Database Security :-

– The goal of database security is the protection of data against threats such as accidental or intentional loss, destruction or misuse.

Introduction to database Security Issues :-

Types of Security:

Database Security addresses many issues, which includes

- Legal and Ethical issues regarding right to access certain information. Some information are private and cannot be accessed legally by unauthorized persons.
- Policy Issues at the governmental, institutional or Corporate level as to what kind of information should not be made publicly available.
- System related Issues – physical hardware level, OS level or DBMS level.
- Multiple Security levels – Top secret, Secret, Confidential and unclassified. Security policy with respect to permitting access to various levels of data must be enforced.

Threats to Databases :-

Threats to databases result in the loss or ~~delegdr~~ degradation of some or all the following Commonly accepted Security goals :

① Loss of Availability :-

– means that data, or the system or both cannot be accessed by the users.

→ This situation may rise due to sabotage of h/w, n/ws or applications.

→ It can seriously cause operational difficulties and affect the financial performance of the an Organization.

→ Database availability refers to making objects available to a human user or a program to which they have a legitimate right.

② Loss of Integrity :-

– Data Integrity refers to the requirement that information be protected from improper modification.

– Integrity is lost if unauthorized changes are made to data by either intentional or accidental acts.

– If the loss of system or data integrity is not corrected, continued use of the contaminated system or corrupted data could result in inaccuracy, fraud or erroneous decisions.

③ Loss of Confidentiality :-

– Database Confidentiality refers to the protection of data from unauthorized disclosure.

– It can range from violation of Data privacy act to jeopardization of national security.

– Unauthorized, unanticipated or unintentional disclosure could result in loss of public confidence, embarassment or legal action against the Organization.

④. Loss of privacy :-

- refers to loss of protecting data from individuals.
- It may lead to blackmail, bribery, stealing of user passwords or legal action against the organization.

⑤ Theft and Fraud affect :-

- affects the database environment.
- It do not necessarily alter data, as is the case of loss of Confidentiality or loss of privacy.

⑥ Accidental losses

- Could be unintentional threats including human error.

Database Security Mechanisms.

DBMS typically includes a database security and authorization subsystem that is responsible for ensuring the security of portions of database against unauthorized access.

- Discretionary Security Mechanisms :
  - Used to grant privileges to users, including the capability to access specific data files, records or fields in a specified mode (such as read, insert, delete or update)
- Mandatory Security Mechanisms :
  - Used to enforce multilevel security by classifying data and users into various security classes and then implementing appropriate security policy.

Control Measures :-

Four main Control measures used to provide security of data in databases. They are

1. Access control
2. Inference Control
3. Flow Control
4. Data Encryption.

Access Control :

- preventing unauthorized persons from accessing the system itself either to obtain information or to make malicious changes to a portion of a db.
- DBMS must include provisions for restricting access to the db system as a whole.
- This fn. is called access control and is handled by creating user accounts and passwords to control the login process by DBMS.

Inference Control :-

- Statistical databases are used to provide Statistical information or summaries of values based on various criteria. (eg) database of population Statistics may provide statistics based on age groups, income levels, education level etc.
- Government Statisticians or market research firms are allowed to access the db. to retrieve some info.

- It is sometimes possible to deduce or infer certain facts concerning individuals from queries.
- This problem is called statistical db security.
- Corresponding control measures are called Inference control measures. are called Inference control measures.

Flow Control :-

- which prevents information from flowing in such a way that it reaches unauthorized users.
- Channels that are pathways for information to flow implicitly in ways that violate the security policy of an Organization are called "Covert Channels"

Data Encryption :-

- which is used to protect sensitive data (Credit card nos) that is transmitted via some type of Communication n/w.
- Encryption can be used to provide additional protection for sensitive data.
- Data is encoded using some coding algorithm. Here unauthorized cannot decrypt, only authorized can decrypt the code and access data.

DBA

— is the Central authority for managing a db System.

— DBA has a DBA account in the DBMS, sometimes called a System or Superuser account. which provides powerful capabilities.

— DBA privileged Commands include Commands for granting & revoking privileges to individual accounts, Users or User groups and perform following actions.

1. Account Creation
2. privilege granting.
3. Privilege revocation
4. Security level assignment

~~Types of Privileges~~:-

Database Access Control & Types of privileges:-

Discretionary Access Control based on Granting & Revoking Privileges

— Enforcing Discretionary access control in a db system is based on granting and Revoking privileges.

Types of Discretionary privileges:-

Authorization identifier — User accounts

— DBMS must provide selective access to each relation in the db based on specific accounts.

There are 2 levels for assigning privileges to use db system.

1. The account level.
2. The relation/table level.

Account level - DBA Specifies the particular privileges that each account holds independently of the relations in the db.

Relation/table level - DBA can control the privilege to access each individual relation or view in the db.

Account level privileges are :

1. Create Schema or Create Table - to create a Schema/relation
2. Create view privilege
3. Alter privilege.
4. Drop / Modify
5. Select privilege.

Relational level privileges ~~are~~: applied to relations or views.

- It provide privileges at the relation and attribute level only.

- Granting & Revoking privilege generally follows an authorization model for discretionary privileges known as "Access Matrix Method", where the rows of a matrix M represent Subjects (users, accounts, pgms) and Columns represent Objects (relations, records, Columns, views, operations) Each position M(i,j) in the matrix represents the types of privileges (read, write, update) that Subject i holds on Object j.

To Control the granting & Revoking of relation Privileges, each Relation R in a database is assigned to an Owner account

DBA can assign an Owner to a whole Schema. Owner account can grant privilege to other users by granting privileges to their accounts.

Following types of privileges.

- Selectact privilege on R
- Modify privilege on R
- Regerences privilege on R

Specifying Privileges Using Views

— Mechanism of views is important discretionary mech.

if Owner A of relation R wants another account B to be able to retrive some fields of R.

A can Create View of R that includes some field and grant Select on V to B.

Revoking Privileges:

Owner of Relation may want to grant the Select privilege to a user for a Specific task and then revoke that privilege once the task is Completed.

Propagation of privileges using the GRANT option.

GRANT CreateTab to A_1.

GRANT INSERT, DELETE ON EMPLOYEE TO A_L;

GRANT update (A_1, A_2) ON R_1 TO User4;

Grant <privilage list> ON <relation or View name> To <user list> [with GRANT option];

REVOKE <privilege list> ON <relation or View name> from <user-list>

Audit Trial at page 7

Mandatory Access Control:-

- In DAC – user has or does not have a certain privilege.
- An additional Security policy is needed that classifies data and users based on security classes. - known as Mandatory Access Control.
- Multilevel security exist in gov, mil, intelligence app/:
- Security class Topsecret (TS), Secret (S), Confidential (C) and Unclassified (U) TS is highest level and U the lowest.

$$TS \geq S \geq C \geq U$$

- Model used for multilevel security is known as Bell-La Padula model which classifies each subject (User, account, pgm) and object (relation, tuple, column, view, operation) into one of security classification TS, S, C or U.

We will refer to clearance of ~~security classifications~~ subject as class(S) and to the classification of an object O as class(O)

Two restrictions are enforced on data access based on subject/object classifications:

1. A Subject S is not allowed read access to an object O unless class(S) $\geq$ class(O) This is known as the Simple security property

2. A Subject S is not allowed to write an object O unless class(S) ≤ class(O). This is known as the Star property (or) * property.

① States that no subject can read an object whose security classification is higher than the subject's security clearance.

② prohibits a subject from writing an object at a lower security classification than the subject's security clearance.

Apparent key :
— is a set of attributes that would have formed the primary key in regular (single level) relation.

Filtering :-
— it is possible to store a single tuples at a ~~higher level~~ classification level and produce the corresponding tuples at a lower level classification through a process known as filtering.

Polyinstantiation :
— where several tuples can have the same apparant key value but have different attribute values for users at different classification levels.

Multilevel relation.

