Web Usage Mining

- An introduction

Introduction

- **Web usage mining**: automatic discovery of patterns in **clickstreams** and **associated data collected or generated (server logs)** as a result of user interactions with one or more Web sites.

- **Goal**: Capture, Model and Analyze the **behavioral patterns and profiles of users** interacting with a Web site.

- The discovered patterns are usually represented as **collections of pages, objects, or resources that are frequently accessed** by groups of users with common interests.

Introduction

- Data in Web Usage Mining:
  - Web server logs
  - Site contents
  - Data about the visitors, gathered from external channels
  - Further application data

- Not all these data are always available.
- When they are, they must be integrated.

- A large part of Web usage mining is about processing **usage/ clickstream data.**
  - After that various data mining algorithm can be applied.

Web server logs

IP Address

Resource Name

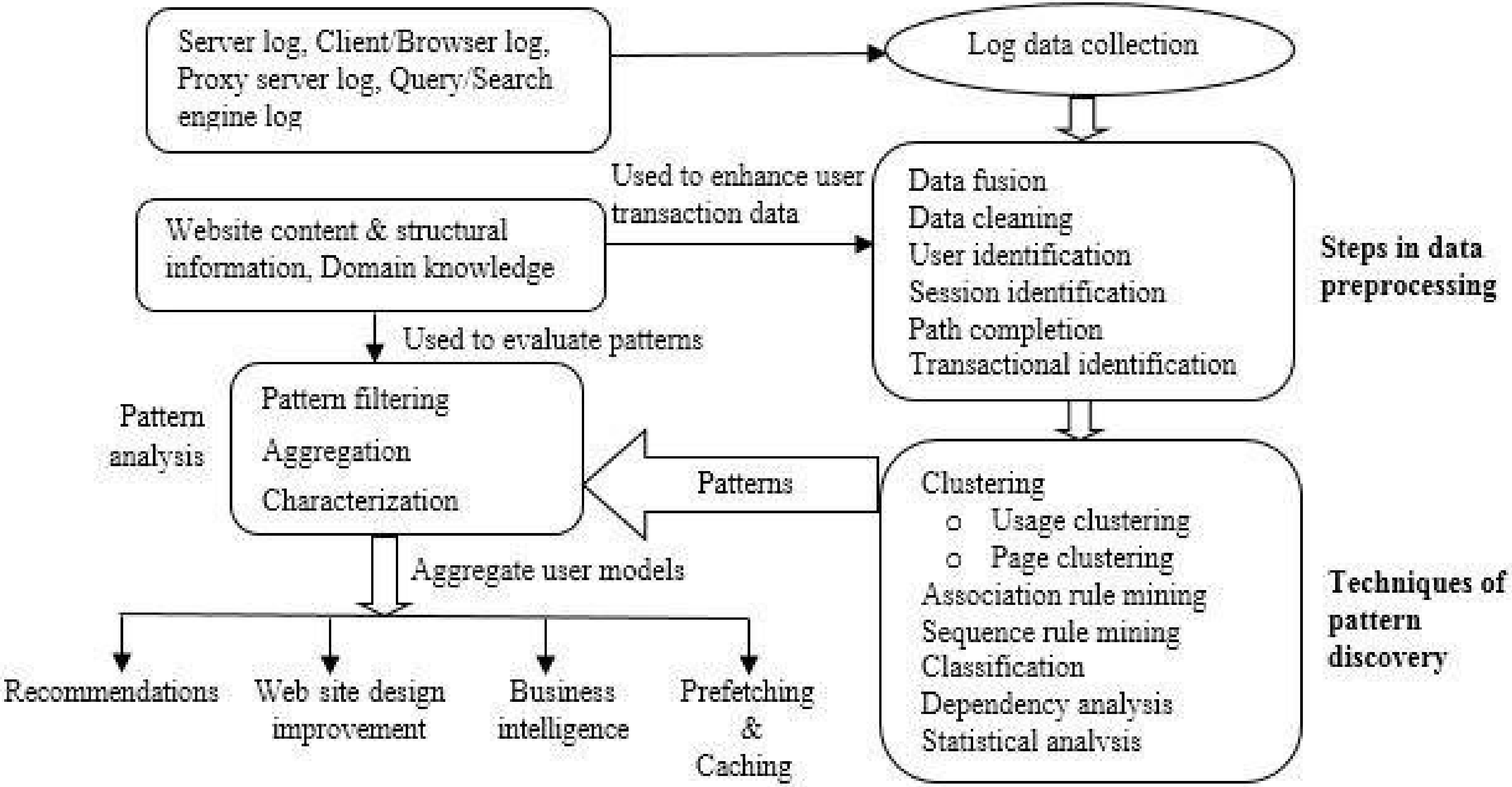
1	2006-02-01 00:08:43 1.2.3.4 - GET /classes/cs589/papers.html - 200 9221 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR+2.0.50727) http://dataminingresources.blogspot.com/
2	2006-02-01 00:08:46 1.2.3.4 - GET /classes/cs589/papers/cms-tai.pdf - 200 4096 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR+2.0.50727) http://maya.cs.depaul.edu/~classes/cs589/papers.html
3	2006-02-01 08:01:28 2.3.4.5 - GET /classes/ds575/papers/hyperlink.pdf - 200 318814 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1) http://www.google.com/search?hl=en&lr=&q=hyperlink+analysis+for+the+web+survey
4	2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/announce.html - 200 3794 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1) http://maya.cs.depaul.edu/~classes/cs480/
5	2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/styles2.css - 200 1636 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1) http://maya.cs.depaul.edu/~classes/cs480/announce.html
6	2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/header.gif - 200 6027 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1) http://maya.cs.depaul.edu/~classes/cs480/announce.html

Operating System Information

Browser type and version

Referrer : Reached to this location due to Google search

Referrer : indicates that the user came to this location from this outside source



Server log, Client/Browser log, Proxy server log, Query/Search engine log

Log data collection

Website content & structural information, Domain knowledge

Used to enhance user transaction data

Data fusion
Data cleaning
User identification
Session identification
Path completion
Transactional identification

Steps in data preprocessing

Used to evaluate patterns

Pattern analysis

Pattern filtering
Aggregation
Characterization

Patterns

Clustering
- Usage clustering
- Page clustering

Association rule mining
Sequence rule mining
Classification
Dependency analysis
Statistical analysis

Techniques of pattern discovery

Aggregate user models

Recommendations

Web site design improvement

Business intelligence

Prefetching & Caching

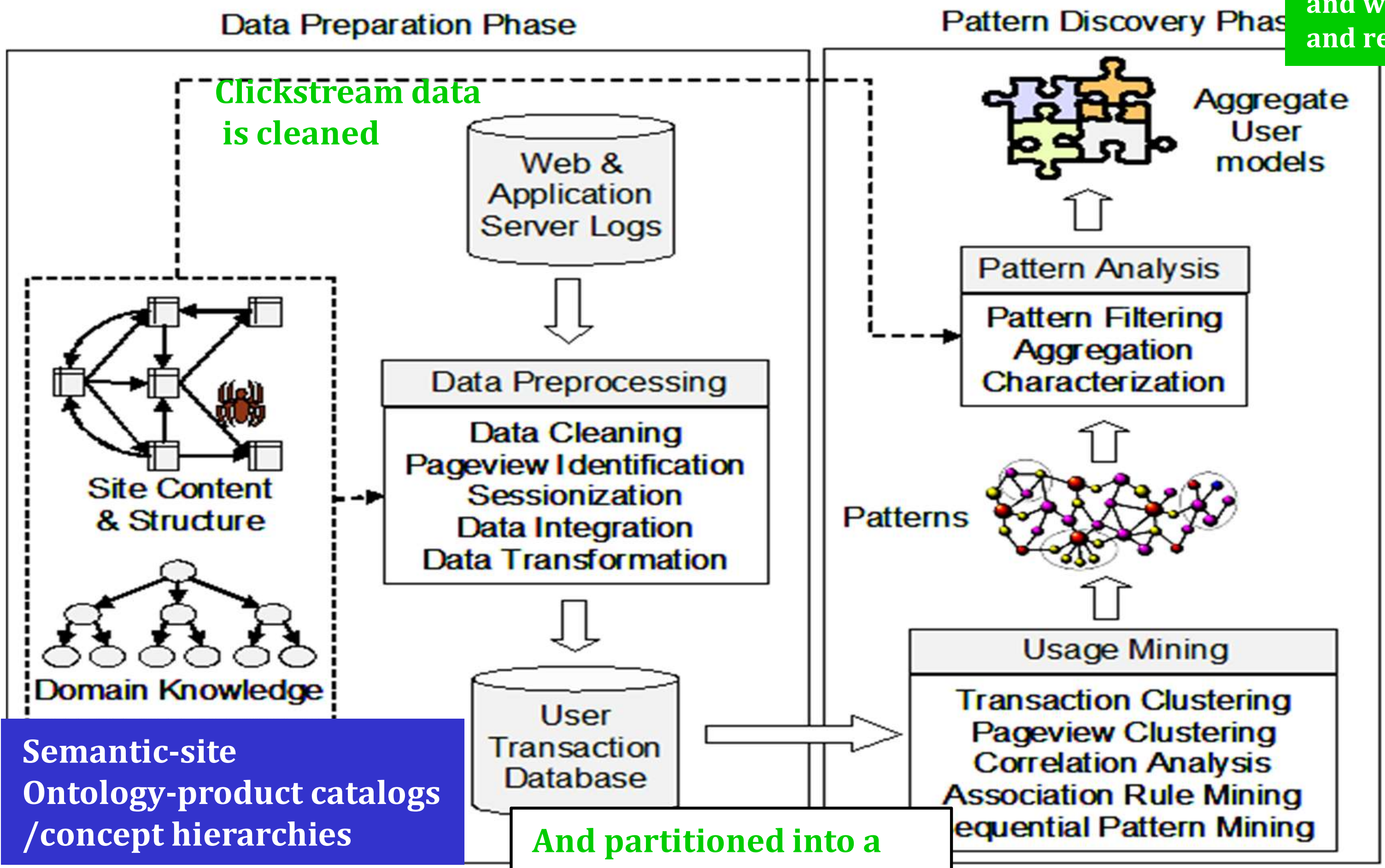
Web usage mining process

3 independent stages

1.Data collection and pre-processing

2. Pattern discovery

3.Pattern Analysis



Clickstream data is cleaned

And partitioned into a set of user Transactions Representing the activities of each user during different visits to the site

Semantic-site Ontology-product catalogs /concept hierarchies

Input to
Recommedation Engines,
Visualisation tools
and web analytics
and report generating tools

User behavior, Summary of statistics On Web resources, Sessions, users

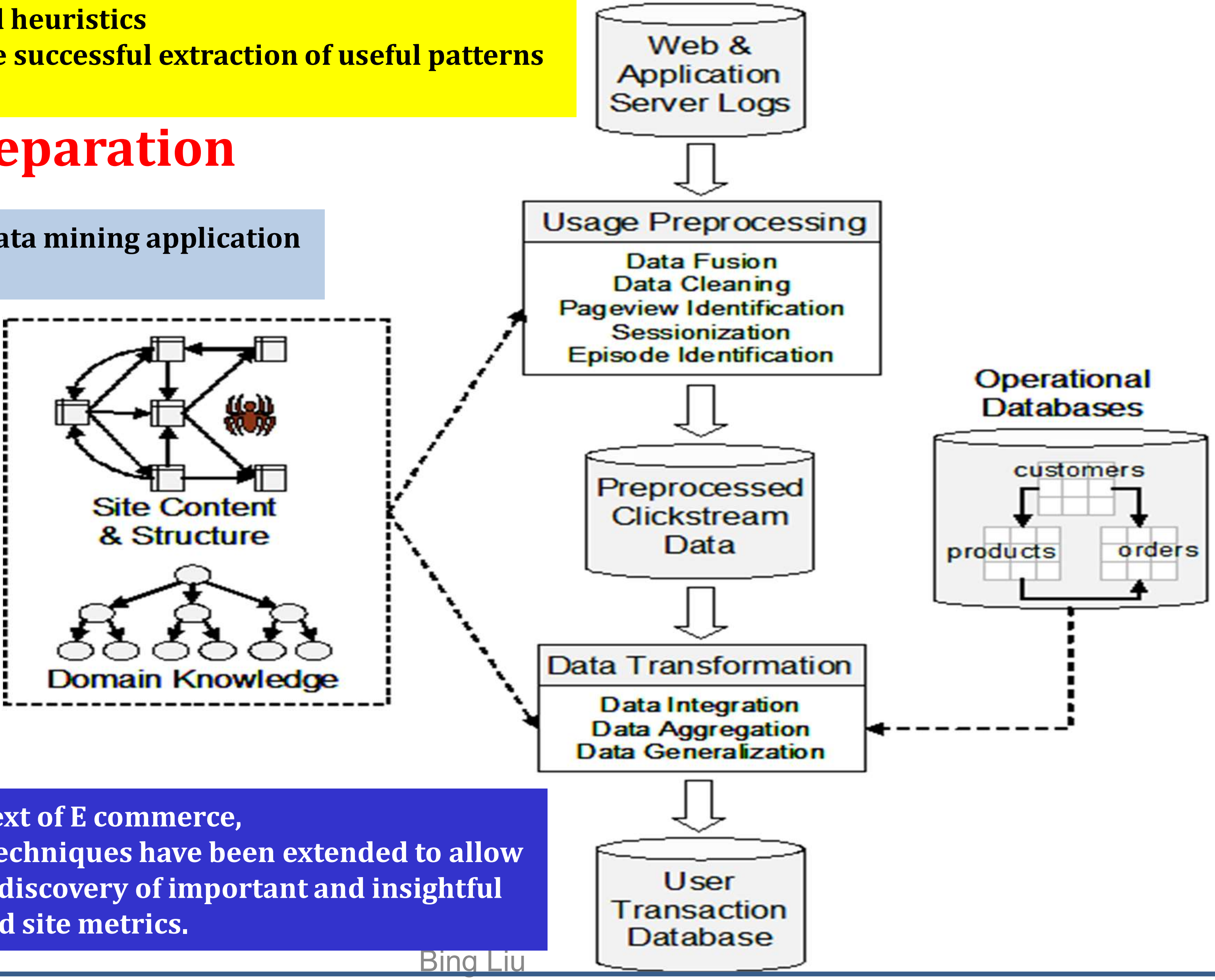
ML
Statistical
Database
operations

Computationally intensive and time consuming process
Need special algorithms and heuristics
This process is critical to the successful extraction of useful patterns
From the data

Data preparation

An important task in any data mining application
Is suitable target dataset.

