DATA MINING.

Page: Date: / /

SPPU Question paper of Dec-2015.

Q.1 a] For each of the following geniers, identify & write the type of data mining task.

i> find all credit applicants who are poor credit risks.

ii> Identify customers with similar buying habits.

iii> find all items which are frequently purchased with milk.

→ i> classification -: classification is the process of finding a model (or function) that describes & distinguishes data class or concepts.

Based on credit applications, customers can be classified in various classes like poor, medium & high credit risk types of customers.

ii> clustering -: clustering analyzes data objects without consulting data objects. clustering can be used to generate class labels for a group of data.

clusters can be formed based on similar types of buying patterns. Then customer belongs to those clusters can be identified.

iii> Association -: Various items which has been frequently purchased with milk can be identified with association data mining task. Based on the support & confidence, milk can be associated with frequent items.

Page: Date: / /

b] consider the following :-

Transactions	Items
t1	Bread, jelly, peanut butter
t2	Bread, peanut butter
t3	Bread, milk, peanut butter
t4	Beer, Bread
t5	Beer, milk

calculate Support & confidence for following association rules -:

i> Bread → peanut butter ii> jelly → milk

iii> Beer → Bread.

→ By considering min support count = 2 & min confidence = 80 %

Step-1 -: (c1) = . Scan for count of each candidate

Itemlist	Sup-count
Bread	4
Jelly	1
peanut butter	3
milk	2
Beer	2

Page: Date: 11

Step 2 - Compare caddidate Support count with min Support count (ie - 2)

Itemlist	Sup count
Bread	4
peanut butter	3
milk	2
Beer	2

Step 3 -:

Itemsets	Sup count
{Bread, peanut Butter}	3
{Bread, milk}	1
{Bread, Beer}	1
{peanut butter, milk}	1
{peanut Butter, Beer}	0
{milk, Beer}	1

Step 4 -: Compare candidate Support count with minimum Support count.

freq Itemset	sup count
{Bread, peanut butter}	3

Step - 5 -: frequent itemset is -: {Bread, peanut butter}.

Page: Date: 11

Association Rule	Support	confidence	conti (%)
Bread → peanut butter	3	3/4	75%
Bread → peanut butter	1	3/3	100%

Minimum confidence threshold is -: 80%.

final rule -: peanut Butter → Bread

Association rule	Support	conti	conti (%)
Bread → peanut butter	3	3/4	75%
jelly → milk	0	0/1	0%
Beer → Bread	1	1/2	50%

c] Consider 10 records given below :-

ID	Income	Credit	class	X_1
1	4	Excellent	h_1	X_4
2	3	good	h_1	X_7
3	2	Excellent	h_1	X_2
4	3	good	h_1	X_7
5	4	good	h_1	X_8
6	2	Excellent	h_1	X_2
7	3	Bad	h_2	X_{11}
8	2	Bad	h_2	X_{10}
9	3	Bad	h_3	X_{11}
10	1	Bad	h_4	X_9

Calculate the prior probabilities of each of class h_1, h_2, h_3, h_4 & probabilities for data points X_1, X_4, X_7, X_8 belonging to the class h_1

→ Assign ten data values for all combination of credit & income.