Employee

Name	Salary	Job performance	Tot. Credit
Smith U	40000 C	Fair S	S
Brown C	80000 S	Good C	S

Original Employee tuples

Name	Salary	Jobperformance	Tot_Credit
Smith U	40000 C	Null C	C
Brown C	Null C	Good C	C.

After filtering for classification C users.

Smith U	NULL U	NULL U	U

after filtering for U users.

Smith U	40000 C	Fair S	S
Smith U	40000 C	Excellent C	C
Brown C	80000 S	Good C	S

Polyinstantiation of Smith tuple

DAC

1) High degree of flexibility Suitable for large variety of domains
2) do not impose any control on how information is propagated and used ~~Once~~
5) Malicious attacks Trojan horses embedded

MAC

1) Ensures high degree of protection.
2) Prevent illegal flow of information.
3) Suitable for military types. appl:
4) Too rigid as it require strict classification of subjects & objects
5) Hence not many appl: DAC is used.

Role Based Access Control:

- RBAC emerged in 1990s for managing & enforcing security in large scale enterprise wide systems.
- permission are associated with Roles. Users are assigned appropriate roles.
- Roles can be created using CREATE ROLE and DESTROY ROLE Commands. GRANT ad REVOKE Commands can then be used to assign ad revoke privileges.
- RBAC is a way to Organize roles to reflect the Organization's lines of authority and responsibility.
- RBAC is used for addressing key requirements of web based applications.
- It has desirable features
  - Flexibility
  - policy neutrality
  - better support for Admin & security management.
- mainly used in web based applications where DAC & MCA lack capabilities.
- Easier deployment. made it a Superset model.

XML Access Control:-

- XML is world wide used in Commercial & Scientific appl/n.
- Security Standards of XML are digital Signatures and encryption Standards

- XML language can play a key role in accen control for E-Commerce appl!:
- Credential is a set of properties concerning a user that are relevant for security purposes.
- XML used for specifying credentials cd accen control policies secure credential submission cd export of accen control policies.

PGP - pretty Good privacy
SSL - Secure Socket layer.

Audit Trails

- a special file or database in which syste automatically keeps track of all operations performed by users on regular data.
  - It is a log of all changes (update, delete insert so on)
  - aids security to database

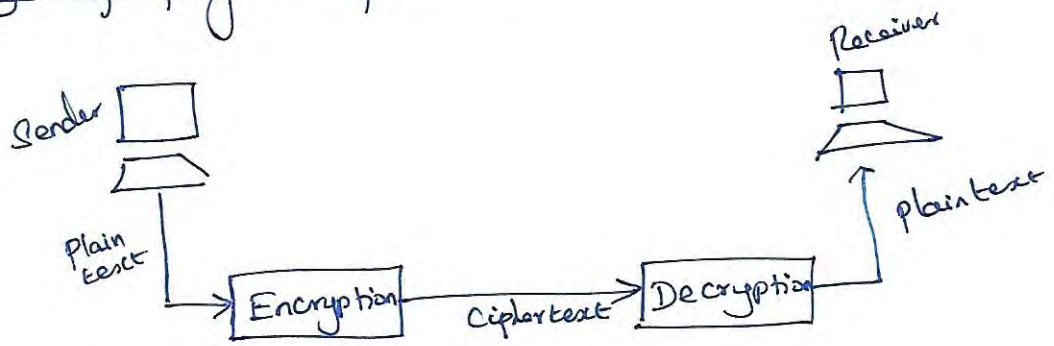
Entries in Audit Trail file

1) request
2) Terminal from which operation was evoked
3) User who evoked the operation.
4) date and time of the "
5) tuples, attributes affected
6) old values
7) New values.

Cryptography

Cryptography means "Secret Writing" in Greek.
It can provide Confidentiality, integrity, authentication and non repudiation of messages.
It can also provide entity authentication.

Cryptography Components

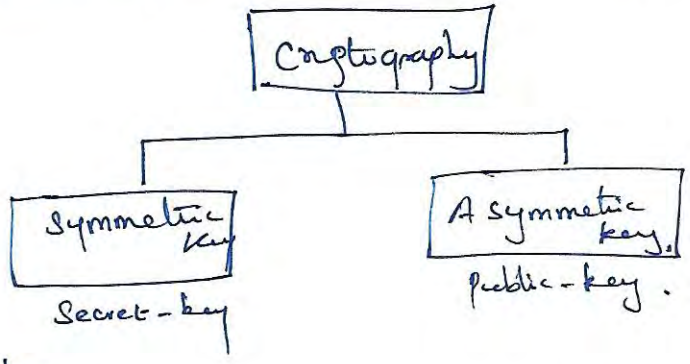


"Encryption" is a means of maintaining secure data in an insecure environment.
"An Encryption algorithm" transforms the plaintext into Cipher text
"A decryption algorithm" " " ciphertext to Plaintext.
"Ciphertext" – is a scrambled message produced as o/p. It depends on the plaintext and the key. For a given msg. 2 different keys will produce two different ciphertexts.

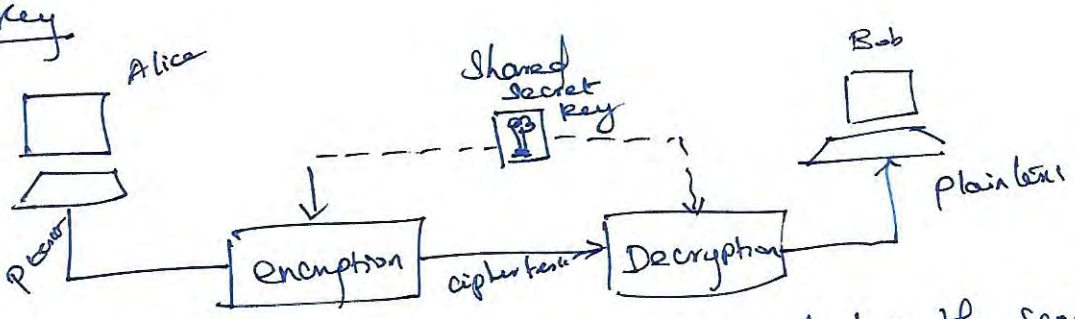
Public & Private keys:

These are a pair of keys that have been selected so that if one is used for encryption, the other is used for decryption. The exact transformation performed by the encryption algorithm depend on the public or private key that is provided as i/p.

Two Categories of Cryptography :

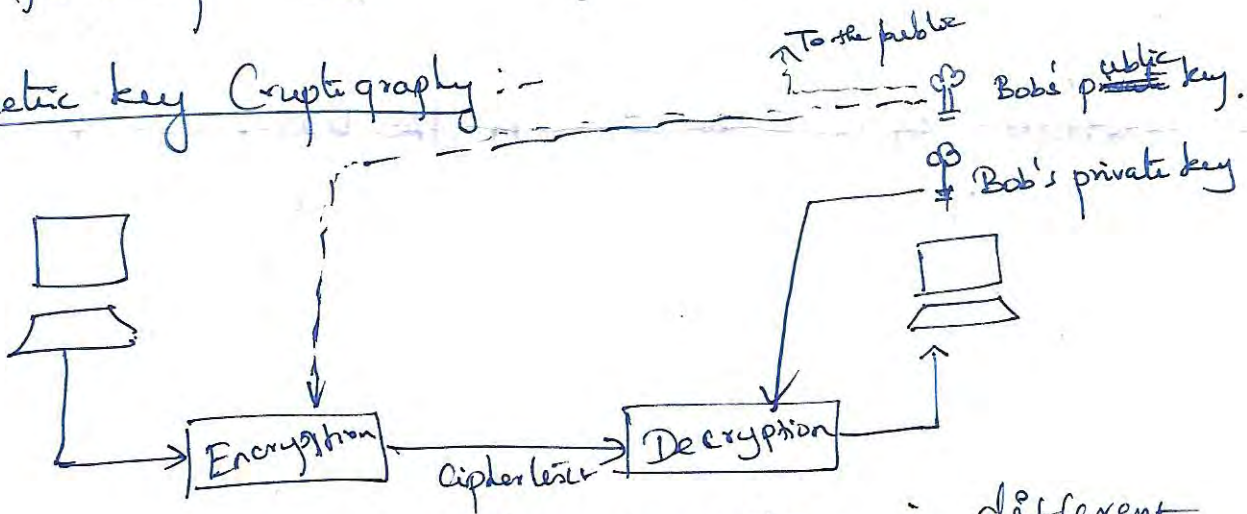


Symmetric key



In Symmetric key, the same key is used by the sender for encryption and receiver for decryption, the key is shared.