It is important step for web usage mining due to characteristics of clickstream data and its relationship to other related data collected from multiple sources and multiple channels


Web & Application Server Logs

Usage Preprocessing
Data Fusion
Data Cleaning
Pageview Identification
Sessionization
Episode Identification

Site Content & Structure

Domain Knowledge

Operational Databases
customers
products
orders

Preprocessed Clickstream Data

Data Transformation
Data Integration
Data Aggregation
Data Generalization

User Transaction Database

In context of E commerce,
These techniques have been extended to allow
For the discovery of important and insightful
User and site metrics.

Sources and Types of Data- Web Server Access Logs and application server logs

Additional data sources needed for data preparation and pattern discovery are

Site files and meta-data

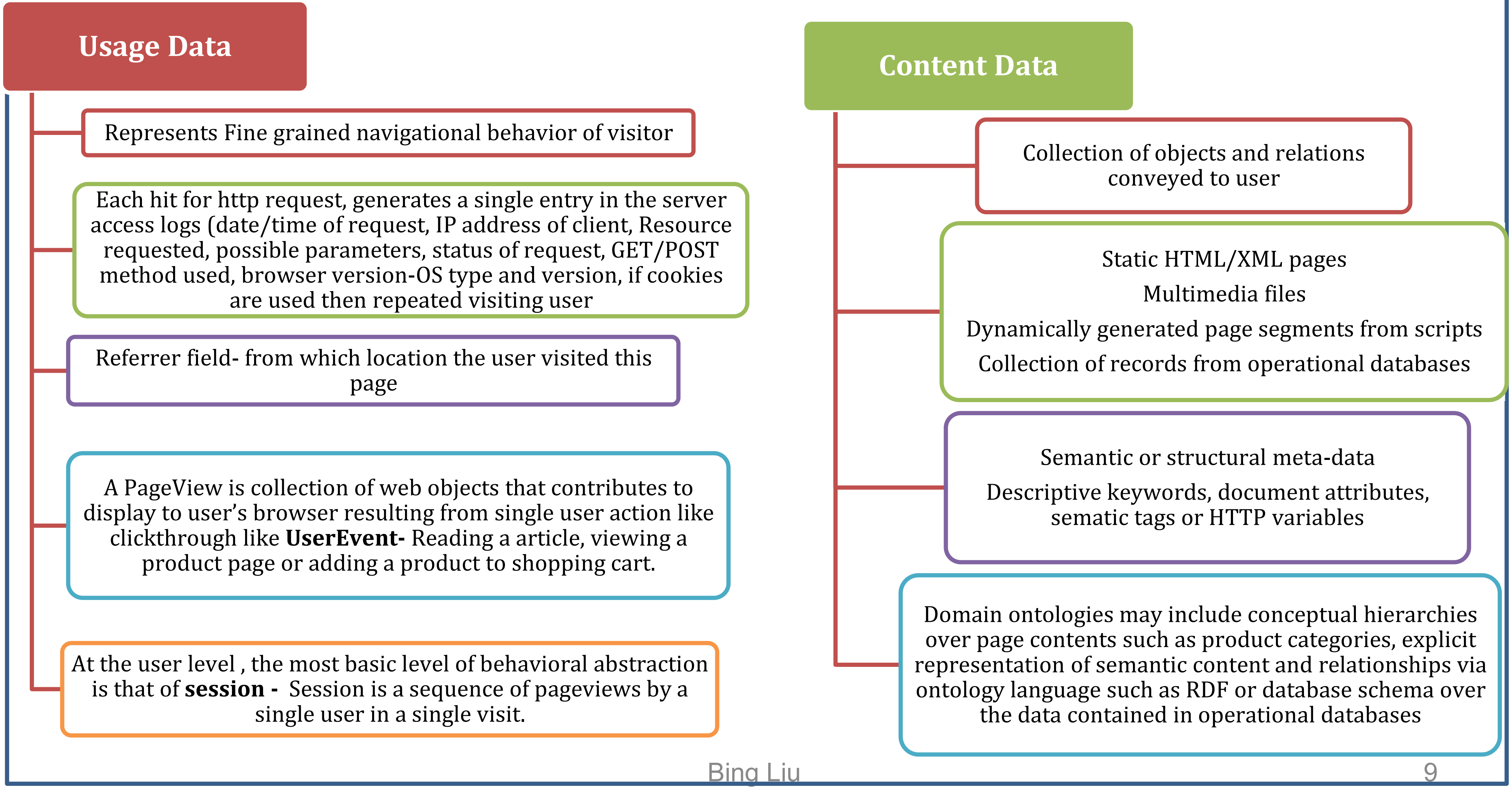
operational databases

application templates

domain knowledge

client-side/proxy level data collection

Sources and Types of Data- Web Server Access Logs and application server logs



Usage Data

- Represents Fine grained navigational behavior of visitor
- Each hit for http request, generates a single entry in the server access logs (date/time of request, IP address of client, Resource requested, possible parameters, status of request, GET/POST method used, browser version-OS type and version, if cookies are used then repeated visiting user
- Referrer field- from which location the user visited this page
- A PageView is collection of web objects that contributes to display to user's browser resulting from single user action like clickthrough like **UserEvent-** Reading a article, viewing a product page or adding a product to shopping cart.
- At the user level , the most basic level of behavioral abstraction is that of **session -** Session is a sequence of pageviews by a single user in a single visit.

Content Data

- Collection of objects and relations conveyed to user
- Static HTML/XML pages

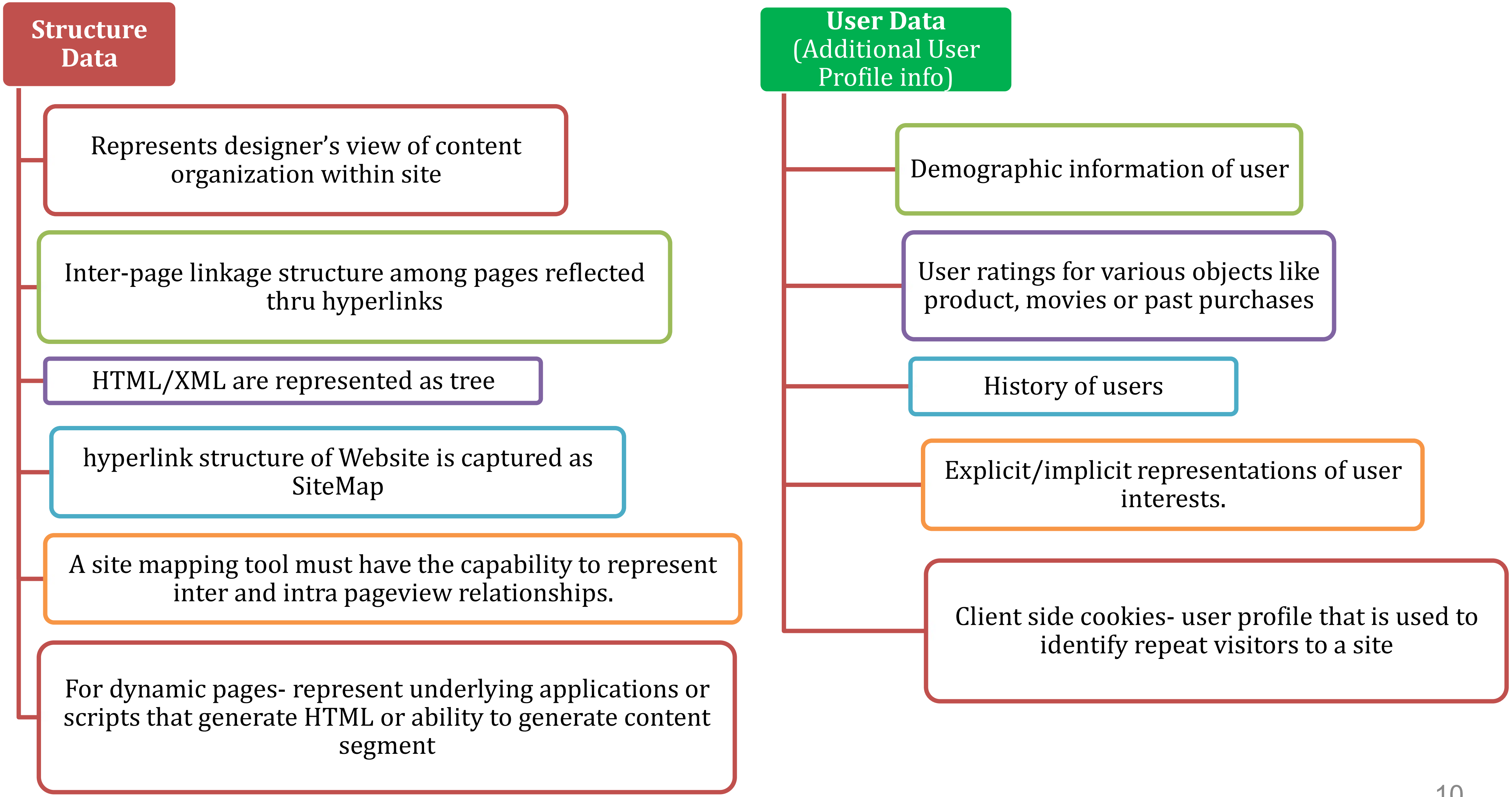
  Multimedia files

  Dynamically generated page segments from scripts

  Collection of records from operational databases
- Semantic or structural meta-data

  Descriptive keywords, document attributes, sematic tags or HTTP variables
- Domain ontologies may include conceptual hierarchies over page contents such as product categories, explicit representation of semantic content and relationships via ontology language such as RDF or database schema over the data contained in operational databases

Data Sources and Its Types



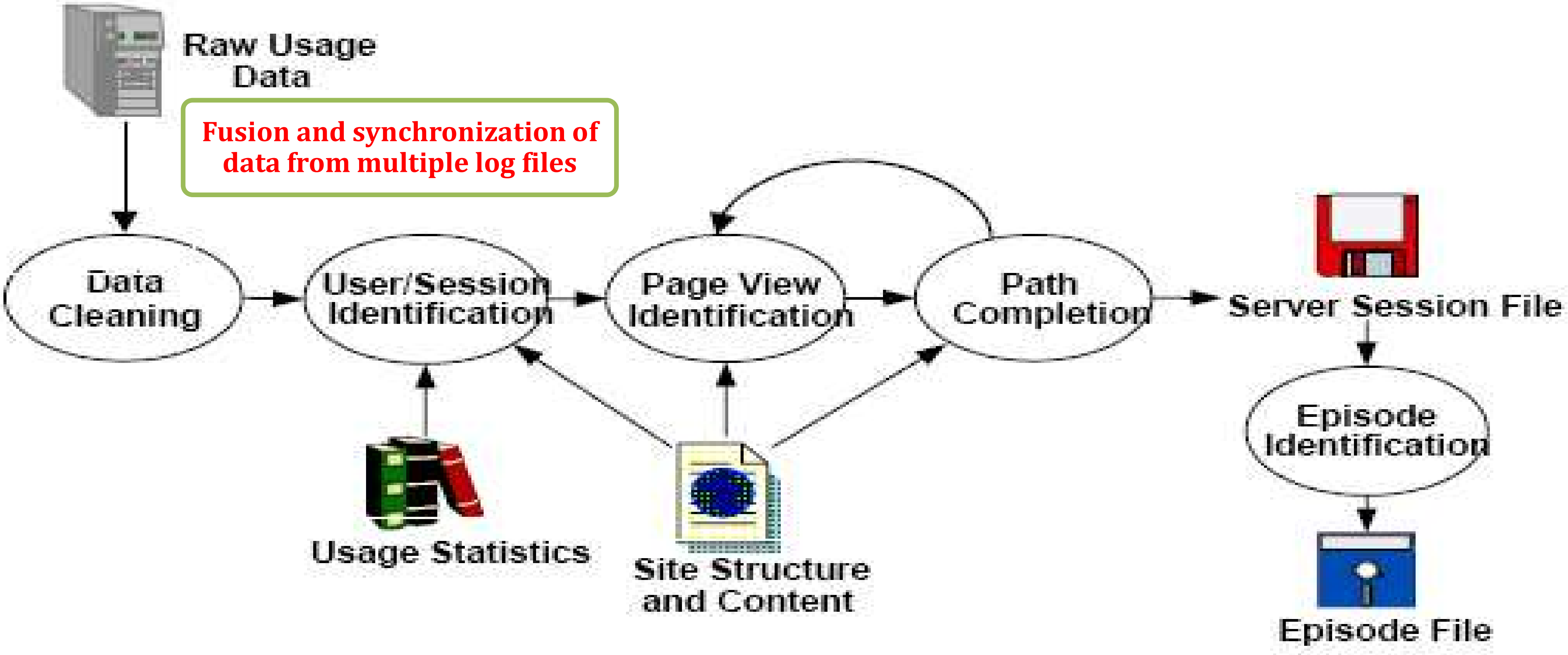
Structure Data

- Represents designer's view of content organization within site
- Inter-page linkage structure among pages reflected thru hyperlinks
- HTML/XML are represented as tree
- hyperlink structure of Website is captured as SiteMap
- A site mapping tool must have the capability to represent inter and intra pageview relationships.
- For dynamic pages- represent underlying applications or scripts that generate HTML or ability to generate content segment

User Data (Additional User Profile info)

- Demographic information of user
- User ratings for various objects like product, movies or past purchases
- History of users
- Explicit/implicit representations of user interests.
- Client side cookies- user profile that is used to identify repeat visitors to a site

