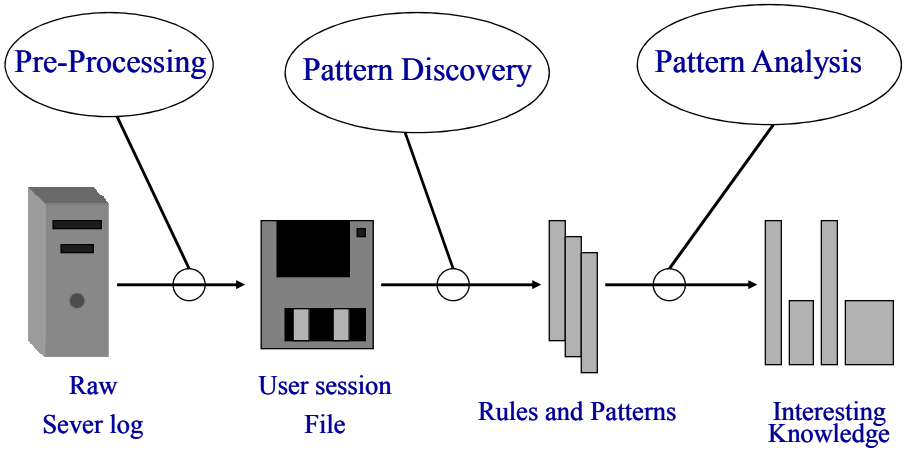
Web Usage Mining (WUM)

The discovery of interesting user access patterns from Web server logs

- Generate simple statistical reports:
  - A summary report of hits and bytes transferre
  - A list of top requested URLs
  - A list of top referrers
  - A list of most common browsers used
  - Hits per hour/day/week/month reports
  - Hits per domain reports
- Learn:
  - Who is visiting you site
  - The path visitors take through your pages
  - How much time visitors spend on each page
  - The most common starting page
  - Where visitors are leaving your site

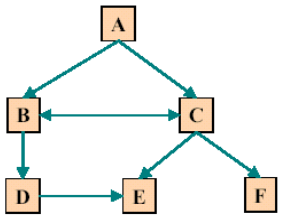
54

Web Usage Mining – Three Phases



55

Sessionization Example



Time	IP	URL	Ref	Agent
0:01	1.2.3.4	A	-	IE5;Win2k
0:09	1.2.3.4	B	A	IE5;Win2k
0:10	2.3.4.5	C	-	IE4;Win98
0:12	2.3.4.5	B	C	IE4;Win98
0:15	2.3.4.5	E	C	IE4;Win98
0:19	1.2.3.4	C	A	IE5;Win2k
0:22	2.3.4.5	D	B	IE4;Win98
0:22	1.2.3.4	A	-	IE4;Win98
0:25	1.2.3.4	E	C	IE5;Win2k
0:25	1.2.3.4	C	A	IE4;Win98
0:33	1.2.3.4	B	C	IE4;Win98
0:58	1.2.3.4	D	B	IE4;Win98
1:10	1.2.3.4	E	D	IE4;Win98
1:15	1.2.3.4	A	-	IE5;Win2k
1:16	1.2.3.4	C	A	IE5;Win2k
1:17	1.2.3.4	F	C	IE4;Win98
1:25	1.2.3.4	F	C	IE5;Win2k
1:30	1.2.3.4	B	A	IE5;Win2k
1:36	1.2.3.4	D	B	IE5;Win2k

56

Sessionization Example

1. Sort users (based on IP+Agent)

Time	IP	URL	Ref	Agent
0:01	1.2.3.4	A	-	IE5;Win2k
0:09	1.2.3.4	B	A	IE5;Win2k
0:10	2.3.4.5	C	-	IE4;Win98
0:12	2.3.4.5	B	C	IE4;Win98
0:15	2.3.4.5	E	C	IE4;Win98
0:19	1.2.3.4	C	A	IE5;Win2k
0:22	2.3.4.5	D	B	IE4;Win98
0:22	1.2.3.4	A	-	IE4;Win98
0:25	1.2.3.4	E	C	IE5;Win2k
0:25	1.2.3.4	C	A	IE4;Win98
0:33	1.2.3.4	B	C	IE4;Win98
0:58	1.2.3.4	D	B	IE4;Win98
1:10	1.2.3.4	E	D	IE4;Win98
1:15	1.2.3.4	A	-	IE5;Win2k
1:16	1.2.3.4	C	A	IE5;Win2k
1:17	1.2.3.4	F	C	IE4;Win98
1:26	1.2.3.4	F	C	IE5;Win2k
1:30	1.2.3.4	B	A	IE5;Win2k
1:36	1.2.3.4	D	B	IE5;Win2k

