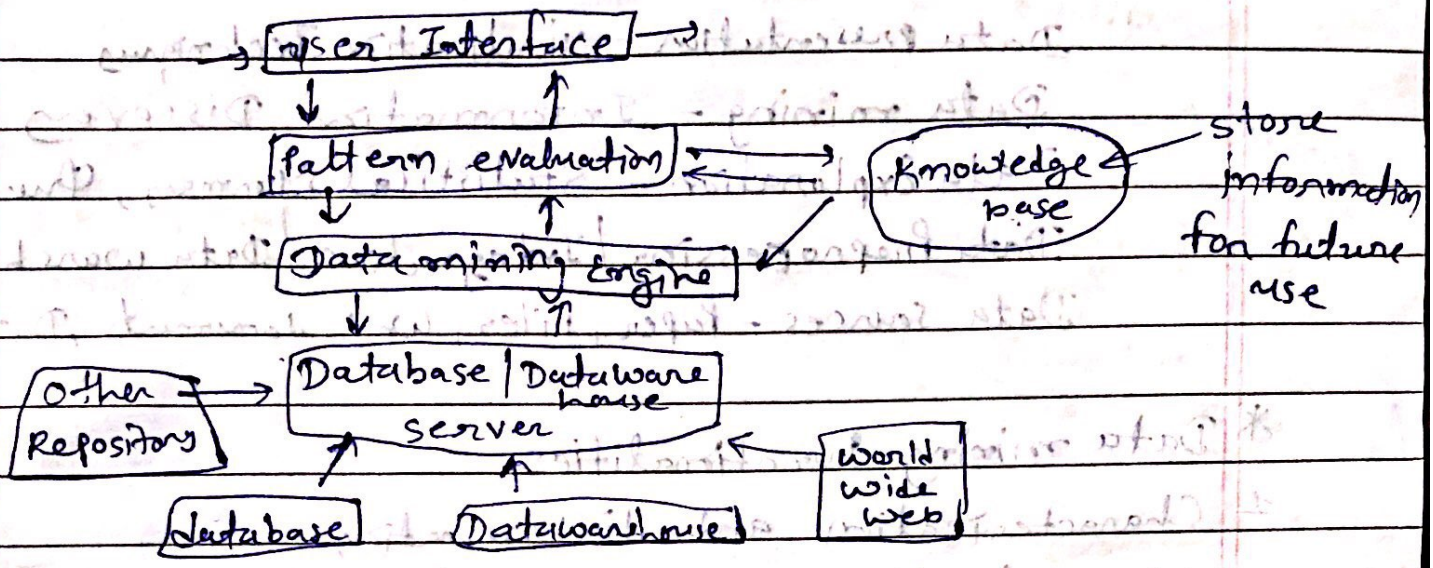
Data mining refers to extracting knowledge from large amount of data



- Data mining - Automated analysis of massive datasets
- Extraction of interesting - non trivial, implicit, previously unknown, potentially useful patterns or knowledge from huge amount of data.

- Knowledge discovery in databases, data archeology, data dredging, information harvesting, business intelligence

* Web Mining Framework

① Data Cleaning ② Data Integration from multiple sources
③ Warehousing the data ④ Data cube construction
⑤ Data selection for data mining ⑥ Data mining
⑦ Presentation of the mining result
⑧ Pattern & Knowledge to be used or stored into knowledge base

Scanned with CamScanner

Business Intelligence

Decision making
Data presentation - visualization Techniques
Data mining - Information Discovery
Data Exploration - Statistical, Summary, Querying, Reporting
Data Preprocessing / Integration / Data warehouse
Data sources - Paper, files, web documents, Database system