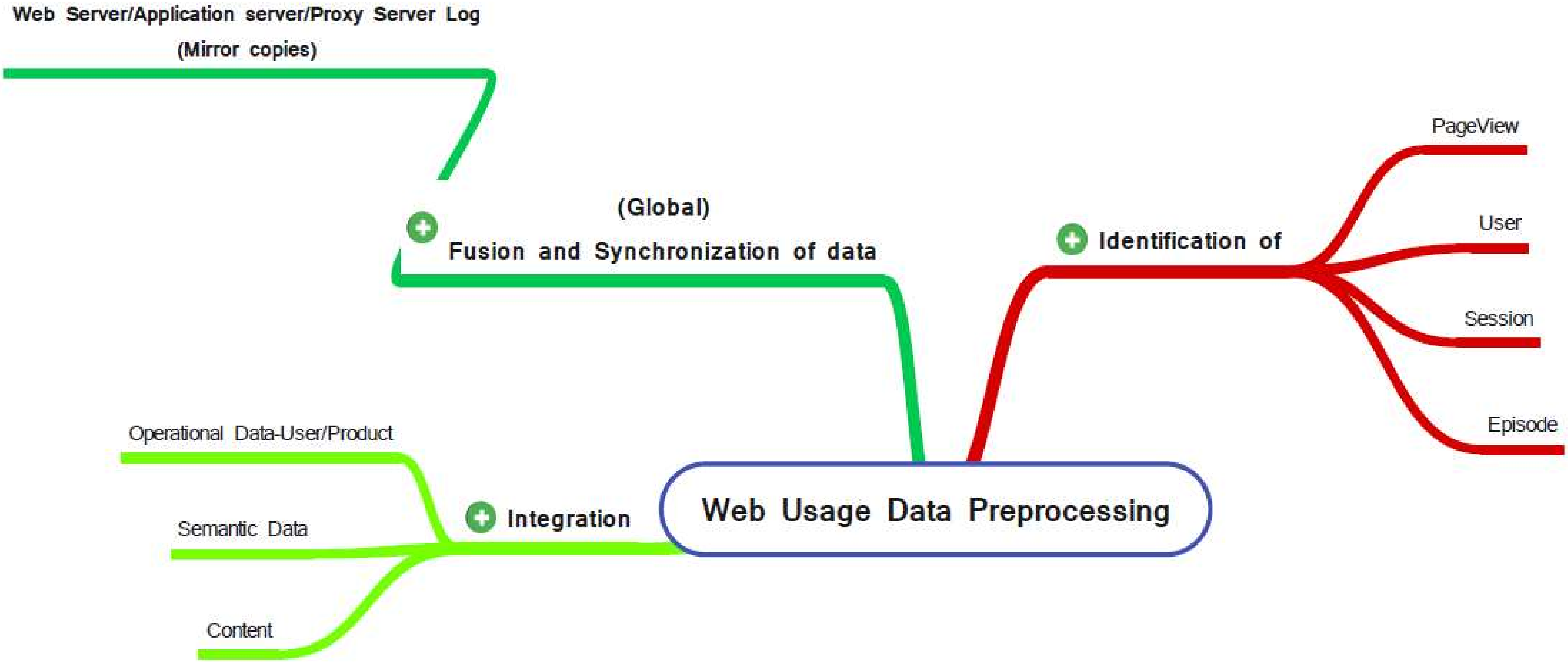
Pre-processing of web usage data



Fusion and synchronization of data from multiple log files

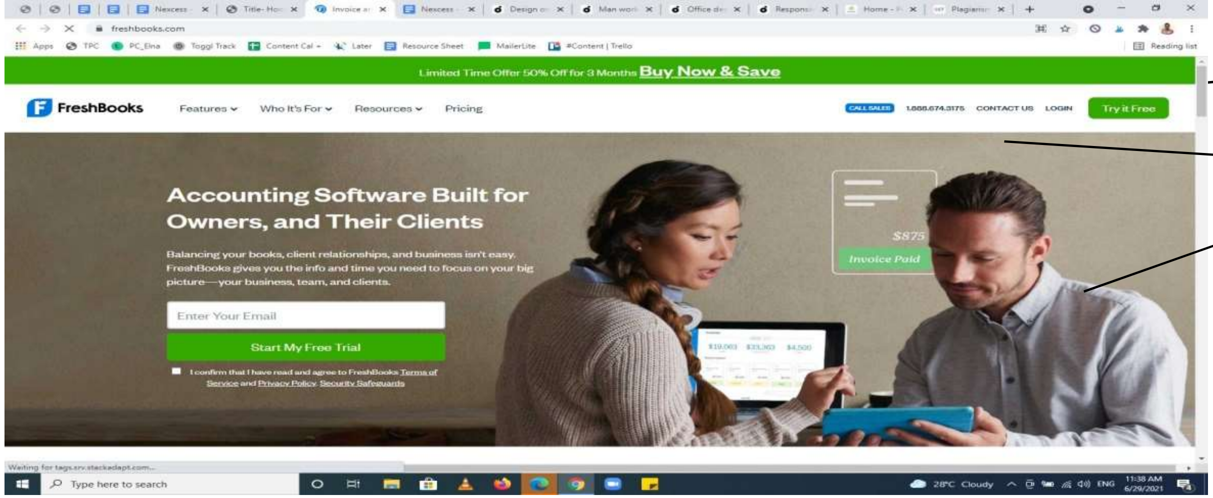
User/Product information from operational databases

Key Tasks in web usage data Pre-processing


Web Server/Application server/Proxy Server Log
(Mirror copies)
(Global)
Fusion and Synchronization of data
Identification of
PageView
User
Session
Episode
Operational Data-User/Product
Semantic Data
Content
Integration
Web Usage Data Preprocessing

web usage data Pre-processing – Data Cleaning **(It is site specific)**

Remove



Accounting Software Built for Owners, and Their Clients

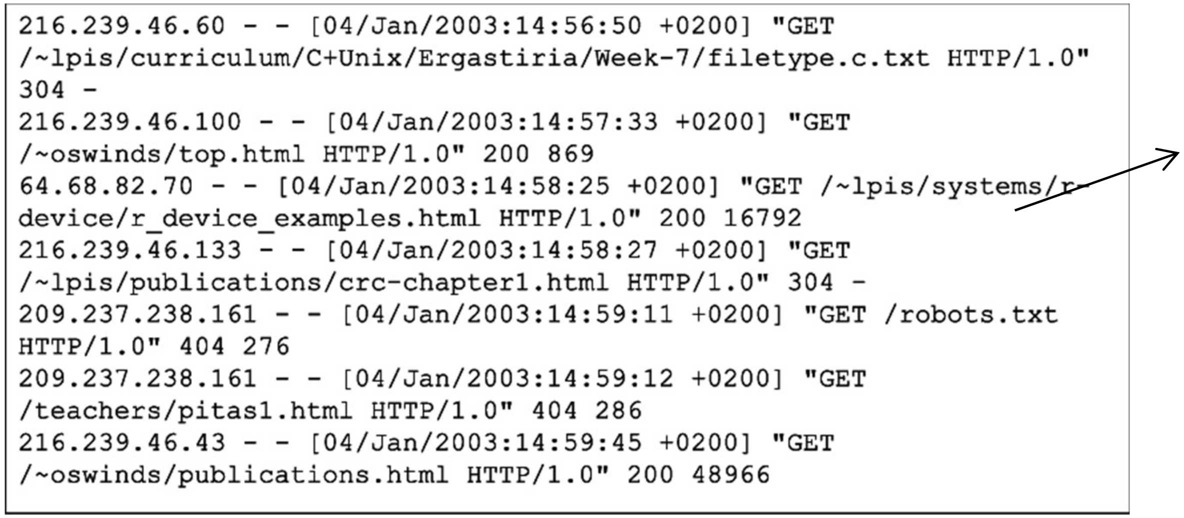
Balancing your books, client relationships, and business isn't easy. FreshBooks gives you the info and time you need to focus on your big picture—your business, team, and clients.

□ I confirm that I have read and agree to FreshBooks Terms of Service and Privacy Policy. Security Safeguards

Reference to external stylesheet

Extraneous reference of embedded object

Images/sound files



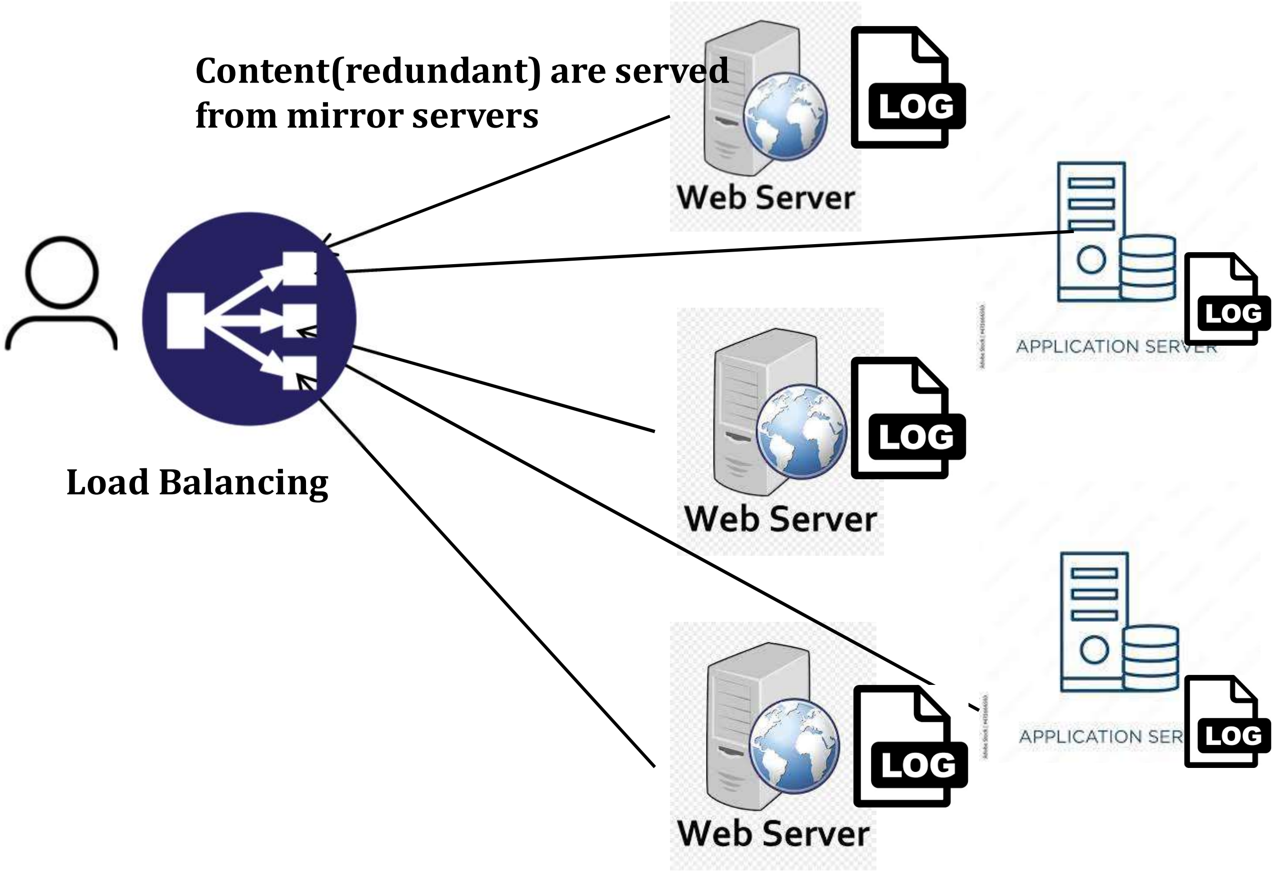
```
216.239.46.60 - - [04/Jan/2003:14:56:50 +0200] "GET
/~lpis/curriculum/C+Unix/Ergastiria/Week-7/filetype.c.txt HTTP/1.0"
304 -
216.239.46.100 - - [04/Jan/2003:14:57:33 +0200] "GET
/~oswinds/top.html HTTP/1.0" 200 869
64.68.82.70 - - [04/Jan/2003:14:58:25 +0200] "GET /~lpis/systems/r-
device/r_device_examples.html HTTP/1.0" 200 16792
216.239.46.133 - - [04/Jan/2003:14:58:27 +0200] "GET
/~lpis/publications/crc-chapter1.html HTTP/1.0" 304 -
209.237.238.161 - - [04/Jan/2003:14:59:11 +0200] "GET /robots.txt
HTTP/1.0" 404 276
209.237.238.161 - - [04/Jan/2003:14:59:12 +0200] "GET
/teachers/pitas1.html HTTP/1.0" 404 286
216.239.46.43 - - [04/Jan/2003:14:59:45 +0200] "GET
/~oswinds/publications.html HTTP/1.0" 200 48966
```

Server log file

Number of bytes transferred
Version of HTTP
Referrer-
References due to crawler navigations-(sometimes 50%)

aa

Data Fusion and Data cleaning



Content(redundant) are served from mirror servers

Load Balancing

Mirror web/application servers

ANALYSIS OF USER BEHAVIOUR

Data fusion refers to merging of log files From multiple web/application servers.

- This may require global synchronization between these servers.
- Heuristic methods used might be ..
  “Referrer” field,
  Sessionization,
  User identification

This step is essential in “inter-site” web usage mining where the analysis of user behavior is performed over the log files of multiple related web sites

Data Fusion and Data cleaning

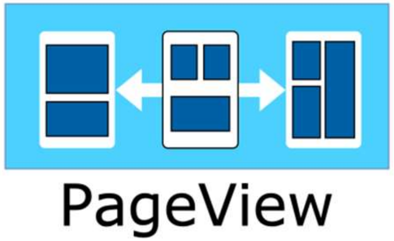
Data cleaning (usually site specific)

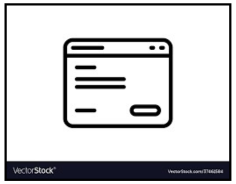
- remove irrelevant references and data fields in server logs (number of bytes transferred or HTTP Protocol version)
- remove erroneous references
- remove useless references of embedded objects, style files, graphics, sound files
- add missing references due to caching (done after sessionization)
- remove references due to spider navigation (50% of references are resulting from crawlers)

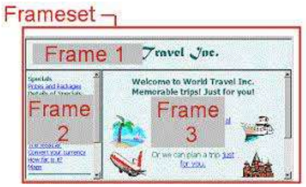
A significant portion of crawler references are from those that either do not identify themselves explicitly (e.g. in agent field) or implicitly; or from those crawlers that deliberately masquerade as legitimate users.

In this case identification of crawler references may require the use of heuristic methods that distinguish typical behaviour of web crawlers from those of actual users.

Pageview Identification







Single HTML file have 1:1 with pageview

Multiple HTML files have n:1 with pageview

For dynamic sites, a page view may present
static template of html
+
Contents generated by application servers based on a set of parameters

At higher level of abstraction Pageview is Collection of related pages (of same category)+objects

Product oriented events For e.g. E-Commerce web site Events like

Product view

Registration

Shopping Cart changes

Purchase

- Identification of pageview is heavily dependent on the
  intra-page structure of the site, as well as
  on the **pagecontents** and
  the underlying **site domain knowledge**.
- Conceptually each pageview can be viewed as a **collection of web objects or resources** representing a specific "user event"
- **e.g. clicking on a link, viewing a product page, adding a product to the shopping cart.**

In order to provide a flexible framework for a variety of data mining activities, a **number of attributes** must be recorded with each pageview.

