Web Usage Mining

- We require high-level task in usage data pre-processing. Some of them are:
  - Data Fusion and Cleaning
  - Pageview Identification
  - User Identification
  - Sessionization
  - Path completion

Web Usage Mining

- In large-scale Web Sites, it is typical that the content served to users comes from multiple Web or Application Servers.
- This may require global synchronization across these servers.
- It is not uncommon for typical log file to contain a significant (sometines as high as 50%) percentage of references resulting from search engine or other crawlers (or spiders).

Web Usage Mining

- Data Fusion
  - A way to synchronize log files across Web and Application Servers.
  - It is performed inter-site in order to obtain the user behavior over the logs files of multiple related Web Sites.
- Cleaning
  - Usually is site-specific.
  - Removal of embedded objects that may not be important for future analysis such as style files, graphics or sound files.
  - Removal of data fields that may not provide useful information such as number of bytes.
  - Removal of references due to crawler or spider navigation.

Web Usage Mining

- Data Fusion example

Web Server Logs 1

Time	IP	URL	Ref	Agent
12:01:00 AM	1.2.3.4	A	-	IE5;Win2k
12:09:00 AM	1.2.3.4	B	A	IE5;Win2k
12:10:00 AM	2.3.4.5	C	-	IE6;WinXP;SP1
12:15:00 AM	2.3.4.5	E	C	IE6;WinXP;SP1

Web Server Logs 2

Time	IP	URL	Ref	Agent
12:12:00 AM	2.3.4.5	B	-	IE6;WinXP;SP1
12:19:00 AM	1.2.3.4	C	A	IE5;Win2k
12:22:00 AM	2.3.4.5	D	B	IE6;WinXP;SP1
12:25:00 AM	1.2.3.4	E	C	IE5;Win2k

AM				
12:25:00 AM	1.2.3.4	C	A	IE6;WinXP;SP2
01:10:00 AM	1.2.3.4	E	D	IE6;WinXP;SP2
01:16:00 AM	1.2.3.4	C	A	IE5;Win2k
01:17:00 AM	1.2.3.4	F	C	IE6;WinXP;SP2
01:36:00 AM	1.2.3.4	D	B	IE5;Win2k

AM				
12:58:00 AM	1.2.3.4	D	B	IE6;WinXP;SP2
01:15:00 AM	1.2.3.4	A	-	IE5;Win2k
01:26:00 AM	1.2.3.4	F	C	IE5;Win2k
01:30:00 AM	1.2.3.4	B	A	IE5;Win2k

Web Usage Mining

Time	IP	URL	Ref	Agent
12:01:00 AM	1.2.3.4	A	-	IE5;Win2k
12:09:00 AM	1.2.3.4	B	A	IE5;Win2k
12:10:00 AM	2.3.4.5	C	-	IE6;WinXP;SP1
12:12:00 AM	2.3.4.5	B	C	IE6;WinXP;SP1
12:15:00 AM	2.3.4.5	E	C	IE6;WinXP;SP1
12:19:00 AM	1.2.3.4	C	A	IE5;Win2k
12:22:00 AM	2.3.4.5	D	B	IE6;WinXP;SP1
12:22:00 AM	1.2.3.4	A	-	IE6;WinXP;SP2
12:25:00 AM	1.2.3.4	E	C	IE5;Win2k
12:25:00 AM	1.2.3.4	C	A	IE6;WinXP;SP2
12:33:00 AM	1.2.3.4	B	C	IE6;WinXP;SP2
12:58:00 AM	1.2.3.4	D	B	IE6;WinXP;SP2
01:10:00 AM	1.2.3.4	E	D	IE6;WinXP;SP2
01:15:00 AM	1.2.3.4	A	-	IE5;Win2k
01:16:00 AM	1.2.3.4	C	A	IE5;Win2k
01:17:00 AM	1.2.3.4	F	C	IE6;WinXP;SP2
01:26:00 AM	1.2.3.4	F	C	IE5;Win2k
01:30:00 AM	1.2.3.4	B	A	IE5;Win2k
01:36:00 AM	1.2.3.4	D	B	IE5;Win2k

Web Usage Mining

- It depend on the intra-page structure of the Web Site, as well as on the page contents and the underlying site domain knowledge.
- A Pageview is an aggregate representation of a collection of Web objects or resources representing an specific User Event.

adding a product to the shopping cart.

