

Data Warehousing & Mining

Q1.a) *Explain ETL Process of Data warehouse in detail:*

Ans: Extraction

- The first part of an ETL process involves extracting the data from the source systems.
- Most data warehousing projects consolidate data from different source systems.
- Each separate system may also use a different data organization/format. Common data source formats are relational databases and flat files, but may include non-relational database structures such as Information Management System (IMS) or other data structures such as Virtual Storage Access Method (VSAM) or Indexed Sequential Access Method (ISAM), or even fetching from outside sources such as through web spidering or screen-scraping.
- Extraction converts the data into a format for transformation processing.

Transform

- The transform stage applies a series of rules or functions to the extracted data from the source to derive the data for loading into the end target.
- Some data sources will require very little or even no manipulation of data. In other cases, one or more of the following transformation types may be required to meet the business and technical needs of the target database:
 - Selecting only certain columns to load (or selecting null columns not to load). For example, if source data has three columns (also called attributes) say roll_no, age and salary then the extraction may take only roll_no and salary. Similarly, extraction mechanism may ignore all those records where salary is not present (salary = null).
 - Translating coded values (*e.g.*, if the source system stores 1 for male and 2 for female, but the warehouse stores M for male and F for female), this calls for automated data cleansing; no manual cleansing occurs during ETL
 - Encoding free-form values (*e.g.*, mapping "Male" to "1" and "Mr" to M)
 - Deriving a new calculated value (*e.g.*, sale_amount = qty * unit_price)
 - Filtering
 - Sorting
 - Joining data from multiple sources (*e.g.*, lookup, merge)
 - Aggregation (for example, rollup — summarizing multiple rows of data — total sales for each store, and for each region, etc.)
 - Generating surrogate-key values
 - Transposing or pivoting (turning multiple columns into multiple rows or vice versa)
 - Splitting a column into multiple columns (*e.g.*, putting a comma-separated list specified as a string in one column as individual values in different columns)
 - Desegregation of repeating columns into a separate detail table (*e.g.*, moving a series of addresses in one record into single addresses in a set of records in a linked *address* table)
 - Lookup and validate the relevant data from tables or referential files for slowly changing dimensions.
 - Applying any form of simple or complex data validation. If validation fails, it may result in a full, partial or no rejection of the data, and thus none, some or all the data is handed over to the next step, depending on the rule design and exception handling. Many of the above transformations may result in exceptions, for example, when a code translation parses an unknown code in the extracted data.