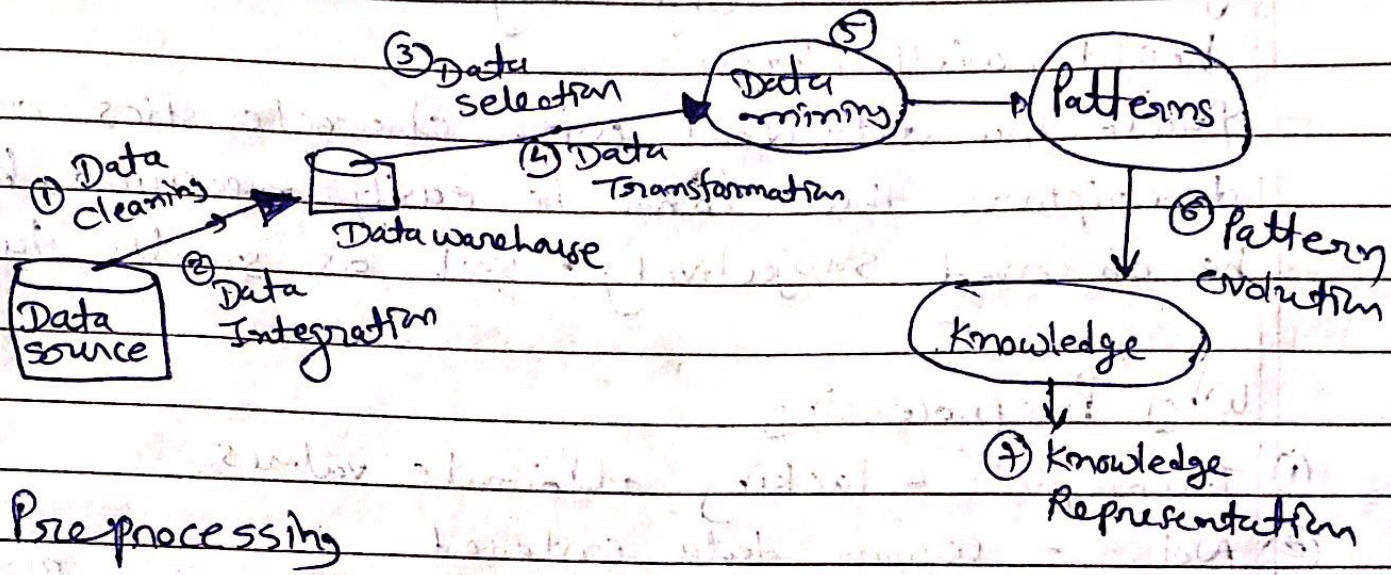
* Data mining Functionalities

1. Characterization and discrimination
   Generalization, summarize and contrast data Characteristics
2. Patterns, association, correlation
3. Classification and Prediction
4. cluster analysis - group data from new classes
5. outlier analysis - useful in fraud detection
6. Trend and evolution analysis
   - Regression Analysis, Periodicity, similarity based Analysis
7. Statistical Analysis

Scanned with CamScanner

वडतालधाम
Dt.: / / | Pg.No.:

Knowledge Discovery from Data



* Preprocessing

Types of Attributes

- Nominal - Ex ID number, eye colari, zip codes
- only name is not sufficient to predict
- Ordinal - Ex ranking, grades, height in cm
- compair value
- Interval - Ex calendendates, temperaturis M celsius
- range of value
- Ratio - Ex temperature
- Computation Can perform

* Properties of attributes value

(1) Distinctness: =, ≠
(2) order < >
(3) Addition : +, −
(4) Multiplication : *, /

Scanned with CamScanner

- Quantitative - data deals with numbers and things you can measure objectively: dimentions such as height, width, length
- Qualitative data deals with characteristics and descriptors that can't be easily measured, but can be observed subjectively - such as smells, tastes, colors

Why Preprocessing

1. Incomplete - lacking attribute values
2. Noisy - wrong data entered
3. Inconsistent - discrepencies in codes Ex: ms 2013 file not open ms2003

* Data Quality

1. Completeness - As long as the data meets the expectations then the data is considered complete
2. Consistency - Data across all systems reflects the same information and are in synch with each other.
3. Conformity - Data follow some standered defnations like data type, size and formate Ex: mobile number - 10 digits
4. Accuracy - degree to which data is correct
5. Integrity - validity of data. all data can be traced & connected to other data
6. Timeliness - information is available when it is expected and needed

Scanned with CamScanner

वडतालधाम
Dt.: / / | Pg.No.:

* Data cleaning
- Parsing - put data in proper formate Ex sir Name Middlename
- Correcting - give correct information
- Standardizing - Data convert into one standard Ex °C
- matching - Remove same tuples,
- Consolidation - make one unit from two dependent Units
- Dealing with missing data - Data is not present
- Dealing with Incorrect & Noisy data