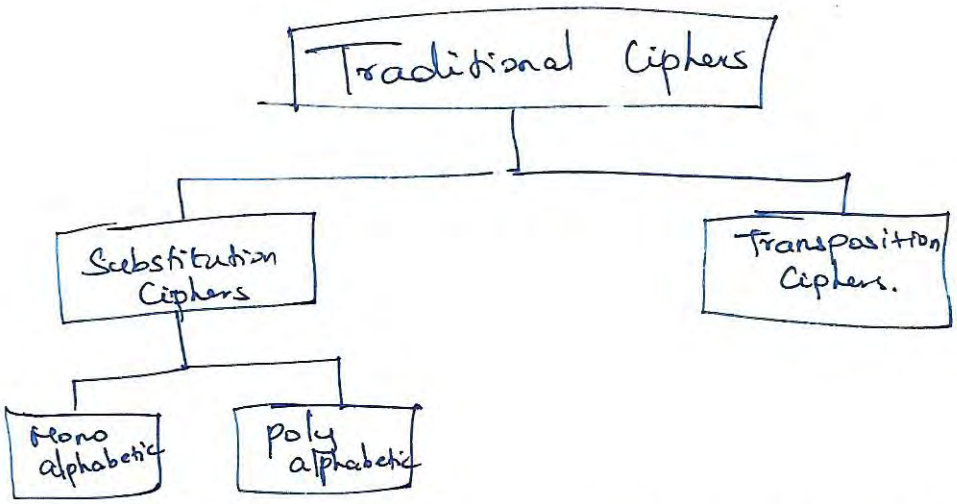
Asymmetric key Cryptography :-



Public key is used for encryption is different from the private key that is used for decryption.

Data Encryption Standard (DES)

- was designed by IBM and adopted by the U.S gov. as the Standard encryption method
- It can provide end-to-end encryption on the Channel b/w the Sender A and Receiver B.



Substitution — replaces one symbol with another

(eg) we replace Character A with D.

or

Replace numbers 3 with 7

Monoalphabetic Cipher

- a Character or a Symbol in the plaintext is always changed to the same character in the ciphertext regardless of its position in the text. (eg) if algorithm says that Character A in plaintext is changed to D every character A is changed to Character D.

Poly alphabetic Cipher: each occurence of a Character can have a different Substitute. (eg) Char. A. Could be Changed to D in the beginning and N in the middle. We can also divide the group of Characters into groups of 3.

Plain Text : HELLO

Cipher text : KHOOR is monoalphabetic.

Plain Text : HELLO

Cipher text : ABNZF not polyalphabetic.

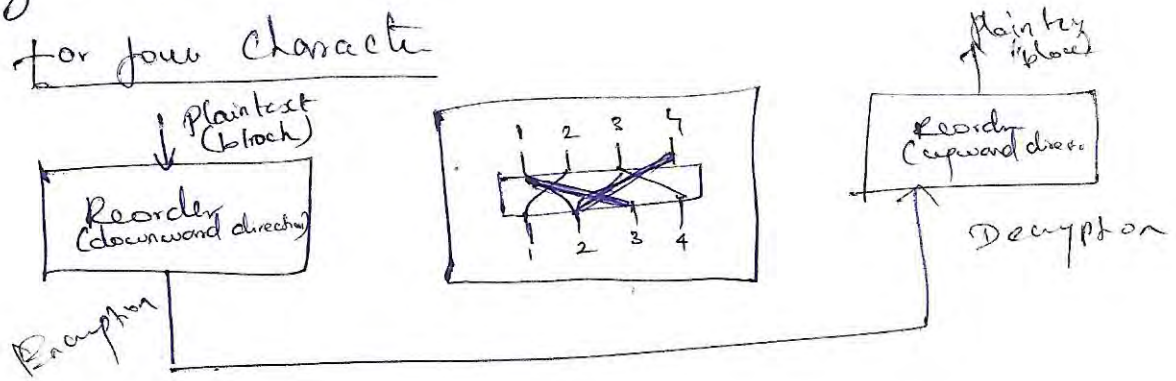
Shift Cipher : Key is: HELLO

WTAAD

Transposition Ciphers :-

— no substitution instead their locations Change. reorders symbols in a block of symbols.

for four characte



Encrypt HELLO MY DEAR

1) Remove Spaces
2) Divide the text into block of four characters
3) Add a bogus char. Z

HELL OMYD EARZ

Create Cipher

ELHLMDOYAZER

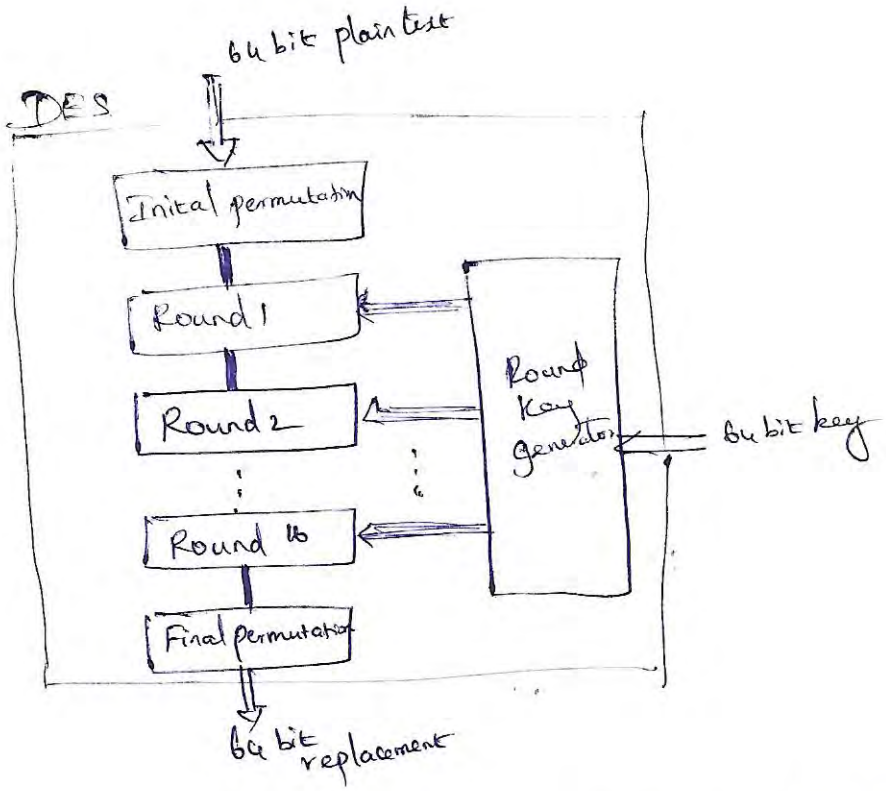
Data Encryption Standard (DES)

— is a System developed by the U.S. government by IBM.

— It can provide end-to-end encryption on the Channel between Sender A and Receiver B.

— DES algorithm is a careful and complex Combination of two of the fundamental building blocks of encryption (Substitution and permutation (transposition)

— It derives its strength from repeated application of these two techniques for a total of 16 cycles.



Questioning the adequacy of DES. NIST (National Institute of Standards) introduced (Advanced Encryption Standards) AES.