Pageview id == URL	Static Pageview type	Other Metadata
	Info page Product View Category View Index Page	Keywords or Product Attributes

Requirement for identification of pageview is
A *PRIORI* SPECIFICATION OF EVENT MODEL BASED ON WHICH USER ACTIONS CAN BE CATEGORISED

Identify sessions (sessionization)

- In Web usage analysis, the data is the sessions of the site visitors. The activities performed by a user from the moment he/she enters the site until the moment he/she leaves it.
- Sessionization is the process of segmenting the user activity record of each user into sessions, each representing a single visit to the site.
- It is Difficult to obtain reliable usage data due to proxy servers and anonymizers, dynamic IP addresses, missing references due to caching, and the inability of servers to distinguish among different visits.
- In case of missing user authentication resulting in absence of embedded session id, we need to rely on heuristics methods of sessionization.
- The goal of sessionization heuristic is to reconstruct from the clickstream data, the actual sequence of actions performed by one user during one visit to the site.

R – conceptual set of real sessions representing real activity of the user on the wen site

h- Sessionization heuristics maps R into set of constructed sessions C_h

Idealistic heuristics h* …. We have R == C_h

Ideal heuristics can reconstruct the exact sequence of user navigation during a session.

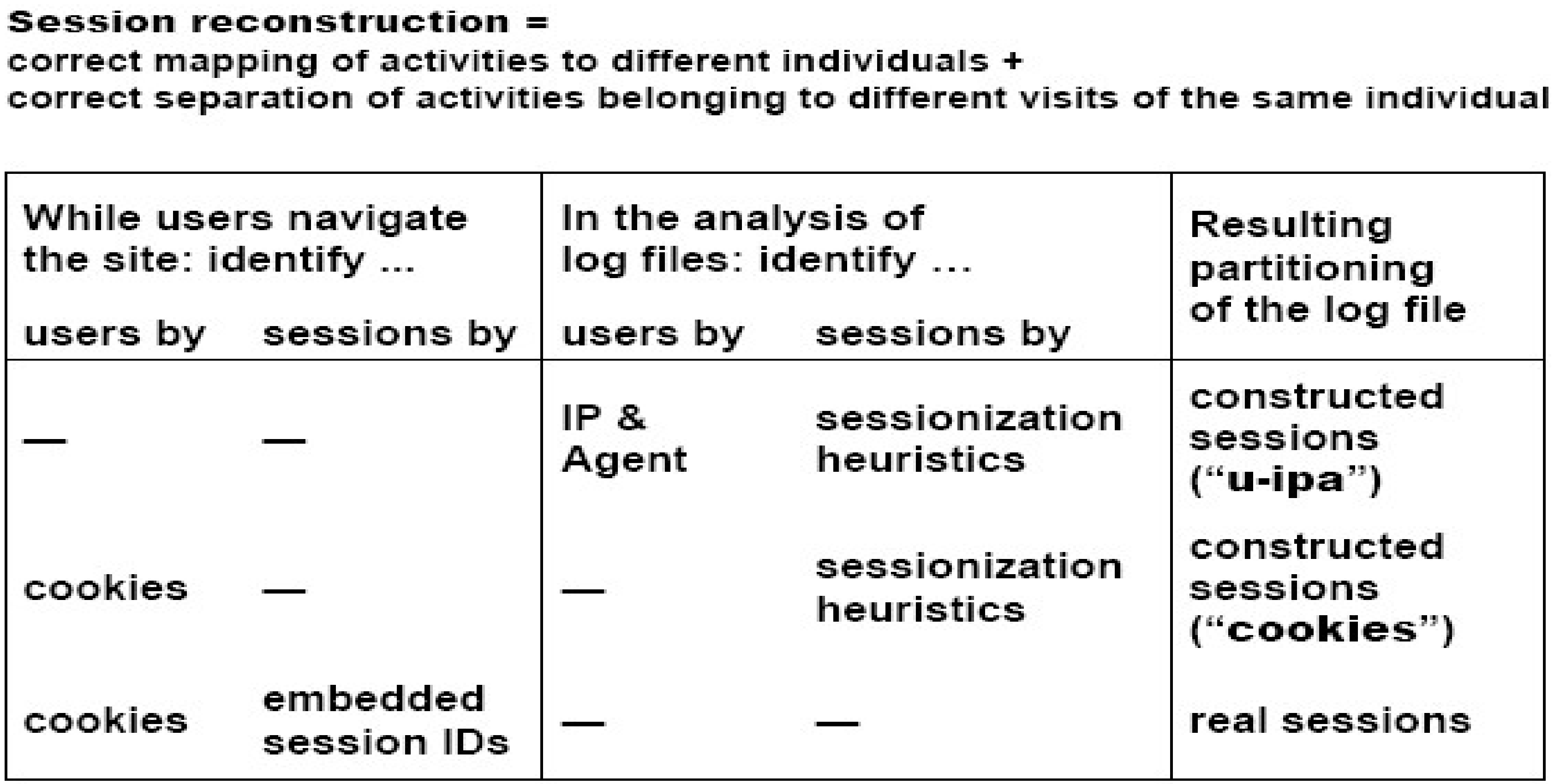
Time Oriented heuristic

- Uses Global or local time-out estimates to distinguish between sessions.

Structure Oriented heuristic

- Uses either static site structure or the implicit linkage structure captured in the referrer fields of server logs.

Sessionization strategies

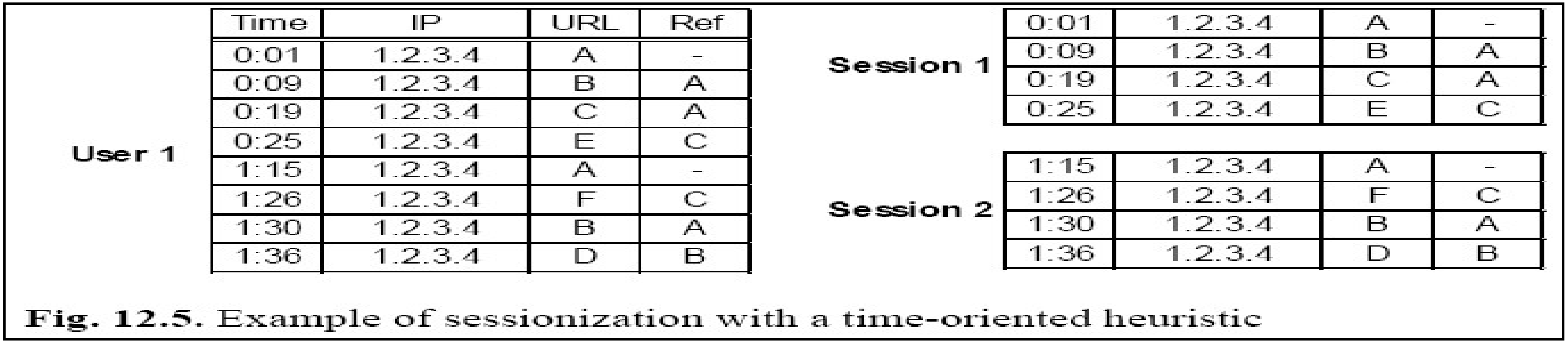


Session reconstruction =
correct mapping of activities to different individuals +
correct separation of activities belonging to different visits of the same individual

While users navigate the site: identify …		In the analysis of log files: identify …		Resulting partitioning of the log file
users by	sessions by	users by	sessions by	
—	—	IP & Agent	sessionization heuristics	constructed sessions (“u-ipa”)
cookies	—	—	sessionization heuristics	constructed sessions (“cookies”)
cookies	embedded session IDs	—	—	real sessions

θ =10 minutes

Sessionization example



User 1

Time	IP	URL	Ref
0:01	1.2.3.4	A	-
0:09	1.2.3.4	B	A
0:19	1.2.3.4	C	A
0:25	1.2.3.4	E	C
1:15	1.2.3.4	A	-
1:26	1.2.3.4	F	C
1:30	1.2.3.4	B	A
1:36	1.2.3.4	D	B

Session 1

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

Session 2

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

Fig. 12.5. Example of sessionization with a time-oriented heuristic

0:01	**1.2.3.4**	**A**	**-**
0:09	**1.2.3.4**	**B**	**A**
0:19	**1.2.3.4**	**C**	**A**
0:25	**1.2.3.4**	**E**	**C**
1:26	**1.2.3.4**	**F**	**C**

1:15	**1.2.3.4**	**A**	**-**
1:30	**1.2.3.4**	**B**	**A**
1:36	**1.2.3.4**	**D**	**B**

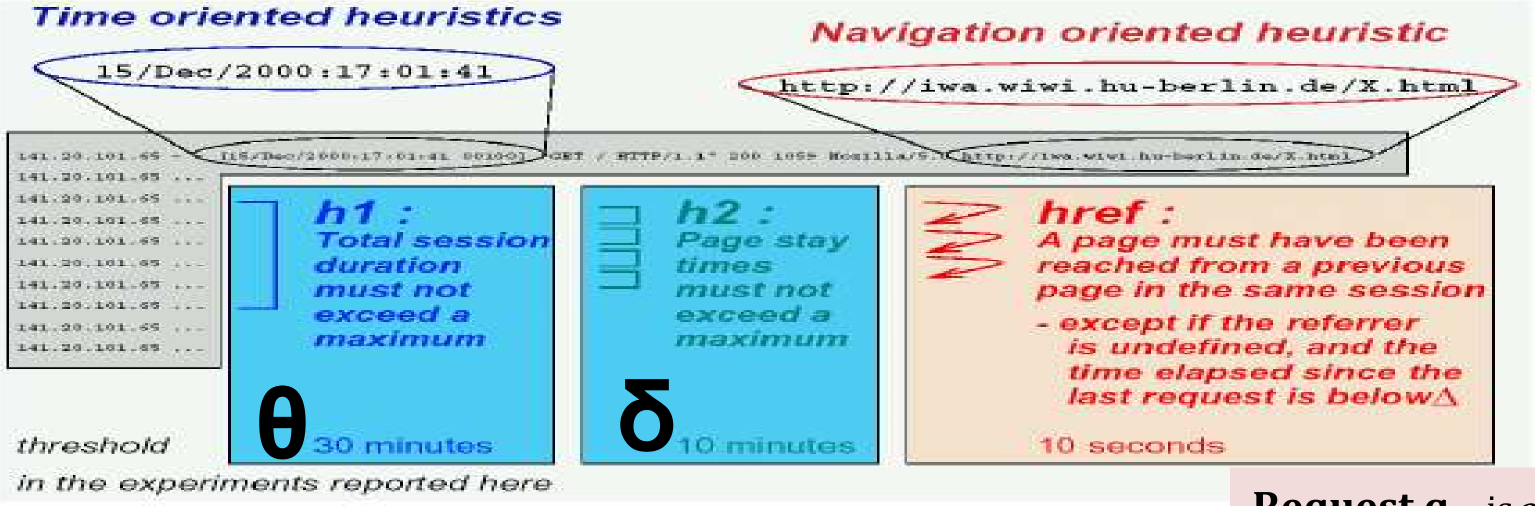
Sessionization with the h-ref heuristic- Once the request for F with timestamp 1:26 is reached, there are two open sessions. A->B->C->E and A

But F is added to the first because its referrer C was invoked in session1. The request for B (with time stamp 1:30) may potentially belong to both open sessions, since its referrer, A, is invoked both in session 1 and in session 2.

Here it is added in session2 as it is the most recently opened session.

Sessionization heuristics

Each heuristic h scans user activity logs to which the web server log is partitioned after user identification and outputs a set of constructed sessions.


Time oriented heuristics
15/Dec/2000:17:01:41
Navigation oriented heuristic
http://iwa.wiwi.hu-berlin.de/X.html
141.20.101.65 - [15/Dec/2000:17:01:41 00100] "GET / HTTP/1.1" 200 1055 Mozilla/5. "http://iwa.wiwi.hu-berlin.de/X.html"
141.20.101.65 ...
141.20.101.65 ...
141.20.101.65 ...
141.20.101.65 ...
141.20.101.65 ...
141.20.101.65 ...
141.20.101.65 ...
141.20.101.65 ...
141.20.101.65 ...
h1 :
Total session duration must not exceed a maximum
θ
30 minutes
h2 :
Page stay times must not exceed a maximum
δ
10 minutes
href :
A page must have been reached from a previous page in the same session
- except if the referrer is undefined, and the time elapsed since the last request is below∆
10 seconds
threshold
in the experiments reported here

