Web Uses mining

Web Mining- Introduction

- **Web mining -** is the application of data mining techniques to discover patterns from the Web
- web mining can be divided into three different types
  - **Web usage mining**
  - **Web content mining** and
  - **Web structure mining**

Web usage mining

- Process of extracting useful information from server logs
- e.g. use Web usage mining is the process of finding out what users are looking for on the Internet
- **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
- Some users might be looking at
  - only textual data,
  - whereas some others might be interested in multimedia data

Contd…

- Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data
- Understands and better serve the needs of Web-based applications
- Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site

Contd...

- Web usage mining itself can be classified further depending on the kind of usage data considered:
  1. Web Server Data:
     - The user logs are collected by the Web server
     - Typical data includes IP address, page reference and access time
  2. Application Server Data:
     - Commercial application servers have significant features to enable e-commerce applications to be built on top of them with little effort
     - A key feature is the ability to track various kinds of business events and log them in application server logs
  3. Application Level Data:
     - New kinds of events can be defined in an application, and logging can be turned on for them thus generating histories of these specially defined events
     - It must be noted, however, that many end applications require a combination of one or more of the techniques applied in the categories above

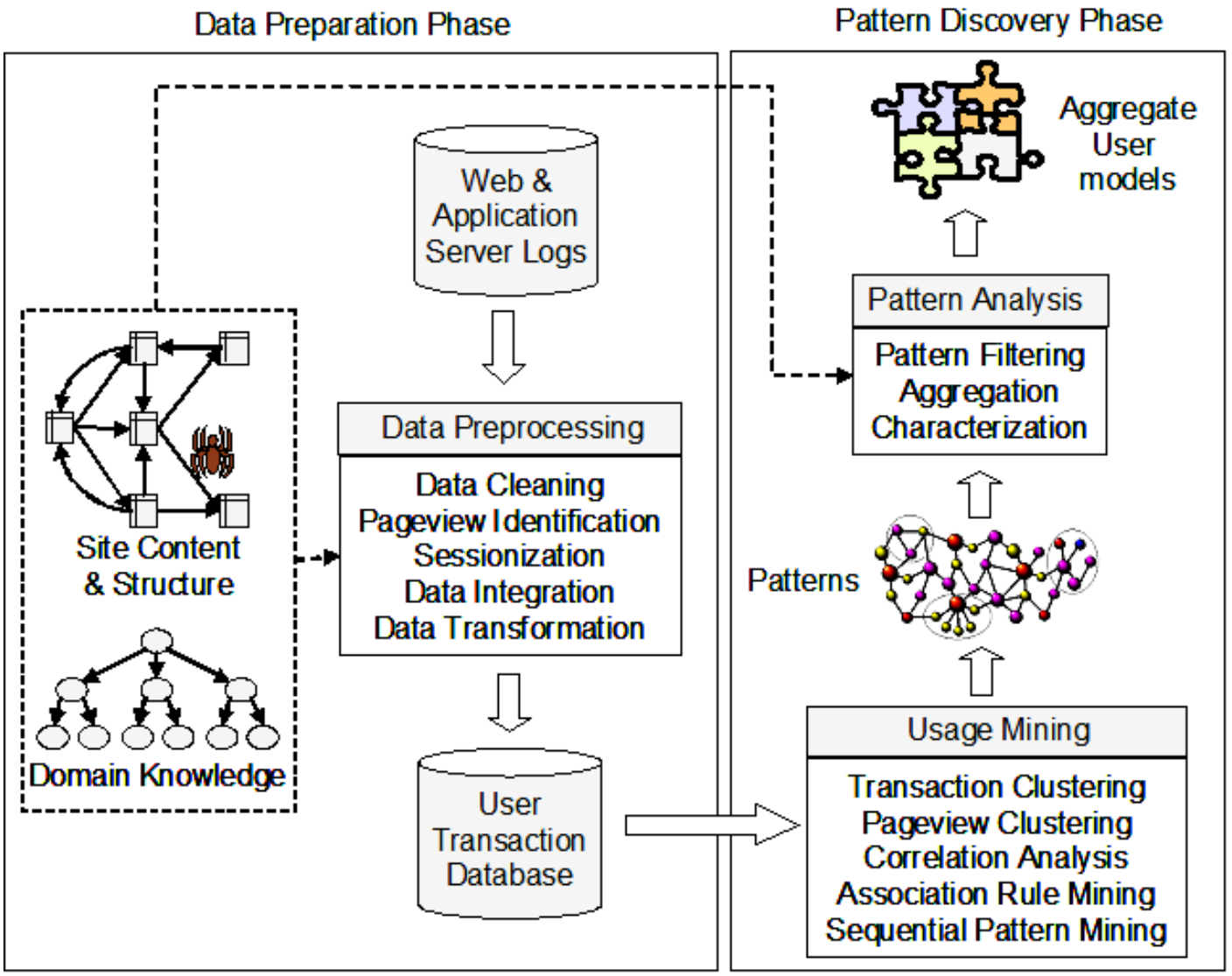
Why Web Usage Mining?