0:01	1.2.3.4	A	-	IE5;Win2k
0:09	1.2.3.4	B	A	IE5;Win2k
0:19	1.2.3.4	C	A	IE5;Win2k
0:25	1.2.3.4	E	C	IE5;Win2k
1:15	1.2.3.4	A	-	IE5;Win2k
1:26	1.2.3.4	F	C	IE5;Win2k
1:30	1.2.3.4	B	A	IE5;Win2k
1:36	1.2.3.4	D	B	IE5;Win2k

0:10	2.3.4.5	C	-	IE4;Win98
0:12	2.3.4.5	B	C	IE4;Win98
0:15	2.3.4.5	E	C	IE4;Win98
0:22	2.3.4.5	D	B	IE4;Win98

0:22	1.2.3.4	A	-	IE4;Win98
0:25	1.2.3.4	C	A	IE4;Win98
0:33	1.2.3.4	B	C	IE4;Win98
0:58	1.2.3.4	D	B	IE4;Win98
1:10	1.2.3.4	E	D	IE4;Win98
1:17	1.2.3.4	F	C	IE4;Win98

Sessionization Example

2. Sessionize using heuristics ($h1$ with 30 min)

0:01	1.2.3.4	A	-	IE5;Win2k
0:09	1.2.3.4	B	A	IE5;Win2k
0:19	1.2.3.4	C	A	IE5;Win2k
0:25	1.2.3.4	E	C	IE5;Win2k
1:15	1.2.3.4	A	-	IE5;Win2k
1:26	1.2.3.4	F	C	IE5;Win2k
1:30	1.2.3.4	B	A	IE5;Win2k
1:36	1.2.3.4	D	B	IE5;Win2k

0:01	1.2.3.4	A	-	IE5;Win2k
0:09	1.2.3.4	B	A	IE5;Win2k
0:19	1.2.3.4	C	A	IE5;Win2k
0:25	1.2.3.4	E	C	IE5;Win2k

1:15	1.2.3.4	A	-	IE5;Win2k
1:26	1.2.3.4	F	C	IE5;Win2k
1:30	1.2.3.4	B	A	IE5;Win2k
1:36	1.2.3.4	D	B	IE5;Win2k

The $h1$ heuristic (with timeout variable of 30 minutes) will result in the two sessions given above.

Sessionization Example

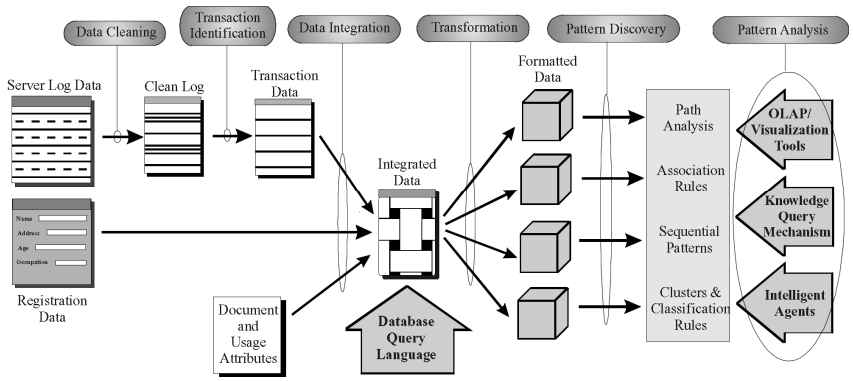
2. Sessionize using heuristics (another example with *href*)

0:22	1.2.3.4	A	-	IE4;Win98
0:25	1.2.3.4	C	A	IE4;Win98
0:33	1.2.3.4	B	C	IE4;Win98
0:58	1.2.3.4	D	B	IE4;Win98
1:10	1.2.3.4	E	D	IE4;Win98
1:17	1.2.3.4	F	C	IE4;Win98

In this case, the referrer-based heuristics will result in a single session, while the *h*1 heuristic (with timeout = 30 minutes) will result in two different sessions.