Chose "Rijndael algorithm" named after Inventors Vincent Rijmen & Joan Daemen.

- AES is a very Complex round cipher.
- AES Designed with 3 key sizes.

Size of block	Number of Rounds	Key sizes
128 bits	10	128 bits
	12	192 bits
	14	256 bits

Public key Encryption : / Asymmetic key :

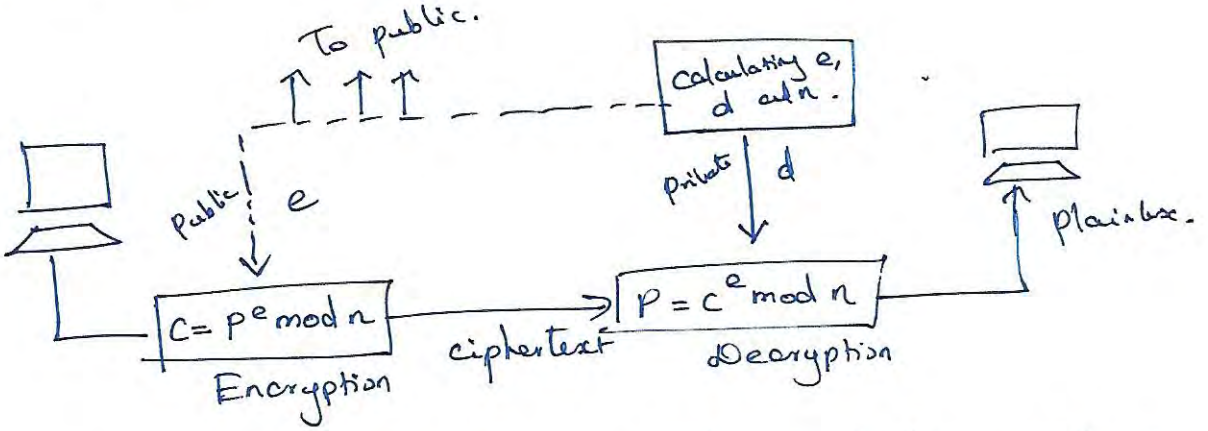
In 1976, Diffie and Hellman proposed a new kind of Crypto system called Public key crypto system.

- It is based on mathematical fns.
- They also Envolve the use of two seperate keys
  - Public key
  - Private key.

RSA public key Encryption Algorithm :-

— introduced by Ron Rivest, Adi Shamir and Len Adleman at MIT — RSA Scheme.

It uses 2 nos. e and d as public & private key.



— RSA algo. operates with modular arithmetic — mod n.

Two keys e & d have a special relationship to each other. How to calculate keys?

Selecting keys :

1) Choose 2 very large prime no. P & Q
(Prime no. divided evenly only by 1 and itself)

2) Multiply the 2 primes to find n.
(i.e) $n = P \times Q$.

3) Calculate another number $\phi = (P-1) \times (Q-1)$

4) Choose a random integer e. such that e and ϕ is relatively prime. Then calculate d so that ~~$d * e = 1 \bmod \phi$~~ $(e \times d) \bmod \phi = 1$.

5) Receiver (Bob) announces e and n to the public, he keeps ϕ & d secret.

For encryption Cipher text is calculated using e and n. Using.

$$C = P^e \pmod n$$

For Decryption, φ and d private

$$P = C^d \pmod n$$

Example:

Prime nos $P = 7$

$q = 11$

$n = P \times q = 77$

$\varphi = (7-1)(11-1) = (6)(10) = 60$

If he Chooses $e = 13$, then

$d \times e = 1 \bmod \varphi$

$d \times 13 = 1 \bmod 60$

$d = 37$

Alice Sends 5 to Bob.

She uses public key 13 to encrypt 5

Plain text 5

$C = 5^{13} = 26 \bmod 77$

Ciphertext = 26

Bob receives the Cipher text 26 and uses the private key 37 to decipher the Ciphertext

Cipher text = 26

$P = 26^{37} = 5 \bmod 77$

plain text = 5

1220703125

In RSA given 2 prime nos. $p = 23, q = 19$ find n and φ Choose $e = 5$ and try to find d such that e and d meet the criteria.

(eg)

$p = 3 \quad q = 11 \quad P = 5$.

$n = 3 \times 11 = 33$

$\phi = (p-1)(q-1) = (2 \times 10) = 20$.

$e = 7$ (factors of $7 = 1, 7$) factors of $20 (1, 2, 4, 5, 10, 20)$

$(d \times e) \bmod \phi = 1$

$d \times 7 \bmod 20 = 1 \quad \therefore d = 3$

~~3~~ $3 \times 7 \bmod 20 = 1$

$21 \% 20 = 1 \quad 1 = 1 \quad d = 3$

if $P = 2$.

$C = 2^7 \bmod 33$

$= 128 \bmod 33 = 29$.

$P = 29^3 \bmod 33$

$P = 2$.

Plain text – 5

$C = 5^7 \bmod 33 \quad (C = P^e \bmod n)$

$= (5^3 \bmod 33)(5^3 \bmod 33)(5 \bmod 33)) \bmod 33$

$C = 14$.

$P = C^d \bmod n$.

$= 14^3 \bmod 33 \quad = 14^3 \bmod 33 = 5$

$P = 5$

$M = 5 \rightarrow [C = 5^7 \bmod 33] \xrightarrow{C=14} [P = 14^3 \bmod 33] \xrightarrow{P=5}$

Digital Signature

– is an example of using encryption techniques to provide authentication services in E-Commerce application.

– Digital signature is a means of associating a mark unique to an individual with the body of text.

- Mark made is unforgettable, Unique so that others can be able to check that the signature comes from the Originator.
- It Consists of a string of symbols.
- If it is same for each message then one can easily Counterfeit by simply Copying Symbols.
- Hence digital signature along with timestamp
- There should also be a Unique Secret no. of the Signer.

Privacy Issues and Preservation :-

- Preserving data privacy is a growing challenge for database security
- In order preserve privacy avoid Central warehouses as a single repository of data.
- If all data is available at a single warehouse violating only one will expose all data.
- Therefore avoid Central warehouses and use distributed datamining algorithms minimizes exchange of data
- (Ex) medical and legal records

Challenges of Database Security :-

Considering the vast growth in volume and speed of threats to databases research effort needs to be devoted to the following issues

1) Data Quality
2) Intellectual property Rights
3) Database Survivability.
   - database in addition to prevent an attack also should detect in the event of occurrence.
   - Containment — take immediate action to eliminate attackers access to prevent further.
   - Damage Assessment — determine the extent of problem including failed fns. and corrupted data.
   - Reconfiguration — allow the operation to continues in a degraded mode while recovery proceeds
   - Repair — Recover Corrupted or lost data (or) reinstall failed system functions to reestablish a normal level of operation.
   - Fault Treatment — Identify the weakness exploited in the attack and take steps to prevent a recurrence.

Statistical Database Security :-

- Used mainly to produce Statistics
- db may Contain Confidential data about individuals which should be protected from User access.

Consider person relation

Name	SSN	Income	Address	city	State	zip	sex	last_degree

Condition can be given

Sex = 'F' and last_degree = "M.S" or Ph.D" ad

City = "Newyork";

can be given.

but name = "Jennifer" cannot be given.

- You can apply only statistical queries
- Stastical db must prohibit queries that reteiere attibute ~~about~~ of individual data, but by allowing only queries that envolve statistical aggregate functions such as COUNT, SUM, MIN, MAX AVERAGE and STANDARD DEVIATION. — called Statistical Queries.
- In some cases it is possible to infer the values of individual tuples. if you are interested to find Salary of "Jane Smith" and you know She has a Ph.D degree and lives in the city Bellaire, Texas. we issue the Statistical Query Q, ~~use~~ with the following condition.