- Explosive growth of E-commerce
  - Provides an cost-efficient way doing business
  - Amazon.com: “online Wal-Mart”
- Hidden Useful information
  - Visitors’ profiles can be discovered
  - Measuring online marketing efforts, launching marketing campaigns, etc.

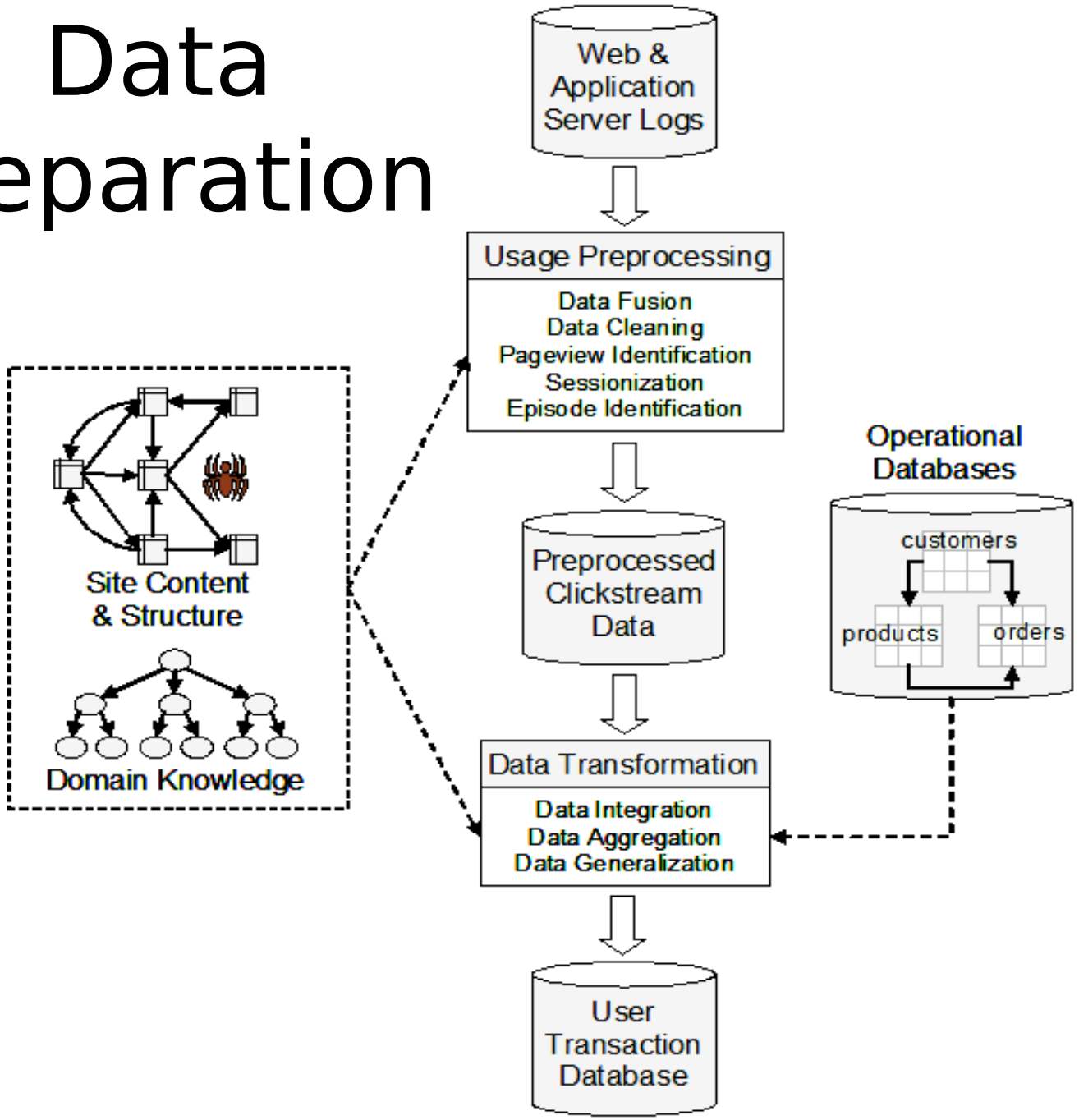
How to perform Web Usage Mining

- Obtain web traffic data from
  - Web server log files
  - Corporate relational databases
  - Registration forms
- Apply data mining techniques and other Web mining techniques
- Two categories:
  - Pattern Discovery Tools
  - Pattern Analysis Tools