* Dealing with missing values
- Ignore the tuple - use when low priority tuple is empty
- Filling Manually
- Using Global Constants - "Null", 'N/A'
- Use attribute mean/median
mean : Normal distribution
median : skewed distribution (Asymmetric)
- Use of data mining Algorithms (decision tree, Clustering)

* Missing Data Reasons
(1) equipment malfunction (2) misunderstanding
(3) inconsistent with other recorded data and thus deleted
(4) Data may not be considered important at the entry time
(5) not register history or changes of the data

Scanned with CamScanner

* Noisy Data - error or variance in measured variable
- Due to ① Instrument ② Data entry problems
③ Data Transmission ④ Technology limitation
⑤ Inconsistency in naming

* Dealing with Noisy Data
(1) Binning
→ ● equal partitioned Bin -
- divide Data into equal parts
→ (a) Bin mean
- replace Data with mean
(b) Bin median
- replace Data with median
(c) Bin boundaries
- replace data who value is near to either of boundary
equi width : Binsize
equi depth : Bin number

* Data Integration
- Give unified view
- sources: multiple Databases, data cubes, flat files
- It is use to make Datawaire house - which is use in Analysis

* Schema Integration
- Problem: customer-id in one data base, C-id in second database. Is both tuple having number? is issue in Integration

॥ वृत्तालये स भगवान् जयतीह साक्षात् ॥

Scanned with CamScanner

* Entity identification Problem
- identify real world entity like Person, mit = Neel
* Detecting and Resolving data value conflicts
- for the same real world entity, attribute values from different sources are different.

* Redundancy in Integration
(1) Object identification - same attribute may have different names in different database
Ex DB: MIT , DB2: MITkuman
(2) Derivable Data: One attribute can be derived from other Ex Age derived from Date of birth
- Redundant attributes may be able to be detected by correlation Analysis and covariance Analysis

$$Cov(A,B) = E((A-\bar{A})(B-\bar{B}))$$
$$= E(A \cdot B) - \bar{A}\bar{B}$$
$$-1 \leq Cov(A,B) \leq 1$$

$Cov(A,B) > 0$ then A, B are +vely correlated
$Cov(A,B) = 0$ " " no relation
$Cov(A,B) < 0$ " " -vely correlated

* Data Integration approaches
(1) Tight Coupling - Data combined from different sources into a single physical location through the process of Extraction, Transformation and Loading

Scanned with CamScanner

② Loose Coupling –
- Data only remains in the actual source database

* Data Reduction
- Dataset much smaller but produce same output
- Complex data set may take very long time to Analysis on dataset so Reducting is require

① Dimention Reduction
Curse of dimention
- when dataset increase, data becomes increasingly sparce
- Density and distance between points is much more so clustering, and analysis become less meaningful
- Possible combination of subspaces will grow exponentially

② Dimentionality Reduction
- Avoid the curse of dimentality
- help eliminate irrelevant features and reduce noise
- Reduce time & space required

② Attribute subset selection
- Redundent attributes - duplicate info (temprature in °c, F°)
- Irrelevent attributes

③ Data Compression
- String compression
- Audio-vidio compression

Scanned with CamScanner

LIC
भारतीय जीवन बीमा निगम
LIFE INSURANCE CORPORATION OF INDIA
बिपीनकुमार अेम. पटेल
सदस्य - अभिकर्ताओ के लिये क्षेत्रिय प्रबंधक क्लब
Bipinkumar M. Patel
Member of the Zonal Manager's Club for Agents
828 83A
सदस्य • अभिकर्ताओं के लिये क्षेत्रीय प्रबंधक क्लब • MEMBER • ZONAL MANAGER'S CLUB FOR AGENTS

Parallel systems

- In parallel systems many operation performed parallel, in which the computational steps are performed sequentially.
- A coarse-grain parallel machine have less number of processor
- massively parallel or fine-grain parallel machine uses thousands of small processor

→ two main measures of performance of a database system ① throughput - number of tasks that can be completed in a given time interval

② Response Time - amount of time it takes to complete a single task from the time it is submitted

* Speedup - Running a given task in less time by increasing the degree of Parallesim

