Chapter 12: Web Usage Mining

- An introduction

Introduction

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

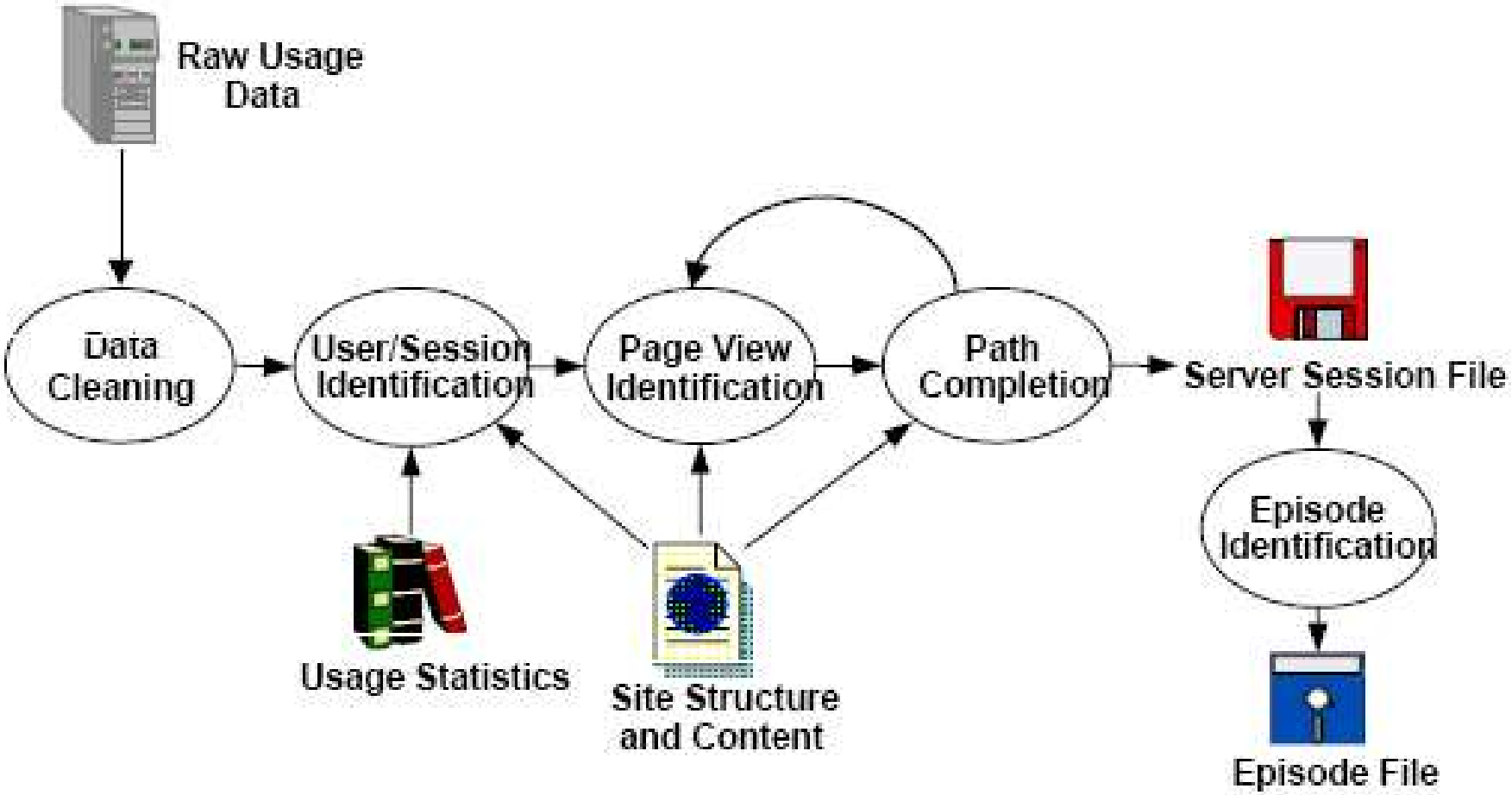
Introduction

- 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

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

Pre-processing of web usage data



Data cleaning

- Data cleaning
  - remove irrelevant references and fields in server logs
  - remove references due to spider navigation
  - remove erroneous references
  - add missing references due to caching (done after sessionization)

Identify sessions (sessionization)