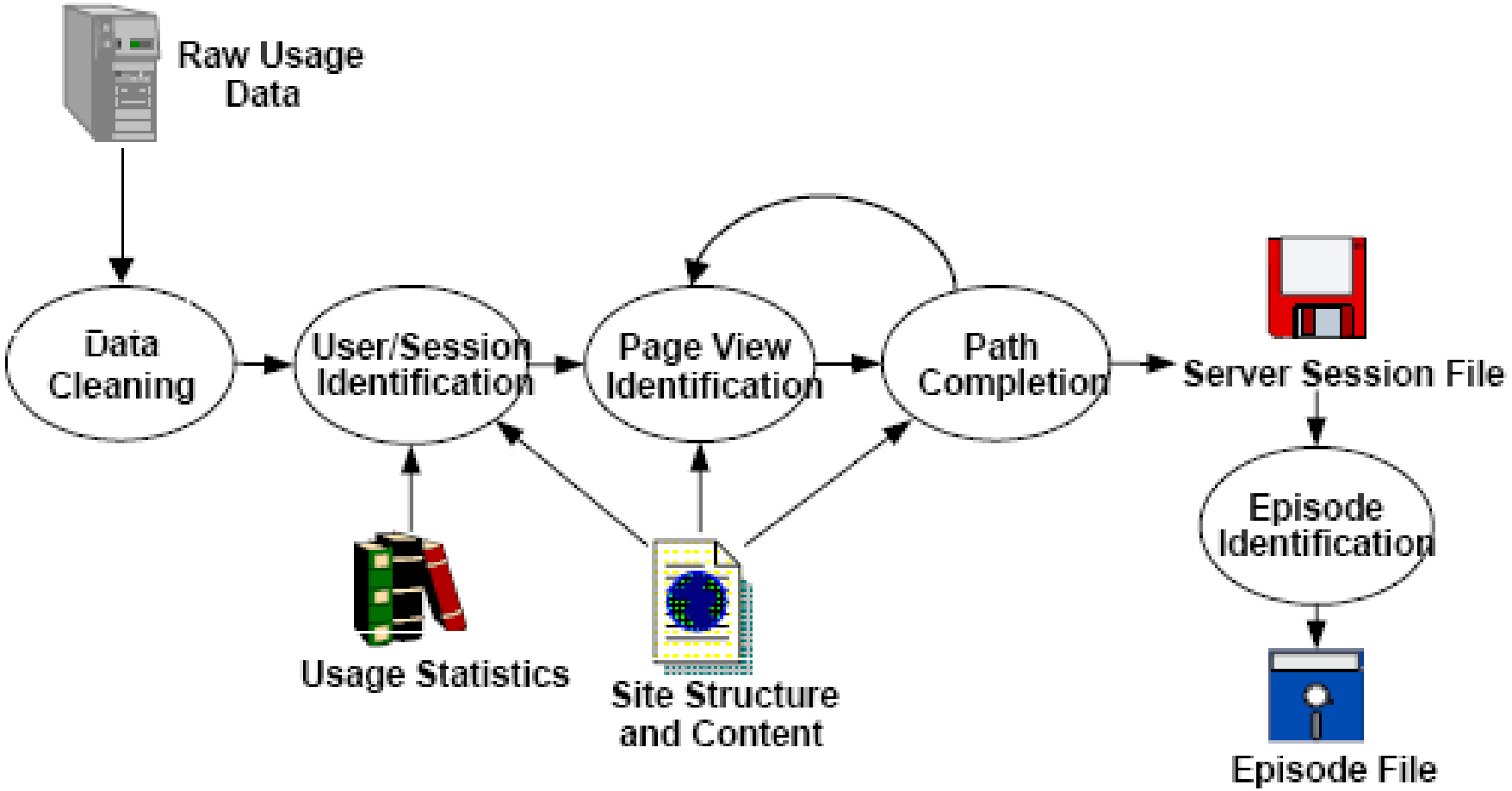
Figure : Web usage mining process



Data preparation



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 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

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

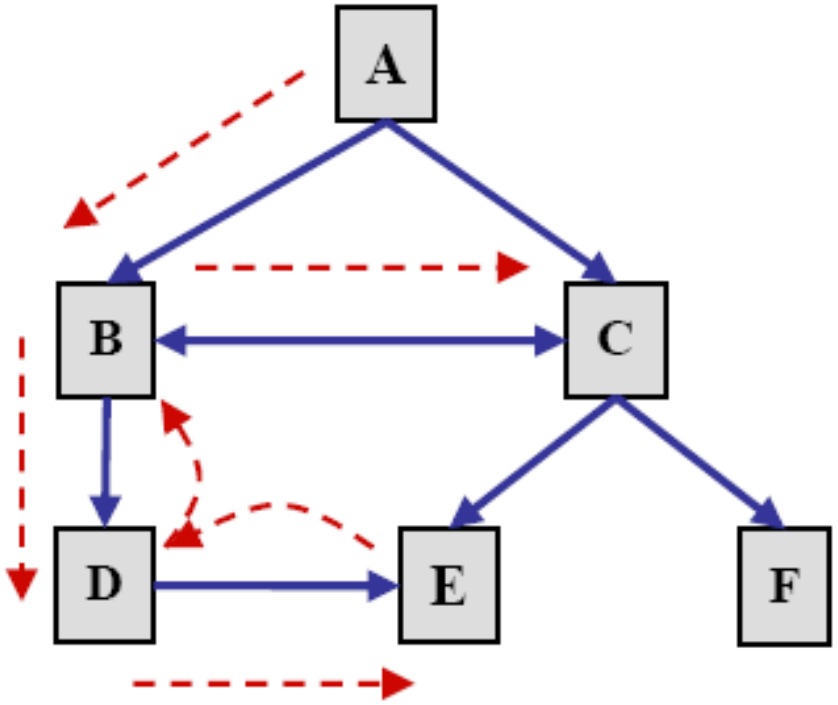
Pageview

- 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.