(last_degree = 'Ph.D' and Sex = 'F' and city = 'Bellaire' and State = 'Texas')

- The possibility of inferring individual information from statistical queries is reduced if no statistical queries are permitted whenever the no. of tuples in the population specified by the selection condition falls below some threshold.
- Prohit queries which retrieves same info. again.
- introduce slight inaccuracies or noise into the results
- Partitioning the database and stored in groups

Distributed Databases & Architecture

Paper Unit IV

Distributed Transactions :-

→ In a distributed system, a transaction initiated at one site can access and update data at several other sites too.

→ Transaction that accesses and updates data only at the site where it is initiated - local Transaction

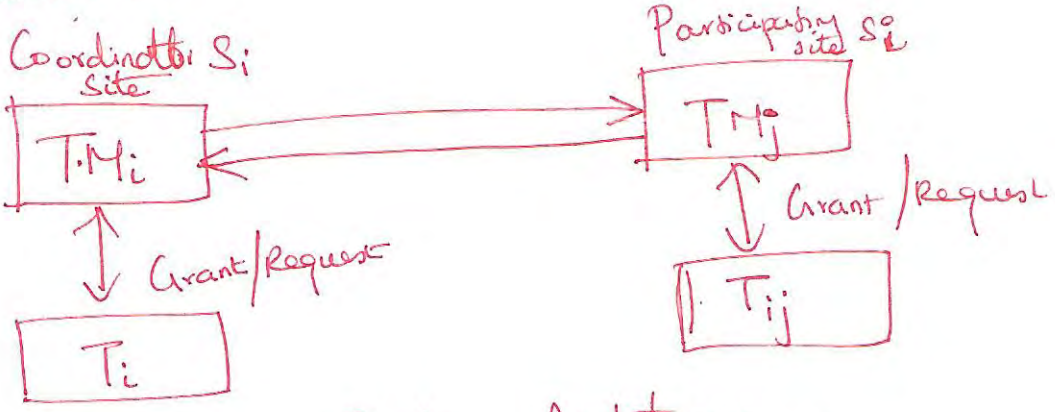
- Transaction that accesses and updates data from several sites or site other than at which it is

initiated is called – global Transaction.

— A transaction that requires data from remote sites is broken into one or more sub transactions.

— A site at which transaction is initiated is called "Co-ordinator site" and sites where subtransactions are executed "participating sites"

— Each site has its own local Transaction manager (TM) that manages the execution of those transactions



System Architecture.

Data mining Techniques / Type of knowledge discovered during DM

— Used to mine different rules and patterns.

The result of mining may be to discover the following type of new information.

- Association Rules — (eg) Whenever a Customer buys a Video player, he or she buys a CD
- Sequential patterns — Suppose a customer buys a camera and within 3 month buys a photographic supplies, then in six months buys an another accessory item.
- Classification trees — (eg) customers may be classified by frequency of visit, types of financing used, amount of purchase — some revealing statistics may be generated.
- Clustering — A given population of events or items can be partioned into sets of similar elements. (eg) an entire population of treatment data on a disease may be divided into groups based on similarity of side effects produced.

(2) categorized into five groups ⟷ most likely to buy / least likely to buy.

→ Categories of users (clusters)

Association Rules :-

- Market Basket Model, Support and Confidence.

— A database is regarded as a collection of transactions, each involving a set of items.

-(Eg) Market - Basket data.

— a set of items a consumer buys in a Supermarket during one visit.

Consider four such transactions.

Transaction-id	Time	items bought
101	6:35	milk, bread, cookies, juice
792	7:38	milk, juice
1130	8:05	milk, eggs
1735	8:40	bread, cookies, juice.

(eg) transaction in Market-Basket Model.

A association rule is of the form $X \Rightarrow Y$ where $X = \{x_1, x_2, \ldots x_n\}$ and $Y = \{y_1, y_2 \ldots y_m\}$ are set of items with x_i and y_j being distinct items for all i and all j.

Association Rule has the form

LHS $\Rightarrow$ RHS where LHS & RHS are sets of items

LHS ∪ RHS is called an Itemset.
Set of items purchased by customers.

Association rule should satisfy some interest measure. Two Common interest measures are:

(1) Support
(2) Confident.

Rule

Support for the rule LHS ⇒ RHS
- it refers how frequently a specific itemset occurs in the database

buys (x, "milk") ⇒ buys (x, "eggs") [1%, 60%]

(a) percentage of transaction in D that Contain LHS ∪ RHS

Support (A ⇒ B) = P(A ∪ B)

Confidence

The Confidence of the Rule LHS ⇒ RHS is Computed as the Support (LHS ∪ RHS) / Support (LHS)

(a) Confidence (A ⇒ B) = P(B/A)

Consider the following 2 Rules

milk ⇒ juice
bread ⇒ juice.

Look (5.1) Support of {milk, juice} is 50%
Support of {bread, juice} is 25%.

Confidence of milk ⇒ juice is 66.7%

Meaning that 3 transactions in which milk occurs, 2 contains juice.

∴ 2/3 = 0.667 ⇒ 66.7%.

Confidence of bread ⇒ juice

Meaning that one of two transactions Containing bread also Contains juice

∴ 1/2 = .50 = 50%.

Rules: Association Rule mining can be viewed as a 2 step process:

1) Find all frequent itemsets: by definition each of these itemsets will occur at least as frequently as a predetermined minimum support Count min. sup

2) Generate Strong association rules from the frequent itemsets, By def, these rules must satisfy min Support and min Confidence.

Def: Itemsets that have a support that exceeds the threshold — large item sets

A Subset of a large itemset must also be large (exceeds min. required support — downward closure

A Superset of a small itemset is also small (i.e does not have enough support) — antimonotonicity

Apriori Algorithm

- Finding frequent itemsets using Candidate generation.
- It employs an iterative approach known as level-wise search, where k-itemsets are used to explore (k+1) itemsets.
- First set of frequent 1 itemsets is found by scanning the database to accumulate the count of each item and collecting those items that satisfy min. support.
- The resulting set is denoted by L1, Next L1 is used to find L2 the set of frequent 2 itemset which is used to find L3 and so on. until no more frequent k-itemsets can be found.
- Finding of each L_k requires one full scan of the database.

Algorithm:-

Input: Database of m transactions, D and a min. support, min_s, represented as a fraction of m.

Output: Frequent itemsets $L_1, L_2, \ldots L_k$

Begin /* steps or statements are numbered */

1. Compute support(ij) = count(ij)/m for each individual item $i_1, i_2 \ldots i_n$ by scanning the db once and counting the no. of transactions that item i_j appears in.
2. The candidate frequent 1-itemset, C_1 will be the set of items $i_1, i_2 \ldots i_n$.
3. The subset of items containing i_j from C_1 where support(i_j) >= min s becomes the frequent 1-itemset, L_1;
4. k = 1 termination = false;

repeat

1. L_{k+1} = ;
2. create the candidate frequent (k+1) itemset, C_{k+1} by combining members of L_k that have k-1 items in common
3. In addition, only consider as element of C_{k+1} those k+1 items such that every subset of size k appears in L_k
4. Scan the db once and compute the support for each member of L_{k+1}; if the support for a member of L_{k+1} >= min s then add that member of L_{k+1};
5. If L_{k+1} is empty then termination = true

else

k = k+1;