	1	2	3
Excellent	X_1	X_3	X_6
good	X_2	X_4	X_5
bad	X_7	X_8	X_{10}

from training data :-

$p(h_1) = 6/10 = 60\%$

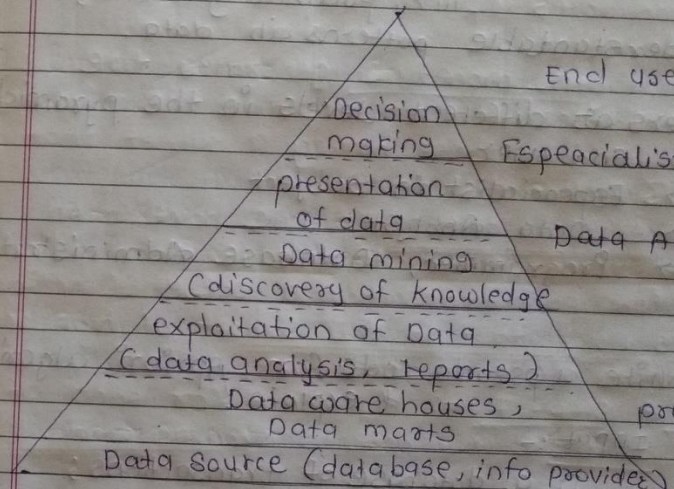
$p(h_2) = 2/10 = 20\%$

$p(h_3) = 1/10 = 10\%$

$p(h_4) = 1/10 = 10\%$

Q-: 2

a] Define data mining. Draw a pyramid showing relationship betn data mining & business intelligence. write type of users at different levels in the pyramid.



Data mining →

1) Data mining means mining or extracting knowledge from large amount of data.

2) Data mining is processing data to identify patterns & establish relationship.

3) Data mining is process of analysing large amounts of data stored in datawarehouse for useful information. which makes use of AI, neural ntw's & advanced statistical tools to reveal trends, patterns & relationships.

4) Data mining is non-trivial process of identifying :-
   i) valid
   ii) Novel
   iii) potentially useful, understandable patterns in data.

- users at different levels in the pyramid :-
  1) End user
  2) Especialist in Subject
  3) Data Analyst
  4) Programmer, Database Administrator

Q:2 b] write a pseudo code for Apriori algorithm & explain.

→ Input :-
D, a database of transactions;
min_sup, the minimum support count threshold.

output :- L, frequent itemsets in D.

method :-

```
(1) L1 = find_frequent_1-itemsets(D);
(2) for (k=2 ; Lk-1 ≠ ϕ ; k++) {
(3) Ck = apriori_gen (Lk-1);
(4) for each transaction t ∈ D
    { // scan D for counts)
(5) Ct = subset (Ck, t)
(6) for each candidate c ∈ Ct
(7)   c.count ++;
(8) }
(9) Lk = {c ∈ Ck | c.count >= min_sup }
(10) }
(11) return L = Uk Lk;

procedure apriori_gen (Lk-1 : frequent (k-1)-itemsets)
1) for each itemset l1 ∈ Lk-1
2) for each itemset l2 ∈ Lk-1
3) if (l1[1] = l2[1]) ∧ (l1[2] = l2[2]) ∧ --- ∧
   (l1[k-2] = l2[k-2]) ∧ (l1[k-1] < l2[k-1])
   then {
4)   c = l1 ⋈ l2;
5)   if has_infrequent_subset (c, Lk-1) then
6)   delete c;
7)   else add c to Ck
8) }
9) return Ck

procedure has_infrequent_subset (c:
   candidate k-itemset;
   Lk-1 : frequent (k-1)-itemsets);
1) for each (k-1) subset s of c
2) if s ∈ Lk-1 then
3)   return TRUE;
4) return false;
```

Explaination of Apriori pseudo code :-

Step-1 :- of apriori finds the frequent 1-itemsets. L1.

Step 2 :- through 10. L_{k-1} is used to generate candidates C_k to find L_k for $K \geq 2$.

Step 3 :- Apriori gen procedure generates the candidates & then uses the apriori property to eliminate those having a subset that is not frequent.

Step 4 :- once all of the candidate have been generated, the database is scanned.

Step 5 :- for each transaction a subset function is used to find all subset of transaction that are candidates.

Step-6 :- & Step-7 :- & count for each of these candidate is accumulated.

Step 9 & step 11 :- finally all candidates satisfying the minimum support from the set of frequent itemsets L.

Q:2 c] Write pseudo code for the construction of Decision tree. State & justify time complexity.

→ **Algorithm** : Generate_Dicision_tree

Input :

- Data partition D, which is set of training tuples & their associated class labels.
- attribute_list, the set of candidate attributes;

output : A Decision tree.