Web Usage Mining

- In order to provide flexible framework for variety of data mining activities a number of attributes must be recorded with each Pageview.
- This attributes include:
  - Pageview id: Normally a URL uniquely representing the Pageview.
  - Static Pageview type: For example, information page, product view, category view, or index page.
  - Other metadata: For example, keywords or product attributes.

Web Usage Mining

- An user can visit a Web Site more than once.
- Not all sites employ cookies, and due to privacy concerns, client-side cookies are sometime disabled by users.
- Proliferation of ISP proxy server which assign rotating IP Addresses to clients.

Web Usage Mining

- The analysis of Web Usage doesn't require knowledge about user's identity, but it is necessary to distinguish among them.
- We refer to the sequence of logged activities belonging to the same user as **user activity record**.
- However, two occurrences of the same IP Address, separated by a sufficient amount of time, might correspond to two different users. (What if two users share the same IP?)
- It's still possible to accurately identify unique users through a combination of IP Addresses and other information such as User Agents and referrers.

Web Usage Mining

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

1:26	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

Web Usage Mining

- User Identification example:

User 1

Time	IP	URL	Ref	Agent
0:01	1.2.3.4	A	-	IE5;Win2k
0:09	1.2.3.4	B	A	IE5;Win2k
0:19	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

User 2

Time	IP	URL	Ref	Agent
0:10	2.3.4.5	C	E	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
0:22	2.3.4.5	D	B	IE6;WinXP;SP1

User 3

Time	IP	URL	Ref	Agent
0:22	1.2.3.4	A	-	IE6;WinXP;SP2
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	C	IE6;WinXP;SP2
1:17	1.2.3.4	F	C	IE6;WinXP;SP2

Web Usage Mining

- Is the process of segmenting the user activity record of each user into sessions.
- Each session represent a single visit to the Web Site.
- Without additional authentication information from users and mechanisms such as session ids, we must rely on heuristic methods.
- 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 Web Site.

Web Usage Mining

- In a conceptual way:
  - R: Set of real sessions. Represent the real activity of the user on the Web Site.
  - h: Sessionization heuristic that attempt to map R into a set of constructed session C_h.
  - For the ideal heuristic h^*, we have $C_{h^*} = R$.
- Generally, sessionization heristics fall into two basic categories:
  - Time-oriented: Apply global or local time-out estimates to distinguish between consecutive sessions.
  - Structure-oriented: Use static site structure or the implicit linkage structure captured in the referrer fields of the server logs.

Web Usage Mining

- Time-oriented example:
  - h1: Total session duration may not exceed a threshold θ. Given t_0, the timestamp for the first request in a constructed session S, the request with a timestamp t is assigned to S if $t - t_0 \leq \delta$.

request assigned to constructed session S, the next request with timestamp t_2 is assigned to S if $t_2 - t_1 \leq \delta$.

Web Usage Mining

- Time-oriented example:

User 1

Time	IP	URL	Ref
12:01:00 AM	1.2.3.4	A	-
12:09:00 AM	1.2.3.4	B	A
12:19:00 AM	1.2.3.4	C	A
12:25:00 AM	1.2.3.4	E	C
01:15:00 AM	1.2.3.4	A	-
01:26:00 AM	1.2.3.4	F	C
01:30:00 AM	1.2.3.4	B	A
01:36:00 AM	1.2.3.4	D	B