Scaleup - Handling larger task by increasing the degree of parallelism

→ Execution time of a task on the larger machine is T_L

" " " " " " smaller " " T_S

$$\text{Speed up} = \frac{T_S}{T_L}$$

→ Linear speedup if the speedup is "N" when the larger system has N times the resources then smaller system

→ If speedup is less then "N" then it's called sublinear speedup

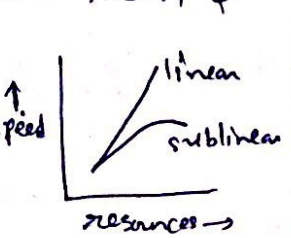
→ Let Φ be a task, Φ_N is "N" times bigger then Φ

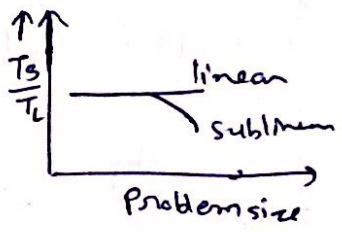
→ Φ on Given Machine m_S is T_S

Φ_N on a parallel machine m_L is T_L

Linear scaleup if task Φ if $T_L = T_S$

sublinear scaleup if $T_L > T_S$





→ Batch scaleup - the size of database increases, and the tasks are large jobs whose run time dep^n on the size of database

→ Transaction scaleup - rate at which transactions are submitted to the database increase and the size of the database increase proportionally to the transaction rate.

कार्यालय : परबडी के पास, पो.समरखा.
निवास : सरदार पार्क सोसायटी, पो.समरखा, ता.जी.आणंद-388360.
फोन : (नि) (02692) 256673 (मो) 98242 32755.
Off. : Nr. Parabadi, Po. Samarkha.
Resi. : Sardar Park Society, Po. Samarkha, Ta. Dist. Anand-388360.
Ph. : (R) (02692) 256673 (M) 98242 32755. E-mail : bipinpatellic@yahoo.com

Scanned with CamScanner

→ A machine that scales up linearly may perform worse than a machine that scales less then linearly, simply because the letter machine is much faster to start off with.

* factor diminish Speedup and Scaleup :-

① startup costs :- Cost with initiating a single process.
- In parallel operation thousands of processes, start up-time may overshadow the actual processing time, speedup - effect

② Interference - process executing parallel often access shared resources, a slowdown process (system bus, disk, locks), speedup scaleup } effect

③ skew - breaking down single task into number of parallel steps, depends on slowest step task.
Ex 100 is divided into 10 task → If one task is of 20 size then output: 5 rather then: 10.

* Interconnection Networks

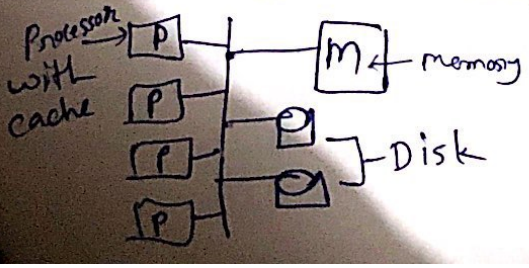
① Bus - suitable for small number of processors
- Bus send and receive data from a single communication channel
- Not work well to scale and increasing parallelism

② Mesh - two dimention each node connected to four adjacent node
- 3D each node connected to 6 adjacent node
- Node which not connected direct communicat via other nodes
- Scales better with increasing parallelism

③ HyperCube - each node is far from each other maximum log(n) links where as in mesh components may be 2(√n - 1) links away
- Thus, communication delay in mesh is High then HyperCube

* Parallel Database Architectures

ⓐ shared memory
- extremely efficient communication between Processors
- One processor can send message to other using memory write
- Maximum 32 or 64 processor due to bus interconnection
- Adding more CPU not work because they wait for work and access memory
- Maintaining Cache coherency become overhead with more CPU



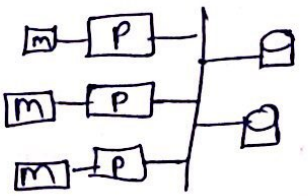
Scanned with CamScanner

LIC
भारतीय जीवन बीमा निगम
LIFE INSURANCE CORPORATION OF INDIA

सदस्य
अभिकर्ताओं के लिये
क्षेत्रीय प्रबंधक क्लब
MEMBER
ZONAL MANAGER'S
CLUB FOR
AGENTS