Data Warehousing & Mining

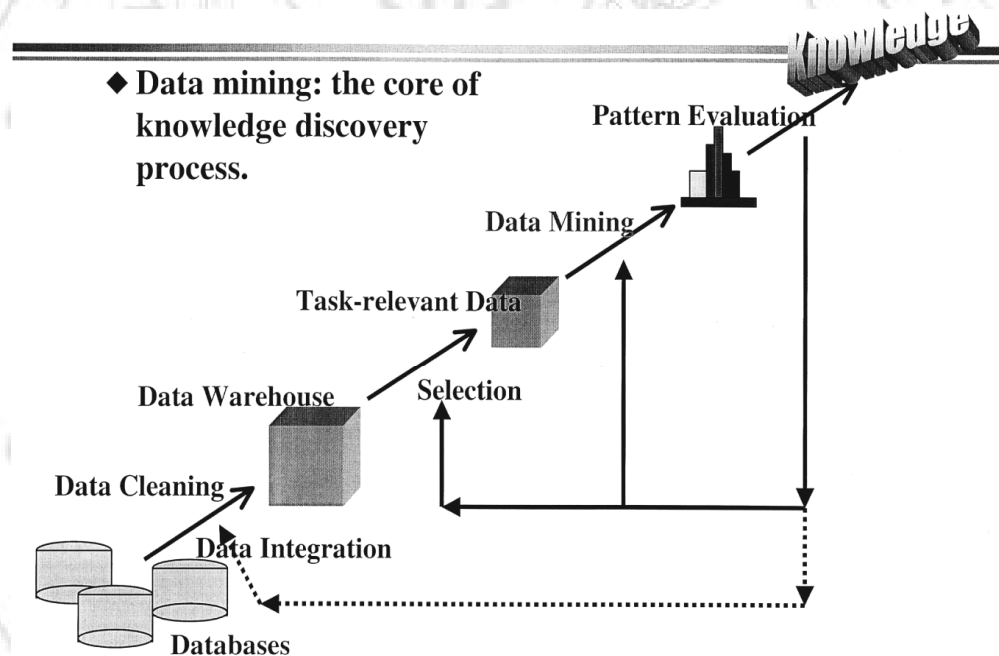
Loading

- The load phase loads the data into the end target, usually the data warehouse (DW).
- Depending on the requirements of the organization, this process varies widely.
- Some data warehouses may overwrite existing information with cumulative, frequently updating extract data is done on daily, weekly or monthly.
- While other DW (or even other parts of the same DW) may add new data in a historicized form, for example, hourly.
- To understand this, consider a DW that is required to maintain sales record of last one year.
- Then, the DW will overwrite any data that is older than a year with newer data.

b) Explain data mining as a step in kdd Give the architecture of typical DM system.

10

Ans:



Source: Han and Kamber (2001)

Selection:

- the data needed for the data mining process may be obtained from many different and heterogeneous data sources
- this first step obtains the data from various databases, files.

Pre-processing

- the data to be used by the process may have incorrect or missing data

Data Warehousing & Mining

- there may be anomalous data from multiple sources involving different data types and metrics.
- There may be many activities performed at this time.

Transformation

- Data from different sources must be converted into a common format for processing.
- Some data may be encoded or transformed into more usable formats.
- Data reduction may used to reduce the number of possible data values being considered.

Data mining

- Based on the data mining tasks being performed, this step applies algorithms to the transformed data to generate the desired results.

Interpretation/evaluation

- How the data mining results are presented to the users is extremely important because the usefulness of the result is dependent on it.
- Various visualization and GUI strategies are used at this last step.

Q2. a) *Explain Dimension table is wide, fact table deep.Explain.*

Explain Star schema and its advantages

10

Ans:

A dimension table is wide.

- Dimension table has many columns or attributes.
- Dimension table have more than 50 attributes.
- Therefore dimension table is wide.

A fact table is deep.

- If you lay the fact table out as 2D table.
- You will note that fact table is narrow with small number of columns, but very deep with a large number of rows.

Star schema:

Star schema is nothing but a combination of dimension and fact table.

There is only one fact table and no. of dimension tables. The fact table contains measures and primary key of each dimension table. That works as foreign key, and connect the dimensions.

Example.

Data Warehousing & Mining

Advantages:

- 1) Easy to understand
- 2) Optimize navigation
- 3) Most suitable for query processing

b) *what is clustering? Explain K-means Clustering algorithm .Suppose the data for clustering is {2,4,10,12,3,20,30,11,25} consider k=2*

Ans: ,

K-means clustering:

K-Means Training starts with a single cluster with its center as the mean of the data. This cluster is split into two and the means of the new clusters are iteratively trained. These two clusters are again split and the process continues until the specified number of clusters is obtained. If the specified number of clusters is not a power of two, then the nearest power of two above the number specified is chosen and then the least important clusters are removed and the remaining clusters are again iteratively trained to get the final clusters.