Session 1

Time	IP	URL	Ref
12:01:00 AM	1.2.3.4	A	-
12:09:00 AM	1.2.3.4	B	A
12:19:00 AM	1.2.3.4	C	A
12:25:00 AM	1.2.3.4	E	C

Session 2

Time	IP	URL	Ref
01:15:00 AM	1.2.3.4	A	-
01:26:00 AM	1.2.3.4	F	C
01:30:00 AM	1.2.3.4	B	A
01:36:00 AM	1.2.3.4	D	B

Web Usage Mining

- Structure-oriented example:
  - h-ref: A request q is added to constructed session S if the referrer for q was previously invoked in S. Otherwise, q is used as the start of a new constructed session. With this heuristic it is possible that a request q may potentially belong to more than one *open* constructed session, q may have been accessed previously in multiple sessions. In this case, q could be added to the most recently opened session that satisfies the condition.

Web Usage Mining

- Structure-oriented example:

User 1

Time	IP	URL	Ref
12:01:00 AM	1.2.3.4	A	-

Session 1

Time	IP	URL	Ref
12:01:00 AM	1.2.3.4	A	-

Session 2

Time	IP	URL	Ref
01:15:00 AM	1.2.3.4	A	-

AM			
12:19:00 AM	1.2.3.4	C	A
12:25:00 AM	1.2.3.4	E	C
01:15:00 AM	1.2.3.4	A	-
01:26:00 AM	1.2.3.4	F	C
01:30:00 AM	1.2.3.4	B	A
01:36:00 AM	1.2.3.4	D	B

AM			
12:19:00 AM	1.2.3.4	C	A
12:25:00 AM	1.2.3.4	E	C
01:26:00 AM	1.2.3.4	F	C

AM			
01:36:00 AM	1.2.3.4	D	B

Web Usage Mining

- Client or proxy side caching can often result in missing access references for those page or objects that have been cached.
- If a user return to 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 client-side.
- Therefore, no request is made to the server and the second reference to A not being recorded on the server log. This is called **Missing References**.

Web Usage Mining

- This Missing References can be heuristically inferred through **a path completion**.
- This relies on the **knowledge of site structure** and **referrer information** form server logs.
- In 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.
- It's possible to recapture the user input through **packet sniffers**, which listen to all incoming and outgoing TCP/IP network traffic on the server side.

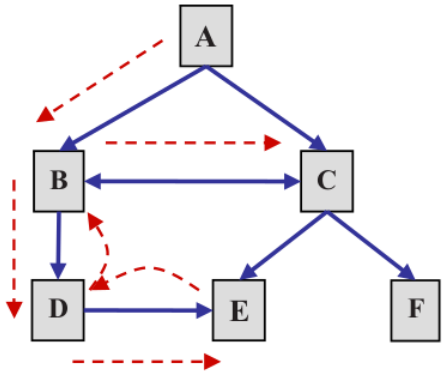
Web Usage Mining

- Path completion example:

Session 1

Time	IP	URL	Ref
01:15:00 AM	1.2.3.4	A	-
01:30:00 AM	1.2.3.4	B	A
01:36:00 AM	1.2.3.4	D	B
01:36:00 AM	1.2.3.4	E	D
01:36:00 AM	1.2.3.4	C	B

- Path completion example:



- Recorded navigation path:
  - $A \rightarrow B \rightarrow D \rightarrow E$
  - $B \rightarrow C$
- Actual navigation path:
  - $A \rightarrow B \rightarrow D \rightarrow E \rightarrow D \rightarrow B \rightarrow C$
- There are many (even infinite) candidate completion. A simple heuristic among candidates paths is to select the one requiring the fewes number of back references.

Data Modeling for WUM

Web Usage Mining

- Usage Data Pre-Processing results in a set of n pageviews P and m user transactions T where:
  - $P = \{p_1, p_2, ..., p_n\}$
  - $T = \{t_1, t_2, ..., t_m\}$
  - T is a subset of P.
- We view each transaction t as an l-length sequence of ordered pairs:
  - $t = \langle (p_1^t, w(p_1^t)), (p_2^t, w(p_2^t)), ..., (p_l^t, w(p_l^t)) \rangle$
  - Where each $p_i^t = p_j$ for some j in $\{1, 2, ..., n\}$ and $w(p_i^t)$ is the weight associated with pageview p_i^t in transaction t, representing its significance.