t0 - timestamp for the first request in the constructed **session S**

t – timestamp for other request is added to **S**

Iff t – t0 <= θ

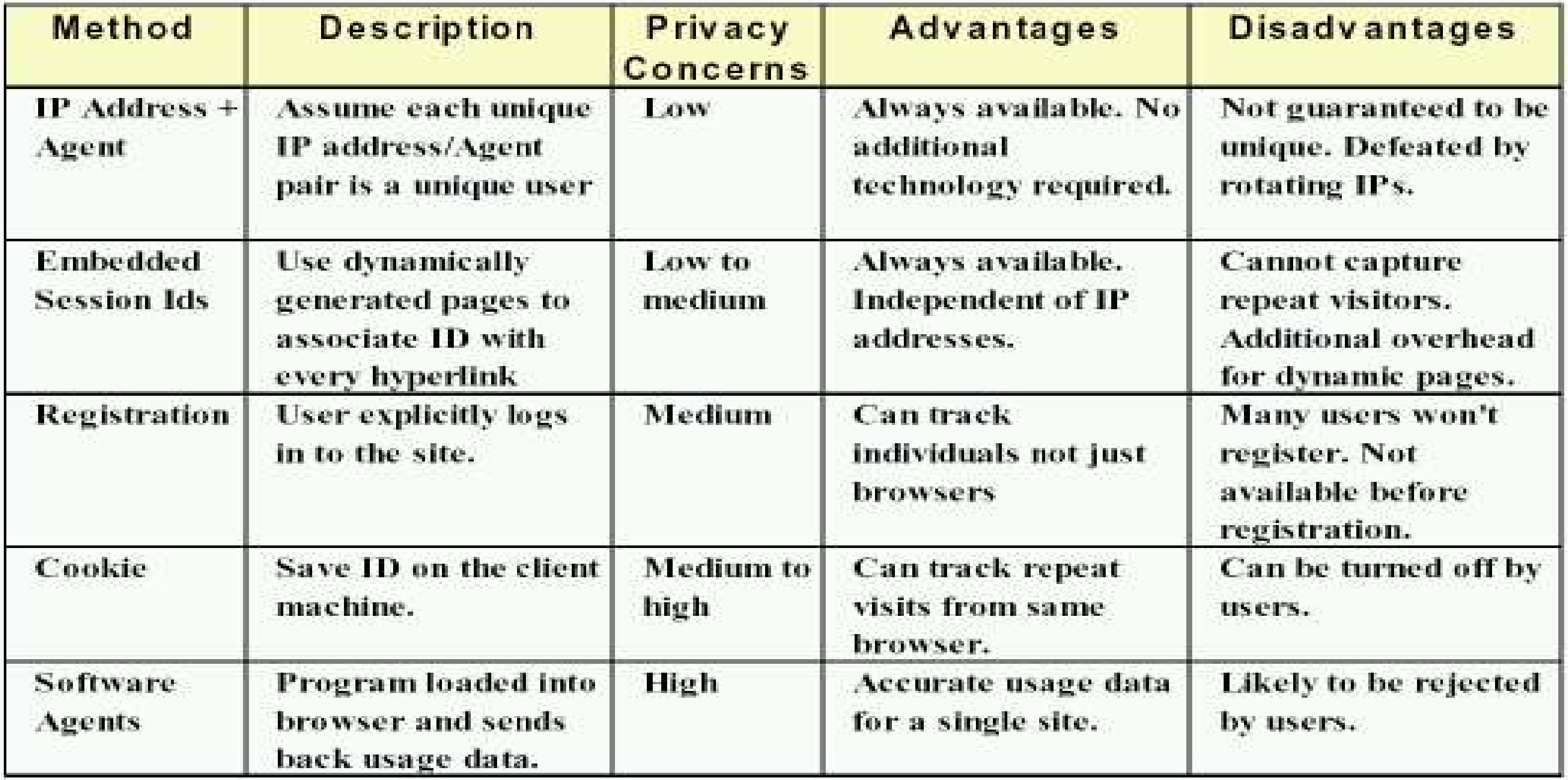
t1 - timestamp for the request assigned **session S**

t2 – timestamp for next request is added to **S**

Iff t2 – t1 <= δ

Request q - is added **session S**
if referrer for q was previously invoked in **S**
otherwise **q** is used as the start of new constructed session.
With this heuristics it is possible that a request q may potentially belong to more than one "**open**" constructed session, since q may have been accessed previously in multiple sessions. In this case additional information can be used for disambiguation. For example q could be added to the most recently opened session satisfying the above condition.

User identification



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.

User identification: an example

It is necessary **to distinguish** among different users.

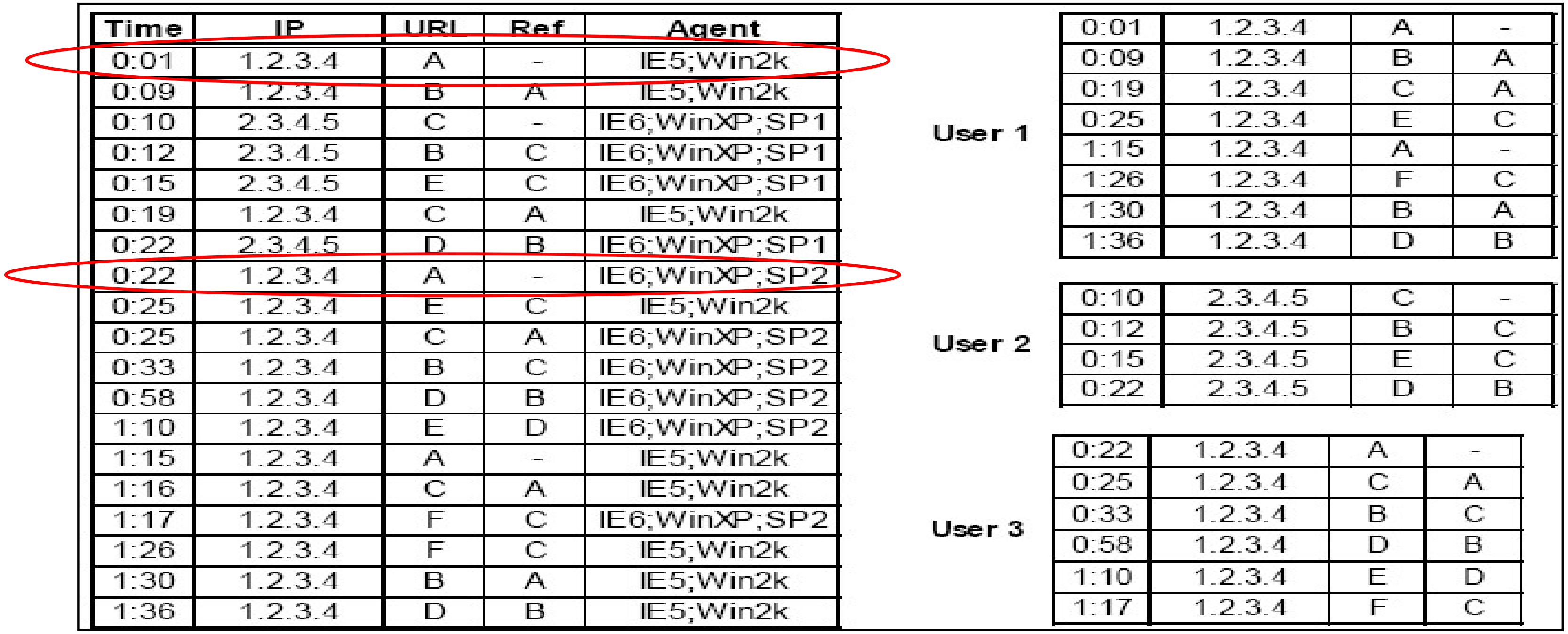
Since the user may visit the site more than once, the **server logs** record multiple sessions for each user.

In the absence of **user authentication** mechanisms – To distinguish users **client side cookies.**

If cookies are disabled due to privacy concerns, check for IP address.

IP address alone is not sufficient as ISP proxy servers assign rotating IP addresses to clients, we may find multiple log entries corresponding to a limited number of proxy server IP addresses from large ISP like America online. **So we may find two occurrences of the same IP Address corresponds to two different users.**

Distinguish Users – IP and Referrer and Time and UserAgent



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	-	IE6;WinXP;SP1
0:12	2.3.4.5	B	C	IE6;WinXP;SP1
0:15	2.3.4.5	E	C	IE6;WinXP;SP1
0:19	1.2.3.4	C	A	IE5;Win2k
0:22	2.3.4.5	D	B	IE6;WinXP;SP1
0:22	1.2.3.4	A	-	IE6;WinXP;SP2
0:25	1.2.3.4	E	C	IE5;Win2k
0:25	1.2.3.4	C	A	IE6;WinXP;SP2
0:33	1.2.3.4	B	C	IE6;WinXP;SP2
0:58	1.2.3.4	D	B	IE6;WinXP;SP2
1:10	1.2.3.4	E	D	IE6;WinXP;SP2
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	IE6;WinXP;SP2
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

User 1

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

User 2

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

User 3

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

Using the combination of IP and Agent fields

We are able to partition the log into activity records for three separate users

User Activity Record : Refer to sequence of logged activities belonging to same user

Pageview

- A pageview is an aggregate representation of a collection of Web objects contributing to the display on a user's browser resulting from a single user action (such as a click-through).
- Conceptually, each pageview can be viewed as a collection of Web objects or resources representing a specific "user event," e.g., reading an article, viewing a product page, or adding a product to the shopping cart.

Episode Identification

- Final step in pre-processing of clickstream data in order to focus on the relevant subsets of pageviews in each user session.
- An Episode is a subset of a session comprised of semantically related pageviews.This task may require the automatic or semi-automatic classification of pageviews into different functional types or into concept classes according to a domain ontology or concept hierarchy.
- In highly dynamic sites, it may also be necessary to map pageviews within each session into "service based" classes according to a concept hierarchy over the space of possible parameters passed to script or database queries.
- For example, the analysis may ignore the quantity and attributes of an item added to the shopping cart and focus only on the action of adding the item to the cart.

Path completion

- Client or proxy-side caching can often result in missing access references to those pages or objects that have been cached.

- For instance,
  - if a user returns to a page A during the same session, the second access to A will likely result in viewing the previously downloaded version of A that was cached on the client-side, and therefore, no request is made to the server.

  - This results in the second reference to A not being recorded on the server logs.

  - These references due to caching can be heuristically inferred through path completion which relies on the knowledge of site structure and referrer information from server logs.

  - In case of dynamically generated pages, form-based applications using HTTP-POST method result in all or part of the user input parameter not being appended to the URL accessed by the user (though in the latter case, it is possible to recapture the user input through packet sniffers which listen to all incoming and outgoing TCP/IP network traffic on the server side)

Missing references due to caching

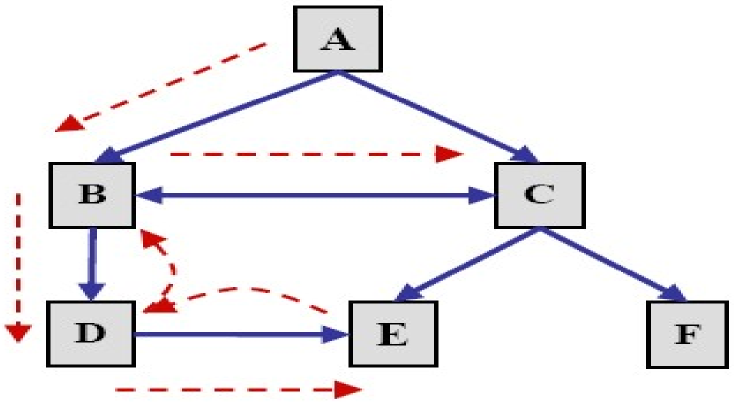
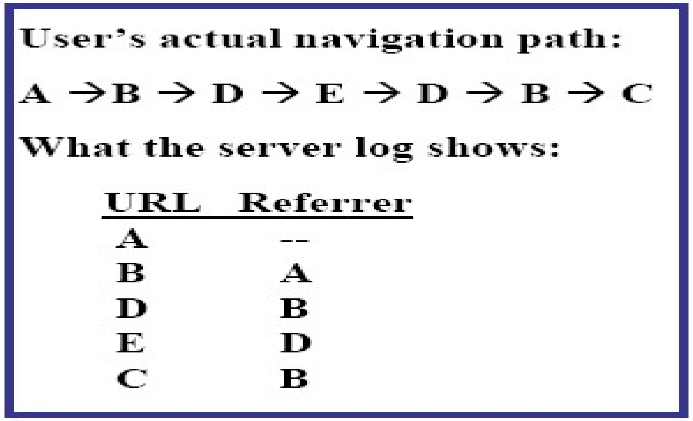


Fig. 12.7. Missing references due to caching.



User's actual navigation path:

A →B → D → E → D → B → C

What the server log shows:

URL	Referrer
A	--
B	A
D	B
E	D
C	B

Graph represents linkage structure of site.

The dotted arrows represent the navigational path followed by a hypothetical user.

After reaching E, user has backtracked using browser back button to page D and then B then navigated to page C. The back references do not appear in log file because these pages were cached on the client side.(thus no explicit server request was made for these pages.)

The log file shows that after a request for E, the next request by the user is for page C with a referrer B.

In other words, there is a gap in the activity record corresponding to user's navigation from page E and page B.

Given the site graph, it is possible to infer the two missing references.(i.e. E->D and D->B from the site structure and the referrer information given above.

It should be noted that there are in general, many (possibly infinite), candidate completions (for example E->D, D->B, B->A, A->B).

A simple heuristic that can be used for disambiguating among candidate paths is to select the one requiring the fewest number of back references.

Path completion

- The problem of inferring missing user references due to caching.
- Effective path completion requires extensive knowledge of the link structure within the site
- Referrer information in server logs can also be used in disambiguating the inferred paths.
- Problem gets much more complicated in frame-based sites.

Integrating with e-commerce events

- The above preprocessing tasks ultimately result in a set of user sessions , each corresponding to a delimited sequence of pageviews. However in order to provide the most effective framework for pattern discovery, data from a variety of other sources must be integrated with the preprocessed clickstream data.

- This is particularly the case in e-commerece applications where the integration of both user data (i.e. demographics, ratings and purchase histories) and product attributes and categories from operational databases is critical. Such data used in conjunction with usage data, in the mining process can allow for the discovery

Integrating with e-commerce events

- Either product oriented or visit oriented
- Used to track and analyze conversion of browsers to buyers.
  - Major difficulty for E-commerce events is defining and implementing the events for a site, however, in contrast to clickstream data, getting reliable preprocessed data is not a problem.
- Another major challenge is the successful integration with clickstream data