Example

Given set

{10, 4, 2, 12, 3, 20, 30, 11, 25, 31}

Step 1

$M_1 = 10$ $M_2 = 4$

$K_1 = \{10, 4\}$

$K_2 = \{2, 12, 3, 20, 30, 11, 25, 31\}$

Step 2

$M_1 = 7$ $m_2 = 16.75$

$K_1 = \{10, 4, 2\}$

$K_2 = \{12, 30, 11, 25, 31\}$

Step 3

$M_1 = 8$ $m_2 = 21.8$

Data Warehousing & Mining

K1= {10, 4, 2, 3}

K2 = {11, 12, 30, 25, 31}

Step 4

M1= 4.75 m2= 21.8

K1= {10, 4, 2, 3, 11}

K2= {12, 30, 25, 31}

Step 5

M1=6 m2=24.5

K1= {10, 4, 2, 3, 11, 12}

K2= {30, 25, 31}

Step 6

M1= 7 m2=28.66

Final answer

K1= {10, 4, 2, 3, 11, 12}

K2 = {30, 25, 31}

Q3. a) Consider the transaction database given below. Use Apriori Algorithm with minimum support count 2, generate the association rules along with its confidence. 10

Ans: Give data

TID	List of Items
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

Item	Support
I1	6
I2	6
I3	5
I4	2
I5	2

Item	Support
I1, I2	4
I1, I3	4
I1, I4	1
I1, I5	2

Data Warehousing & Mining

Item	Support
I1, I2, I3	4
I1, I2, I4	4
I1, I2, I5	1
I1, I3, I5	1

Item	Support
I1, I2, I3, I4	4
I1, I2, I4, I5	4

Minimum support : 2

Confidence : 70%

$\frac{I1, I2}{I1} = 4 / 6 = 0.6 = 60\%$ Less than min confidence

I1

$\frac{I1, I3}{I1} = 4 / 6 = 0.6 = 60\%$ Less than min confidence

I1

$\frac{I1, I3}{I3} = 4 / 5 = 0.8 = 80\%$ selected

I3

$\frac{I1, I5}{I5} = 2 / 2 = 1 = 100\%$ selected

I5

$\frac{I1, I5}{I1} = 2 / 6 = 0.3 = 30\%$ Less than min confidence

I1

So, association rules are:

I1 → I3

I1 → I5

If minimum confidence is 60% then the association rules are:

I1 → I2

I1 → I3

I1 → I5

Data Warehousing & Mining

b) Explain characteristics of the data present in the data warehouse.

Ans :

Characteristics of data warehouse:

1) subject oriented data

In the data set for an order processing application, we keep the data for that particular application. These data sets provide the data for all the functions for entering orders, checking stocks. But these data sets contain only the data that is needed for those functions relating to this particular application.

2) Integrated data

For proper decision making, you need to pull together all the relevant data from the various applications. The data in DW comes from several operational systems. These are disparate applications, so the operational platform and operating system could be different. Data from internal operational system, for many enterprises, data from outside source is likely to be very important.

3) Time variant

The stored data contains the current values. In an account receivable system the balance is the current outstanding balance in the customer's account. The DW is meant for analysis and decision making. Data is stored as snapshots over past and current periods. Every data structure in the DW contain time element. You will find historical snapshots of the operational data in DW.

4) Non-volatile data

Data extracted from the various operational systems and data obtained from outside sources are transformed, integrated and store in the DW. The data in the DW is not intended to run the day to day business.

Q4. a) Explain HITS algorithm.

10

Ans:

```

1  G := set of pages
2  for each page p in G do
3  p.auth = 1 // p.auth is the authority score of the page p
4  p.hub = 1 // p.hub is the hub score of the page p
5  function HubsAndAuthorities(G)
6  for step from 1 to k do // run the algorithm for k steps
7  for each page p in G do // update all authority values first
8  for each page q in p.incomingNeighbors do // p.incomingNeighbors is the
  set of pages that link to p
9  p.auth += q.hub
10 for each page p in G do // then update all hub values
11 for each page r in p.outgoingNeighbors do // p.outgoingNeighbors is the set
  of pages that p links to
12 p.hub += r.auth