Web Usage Mining

- Weights can be determined in **number of ways**, in part based on the type of analysis or the intended personalization framework.
- In most WUM tasks, the weights are either binary, representing the existence or non-existence of a pageview in the transaction.
- Also, it can be a **function** of the duration of the pageview in the users session.
  - Usually the time spent by a user on the last pageview in the session **is not available**.
  - A commonly used option is to set it as the mean time duration for the page taken across all sessions in which the pageview doesn't occur as the last one.
  - It's common in practice to use a normalized value of page duration instead of a raw time duration in order to account for user variances.

Web Usage Mining

- For many practical applications, we can represent each user transaction as a vector over the n-dimensional space of pageviews.
- Given the transaction t above, the transaction vector $\boldsymbol{t}$ is given by:

$$\boldsymbol{t} = \left(w_{p_1}^t, w_{p_2}^t, ..., w_{p_n}^t\right)$$

otherwise.

- Then, the set of all user transactions can be viewed as an $m \times n$ **user-pageview matrix (UPM)**.

Web Usage Mining

- UPM example:

	A.html	B.html	C.html	D.html	E.html
user1	1	0	1	0	1
user2	1	1	0	0	1
user3	0	1	1	1	0
user4	1	0	1	1	1
user5	1	1	0	0	1
user6	1	0	1	1	1

Web Usage Mining

- It's possible to **integrate** other sources of knowledge, such as semantic information from content of Web pages (textual features).
- Each pageview p can be represented as a r-dimensional feature vector, where r is the total number of extracted features (words or concepts) from the site in a global dictionary.
- This vector $\boldsymbol{p}$ can be given by:
  - $\boldsymbol{p} = (\text{fw}^p(f_1), \text{fw}^p(f_2), ..., \text{fw}^p(f_r))$
  - Where $\text{fw}^p(f_j)$ is the weight of the jth feature (f_j) in the pageview p, for $1 \leq j \leq r$.
- For the whole collection of pageviews in the site, we then have an $n \times r$ **pageview-feature matrix** $\text{PFM} = \{\boldsymbol{p}_1, \boldsymbol{p}_2, ..., \boldsymbol{p}_n\}$

Web Usage Mining

- PFM example:

	web	data	mining	business	intelligence	marketing	commerce	search	information	retrieval
A.html	0	0	0	1	1	1	0	1	1	1
B.html	0	1	1	1	1	1	1	0	0	0
C.html	1	1	1	0	0	0	1	1	1	1
D.html	1	1	1	0	0	0	0	0	1	1
E.html	1	0	0	0	1	1	0	0	1	1

- PFM transpose is called Term-pageview matrix.
- The result is a new matrix $\text{TFM} = \{\boldsymbol{t}_1, \boldsymbol{t}_2, ..., \boldsymbol{t}_m\}$
- Each $\boldsymbol{t}_i$ is a r-dimensional vector over the feature space.

Web Usage Mining

- Term-pageview matrix

web	0	0	1	1	1
data	0	1	1	1	0
mining	0	1	1	1	0
business	1	1	0	0	0
intelligence	1	1	0	0	1
marketing	1	1	0	0	1
ecommerce	0	1	1	0	0
search	1	0	1	0	0
information	1	0	1	1	1
retrieval	1	0	1	1	1

Web Usage Mining

- This integration may involve the transformation of user transaction into content-enhanced transactions containing the semantic features of the pageviews.
- The idea is to represent each user session as a vector of semantic features rather than as a vector over pageviews.
- This allows user's session to reflect not only the pages visited, but also the significance of various concepts or context features that are relevant to the user's interaction.

Web Usage Mining

- There is several ways to accomplish this transformation. The resulting matrix is called **content-enhanced transaction matrix**.
- The most direct approach involves mapping each pageview transaction in a transaction to one or more content features (UPM X PFM).