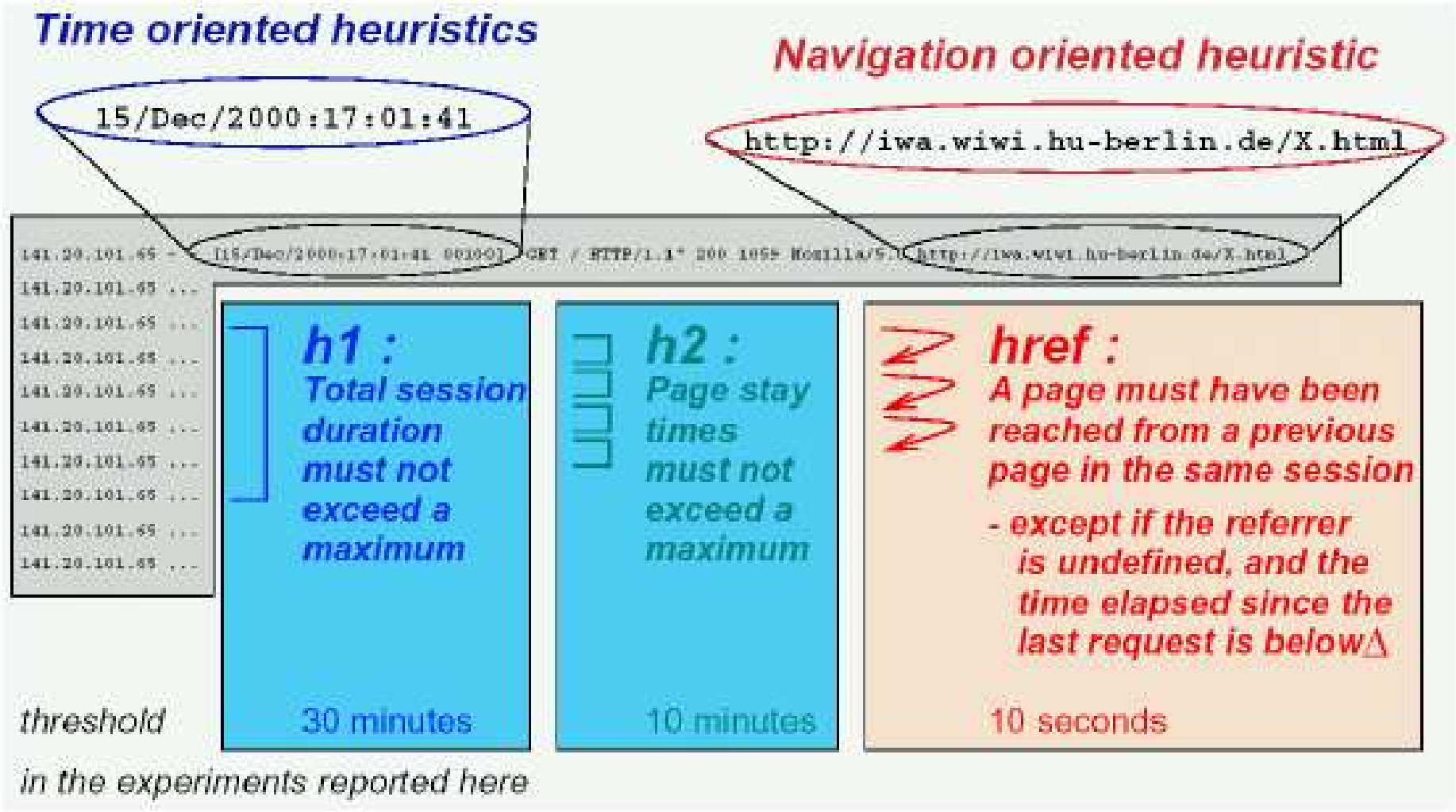
- In Web usage analysis, these data are the sessions of the site visitors: the activities performed by a user from the moment she enters the site until the moment she leaves it.
- Difficult to obtain reliable usage data due to proxy servers and anonymizers, dynamic IP addresses, missing references due to caching, and the inability of servers to distinguish among different visits.