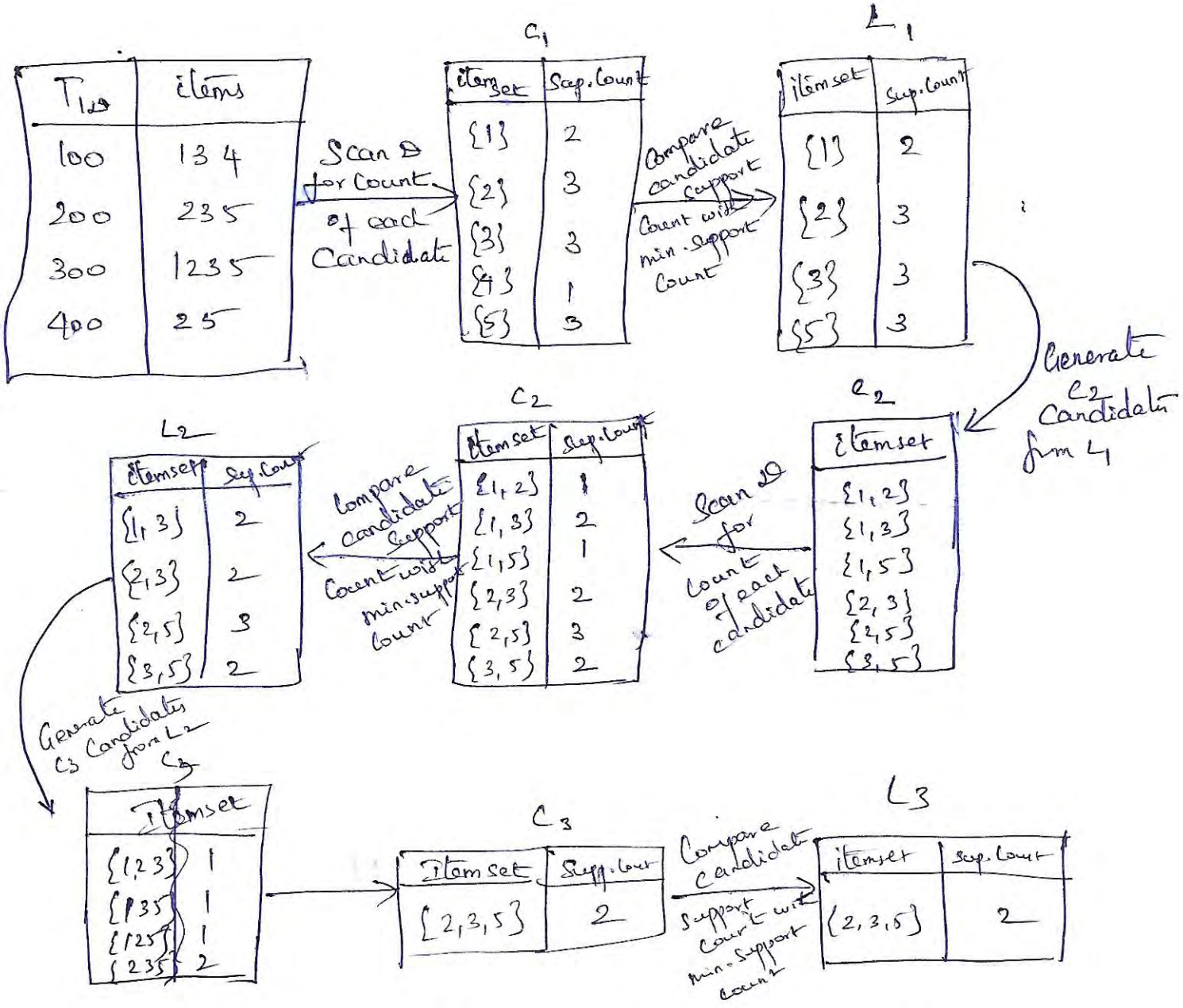
Until termination

End;

(Eg) There are four transactions in this database |D| = 4 and min. support count is taken as 2
Use apriori algorithm for finding frequent itemset in D.

Transactional db. D.

Tid	items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5



Since we have only one itemset in C_3 the algorithm terminates, having found all of frequent Itemset.

Sampling Algorithm :

- Idea is to select a small sample, one that fits in main memory of the db transactions and to determine frequent itemsets from that sample.
- If those frequent itemsets form a superset of the frequent itemsets for the entire db.
- In rare cases some frequent itemsets may be missed and a second scan of the db is needed.
- To decide whether any frequent itemsets have been missed the concept the concept of negative border is used.
- negative border of frequent item set contains the closest itemsets that could also be frequent.

(Eg) Consider set of items $I = \{A, B, C, D, E\}$

Let the combined frequent item sets of size 1 to 3 be $S = \{\{A\}, \{B\}, \{C\}, \{D\}, \{AB\}, \{AC\}\{BC\}\{AD\}\{BD\}\{ABC\}\}$

Negative Border $\{\{E\}, \{BD\}, \{ACD\}\}$
- $\{E\}$ ↓ Not Contained in S
- $\{BD\}$ ↖ Not Contained in S
- $\{ACD\}$ ↘ AC AD Contained in S.

Negative Border is used to ensure that no large itemsets are missed from analysing sample data

Frequent Pattern Tree Algorithm :-

- Apriori generates and test a very large candidate itemsets.
- FP growth alliminates the generation of large no. of candidate itemsets.

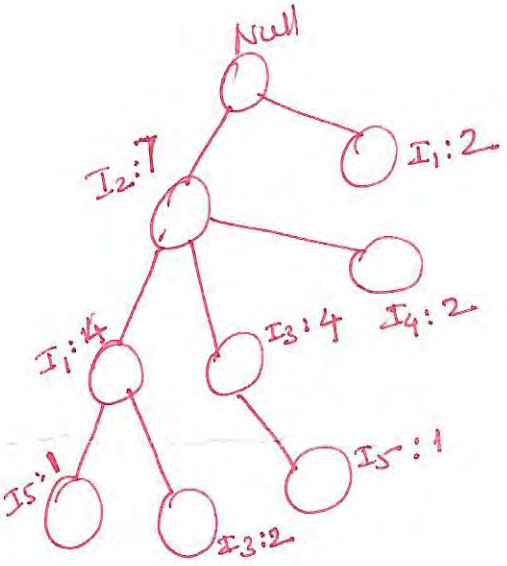
→ The Algorithm first produces a Compressed of db in terms of an FP tree (frequent pattern tree)

→ FP tree stores relevant itemset information and allows for efficient discovery of frequent itemsets

– Actual mining adopts divide and conquer strategy – decomposed into smaller sub tasks.

(21)

Tid	List of items
T_{100}	I_1, I_2, I_5
T_{200}	$I_2 I_4$
T_{300}	$I_2 I_3$
T_{400}	$I_1 I_2 I_4$
T_{500}	$I_1 I_3$
T_{600}	$I_2 I_3$
T_{700}	$I_1 I_3$
T_{800}	$I_1 I_2 I_3 I_5$
T_{900}	$I_1 I_2 I_3$.



item id	Support Count	Node_link
I_2	7	–
I_1	6	
I_3	6	
I_4	2	
I_5	2	

Partition Algorithm :-

- Support of all of them can be tested in one scan by using a partitioning technique
- partitioning divides the database into non-overlapping subsets, these are considered as seperate databases and large itemsets for the partition called "local frequent itemsets" are generated in one pass.
- Then Apriori algorithm is used efficiently on each partition. Partitions are choosen in such a way that they can fit in main memory.
- At the end of pass one, Union of all frequent itemsets, this forms "global candidate frequent itemsets"
- If min. Support ad Confidence are more, then goes for pass 2.
- It is naturally used in parallel and DDBs. for better efficiency.

Classification :

- Classification is the process of learning a model that describes different classes of data.

(Eg) Banking Application – Customers who apply for a Credit card may be classified as poor risk, fair risk or good risk.

- First ~~step~~ step in learning a model is accomplished by using a training set of data that has already been classified
- Each record in a training ~~attribute~~ data contains an attribute called the class label. – which indicates which class the record belongs to.