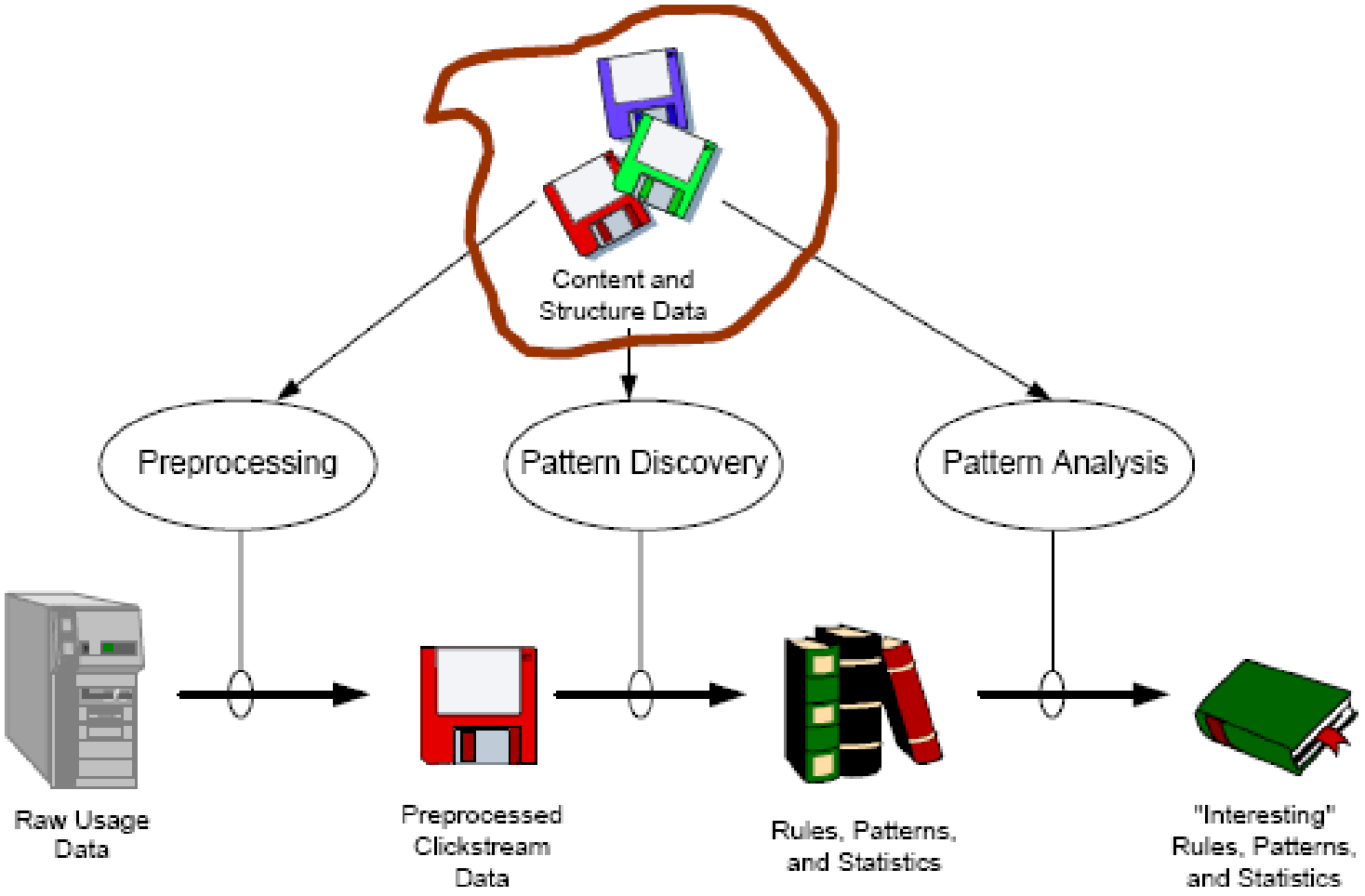
Path completion

- 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.