59

The Web Usage Mining Process



– General Architecture for the WEBMINER –

60

Improve web site navigation through Web Usage Mining

Web Log Mining

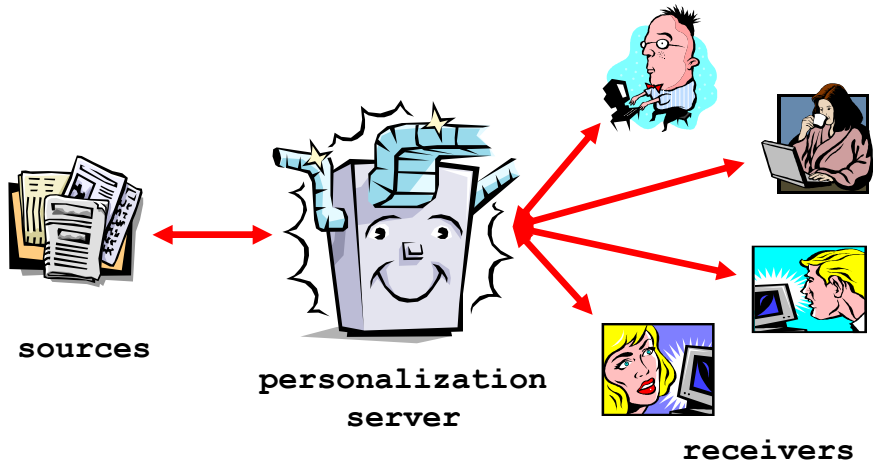
61

Summary report

- 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

62

Personalized information access



Cookies

- **Cookies can be useful**
  - **used like a staple to attach multiple parts of a form together**
  - **used to identify you when you return to a web site so you don't have to remember a password**
  - **used to help web sites understand how people use them**
- **Cookies can be harmful**
  - **used to profile users and track their activities *without their knowledge*, especially across web sites**

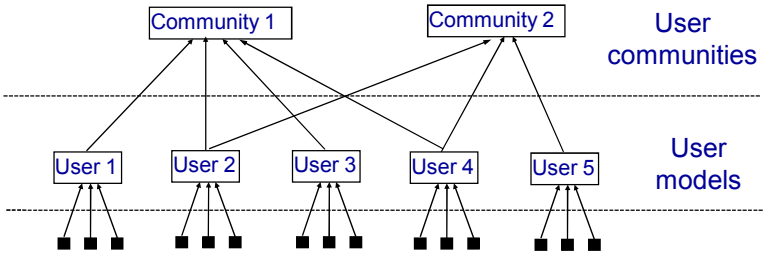
Personalization v. intelligence

- Better service for the user:
  - **Reduction of the information overload.**
  - **More accurate information retrieval and extraction.**
  - **Recommendation and guidance.**
- Customer relationship management:
  - **Customer segmentation and targeted advertisement.**
  - **Customer attraction and retention.**
  - **Service improvement (site structure and content).**

User modeling

- Basic elements:
  - Constructing models that can be used to adapt the system to the user's requirements.
  - Different types of requirement: interests (sports and finance news), knowledge level (novice - expert), preferences (no-frame GUI), etc.
  - Different types of model: personal – generic.
- Knowledge discovery facilitates the acquisition of user models from data.

Learning user models



Observation of the users interacting with the system.

Knowledge discovery process

Data collection

Data pre-processing

Pattern discovery

Knowledge post-processing

Collection of usage data by the server and the client.

Data cleaning, user identification, session identification

Construction of user models

Report generation, visualization, personalization module.

Usage data sources

- Server-side data: access logs in Common or Extended Log Format, user queries, cookies.
- Client-side data: Java and Javascript agents.
- Intermediary data: proxy logs, packet sniffers.
- Registration forms: personal information and preferences supplied by the user.
- Demographic information: provided by census databases.

Pre-processing usage data

- Cleaning:
  - Log entries that correspond to error responses.
  - Trails of robots.
  - Pages that have not been requested explicitly by the user (mainly image files, loaded automatically). Should be domain-specific.
- User identification:
  - Identification by log-in.
  - Cookies and Javascript.
  - Extended Log Format (browser and OS version).
  - Bookmark user-specific URL.
- User session/Transaction identification in log files:
  - Time-based methods, e.g. 30 min silence interval. Problems with cache. Partial solutions: special HTTP headers, Java agents.

Collection and preparation

- Problems:
  - Privacy and security issues:
    - The user must be aware of the data collected.
    - Cookies and client-side agents are often disabled.
  - Caching on the client or an intermediate proxy causes data loss on the server side.
  - Registration forms are a nuisance and they are not reliable sources.
  - User and session identification from server logs is hard.
  - Different data required for different user models.

Personalized assistants

- **Personalized crawling** [Liebermann et al., ACM Comm., 2000]:
  - The system knows the user (log-in).
  - It uses heuristics to extract "important" terms from the Web pages that the user visits and add them to thematic profiles.
  - Each time the user views a page, the system:
    - searches the Web for related pages,
    - filters them according to the relevant thematic profile,
    - and constructs a list of recommended links for the user.
  - The Letizia version of the system searches the Web locally, following outgoing links from the current page.
  - The Powerscout version uses a search engine to explore the Web.