```

Data Warehousing & Mining

b) *Define data warehouse.*

Explain the architecture of Data warehouse with suitable block diagram

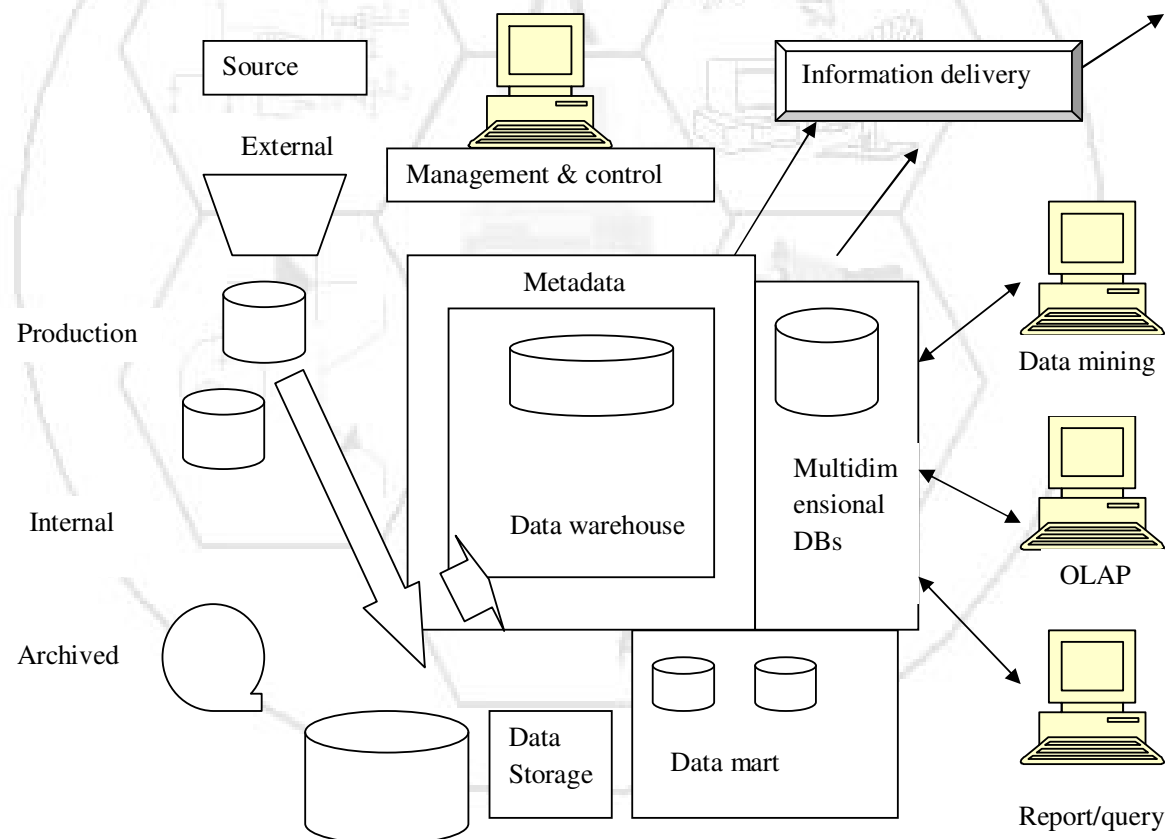
10

Ans :

Data warehouse: It is an information environment that

- Provides an integrated and total view of the enterprise
- Make the historical information easily available for decision making
- Make decision support transaction possible without hindering operational system
- Presents a flexible and interactive source of strategic information.

Data warehouse architecture.



- Data staging components serves as the next building blocks.
- Data storage components that manages the DW.
- These components not only store and manages the data warehouse

Data Warehousing & Mining

- It also keep track of the data by means of the metadata repository

Source data components

1. Production data:

- Data comes from the various operational system
- Based on the information requirements in DW, you choose segments of data from the different operational system
- Data is supported by different database system & O.S.
- This is data from many vertical applications.

2. Internal data:

- In every organization user keep their “Private” spreadsheets, documents, customer profiles.
- This is internal data, part of which could be useful in a data warehouse.
- Profiles of individual customers become very important for consideration.

3. Archived data:

- You periodically take the old data and store it in archived files.
- Some data is archived after a year
- Some times data is left in the operational system database for as long as 5 years.