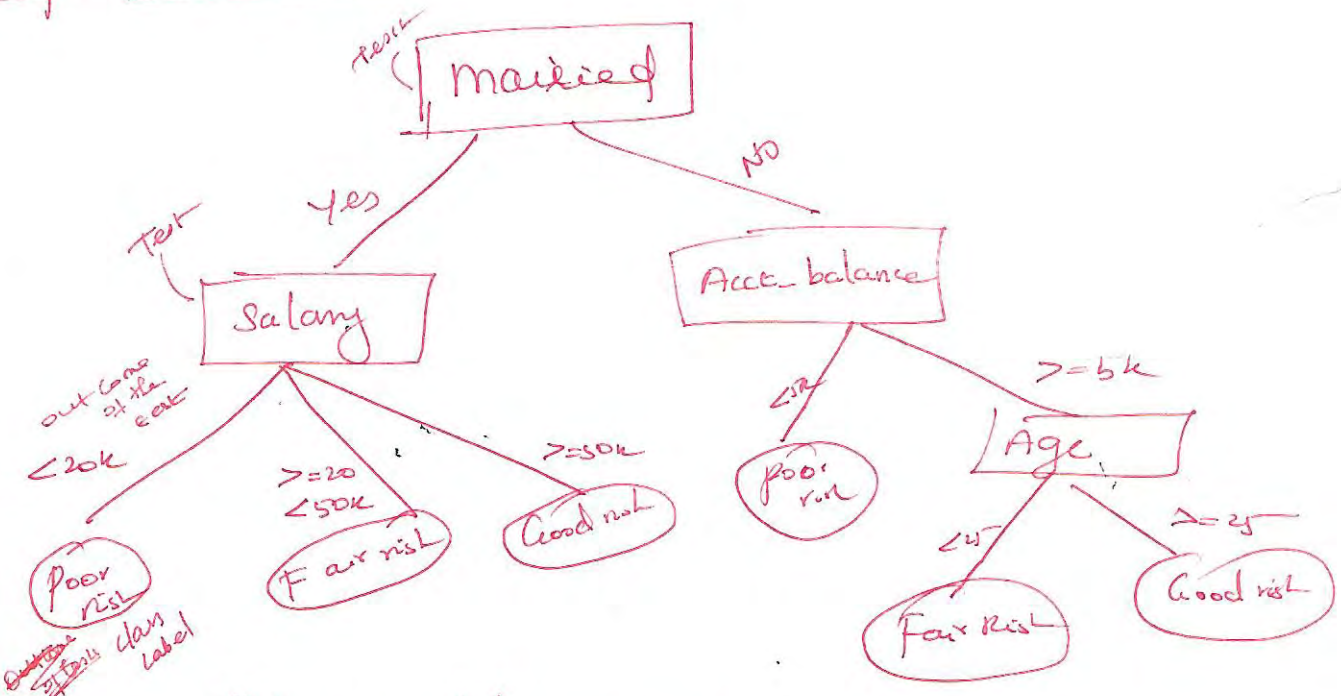
Learning
- Supervised learning – training data are accompanied by labels indicating class of Observation. New data classified based on training set
- Unsupervised " → class label of training data is Unknown

- Model is produced in the form of decision tree. It is a graphical representation of the description of each class. (or) representation of the classification Rules.

RID	Married	Salary	Acct_balance	Age	Loan worthy
1	no	>=50k	<5k	>=25	Yes
2	yes	>=50k	>=5k	>=25	Yes
3	yes	20k..50k	<5k	<25	no
4	no	<20k	>=5k	<25	no
5	no	<20k	<5k	>=25	no
6	yes	20k..50k	>=5k	>=25	Yes

Sample training data for classification algorithm

Example Decision tree



Attribute Selection Measures:-

- Information gain
- Gain Ratio.

The expected information needed to classify a tuple in D is given by

$$Info(D) = -\sum_{i=1}^{m} P_i \log_2(P_i)$$

P_i – is the probability that an arbitrary tuple in D belongs to class C_i and is estimated by $|C_i, D|/|D|$

The expected information required to classify a tuple from D based on the partitioning A is calculated by

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$$

where $|D_j|/|D|$ is a weight of the j^{th} partition.

Information Gain

$$Gain(A) = Info(D) - Info_A(D)$$

$Info(D)$

$$Info(D) = -\frac{3}{6}\log_2\left(\frac{3}{6}\right) - \frac{3}{6}\log_2\left(\frac{3}{6}\right) =$$

$$= 1.$$

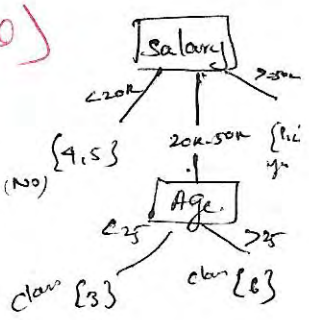
$$Info_{married}(D) = \frac{3}{6}\left(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{2}{3}\right) + \frac{3}{6}\left(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right)$$

$$= 0.92$$

$$Gain(married) = Info(D) - Info_{married}(D)$$

$$1 - 0.92 = 0.08.$$

Gain(salary) = 0.67
Gain(Age) = 0.08
Gain(Acc balance) = 0.46.



Clustering:

— is the process of grouping the data into classes or clusters, so that object within a cluster have high similarity

— Cluster is a collection of data objects they are similar to one another.

— eg. of Un supervised learning

— its learning by observation rather than learning by examples.

— Groups are usually disjoint groups.

Cluster Analysis App/:

1) Pattern Recognition
2) Data Analysis
3) Image processing

In Clustering when data is numeric the Euclidean distance can be used to measure Similarity.

Suppose take 2 n dimensional datapoints (records) r_j & r_k, the Euclidean distance between two points are given by

$$\text{Distance}(r_j, r_k) = \sqrt{(r_{j1} - r_{k1})^2 + (r_{j2} - r_{k2})^2 + \dots + (r_{jn} - r_{kn})^2}$$

The Smaller the distance between two points, the greater the similarity

A Classic Clustering Algorithm is "K-means Algorithm"

Centroid-Based Technique: The K means Method

The K-means algorithm takes the input Parameter k and partitions a set of n objects into k clusters so that intracluster similarity is high but intercluster similarity is low.

Cluster Similarity is measured in regard to the mean value of the Object in the cluster, it is also called Centroid or Center of gravity:

K-means algorithm has four steps

1) Partition Objects into k nonempty subsets
2) Compute seed points as the Centroids of the clusters (ie) mean point of the cluster
3) Assign each object to the cluster with nearest Seed point
4) Go back to Step 2, stop when no more new data is found.

Algorithm aims at minimizing ~~the~~ an objective fn.

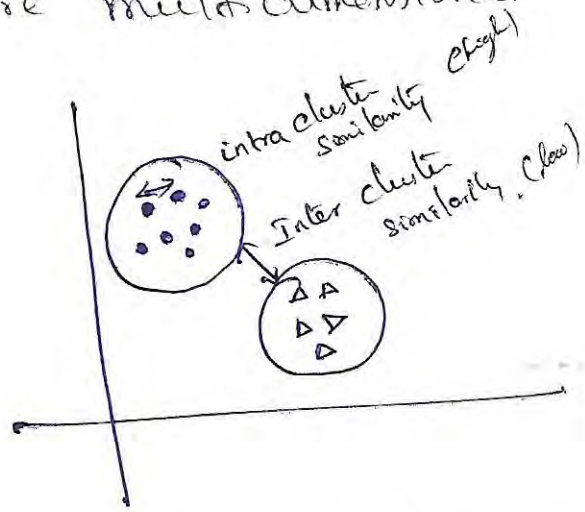
So, mean squared error fn is calculated as

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2$$

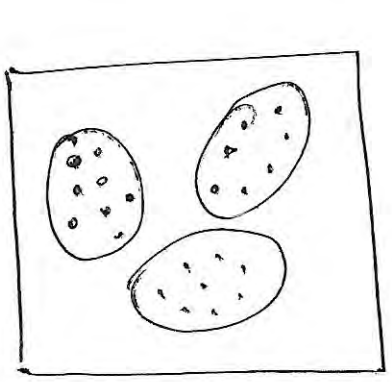
Where E is the sum of the square error for all objects in the data set.

p is the point in space representing a given object.

m_i is the mean of cluster C_i (both p and m_i are multidimensional

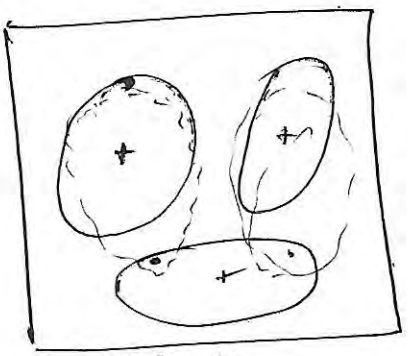


Let $k=3$, user would like the objects to be partitioned into 3 clusters

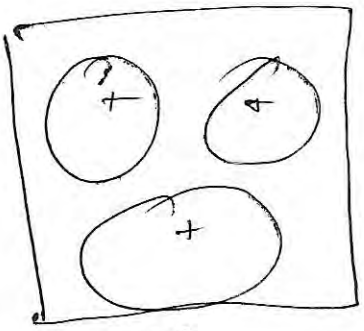


(a)

Mean of the cluster (+)



(b)



(c)

- reassigning objects to clusters to improve partitioning is called iterative relocation

Information Retrieval. (IR)

→ IR refers to the querying of unstructured textual data.

