

Unit 5 Web Mining & Information Extraction

Part 2: Web Usage Mining

Geraldine Gray / Markus Hofmann

Overview

- Web mining part 1 focused on:
 - Web content mining
- In this presentation we focus on
 - Web usage mining: how users use our web sites based on log data and transaction processing / customer tables.

Web Usage Mining

- Web usage Mining is the automated discovery and analysis of patterns in **click stream** or other **user access data**
- Aims at the discovery of user access patterns from web usage logs.
- Generally every click is recorded (known as **clickstream**)
- Often **additional information of a user** is known such as
 - **Personal details** or
 - **General demographic information (e.g. population/regions average salary, average age, ethnicity)**
 - A lot of work (as usual) is put into the pre-processing of the data to get it into the correct format.

Web Usage Mining

- Big advantages are that you
 - Have detailed information about the structure of your website
 - Have domain knowledge
 - Have access to operational databases
 - Have access to other data on the www
 - Have access to web server logs
 - May be able to identify specific users and link them back to your operational records.

Web Usage Mining - Attributes

- In general, log files include:
 - IP address
 - ClientID
 - UserID
 - Date, Time & Time zone
 - Request
 - Status
 - Bytes
- And may additionally include:
 - Protocol
 - Method
 - Referrer
 - ClientBrowser/Agent

Field	Meaning
219.144.222.253	Users' IP address (UIP)
[16/Aug/2004...	The date and time of the request (Date)
GET	The method of the request (Method)
/images/1_r3...	The URL of the current request (URI)
HTTP/1.1	The version of transport protocol (Version)
200	The HTTP status code returned to the client (Status)
418	The content-length of the page transferred (Bytes)
http://202.11...	The URL requested just before (ReferURI)
Mozilla/4.0 (...)	Browser & OS (BrowserOS)

There are a number of log formats:

Common Log Format — supported by all web servers.

Extended Log Format — provides flexibility on what fields are recorded (used by IIS and others).

Other proprietary formats

Log File Examples – Web Log

- RapidMinerResources:

130.243.175.92 - - [13/Nov/2011:00:01:44 +0000] "GET /
stylesheet.php?cssid=59&mediatype=screen HTTP/1.1" 200 7843
"http://rapidminerresources.com/index.php?page=missing-values---
basic" "Opera/9.80 (X11; Linux i686; U; en) Presto/2.7.62 Version/
11.01"

130.243.175.92 - - [13/Nov/2011:00:01:44 +0000] "GET /
stylesheet.php?cssid=61&mediatype=screen HTTP/1.1" 200 2713
"http://rapidminerresources.com/index.php?page=missing-values---
basic" "Opera/9.80 (X11; Linux i686; U; en) Presto/2.7.62 Version/
11.01"

130.243.175.92 - - [13/Nov/2011:00:01:44 +0000] "GET /uploads/
1_ITB_main_campus.jpg HTTP/1.1" 200 15367 "http://
rapidminerresources.com/index.php?page=missing-values---basic"
"Opera/9.80 (X11; Linux i686; U; en) Presto/2.7.62 Version/11.01"