4. External data:

- Most executives depend on data from external sources for higher percentage of the information they use.
- You need external sources
- Data from outside sources do not conform to your formats.

Data staging components:

- After extracted data from various operational system and from external sources, you have to prepare the data for storing in data warehouse
- The extracted data coming from several disparate sources need to be changed, converted and made ready in a format that is suitable to be stored for querying and analysis
- 3 major functions need to perform for getting the data ready.
 1. Extract the data
 2. transform the data
 3. Load the data in DW storage.
- Separate place or component is required to perform the data preparation
- When we implement an operational system, we are likely to pickup data from different sources
- Move the data into the new operational system dbs, and run the conversions.

Data Warehousing & Mining

Q5.a) *Distinguish between :*

10

1) *Top-Down & Bottom-Up approach*

2) *Explain Difference between OLTP & OLAP.*

Ans:

Top-down approach:

- The top-down design methodology generates highly consistent dimensional views of data across data marts since all data marts are loaded from the centralized repository
- Generating new dimensional data marts against the data stored in the data warehouse is a relatively simple task

Advantages:

Your organization realizes a focused use of resources from the individual managed application.

- The first implementation becomes a showcase for the identity management solution.
- When the phases are completed for the managed application, you have implemented a deeper, more mature implementation of the identity management solution.
- Operation and maintenance resources are not initially impacted as severely as with the bottom-up approach.

Disadvantages:

- The top-down methodology is that it represents a very large project with a very broad scope.
- The solution provides limited coverage in the first phases.
- A minimal percentage of user accounts are managed in the first phases.
- You might have to develop custom adapters at an early stage.
- The support and overall business will not realize the benefit of the solution as rapidly.
- The implementation cost is likely to be higher.

Bottom-up design

- *bottom-up* approach data marts are first created to provide reporting and analytical capabilities for specific business processes

Data marts contain, primarily, dimensions and facts

- Facts can contain either atomic data and, if necessary, summarized data.

Advantages:

- User and business awareness of the product. Benefits are realized in the early phases.
- You can replace many manual processes with early automation.

Data Warehousing & Mining

- You can implement password management for a large number of users.
- You do not have to develop custom adapters in the early phases.
- Your organization broadens identity management skills and understanding during the first phase.
- Tivoli Identity Manager is introduced to your business with less intrusion to your operations.

Disadvantages:

The organizational structure you establish might have to be changed in a later roll-out phase.

- Because of the immediate changes to repository owners and the user population, the roll-out will have a higher impact earlier and require greater cooperation.
- This strategy is driven by the existing infrastructure instead of the business processes.

Difference between OLAP & OLTP

Characteristics	OLAP	OLTP
1. Analytical capabilities	moderate	Very low
2. Data for single session	Small to medium	Very limited
3. Size of result	Large	Small
4. Response time	Fast to moderate	Very fast
5. data granularity	Detail and summary	Detail
6. data currency	Current & historical	current
7. Access Method	Predefined & ad-hoc	Predefined
8. Basic motivation	Provide Information	Collect & input data
9. optimization of database	For analysis	For transaction
10. scope of user interaction	Through out data content	Single transaction

Data Warehousing & Mining

b) Explain partitioning methods for clustering.

10

Ans:

Partitioning methods for clustering.

- Nonhierarchical
- Creates clusters in one step as opposed to several steps.
- Since only one set of clusters is output, the user normally has to input the desired number of clusters, k .
- Usually deals with static sets.

Partitioning method for clustering.

- K-means clustering
- Nearest neighbor
- K-medoids

K-means:

- Initial set of clusters randomly chosen.
- Iteratively, items are moved among sets of clusters until the desired set is reached.
- High degree of similarity among elements in a cluster is obtained.
- Given a cluster $K_i = \{t_{i1}, t_{i2}, \dots, t_{im}\}$, the *cluster mean* is
- $m_i = (1/m)(t_{i1} + \dots + t_{im})$

K-Means Algorithm

Input:

$D = \{t_1, t_2, \dots, t_n\}$ // Set of elements
 A // Adjacency matrix showing distance between elements.
 k // Number of desired clusters.