Web Usage Mining

- Content-Enhanced Transaction Matrix example:

UPM

	A.html	**B.html**	**C.html**	**D.html**	**E.html**
user1	1	0	1	0	1
user2	1	1	0	0	1
user3	0	1	1	1	0
user4	1	0	1	1	1
user5	1	1	0	0	1
user6	1	0	1	1	1

Web Usage Mining

- Content-Enhanced Transaction Matrix example:

	web	data	mining	business	intelligence	marketing	commerce	search	information	retrieval
A.html	0	0	0	1	1	1	0	1	1	1
B.html	0	1	1	1	1	1	1	0	0	0
C.html	1	1	1	0	0	0	1	1	1	1
D.html	1	1	1	0	0	0	0	0	1	1
E.html	1	0	0	0	1	1	0	0	1	1

Web Usage Mining

- Content-Enhanced Transaction Matrix example:

	web	data	mining	business	intelligence	marketing	commerce	search	information	retrieval
user1	2	1	1	1	2	2	1	2	3	3
user2	1	1	1	2	3	3	1	1	2	2
user3	2	3	3	1	1	1	2	1	2	2
user4	3	2	2	1	2	2	1	2	4	4
user5	1	1	1	2	3	3	1	1	2	2
user6	3	2	2	1	2	2	1	2	4	4

Discovery and Analysis of Web Usage Patterns

- The satistical analysis of pre-processed session data constitutes the most common form of analysis.
- Here, data is aggregated by predetermined unit such as: days, sessions, visitors or domains.
- Standard statistical techniques can be used on this data to gain knowledge about visitor behavior.
- This knowledge could be used for improve the system performance and supporting marketing decisions.

Discovery and Analysis of Web Usage Patterns

- Reports based on this type of analysis may include:
  - Information about most frequently accessed pages.
  - Average view time of page.
  - Average length of path through a site.
  - Common entry and exit points.

Discovery and Analysis of Web Usage Patterns

- Clustering is a data mining technique that groups together a set of items having similar characteristics.
- In the Usage domain, there two kinds of interesting clusters that can be discovered:
  - User clusters
  - Page clusters

- Clustering of users records, such as sessions or transactions.
- Tends to establish groups of users **exhibiting similar browsing patterns**.
- Is useful for inferring **user demographics**.
- Commonly used to:
  - Perform market segmentation in e-commerce.
  - Provide personalized Web content to the users with similar interests.
  - Create Web-based user communities, reflecting similar interest of groups of users.
  - Learn user models that can be used to provide dynamic recommendations in Web personalization applications.

Discovery and Analysis of Web Usage Patterns

- Given the mapping of user transactions into a multi-dimensional space as vectors of pageviews, standard clustering algorithms could be applied, such as k-mean, partition this space into groups of transactions that are **close to each other**.
- Transaction cluster can represent user or visitor segments based on their navigational behavior or other attributes.
- However, these are **not an effective** means of capturing the aggregated view of common user patterns for further analysis, due to its magnitude.

Discovery and Analysis of Web Usage Patterns

- One approach in creating and aggregate view of each cluster is to compute the **centroid** of each cluster (mean vector).
- The dimension value for each pageview in the mean vector is computed by finding the ratio of the sum of the pageview weights across transactions to the total number of transactions in the clusters.
- Given a pageview p in a cluster centroid, the centroid dimension value of p provides a measure of its **significance** in the cluster.
- Pageviews in the centroid can be sorted according to these weights and lower ones can be filtered out.