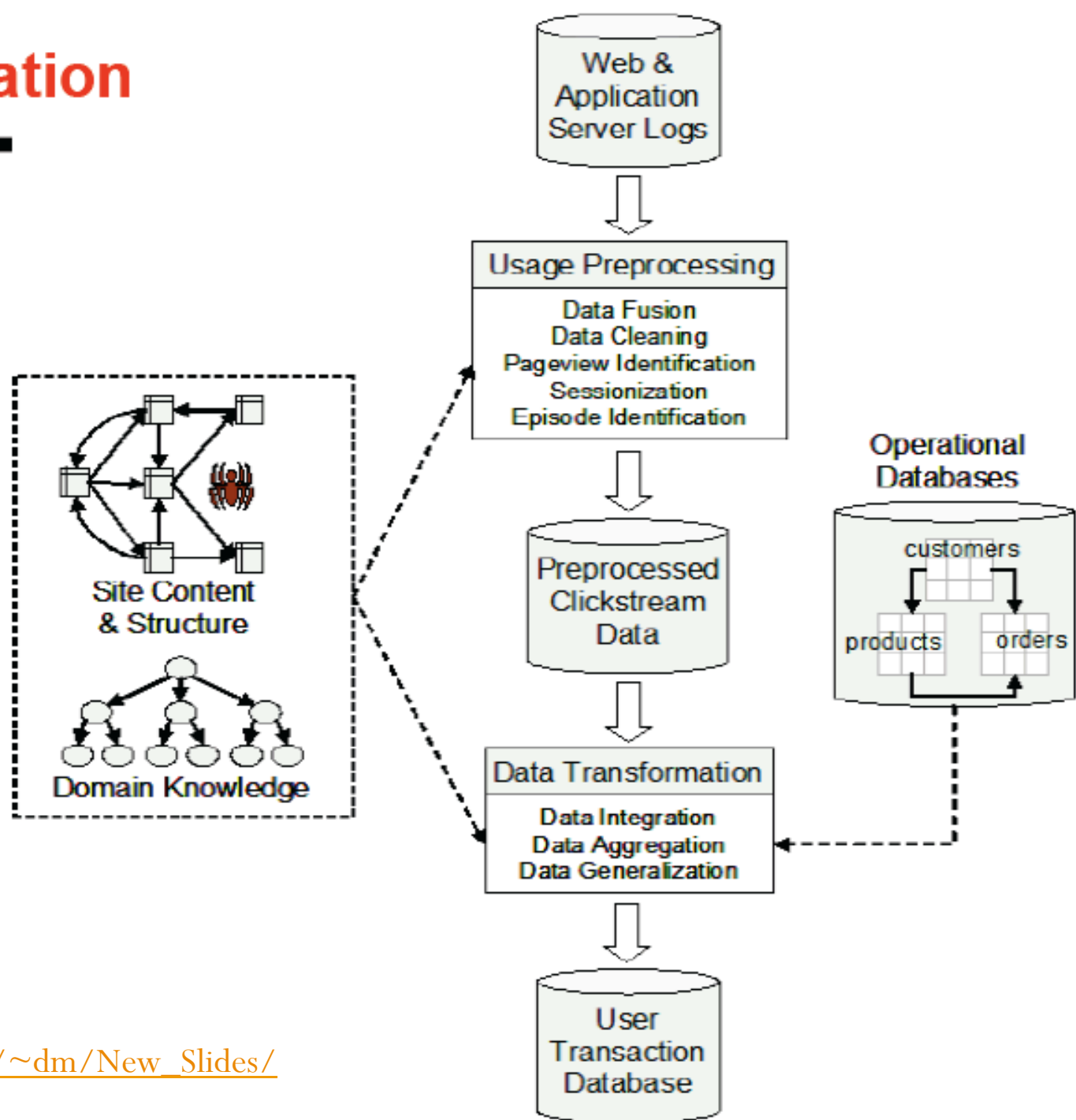
Log File Examples – CGIError Log

- PHP Warning: session_start() [PHP Warning: session_start() [

Information Available from Server Log

- **Click-through rate or CTR** – divide the number of users who clicked on an ad by the number of times the ad was delivered (impressions).
- **The Customer Conversion Rate** - take the number of people who buy something from your site in a week divided by the total number of unique visitors to your site.
- **New Customer Acquisition Rate** – the number of visitors that become customers. Will need additional information from company database to distinguish between new and existing customers.
- **Acquisition Cost**: the cost of acquiring a new customer versus the life time value of that customer (i.e. how much does an ad cost versus what type of customer is it likely to bring).
- **Click Streams**: typical paths taken by users through the website; or typical paths taken by users who make a purchase versus users who don't.
- **Entry and Exit pages.**
- **Referrer pages.**

Data preparation



Images from:

http://www.dais.unive.it/~dm/New_Slides/10_WUM.pdf

Key elements of pre-processing

1. Data Fusion and Cleaning

- In distributed environments, logs are often kept on different web servers. This stage merges these log files.
 - Must ensure that timestamps include time zone
- Cleaning is site specific but could include removal of unimportant objects (e.g. images, css), identification and removal of crawler related traffic, removing records with failed HTTP status codes.

2. Page view Identification

- Identification of page views depends on the intra page structure as each page generally consists of multiple files.