Output:

 K // Set of clusters.

K-Means Algorithm:

assign initial values for means m_1, m_2, \dots, m_k ;

repeat

 assign each item t_i to the cluster which has the closest mean ;

calculate new mean for each cluster;

until convergence criteria is met;

Nearest Neighbor

- Items are iteratively merged in the existing clusters that are closest.
- Incremental
- Threshold, t , used to determine if items are added to existing clusters or a new cluster is created.

Data Warehousing & Mining

Input:

$D = \{t_1, t_2, \dots, t_n\}$ // Set of elements

A // Adjacency matrix showing distance between elements.

Output:

K // Set of clusters.

Nearest Neighbor Algorithm:

$K_1 = \{t_1\};$

$K = \{K_1\};$

$k = 1;$

for $i = 1$ to n do

find the t_m in some cluster K_m in K such that $dis(t_i, t_m)$ is the smallest;

if $dis(t_i, t_m) \leq t$ then

$K_m = K_m \cup t_i$

else

$k = k + 1;$

$K_k = \{t_i\};$

K-medoid:

■ **Partitioning Around Medoids (PAM) (K-Medoids)**

■ Handles outliers well.

■ Ordering of input does not impact results.

■ Does not scale well.

■ Each cluster represented by one item, called the *medoid*.

■ Initial set of k medoids randomly chosen.

Input:

$D = \{t_1, t_2, \dots, t_n\}$ // Set of elements

A // Adjacency matrix showing distance between elements.

k // Number of desired clusters.

Output:

K // Set of clusters.

PAM Algorithm:

arbitrarily select k medoids from D ;

repeat

for each t_h not a medoid do

for each medoid t_i do

calculate TC_{ih} ;

find i, h where TC_{ih} is the smallest;

if $TC_{ih} < 0$ then

replace medoid t_i with t_h ;

until $TC_{ih} \geq 0$;

for each $t_i \in D$ do

assign t_i to K_j where $dis(t_i, t_j)$ is the smallest over all medoids;

Data Warehousing & Mining

Q6. a) Explain Different OLAP operations.

10

Ans:

Different OLAP operations. The analyst can understand the meaning contained in the databases using multi-dimensional analysis. By aligning the data content with the analyst's mental model, the chances of confusion and erroneous interpretations are reduced. The analyst can navigate through the database and screen for a particular subset of the data, changing the data's orientations and defining analytical calculations. The user-initiated process of navigating by calling for page displays interactively, through the specification of slices via rotations and drill down/up is sometimes called "slice and dice". Common operations include slice and dice, drill down, roll up, and pivot.

Slice: A slice is a subset of a multi-dimensional array corresponding to a single value for one or more members of the dimensions not in the subset.

Example:

Dice: The dice operation is a slice on more than two dimensions of a data cube (or more than two consecutive slices).

Example:

Drill Down/Up: Drilling down or up is a specific analytical technique whereby the user navigates among levels of data ranging from the most summarized (up) to the most detailed (down).

Example:

Roll-up: A roll-up involves computing all of the data relationships for one or more dimensions. To do this, a computational relationship or formula might be defined.

Example:

Pivot: This operation is also called rotate operation. It rotates the data in order to provide an alternative presentation of data - the report or page display takes a different dimensional orientation.