Discovery and Analysis of Web Usage Patterns

- More formally:
  - Given a transaction cluster cl.
  - We can construct the aggregate profile $\mathrm{pr}_{\mathrm{cl}}$ as a set of **pageview-weight** pairs by computing the centroid of cl:

$$\mathrm{pr}_{\mathrm{cl}} = \{(p, \mathrm{weight}(p, \mathrm{pr}_{\mathrm{cl}})) \mid \mathrm{weight}(p, \mathrm{pr}_{\mathrm{cl}}) \geq \mu\}$$

Discovery and Analysis of Web Usage Patterns

- Where:
  - The significance weight, $\mathrm{weight}(p, \mathrm{pr}_{\mathrm{cl}})$, of the page p within the aggregate profile $\mathrm{pr}_{\mathrm{cl}}$ is given by:

$$\mathrm{weight}(p, \mathrm{pr}_{\mathrm{cl}}) = \frac{1}{|\mathrm{cl}|} \sum_{s \in \mathrm{cl}} w(p, s)$$

- $w(p, s)$ is the weight of page p in transaction vector s of cluster cl.
- μ: threshold used to focus only on those pages in the cluster that appear in a sufficient number of vectors in that cluster.

Discovery and Analysis of Web Usage Patterns

- Each profile can be represented as a vector in the original n-dimensional space of pageviews.
- This aggregate representation can be used directly for **predictive modeling** and in applications such as **recommender systems**.
- One recommender system application example:
  - Given a new user u, who has accessed a set of pages P_u so far, we can measure the similarity of P_u to the discovered profiles, and recommend to the user those pages in matching profiles which have not yet been accessed by the user.

Discovery and Analysis of Web Usage Patterns

User clusters example:

- Assuming the data has already been clustered.
- Pages B and F are the most significant pages characterizing the common interest of users in this segment.
- Page C only appears in one transaction and might be removed given a threshold greater than 0.25.

	A	B	C	D	E	F
user1	0	0	1	1	0	0
user4	0	0	1	1	0	0
user7	0	0	1	1	0	0
user0	1	1	0	0	0	1
user3	1	1	0	0	0	1
user6	1	1	0	0	0	1
user9	0	1	1	0	0	1
user2	1	0	0	1	1	0
user5	1	0	0	1	1	0
user8	1	0	1	1	1	0

Aggregated Profile for Cluster 1	
Weight	Pageview
1.00	B
1.00	F
0.75	A
0.25	C

Discovery and Analysis of Web Usage Patterns

- Clustering of pages can be performed based on the **usage data** (user sessions or transaction data) or based on the **content features** associated with pages (keywords or product attributes).
- The result may be collections of pages **related to the same topic or category**.
- Some examples are:
  - Items that are commonly accessed or puchased together can be organize into groups.
  - HTML pages that suggest related hyperlinks to users according to their past history of navigational or purchase activities.

Discovery and Analysis of Web Usage Patterns

- Association rule discovery and statistical correlation analysis can find groups of items or pages that are commonly accessed or purchased together.
- This enables Web sites to organize the site content more efficiently, or to provide effective cross-sale product recommendations.

transactions, satisfying a user specified minimum support threshold.

- Example:
  - If a site does not provide a direct linage between two page A and B, the discovery of a rule, $A \rightarrow B$, would indicates that providing a direct hyperlink from A to B might aid users in finding the intended information.

Discovery and Analysis of Web Usage Patterns

- Attempts to find **inter-session patterns** such that the presence of a set of items is followed by another item in a time-ordered set of sessions or episodes.
- With this, Web marketers can predict future visit patterns which will be helpful in placing advertisements aimed at certain user groups.
- Can be used to capture **frequent navigational paths** among user trails.
- Usually, this is represented using Markov Chains models.

Discovery and Analysis of Web Usage Patterns

- In Web domain, one is interested in developing a profile of users belonging to a particular class or category.
- This requires extraction and selection of features that best describe the properties of a given class, and then use a supervised learning algorithm.
- It is also possible to use previously discovered clusters and association rules for classification of new users.
- These techniques play an important role in Web analytics applications for modeling the users according to various predefined metrics.

Discovery and Analysis of Web Usage Patterns

- These techniques play an important role in Web analytics applications for modeling the users according to various predefined metrics, such as:
  - Given a set of user transactions, the sum of purchases made by each user within a specified period of time can be computed.
  - A classification model can then be build based on this enriched data in order to classify users into whose have a high propensity to buy and those who do not.