Key elements of pre-processing

3. User Identification

- User activity record refers to the sequence of logged activities belonging to the same user
- Cookies are often used to separate users
- IP addresses alone are often not sufficient but can be used in combination with other data such as the Agent who accessed the web server (e.g. IE8;Win7;SP1), the operating systems

4. Sessionisation

- Segmenting the user activity record of each user into sessions, each representing a single visit to the site

User Identification options:

Method	Description	Privacy Concerns	Advantages	Disadvantages
IP Address + Agent	Assume each unique IP address/Agent pair is a unique user	Low	Always available. No additional technology required.	Not guaranteed to be unique. Defeated by rotating IPs.
Embedded Session Ids	Use dynamically generated pages to associate ID with every hyperlink	Low to medium	Always available. Independent of IP addresses.	Cannot capture repeat visitors. Additional overhead for dynamic pages.
Registration	User explicitly logs in to the site.	Medium	Can track individuals not just browsers	Many users won't register. Not available before registration.
Cookie	Save ID on the client machine.	Medium to high	Can track repeat visits from same browser.	Can be turned off by users.
Software Agents	Program loaded into browser and sends back usage data.	High	Accurate usage data for a single site.	Likely to be rejected by users.

Key elements of pre-processing

Examples of rules for session identification:

1. The different IP addresses distinguish different users;
2. If the IP addresses are same, the different browsers and operation systems indicate different users;
3. If all of the IP address, browsers and operating systems are same, the referer information should be taken into account. The ReferURI field is checked, and a new user session is identified if the URL in the ReferURI field hasn't been accessed previously, or there is a large interval (usually more than 10 seconds [4]) between the accessing time of this record and the previous one if the ReferURI field is empty;

OR . . .
Use cookies to
store sessionID/
customerID.

Key elements of pre-processing

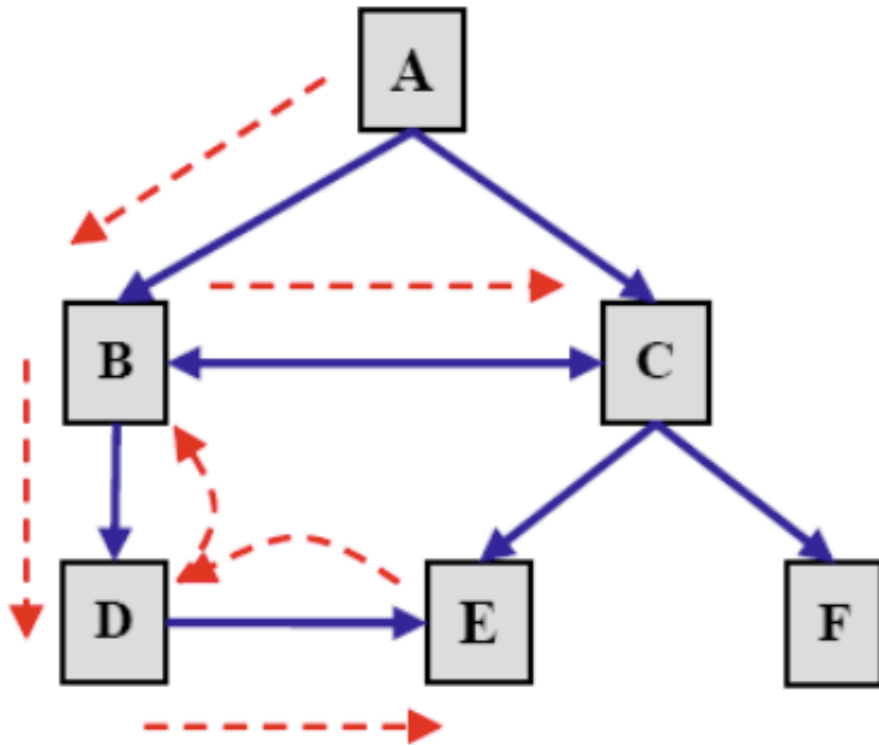
5. Path Completion

- Client or proxy-side caching can result in missing access references to previously cached objects
- Knowing the structure of the site it is possible to fill in the missing values.
- Simple heuristic can be applied to fill in the gaps, e.g.
 - Look at the referrer URI
 - Look at the time interval, and on average how long that user spends on content pages.
 - Fill missing pages using the shortest path (via viewed pages) to next new page requested.

6. Data Integration

- Integration of pre-processed data with data coming from other sources

Path completion example . . .



User's actual navigation path:

$A \rightarrow B \rightarrow D \rightarrow E \rightarrow D \rightarrow B \rightarrow C$

What the server log shows:

<u>URL</u>	<u>Referrer</u>
A	--
B	A
D	B
E	D
C	B

Knowledge of site structure is required to infer missing pages.