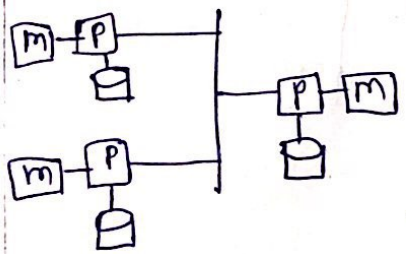
बिपीनकुमार अेम. पटेल
सदस्य - अभिकर्ताओ के लिये क्षेत्रिय प्रबंधक क्लब
Bipinkumar M. Patel
Member of the Zonal Manager's Club for Agents
828 83A

② Shared Disk - all disks directly via interconnection network, attach with processor



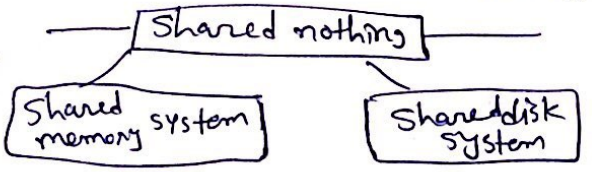
→ memory bus is not bottleneck here
→ fault tolerance ↑
→ If one CPU fail then other can do that task due to disk is shared. - RAID Architecture is used in Disk
→ Prob^m - scalability
→ Disk subsystem is now a bottlenack; when database makes a large number of accesses to disks
→ large number of CPU can use but communication between CPU's slow

③ Shared Nothing ⊢ One processor may communicate with other by a high-speed interconnection Network



→ For I/O, no interconnection Network require so it is easily scaleable
→ draw back - costs of communication & non-local disk
→ software interaction require at both ends

④ Hierarchical ⊢ Combination of above three



→ Reduce the complexity of programming such system have yielded "Distributed Virtual memory" - logically divide single memory between systems
- Non Uniforme Memory Architecture - One memory divide different amount

* Distributed Systems ⊢
→ The computers in distributed system are referred to by a number of different name, such as sites or nodes.
→ A local transaction, is one that access data only from sites where the transaction was initiated
→ A global transaction, one that either access data in a site different from the one at which the transaction was initiated, or access data in different sites

कार्यालय : परबडी के पास, पो.समरखा.
निवास : सरदार पार्क सोसायटी, पो.समरखा, ता.जी.आणंद-388360.
फोन : (नि) (02692) 256673 (मो) 98242 32755.

Offi. : Nr. Parabadi, Po. Samarkha.
Resi. : Sardar Park Society, Po. Samarkha, Ta. Dist. Anand-388360.
Ph. : (R) (02692) 256673 (M) 98242 32755. E-mail : bipinpatellic@yahoo.com

Scanned with CamScanner

Advantages of Distributed

① Sharing data – One site may be able to access the data residing at other sites.

② Autonomy – each site is able to retain a degree of control over data that are stored locally. In distribution system, there is a global database administrator responsible for entire system. Each local database administrator have local autonomy

③ Availability – If one site fails in system, the remaining sites may be able to continue operating. If data items are replicated in several sites, then 24 x 7 file available

– Easily find failure in system, Although recovery from failure in distributed system is diffiadt

* Two phase (2PC) Protocol use in distributed system ⊢

→ For example bank have four center in one city and two person simultenously use one account then according 2PC if all branch are ready then and then either of person can commit his action. (Soln → Atomicity)

→ For concurrency control – locking system is implemented

Disadvantage of Distributed

① Software – development cost ② Greater potential for bugs

③ Increased processing overhead

* Network Types ⊢

– local area Networks – CPU allocated in same geographical Area

– wide area Network – CPU allocated distributed over Area

– Distributed operating system require

* LAN ⊢ numerous small computers are benificial then one single large system. It is used in office. Channel is twisted pair, coaxial cable, fiber optics, wireless

– Storage Area Network is between disk and LAN. This system gives highly availability and scalability. RAID organization is used

* WAN ⊢

– WAN, have significant Latency – speed of Light delay, Queuing delay

Discontinuous Connection – Mobile Internet

Continuous connection – Wired Internet

Scanned with CamScanner