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

Sessionization heuristics



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

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

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

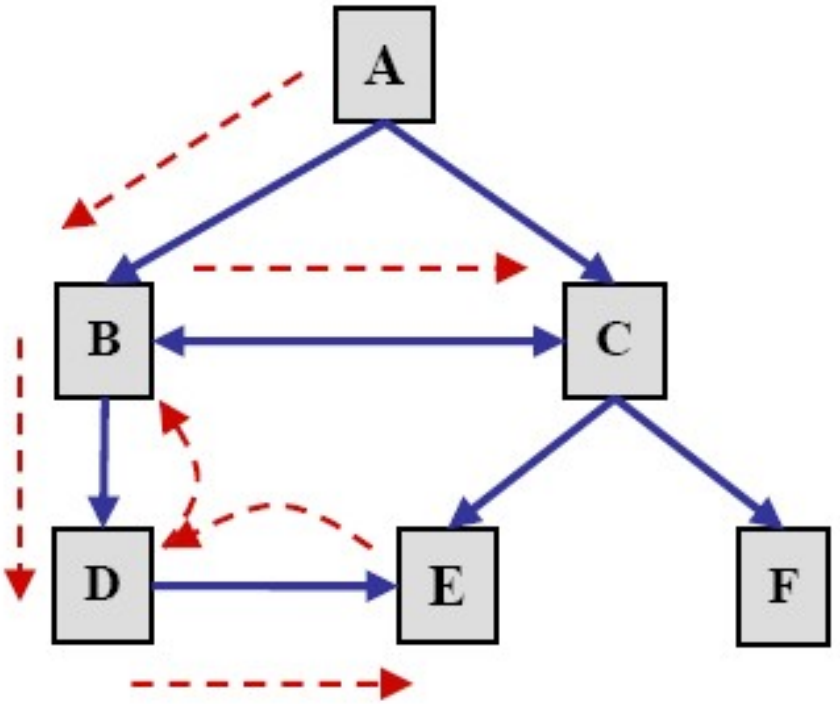
Pageview

- 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.

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.

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

Fig. 12.7. Missing references due to caching.