→ Querying based on keywords and ranking documents on the basis of their relevance to the query

→ IR is the process of searching for document information within the document / within the database

→ The process of IR consists of locating relevant documents based on User i/ps.

→ Retrieve a particular set of documents by set of words or keywords.

→ IR System locates & returns the document whose associated keywords match with the given keywords.

(eg) keywords "Computer memory" & "memory System" used to locate documents describing memory system of the Computer.

→ IR based on keywords are not only used for retrieving textual data but also retrieving other types of data / objects such as video or audio data / object.

(eg) A Song can be associated with several related keywords such as title, movie name & actors.

→ IR Systems Support Search based on query expressions using keywords & logical operators 'AND', 'OR' and 'NOT'.

(eg) A User could Query for "Cars AND bikes" or "Cars OR bikes" or "Cars NOT Bikes"

→ A Query Containing keywords without any of the logical Connectives is assumed to have 'AND' implicitly Connecting the keywords.

→ Full text retrieval in searching the full text of a document for the presence of keyword or keywords.

→ All the words in the document (also called term) are considered ~~as~~ to be keywords.

→ IR locates & returns all the documents relevant to query or contain all the keywords in the query.

→ Complex IR System estimate relevance of documents to a query so that the documents can be shown in order of estimated relevance.

→ They use information about no. of occurrences of the keyword, hyperlink information etc to estimate relevance

Relevance Ranking :-

→ IR Systems estimate relevance of documents to a query & return only highly ranked documents as answers.

Ranking Using TF-IDF

* Term Frequency – Inverse Document Frequency (TF-IDF) is a numerical statistic that is intended to reflect how important a word is to a document in a collection.

→ It is often used as a weighting factor in IR and text mining.

→ The TF-IDF value increases proportionally to the no. of times a word appears in the document,

→ Variations of TF-IDF weighting scheme are often used by search engine as a central tool in scoring & ranking a document's relevance given a user query.

IDF of a term t is defined as

$\frac{1}{a(t)}$, where $a(t)$ → total no. of documents that contains the term t.

The relevance of a document 'd' to a query 'Q' is defined as

$$R(d, Q) = \sum_{i=Q} a(d, t) * IDF(t)$$

$a(d, t)$ → represents TF (no. of times term t appears in the document d)

Similarity Based Retrieval :-

→ Users can retrieve documents that are similar to a given document d.

→ Similarity between 2 documents may be on the basis of common terms

→ The document that is most similar to the query is ranked highest.

Popularity Ranking :-

→ Find pages that are popular and to rank them higher than other pages that contain the specified keywords.

→ Most searches are intended to find information from popular pages ranking such pages higher will

Produce more relevant matches.

Page Rank :-

→ It is an algorithm used by Google Search to rank websites in their search engine results.
→ It is a way of measuring the importance of website pages.
→ It works by counting the no. & quality of links to a page to determine a rough estimate of how important the website is

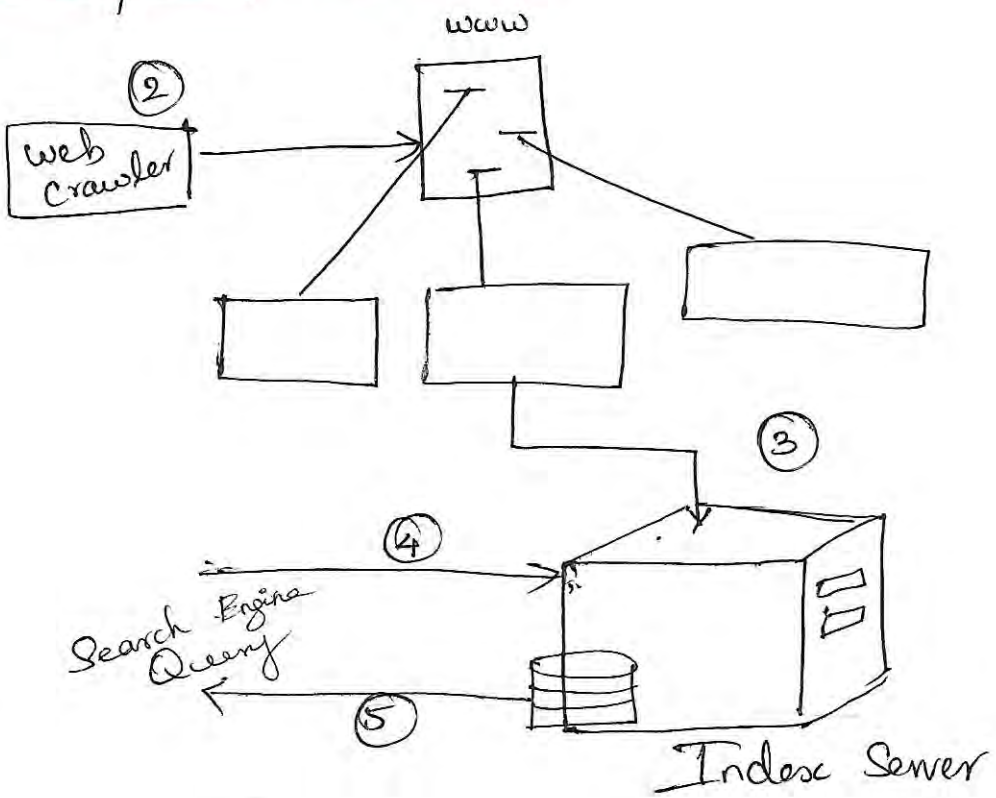
Crawling & Indexing the Web :-

→ For a Search to be successful, web search engines must find all the relevant documents in extremely large no. of documents.
→ documents have links to other documents help in locating relevant documents for a given Search.
→ Web Search engines Crawl the web to locate & gather information found in the documents to a Combined Index.

A Web crawler is a software application that automatically & systematically browses

the www, typically for the purpose of web indexing.

→ A web crawler may also be called a web spider, an ant or an automatic indexer.



"Web Crawler in Action"

(1) Web crawler identifies all the hyperlinks in the pages and adds them to the list of URL's to visit.

(2) The web crawler collects documents from the web to build a searchable index for the search engine.

(3) The web crawler continuously populate the index servers where the content is similar to the index in the books.

④ The webserver sends the query to the index servers that tell which pages contain the words that match the query. The query travels to the document servers, which retrieve the stored documents.

⑤ Snippets are generated to describe each search result.
- The search results are refered to the user.

Object Oriented Database (OODB)

- Systems are usually associated with applications that draw their strength from intuitive GUI, Power modelling techniques and advanced database management capabilities.

Characteristics

1. Maintain direct correspondance b/w realworld and database object so that objects do not loose integrity and identity.
2. OODBs provide a Unique system generated object identifier (OID) for each object

9. OODBs are extensible — capable of defining new data types