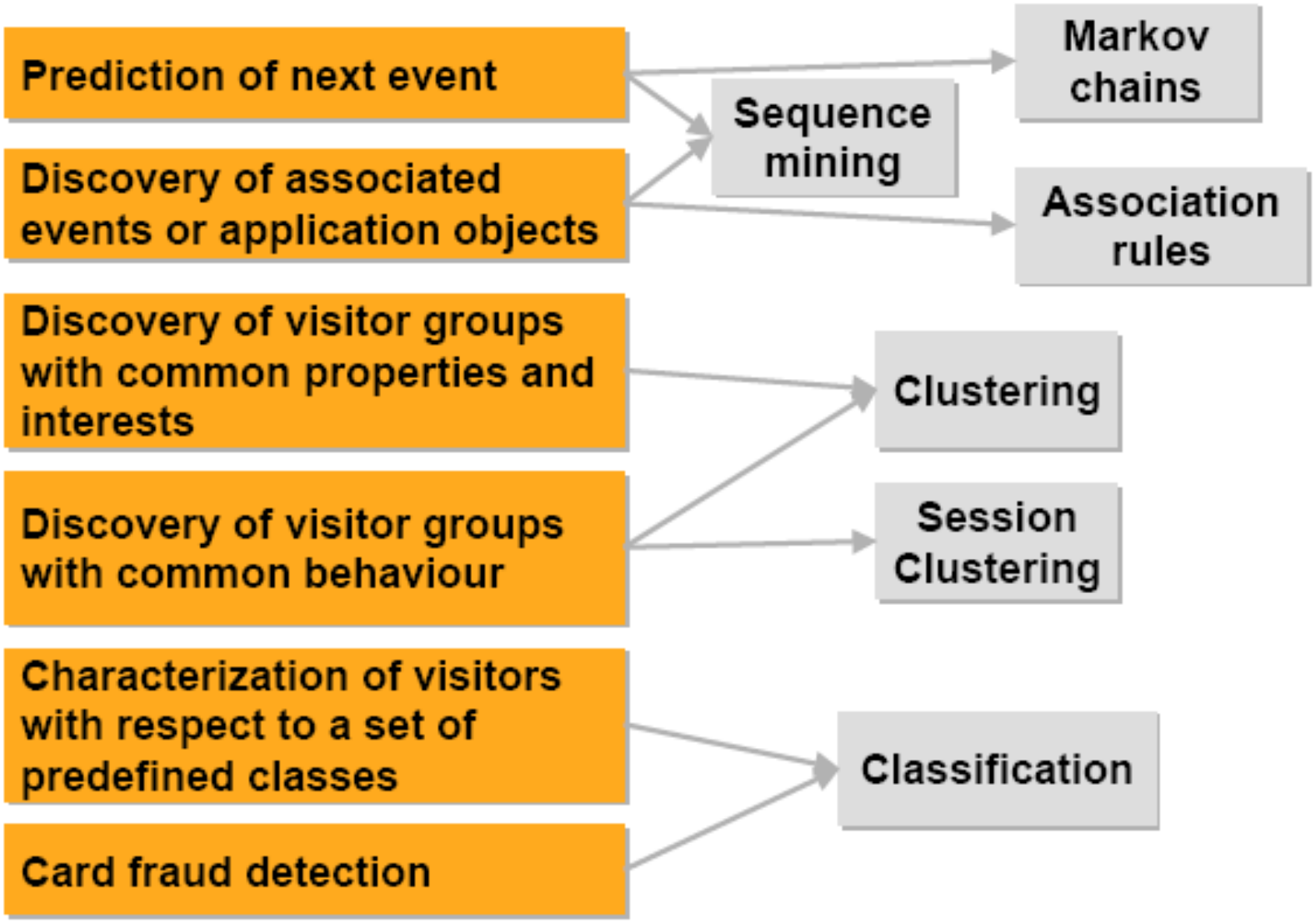
Integrating with e-commerce events

- 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

Web usage mining process



Some usage mining applications



Advantages : Web Usage mining

- Allows companies to produce productive information pertaining to the future of their business function ability
- Some of this information can be derived from the collective information of lifetime user value, product cross marketing strategies and promotional campaign effectiveness
- The usage data that is gathered provides the companies with the ability to produce results more effective to their businesses and increasing of sales

Contd...

- To develop marketing skills that will out-sell the competitors and promote the company's services or product on a higher level
- To aid in e-businesses whose business is based solely on the traffic provided through search engines
- The use of this type of web mining helps to gather the important information from customers visiting the site.
- This enables an in-depth log to complete analysis of a company's productivity flow
- E-businesses depend on this information to direct the company to the most effective Web server for promotion of their product or service

Contd...

- To provide the best access routes to services or other advertisements
- When a company advertises for services provided by other companies, the usage mining data allows for the most effective access paths to these portals
- To provide the companies with the information needed to provide an effective presence to their customers
- This collection of information may include user registration, access logs and information leading to better Web site structure, proving to be most valuable to company online marketing
- These present some of the benefits for external marketing of the company’s products , services and overall management

Contd...

- To provide information for improvement of communication through intranet communications
- Developing strategies through this type of mining will allow for intranet based company databases to be more effective through the provision of easier access paths
- The projection of these paths helps to log the user registration information giving commonly used paths the forefront to its access
- To keep a record of fraudulent payments which can all be researched and studied through data mining
- This information can help develop more advanced and protective methods that can be undertaken to prevent such events from happening

Contd…

- To foster marketing of businesses and a direct impact to the success of their promotional strategies and internet traffic.
- This information is gathered on a daily basis and continues to be analyzed consistently
- Analysis of this pertinent information will help companies to develop promotions that are more effective, internet accessibility, inter-company communication and structure, and productive marketing skills through web usage mining

Disadvantages

- Invasion of privacy
- Misuse of data(collected for different purpose n misused for different purpose)
- some mining algorithms might use controversial attributes like sex, race, religion, or sexual orientation to categorize individuals

Applications: Web Usage mining

Major application areas for web usage mining

- Personalization
- System improvement
- Site modification
- Business intelligence
- Usage characterization