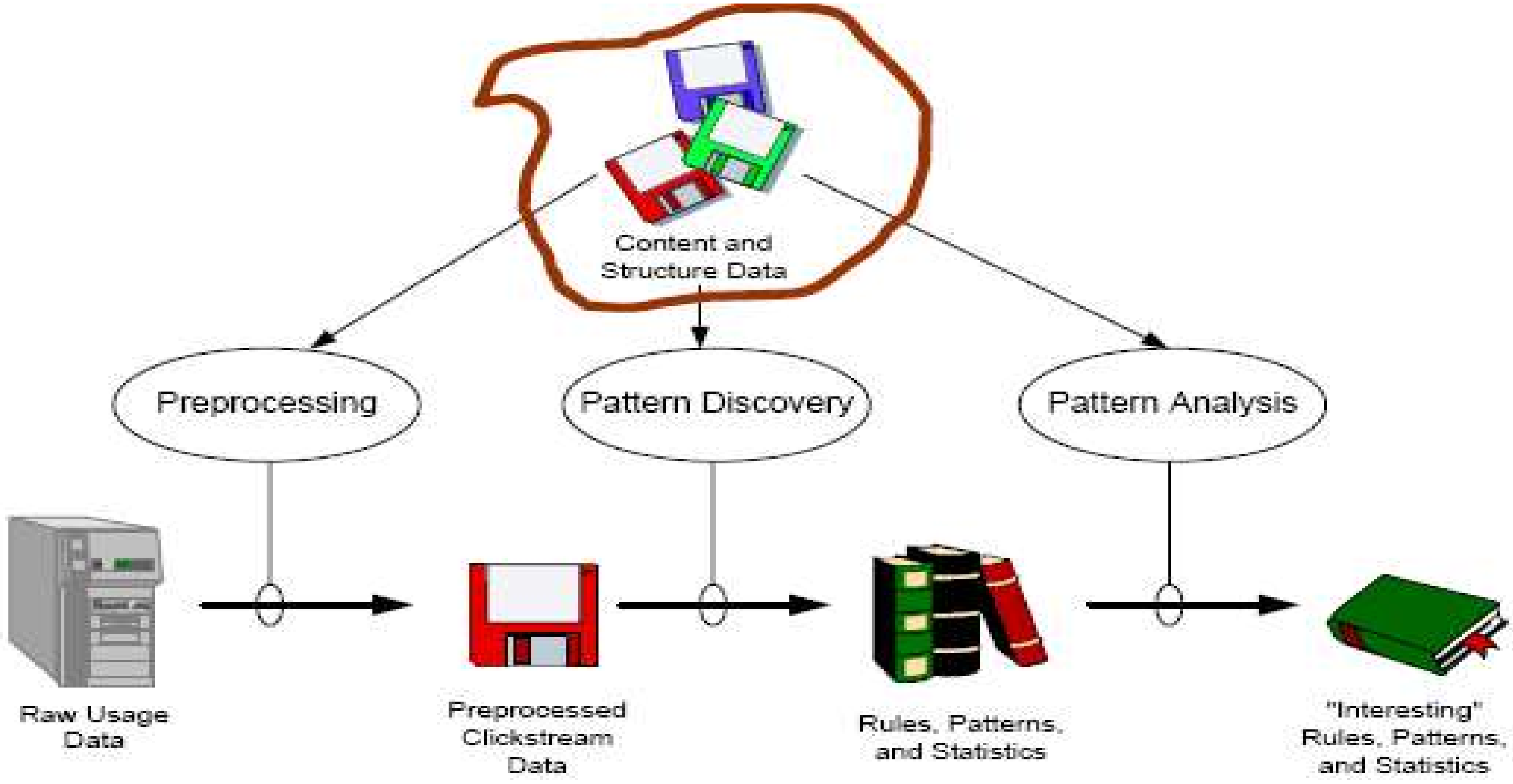
Product-Oriented Events

- Product View
  - Occurs every time a product is displayed on a page view
  - Typical Types: Image, Link, Text
- Product Click-through
  - Occurs every time a user “clicks” on a product to get more information

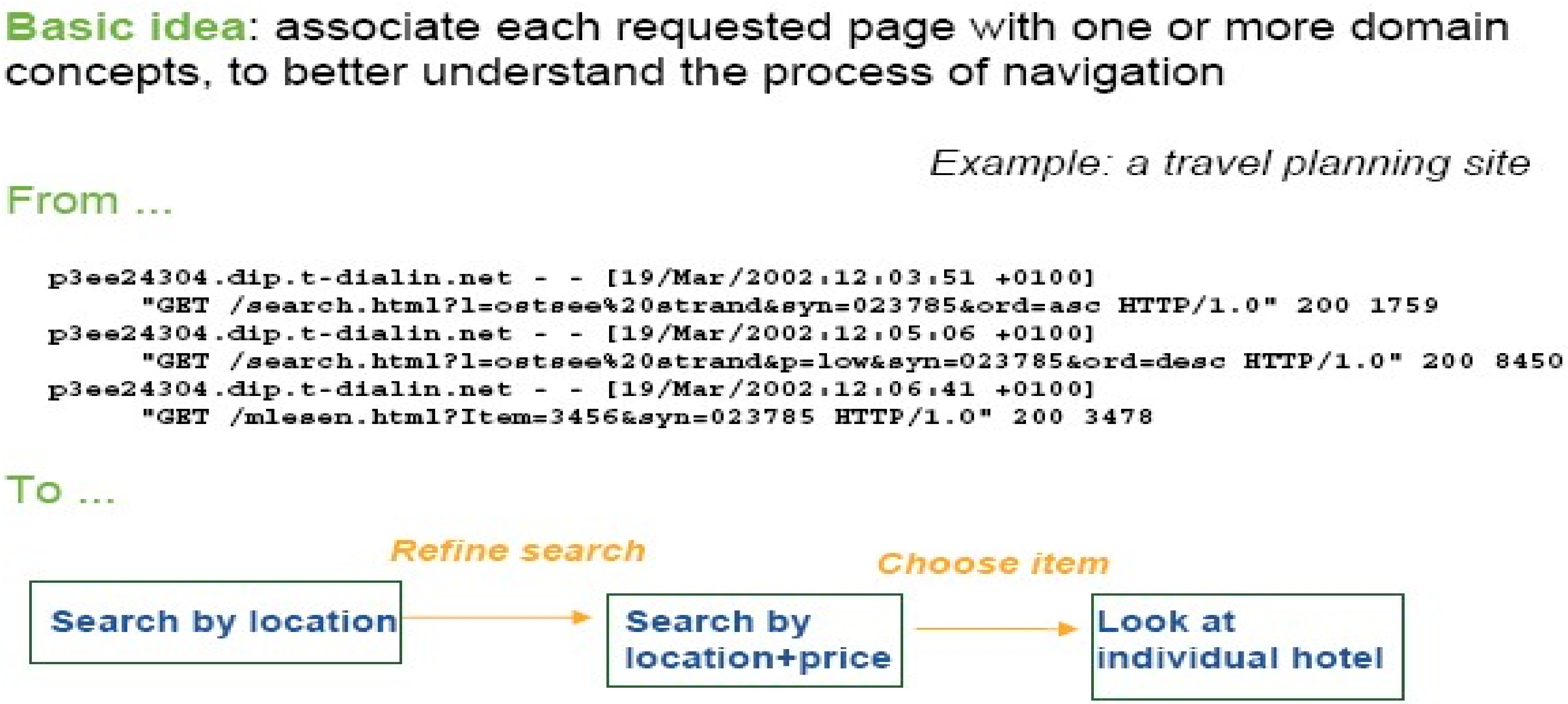
Product-Oriented Events

- Shopping Cart Changes
  - Shopping Cart Add or Remove
  - Shopping Cart Change - quantity or other feature (e.g. size) is changed
- Product Buy or Bid
  - Separate buy event occurs for each product in the shopping cart
  - Auction sites can track bid events in addition to the product purchases

Web usage mining process


Content and Structure Data
Preprocessing
Pattern Discovery
Pattern Analysis
Raw Usage Data
Preprocessed Clickstream Data
Rules, Patterns, and Statistics
"Interesting" Rules, Patterns, and Statistics

Integration with page content



Basic idea: associate each requested page with one or more domain concepts, to better understand the process of navigation

Example: a travel planning site

From ...

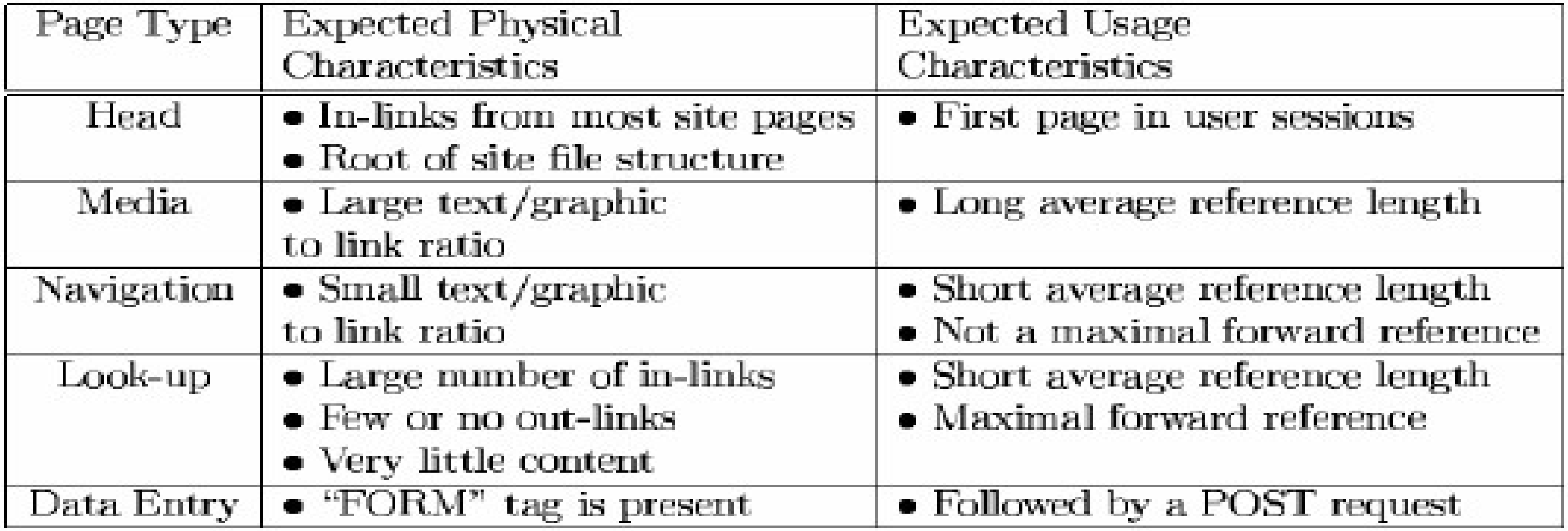
```
p3ee24304.dip.t-dialin.net - - [19/Mar/2002:12:03:51 +0100]
     "GET /search.html?l=ostsee%20strand&syn=023785&ord=asc HTTP/1.0" 200 1759
p3ee24304.dip.t-dialin.net - - [19/Mar/2002:12:05:06 +0100]
     "GET /search.html?l=ostsee%20strand&p=low&syn=023785&ord=desc HTTP/1.0" 200 8450
p3ee24304.dip.t-dialin.net - - [19/Mar/2002:12:06:41 +0100]
     "GET /mlesen.html?Item=3456&syn=023785 HTTP/1.0" 200 3478
```

To ...

Search by location → *Refine search* → Search by location+price → *Choose item* → Look at individual hotel

Integration with link structure

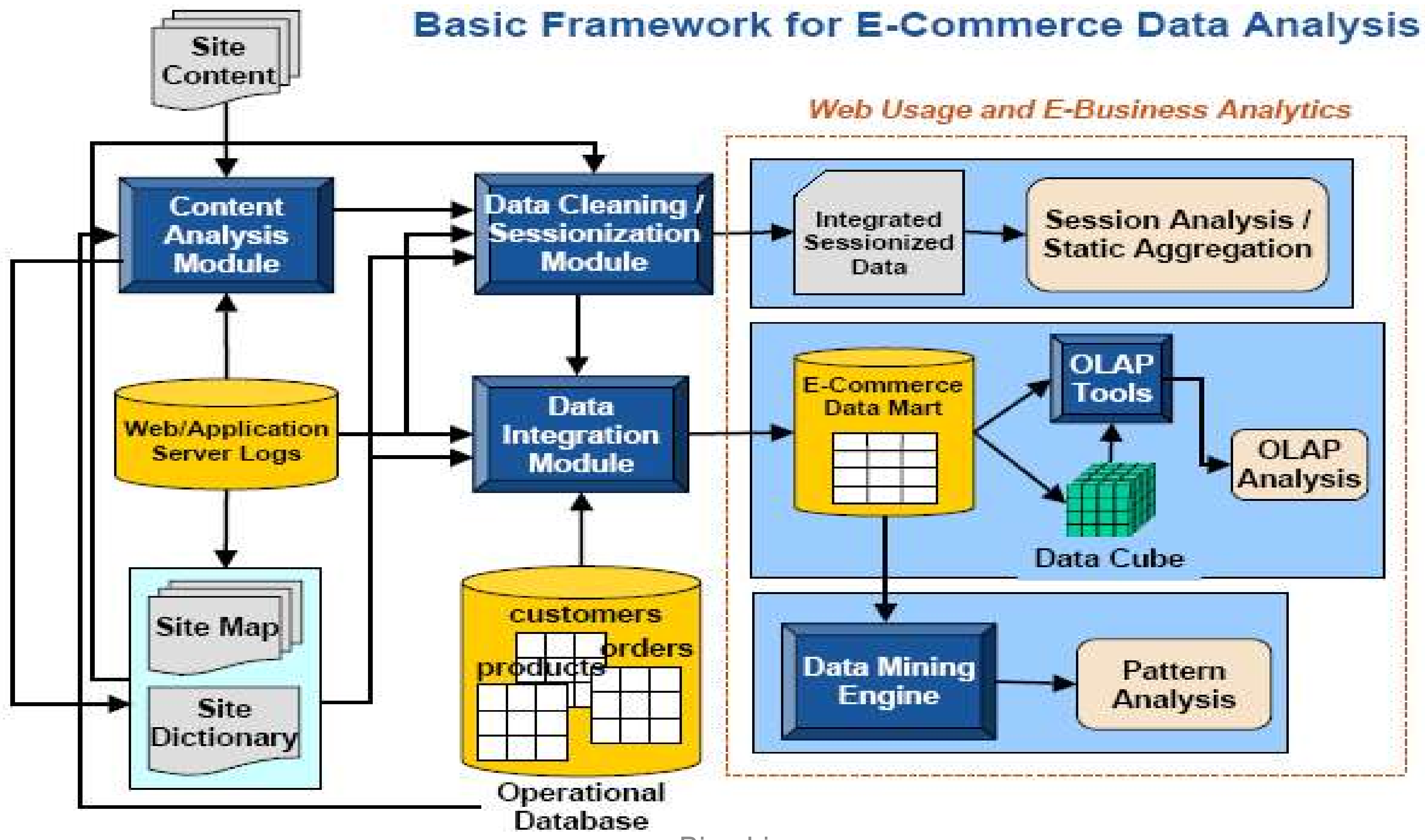
Page type defined by hyperlink structure bears information on function, or the designer's view of how pages will be used [from Cool00]:



Page Type	Expected Physical Characteristics	Expected Usage Characteristics
Head	• In-links from most site pages • Root of site file structure	• First page in user sessions
Media	• Large text/graphic to link ratio	• Long average reference length
Navigation	• Small text/graphic to link ratio	• Short average reference length • Not a maximal forward reference
Look-up	• Large number of in-links • Few or no out-links • Very little content	• Short average reference length • Maximal forward reference
Data Entry	• "FORM" tag is present	• Followed by a POST request

- can be assigned manually by the site designer,
- or automatically by using classification algorithms
- a classification tag can be added to each page (e.g., using XML tags).