Results

When the six preprocessing steps are complete, the following can be generated from log files for further analysis:

- User page view matrix
 - Lists users and pages they've viewed
- Term pageview matrix
 - Terms are displayed on the vertical axis and pages on the horizontal axis
 - The transpose of this will be the pageview feature matrix
- Content Enhanced transaction matrix
 - Displays users on the vertical axis and terms on the horizontal axis

Page view matrix:

	page A	page B	page C	page D	page E
user 0	15	4	1	0	0
user 1	2	0	25	0	0
user 2	200	1	0	0	3
user 3	56	0	0	4	4
user 4	0	0	23	50	0
user 5	0	0	5	3	0

Term Pageview matrix:

	food	news	car	house	party	sky
page A	0	1	1	0	0	0
page B	1	0	0	1	0	0
page C	1	1	0	0	0	0
page D	0	0	1	0	0	1
page E	0	0	0	1	1	0

Facilitates:

- 1. Clustering of visitor segments
- 2. Clustering of page views
- 3. Sequential rules analysis of the order in which pages are viewed

Content enhanced transaction matrix:

Combine to determine user interests:

	food	news	car	house	party	sky
user 0	1	1	1	1	0	0
user 1	1	1	0	0	0	0
user 2	0	1	1	1	1	0
user 3	0	1	2	1	1	1
user 4	1	1	1	0	0	1
user 5	1	1	0	0	0	0






Web usage mining on Rapidminer:

There are a number of operators on the Web Mining package:

1. Read Server Log

- Accepts a number of log formats
- Identifies sessions
- Filter web log: e.g. removes requests for images

Note: RapidMiner 5.2 is missing the jar file needed for this operator. Earlier and later version work OK.

config file	<input type="text" value="rRepository/logs/config.xml"/> 
log dir	<input type="text" value="idMinerRepository/logs/log"/> 
<input type="checkbox"/> <i>dns lookup</i>	
<i>robot filter</i>	<input type="text"/>
<i>filetype filter</i>	<input type="text" value="gif png css jpg txt"/>
<input checked="" type="checkbox"/> <i>only HTTP 200</i>	
<i>browser matcher</i>	 Edit List (0)...
<i>os matcher</i>	 Edit List (3)...
<i>language matcher</i>	 Edit List (0)...
<i>session timeout</i>	<input type="text" value="400000"/>

Web server Log file: initial cleaning

row no.	session	ip	agent	uri	referer	time	os_name ▲	browser	language
144	s33	146.107.217.2	Mozilla/5.0 (I	/	http://www-	17888500	mac	other	other
145	s33	146.107.217.2	Mozilla/5.0 (I	/menue.htm	http://icdm0	17888500	mac	other	other
146	s33	146.107.217.2	Mozilla/5.0 (I	/hauptseite.h	http://icdm0	17888500	mac	other	other
147	s33	146.107.217.2	Mozilla/5.0 (I	/Organisation	http://icdm0	17888500	mac	other	other
148	s33	146.107.217.2	Mozilla/5.0 (I	/Organisation	http://icdm0	17888500	mac	other	other
170	s39	208.39.140.82	Mozilla/4.0 (C	/	http://www.c	17888613	mac	IE	other
171	s39	208.39.140.82	Mozilla/4.0 (C	/hauptseite.h	?	17888613	mac	IE	other

- How would you analyse the click stream behaviour for each session?