Pseudo code :-

```
1) create a node N;
2) if tuples in D are all the same class C, then
3) return N as a leaf node labeled with the
   class C;
4) if attribute_list is empty then
5) return N as a leaf node labeled with the
   majority class in D.;
6) apply Attribute_selection_method (D, attri-
   bute_list) to find the "best" splitting_criteria
7) label node N with splitting_criteria
8) if splitting_attribute is discrete-valued &
   multiway splits allowed then
9) attribute_list ← attribute_list - splitting
   attribute
10) for each outcome j of splitting_criteria
```

11) Let Dj be the set of data tuples in D satisfying outcome j;
12) If Dj is empty then
13) attach a leaf labeled with the majority class in D to node N;
14) else attach the node returned by Generate decision tree to node N;
endfor
15) return N;

Time Complexity -:

For normal simple decision tree the time complexity is O(N D^2).
where, D is no of features.

A single level decision tree would be O(N D)

Q-3

a) Using K-means clustering, cluster the following data into 2-groups. (clusters).

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

⇒ Step 1 -: from given data, randomly assign alternative values to each cluster.

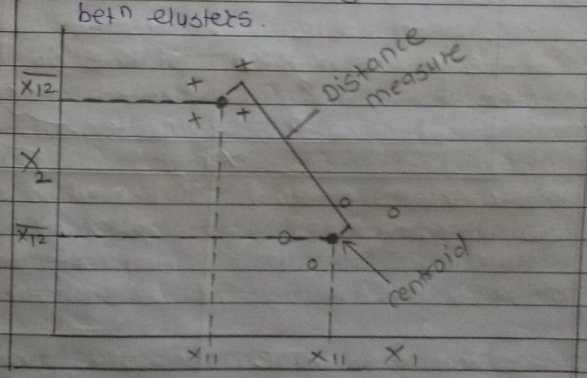
Step 2 -: No of cluster = 2. therefore,
K_1 = {2,10,3,30,25}, mean = 14
K_2 = {4,12,20,11}, mean = 11.75

Step 3 -: Reassign
K_1 = {20,30,25}, mean = 25
K_2 = {2,4,10,12,3,11}, mean = 7.

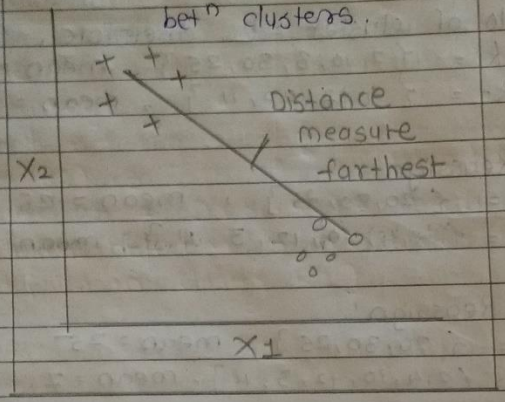
Step 4 -: Reassign.
K_1 = {20,30,25}, mean = 25
K_2 = {2,4,10,12,3,11}, mean = 7.

Final Ans -: K_1 = {2,3,4,10,11,12}
K_2 = {20,30,25}

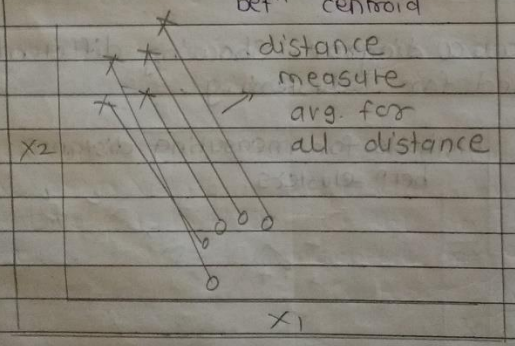
b] Draw a diagram showing different approaches used for clustering.



methods for measuring distance bet^n clusters.