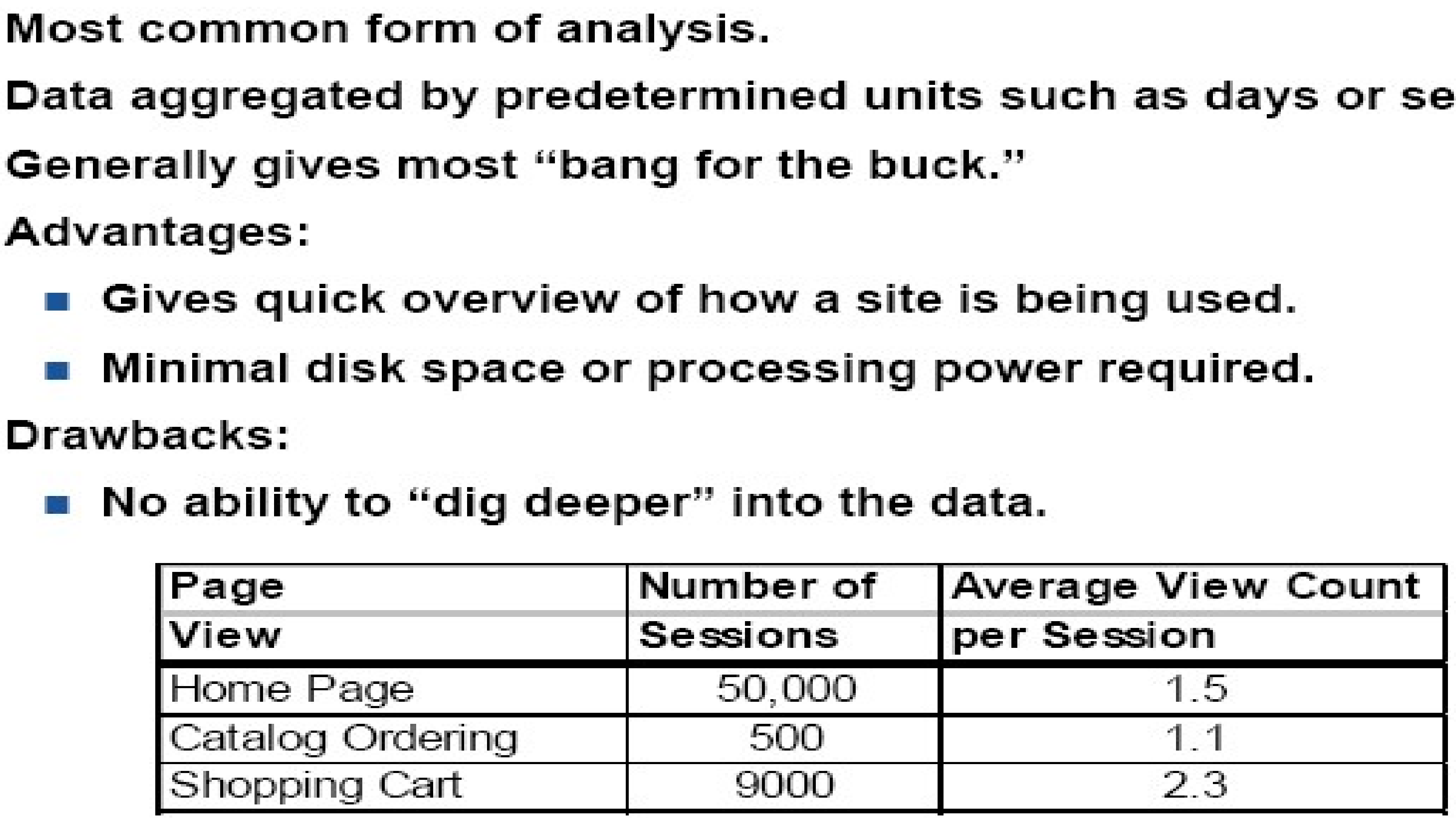
E-commerce data analysis



Session analysis

- Simplest form of analysis: examine individual or groups of server sessions and e-commerce data.
- Advantages:
  - Gain insight into typical customer behaviors.
  - Trace specific problems with the site.
- Drawbacks:
  - LOTS of data.
  - Difficult to generalize.

Session analysis: aggregate reports



Most common form of analysis.

Data aggregated by predetermined units such as days or sessions.

Generally gives most “bang for the buck.”

Advantages:

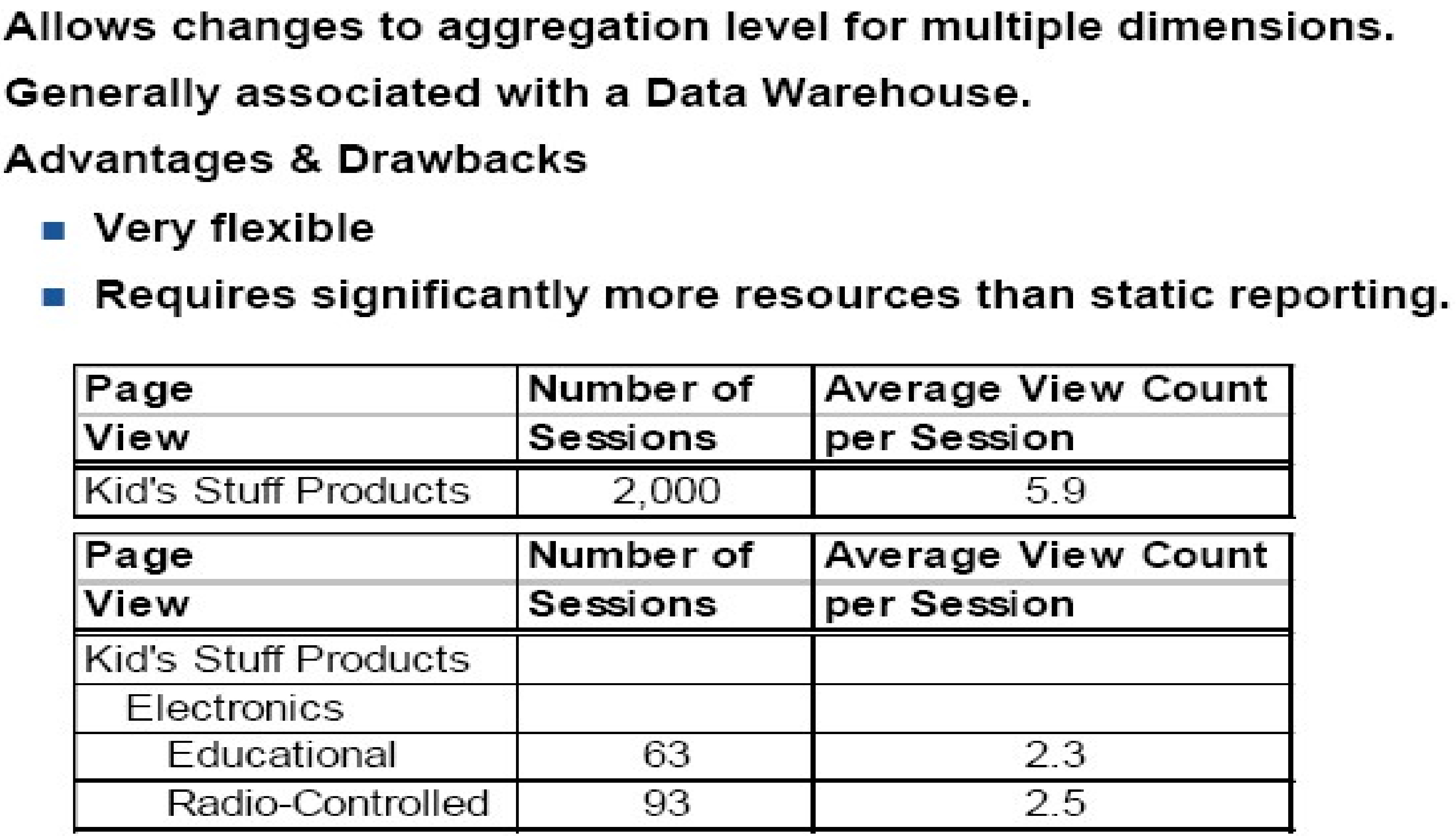
- Gives quick overview of how a site is being used.
- Minimal disk space or processing power required.

Drawbacks:

- No ability to “dig deeper” into the data.

Page View	Number of Sessions	Average View Count per Session
Home Page	50,000	1.5
Catalog Ordering	500	1.1
Shopping Cart	9000	2.3

OLAP



Allows changes to aggregation level for multiple dimensions.

Generally associated with a Data Warehouse.

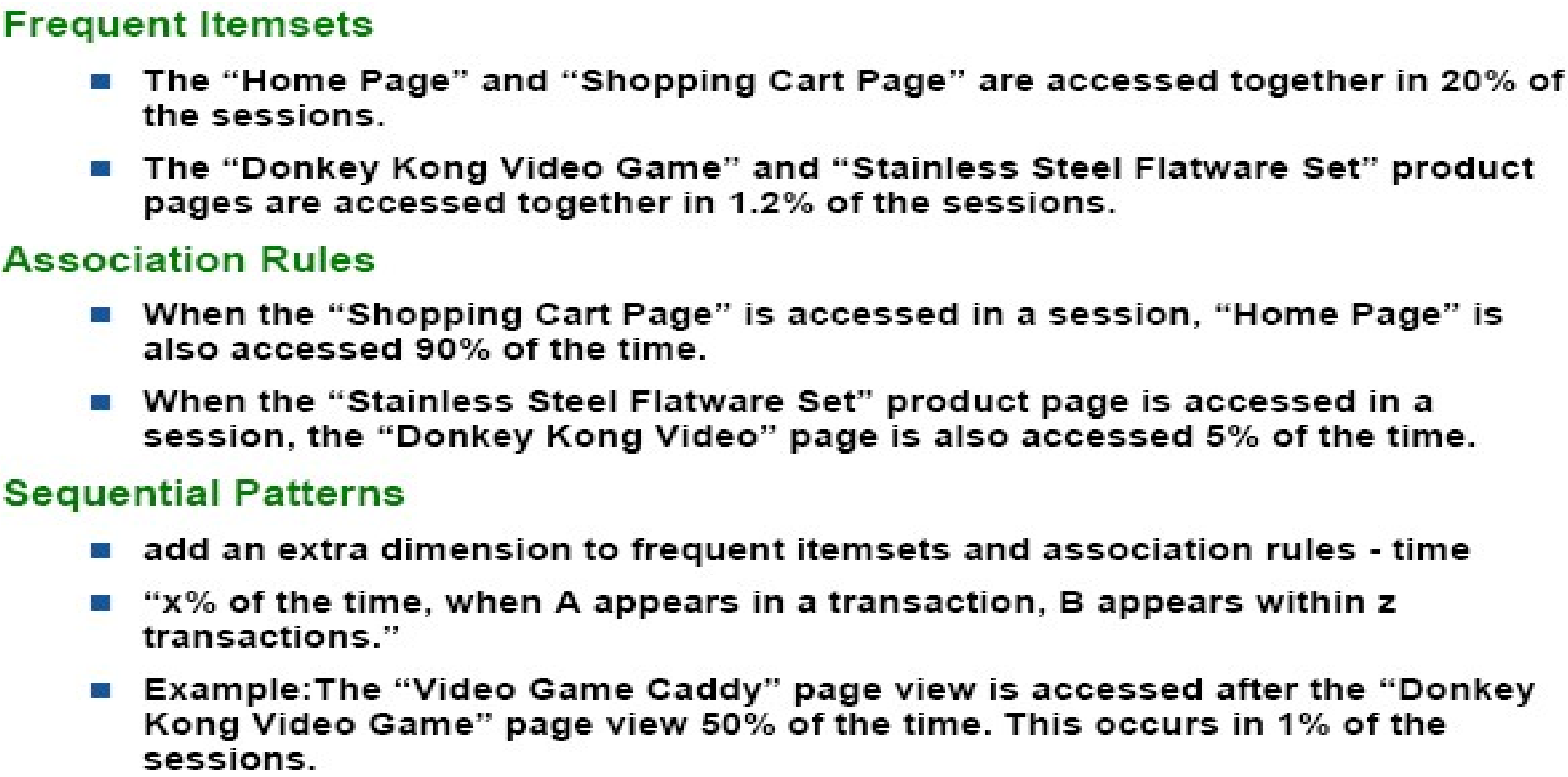
Advantages & Drawbacks

- Very flexible
- Requires significantly more resources than static reporting.

Page View	Number of Sessions	Average View Count per Session
Kid's Stuff Products	2,000	5.9

Page View	Number of Sessions	Average View Count per Session
Kid's Stuff Products		
Electronics		
Educational	63	2.3
Radio-Controlled	93	2.5

Data mining (cont.)



Frequent Itemsets

- The "Home Page" and "Shopping Cart Page" are accessed together in 20% of the sessions.
- The "Donkey Kong Video Game" and "Stainless Steel Flatware Set" product pages are accessed together in 1.2% of the sessions.

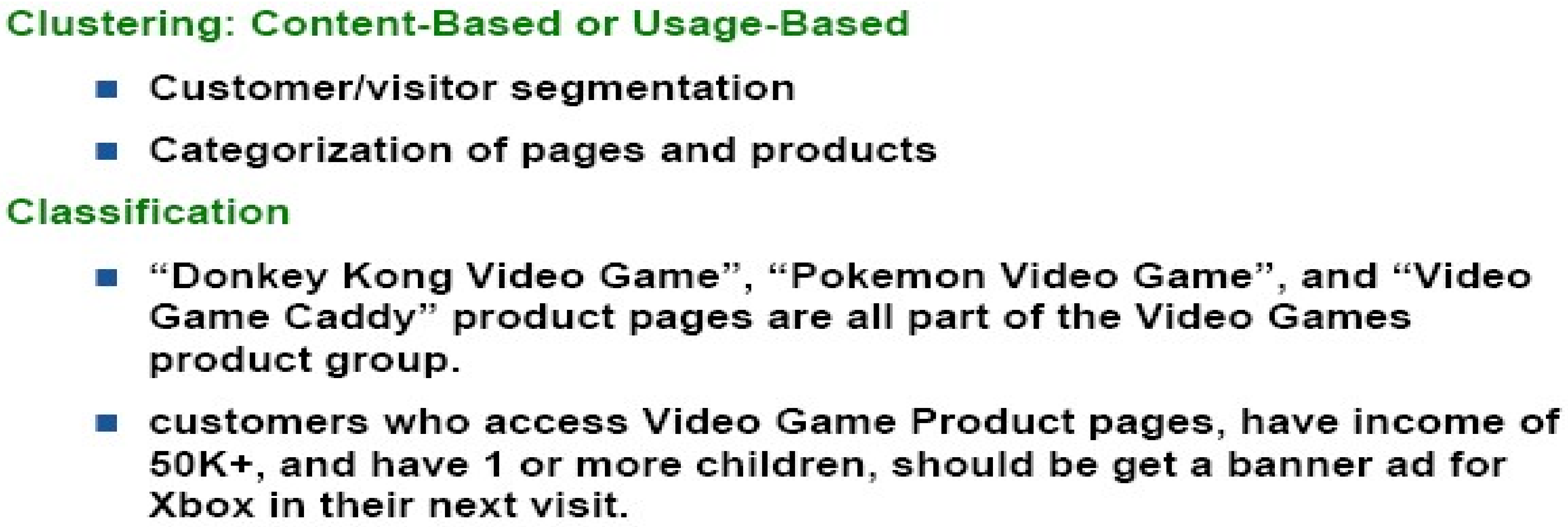
Association Rules

- When the "Shopping Cart Page" is accessed in a session, "Home Page" is also accessed 90% of the time.
- When the "Stainless Steel Flatware Set" product page is accessed in a session, the "Donkey Kong Video" page is also accessed 5% of the time.