Example:

Data Warehousing & Mining

b) Given the training data for height classification, Classify the tuple $t = \langle \text{rohit, M, 1.9m} \rangle$ using bayessian classification

10

Ans :

Probability Table for given data :

Gender		Count			Probability		
		Short	Medium	Tall	Short	Medium	Tall
Gender	M	1	2	3	1/4	2/8	3/3
	F	3	6	0	3/4	6/8	0/3
Height	0-1.6	2	1	0	2/4	1/8	0
	1.6-1.7	2	0	0	2/4	0	0
	1.7-1.8	0	0	3	0	3/8	0
	1.8-1.9	0	0	3	0	3/8	0
	1.9-2.0	0	1	0	0	1/8	0
	2.0-2.1	0	0	1	0	0	1/3
	2.1-2.2	0	0	2	0	0	2/3
	2.2- ∞	0	0	0	0	0	0

Male:

$$P(\text{short}) = 1/4 = 0.25$$

$$P(\text{medium}) = 2/8 = 0.25$$

$$P(\text{tall}) = 3/3 = 1$$

Given Tuple is: $t = \langle \text{Rohit, M, 1.9 M} \rangle$

$$P(t/\text{short}) = 1/4 \times 0 = 0$$

$$P(\text{medium}) = 0.25 \times 1/8 = 0.03125$$

$$P(\text{tall}) = 1 \times 0 = 0$$

$$\text{Likelihood of being short} = 0.25 \times 0 = 0$$

$$\text{Likelihood of being Medium} = 0.25 \times 0.03125 = 0.0078125$$

$$\text{Likelihood of being Tall} = 1 \times 0 = 0$$

Data Warehousing & Mining

$$P(t) = 0 + 0.0078125 + 0 = 0.0078125$$

Actual Probabilities of each event:

$$P(\text{short} | t) = \frac{0 \times 0.25}{0.0078125} = 0$$

$$P(\text{medium} | t) = \frac{0.03125 \times 0.25}{0.0078125} = 1$$

$$P(\text{tall} | t) = \frac{1 \times 0}{0.07437} = 0$$

Therefore, based on these probabilities, we classify the new tuple

< Rohit, M, 1.9 M> as **Medium** because it has highest probability.

Q7. Write short note on(any four).

20

a) Web Structure Mining

Ans:

Web structure mining

- Mine structure (links, graph) of the Web
- Techniques
 - PageRank
 - CLEVER
- Create a model of the Web organization.
- May be combined with content mining to more effectively retrieve important pages.

Page rank:

- Used by Google
- Prioritize pages returned from search by looking at Web structure.
- Importance of page is calculated based on number of pages which point to it –

Backlinks.

- Weighting is used to provide more importance to backlinks coming from important pages.

$$\text{PR}(p) = c (\text{PR}(1)/N_1 + \dots + \text{PR}(n)/N_n)$$

Data Warehousing & Mining

- PR(i): PageRank for a page i which points to target page p.
- Ni: number of links coming out of page i

Clever:

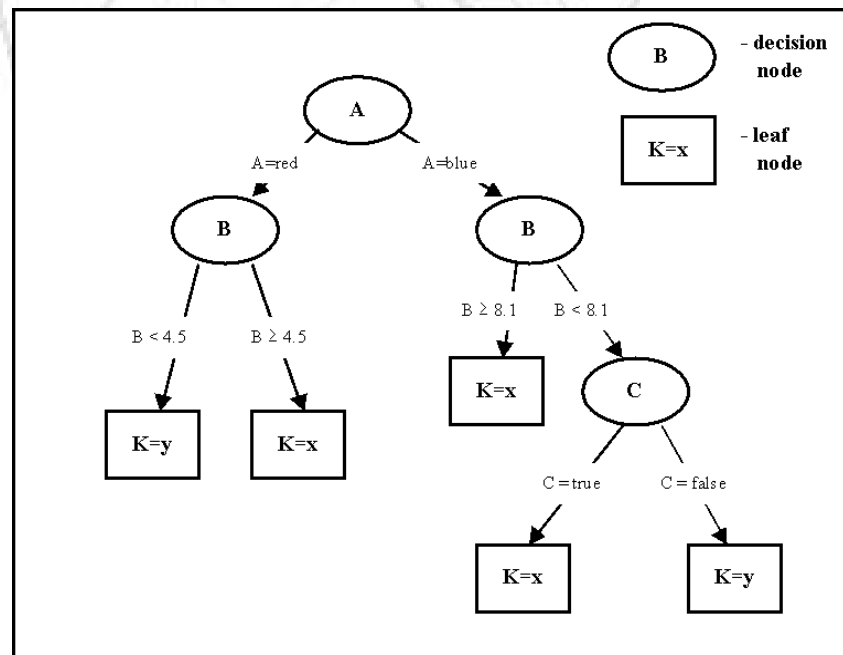
- Identify authoritative and hub pages.
- **Authoritative Pages** :
 - Highly important pages.
 - Best source for requested information.
- **Hub Pages** :
 - Contain links to highly important pages.