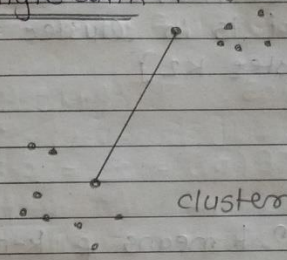


methods for measuring distance bet^n centroid



c] Many clustering algorithms need to determine the distance bet^n two clusters. write the formula to determine the distance bet^n two given clusters K_1 & K_2. using single link, complete link & Avg. methods

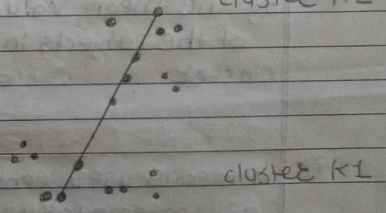
→ Single link :-



In single linkage method, ~~D(A,B)~~ is $D(K_1, K_2)$ is computed as,

$D(K_1, K_2) = \text{Min}\{ d(i,j) :$ where object i is in cluster K_1 & object j is in cluster K_2 $\}$.

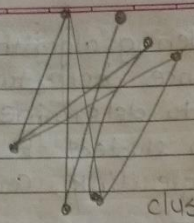
Complete link :-



In a complete linkage method $D(K_1, K_2)$ is computed as $D(K_1, K_2) = \{\max\{d(i,j) :$ where i is in cluster K_1 & object j is in cluster $B\}$.

Page: Date: 1 1

Average link:-



In the Avg linkage method, $D(K_1, K_2)$ is computed as = $D(K_1, K_2) = \text{mean}\{ d(i,j)$ where object i is in cluster K_1 & object j is in cluster $K_2 \}$.

Q:- 4

q] Differentiate betⁿ k-means & k-mediod clustering algorithm.

K-means Algorithm	K-mediod algorithm
1> Each cluster is represented by the mean value of the objects in the cluster.	1> Each cluster is represented by one of the objects located near the center of the cluster.
2> k-means uses centroid as representatives	2> k-mediod uses the objects in the datasets themselves as representatives
3> less robust to noise	3> It is more robust to noise
4> Sum of squared ~~Full~~ Euclidean distance-type metric to evaluate variance that we uses in k-means.	4> There is an advantage to using the pairwise distance measure in the k-medoid algorithm.

Page: Date: 1 1

Q:- 5

q] Consider the following 4 documents D_1, D_2, D_3 & D_4. for each document D_i, various terms that occure in document D_i, is provided.

D_1 = { To do is to be to be is to do }
D_2 = { To be or not be I am what I am }
D_3 = { I think therefore I am }
D_4 = { Do dodo da da da let it be let it be }

Write the equations & calculate the term frequency "tf" inverse document freq" "idf & "tf idf" for the terms "to" & "do" for the documents D_1 & D_2.

⇒ $TF(t) = \frac{\text{No. of times term t appear in a document}}{\text{total no of terms in document}}$

IDF: $1 + \log\left(\frac{\text{total no. of document}}{\text{no. of document with term t in t}}\right)$

TF-IDF : TF (doc) * idf (doc) for 'to'

1) $TF(D_1) = \frac{4}{10} = 0.05$. $TF(D_4) = \frac{0}{12} = 0$.

for 'do' $TF(D_1) = \frac{2}{10} = 0.2$.

$TF(D_4) = \frac{3}{12} = 0.25$.

2) $IDF = 1 + \log\left(\frac{4}{2}\right) = 1 + \log(2) = 1.3010$.

$IDF('do') = 1 + \log\left(\frac{4}{2}\right)$

$= 1 + \log(2) = 1.3010$.

for 'to'

TF - idf = 0.05 * 1.3010

D_1 = 0.06505

TF-idF (D_4) = 0 x 1.3010

D_4 = 0.

for ('do')

Tf-idF (D_1) = 0.2 * 1.3010

= 0.2602

D_1 = 0.2602.

Tf-IDf (D_4) = 0.25 * 1.3010

= D_4 = 0.32525.

c] Enumerate various text operations (also called preprocessing) that are used by an information retrieval system.