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

- 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

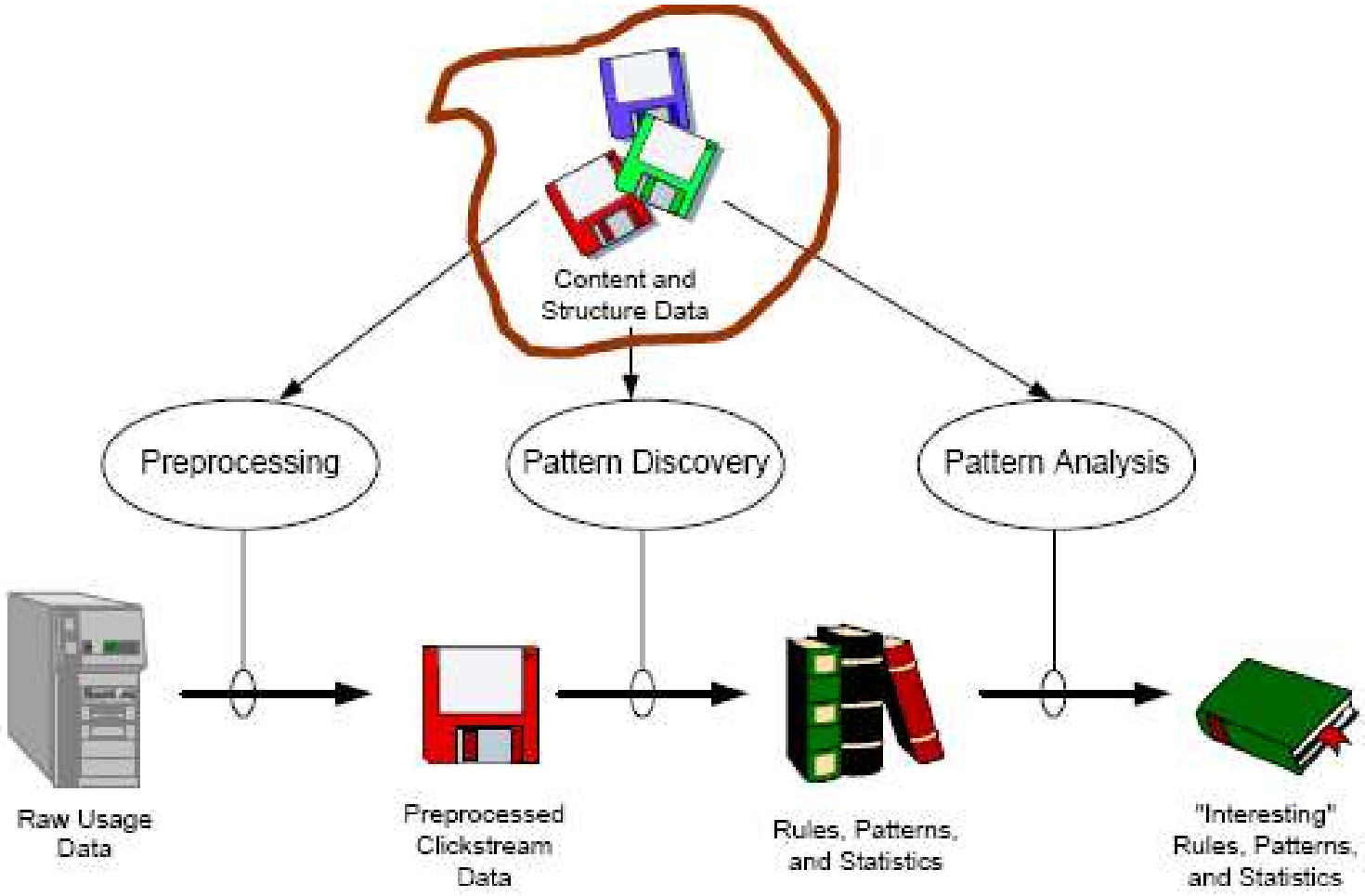
Product-Oriented Events

- 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



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 /mlessen.html?Item=3456&syn=023785 HTTP/1.0" 200 3478
```

To ...