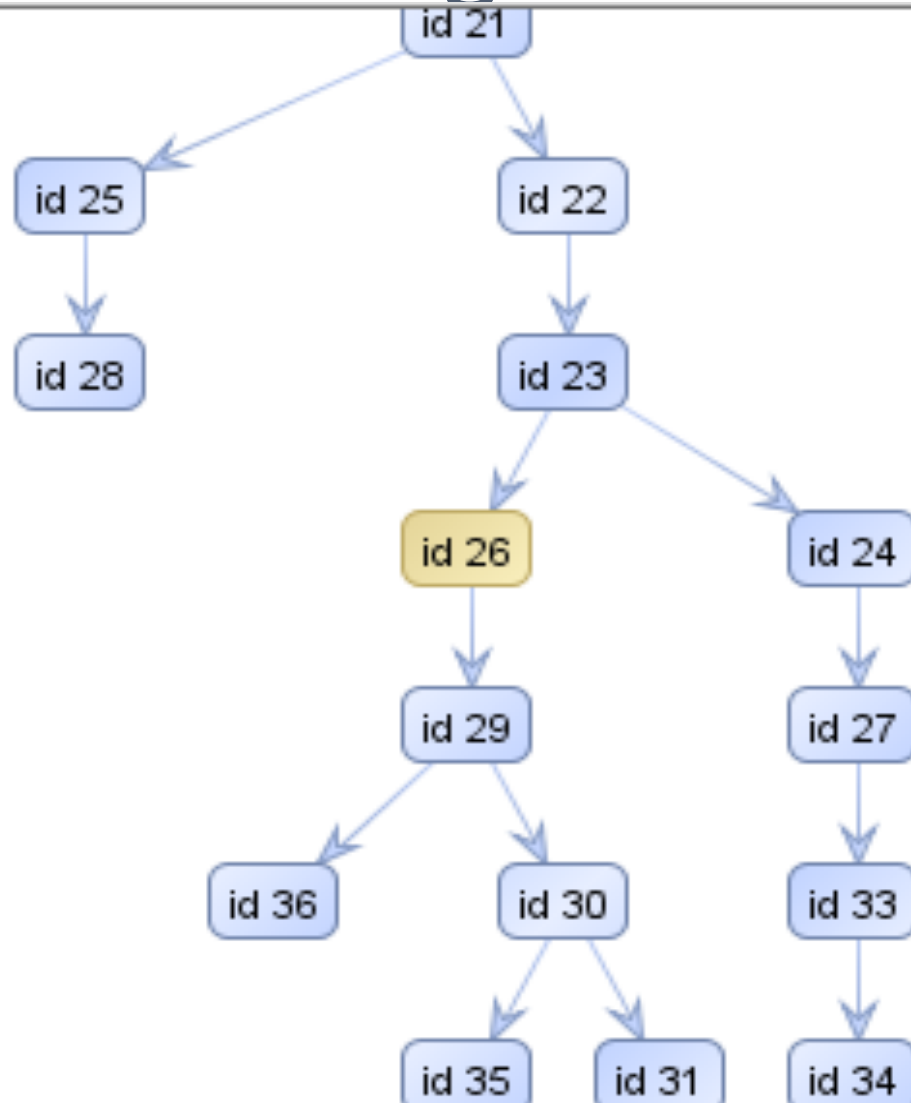
Transform Log to Session

row no.	/Organisati...	/Awards.html	/hauptseite....	/RelatedLin...	/Richtlinien...	/menue.html	/index.html	/Daten.html	/Organisati...	/Panel.html	/Links.html	
1	0	0	1	0	0	1	0	1	0	0	0	1
2	1	0	1	0	0	1	0	0	1	1	1	1
3	0	0	1	0	0	1	0	0	0	0	0	1
4	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	1	0	0	1	1	0	0	0	0	1
6	0	0	1	0	0	1	0	1	0	0	0	1
7	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	1	0	0	1	0	0	0	0	0	1

1 row per session

- Distribution graph on each column shows number of clicks per session for each page
- Aggregates would give total counts for each page
- Market Basket Analysis gives click stream behaviour
- Clusters could be used to identify different user groups, and their behaviour on the site.
- For purchase related statistics, you would need to identify, and label, pages which signify a purchase was made during a session, or combining a session user with their customer data.

Clustering web site visitors



/Themen.html
/Organisation1.html
/Organisation2.html
/Panel.html
/Sponsoren.html
/Steering.html
/Richtlinien.html
/Benefits.html
/Submission.html
/Eingeladen.html

More Rapidminer processes. . .

Take a look at some Rapid Miner processes on moodle illustrating web mining:

- weblog_cluster: clusters web pages. Pages are similar if they were accessed in the same sessions.
- weblog_fpgrowth: which pages are accessed together
- weblog_typicalPath: what is the order in which pages are access
- Weblog-transitionMatrix: if on page X, what is the likelihood of moving to page Y in the website.
- Request statistics: page statistics

Web Usage Mining - Summary

- Domain knowledge is required
- Can be very powerful
 - E.g. Amazon.co.uk
- Future focus lies on improvements of techniques and architectures to integrate more data

Web Structure Mining

- Web structure mining discovers useful information and knowledge by analysing hyperlinks which represent the structure of the web
- For example, analysing links lets us derive information about most important web pages, gateway pages, communities of users who share common interest
- We are not covering this topic any further. . . .

E-commerce data analysis

Basic Framework for E-Commerce Data Analysis