b) **Decision Tree based classification approach.**

Ans:

Decision tree based classification approach

- Decision Tree Classification generates the output as a binary tree-like structure, which gives fairly easy interpretation to the marketing people and easy identification of significant variables for the churn management
- A Decision Tree model contains rules to predict the target variable.
- A decision tree is a class discriminator that recursively partitions the training set until each partition consists entirely or dominantly of examples from one class.
- Each non-leaf node of the tree contains a split point that is a test on one or more attributes and determines how the data is partitioned.
- The tree is built by recursively partitioning the data.
- The initial lists created from the training set are associated with the root of the decision tree.



c) **Crawler**

Ans:

Crawlers:

- A robot is a program that traverses the hypertext structure in the web.
- The pages that crawlers start with are referred to as the seed URL.
- By starting at one page, all links from it are recorded and saved in a queue.
- These new pages are in turn searched and their links are saved.
- A crawler may visit a certain number of pages and then stop, build index and replace the existing index.
- This type of crawler is referred as a periodic crawler because it is activated as periodically.
- Crawlers are used to facilitate the creation of indices used by search engines
- They allow the indices to be kept relatively up to date with little human interaction

d) **Metadata:**

Ans:

- **Metadata** is loosely defined as data about data.
- Metadata is a concept that applies mainly to electronically archived data and is used to describe the
 - a) Definition,
 - b) Structure and
 - c) Administration of data files with all contents in context to ease the use of the captured and archived data for further use.

Types of metadata :

1. **Structural metadata:** it is used to describe the structure of computer systems such as tables, columns and indexes.

2. **Technical metadata:** correspond to internal metadata

3. **business metadata :**

- This type of metadata stores business definitions of the data, it contains high-level definitions of all fields present in the data warehouse, information about cubes, aggregates, data marts.
- Business metadata is mainly addressed to and used by the data warehouse users, report authors (for ad-hoc querying), cubes creators, data managers, testers, analysts.

Data Warehousing & Mining

e) *Web personalization*:

Ans:

Web site personalization can be defined as the process of customizing the content and structure of a Web site to the specific and individual needs of each user taking advantage of the user's navigational behaviour. The steps of a Web personalization process include:

- (a) The collection of Web data,
- (b) The modelling and categorization of these data (pre-processing phase),
- (c) The analysis of the collected data, and
- (d) The determination of the actions that should be performed.

Content-based filtering systems are solely based on individual users' preferences. The system tracks each user's behaviour and recommends items to them that are similar to items the user liked in the past.

Collaborative filtering systems invite users to rate objects or divulge their preferences and interests and then return information that is predicted to be of interest to them. This is based on the assumption that users with similar behaviour (e.g. users that rate similar objects) have analogous interests.

- *Rule-based filtering* the users is asked to answer a set of questions.
- These questions are derived from a decision tree, so as the user proceeds to answer them, what he finally receives as a result (e.g. a list of products) is tailored to his needs.
- Content-based, rule-based, and collaborative filtering may also be used in combination, for deducing more accurate conclusions.

In this work we focus on *Web usage mining*. This process relies on the application of statistical and data mining methods to the Web log data, resulting in a set of useful patterns that indicate users' navigational behaviour. The data mining methods that are employed are: association rule mining, sequential pattern discovery, clustering, and classification. This knowledge is then used from the system in order to personalize the site according to each user's behavior and profile.