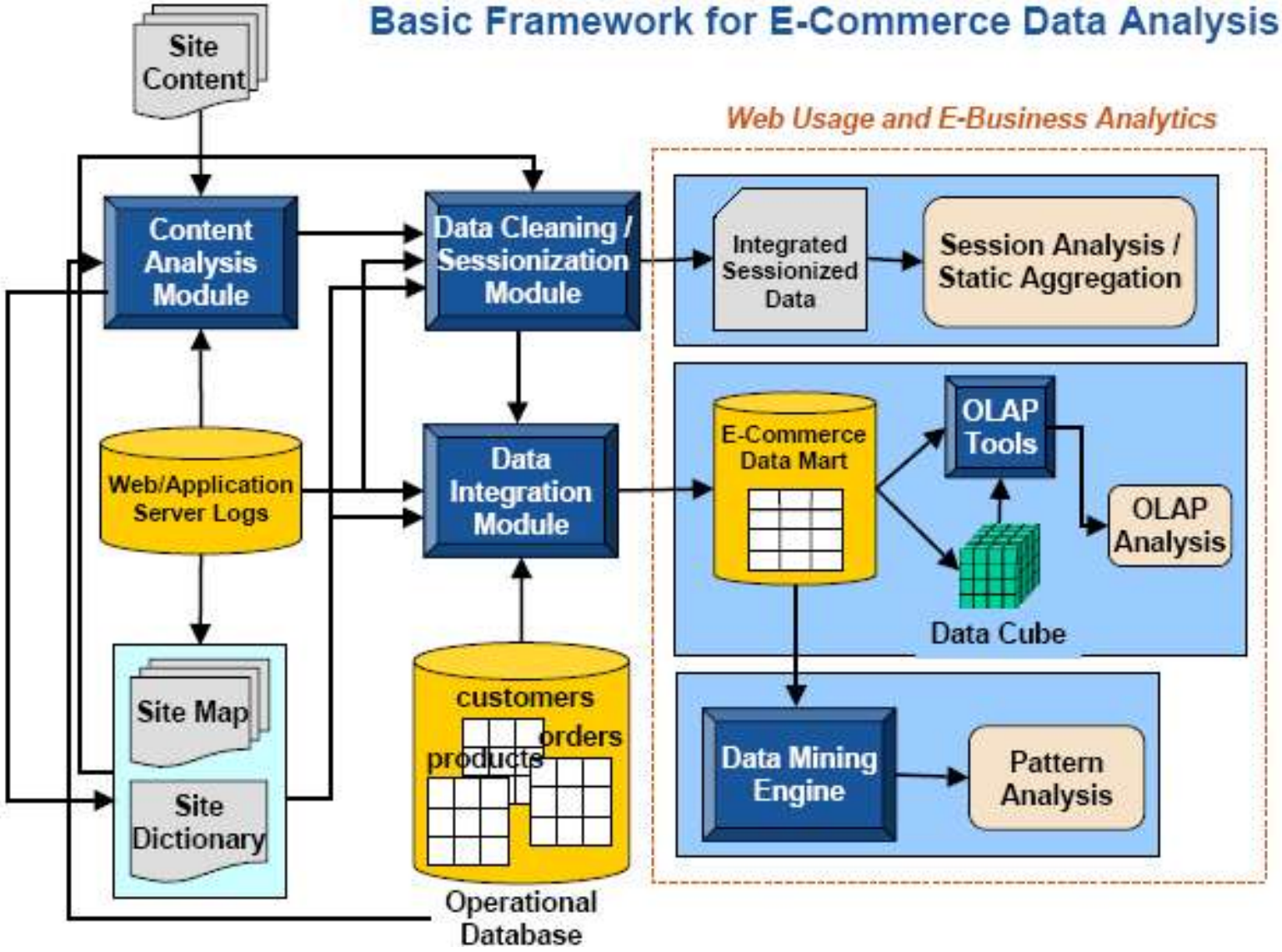
Integration with link structure

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

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.)

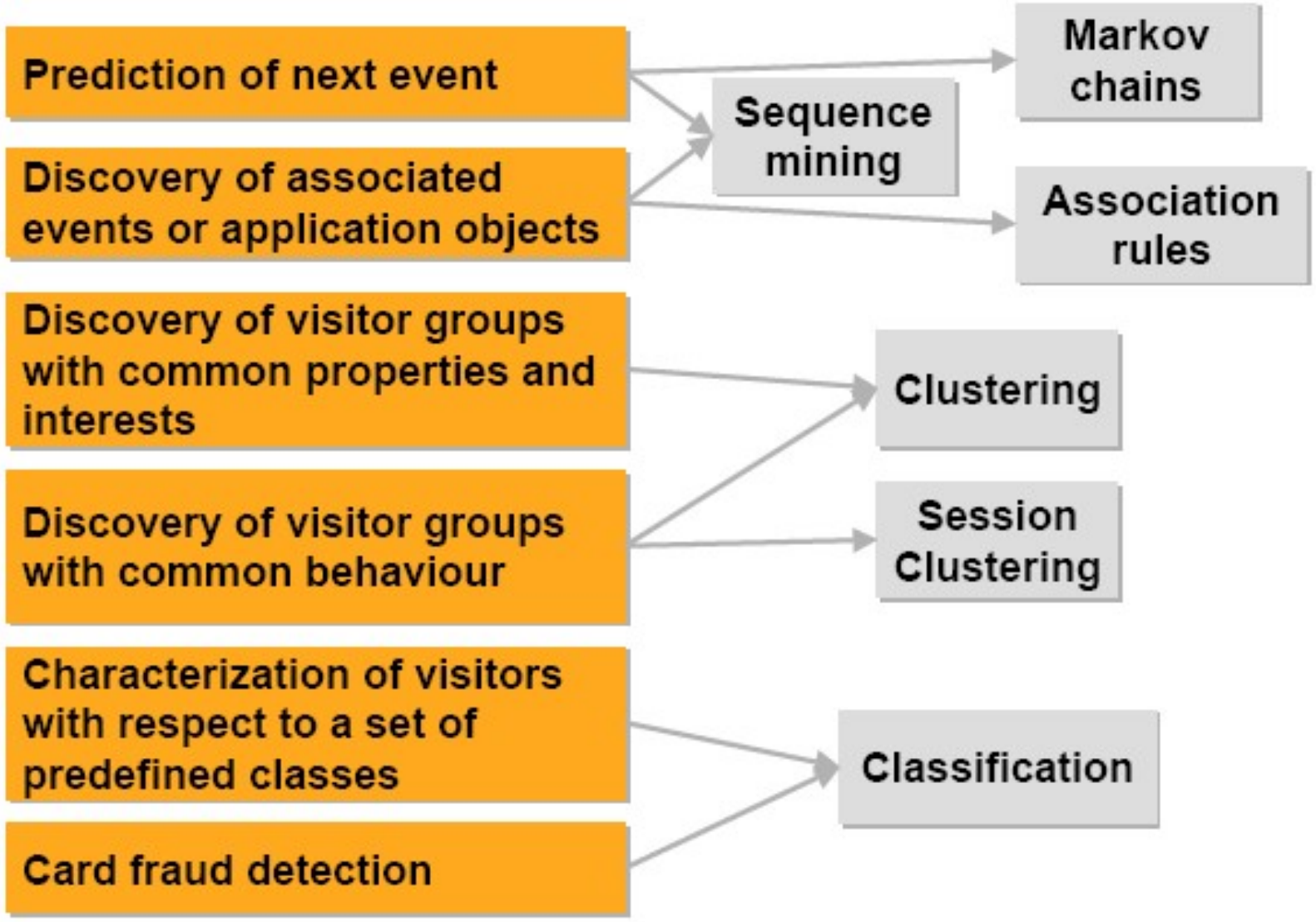
Clustering: Content-Based or Usage-Based

- 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



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.