Sequential Patterns

- add an extra dimension to frequent itemsets and association rules - time
- "x% of the time, when A appears in a transaction, B appears within z transactions."
- Example:The "Video Game Caddy" page view is accessed after the "Donkey Kong Video Game" page view 50% of the time. This occurs in 1% of the sessions.

Data mining (cont.)



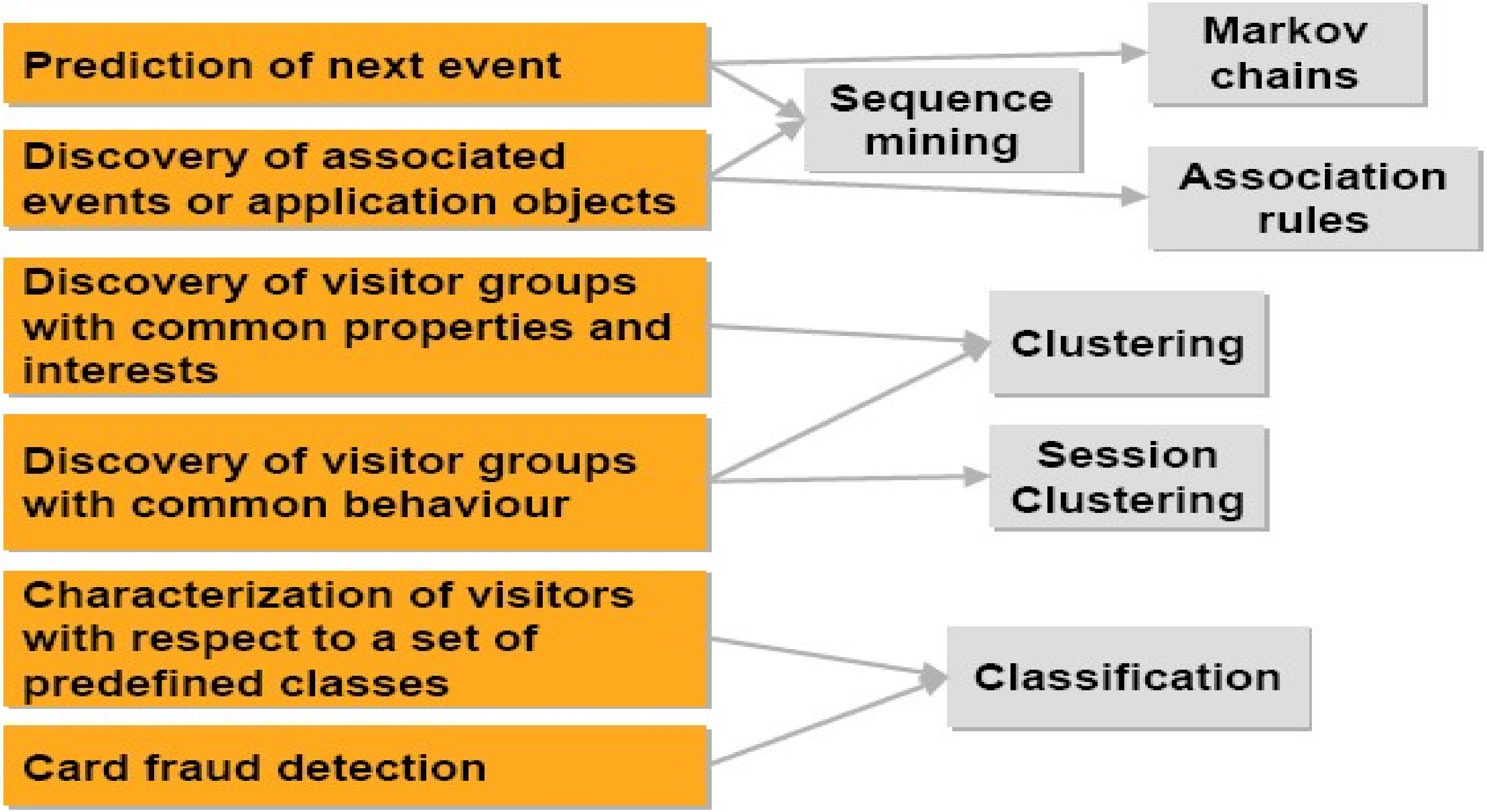
Clustering: Content-Based or Usage-Based

- Customer/visitor segmentation
- Categorization of pages and products

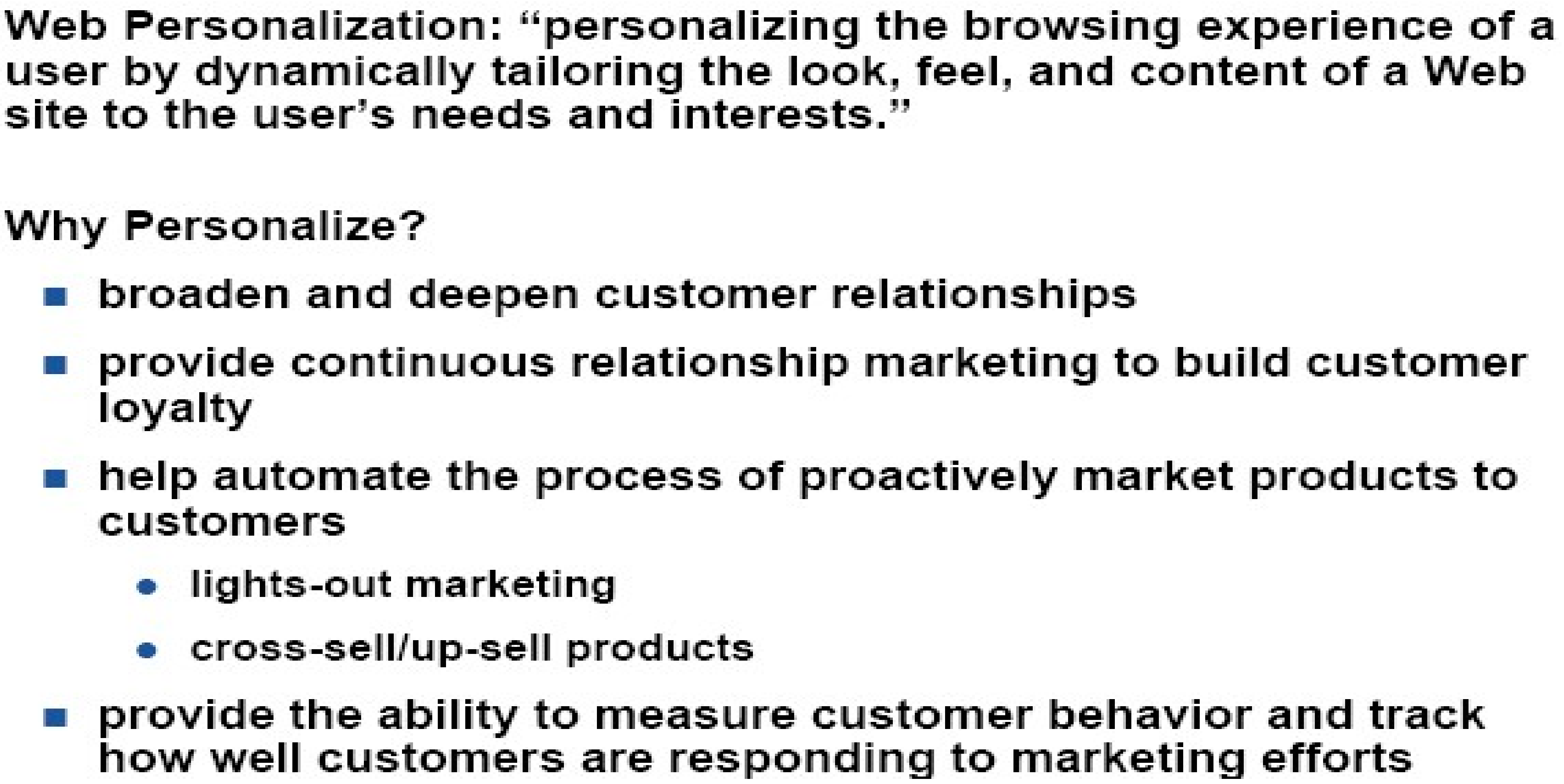
Classification

- “Donkey Kong Video Game”, “Pokemon Video Game”, and “Video Game Caddy” product pages are all part of the Video Games product group.
- customers who access Video Game Product pages, have income of 50K+, and have 1 or more children, should be get a banner ad for Xbox in their next visit.

Some usage mining applications


Prediction of next event
Discovery of associated events or application objects
Discovery of visitor groups with common properties and interests
Discovery of visitor groups with common behaviour
Characterization of visitors with respect to a set of predefined classes
Card fraud detection
Markov chains
Sequence mining
Association rules
Clustering
Session Clustering
Classification

Personalization application

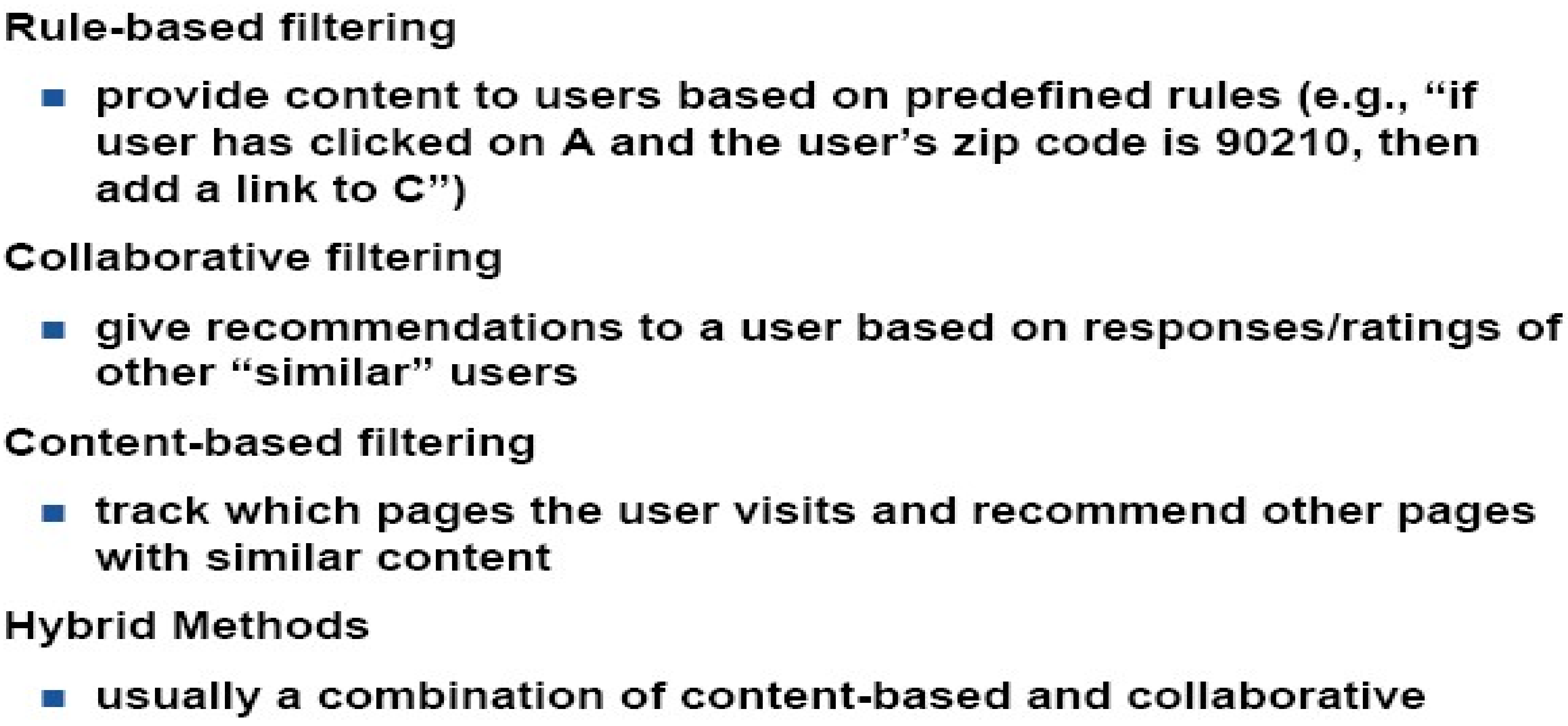


Web Personalization: “personalizing the browsing experience of a user by dynamically tailoring the look, feel, and content of a Web site to the user’s needs and interests.”

Why Personalize?

- broaden and deepen customer relationships
- provide continuous relationship marketing to build customer loyalty
- help automate the process of proactively market products to customers
  - lights-out marketing
  - cross-sell/up-sell products
- provide the ability to measure customer behavior and track how well customers are responding to marketing efforts

Standard approaches



Rule-based filtering

- provide content to users based on predefined rules (e.g., “if user has clicked on A and the user’s zip code is 90210, then add a link to C”)

Collaborative filtering

- give recommendations to a user based on responses/ratings of other “similar” users

Content-based filtering

- track which pages the user visits and recommend other pages with similar content

Hybrid Methods

- usually a combination of content-based and collaborative

Summary

- Web usage mining has emerged as the essential tool for realizing more personalized, user-friendly and business-optimal Web services.
- The key is to use the user-clickstream data for many mining purposes.
- Traditionally, Web usage mining is used by e-commerce sites to organize their sites and to increase profits.
- It is now also used by search engines to improve search quality and to evaluate search results, etc, and by many other applications.