→ Text operations reduce the complexity of document representation

Text operations :- Some search engines on the web are giving up text operations entirely & simply indexing all the words in the text.

Text compression :- Text compression is about finding ways to represent the text in fewer bits or bytes. Compression method creates a reduced representation by identifying & using structures that exists in the text.

A major obstacle for storing text in compressed form is need for IR systems to access text randomely.

following are some text operations that are used by an info. retrieval system :-

1) Lexical Analysis of the text :- 1) Is a process of converting a string of characters into stream of words.

2) Thus one of the major objective of lexical analysis phase is the identification of the words in the text.

2) Elimination of stopwords :-

1) A word which occures in 80% of the documents in the collection is useless for

purpose of retrival. Such words are frequently referred to as stopwords & are normally filtered out as potential index terms

2) Articles, prepositions, & conjuctions are natural candidates for a list of stopwords

3) elimination of stopwords has an additional important benefits.

3) Stemming :-

1) the user specifies a word in a query but only a variant of this word is present in a relevent document.

2) this problem can be partially overcome with the substitution of words by their respective stems

3) A stem is portion of word which is left after the removal of its affixes

4) stems are thought to be useful for improving retrieval performance because they reduce variants of the some root word to a common concept.

4) Index term selection :-

1) Distinct automatic approaches for selecting index terms can be used.

2) A good approach is the identification of noun groups.

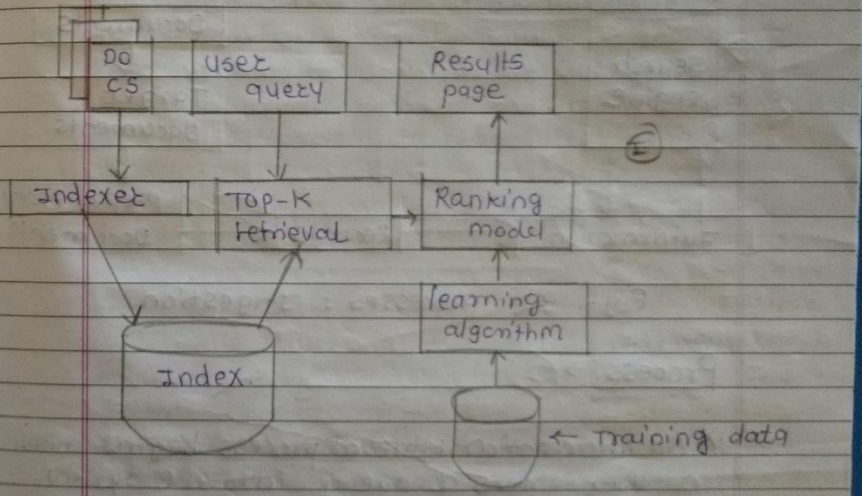
3) It is common to combine two or three nouns in single component it makes sense to cluster nouns which appear nearby inthe text into single indexing component.

Q:6

a] Draw a neat diagram showing retrieval process of an IR system & briefly describe its components.

→

1) IR is concerned with retrieving texual records, not data items like relational databases with finding patterns like data mining.



2) IR finds document with text that contains the words, information adjacent to retrieval.

3) IR focuses on finding the most appropriate or relevant records to the users request.

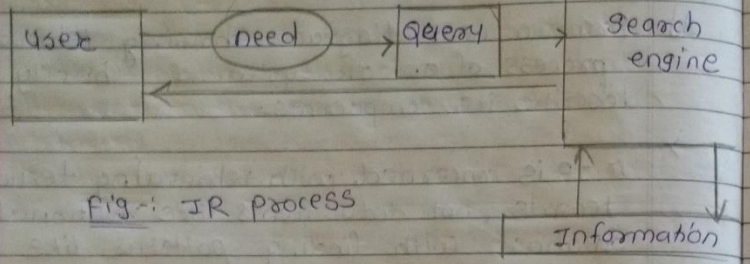


Fig:- IR Process

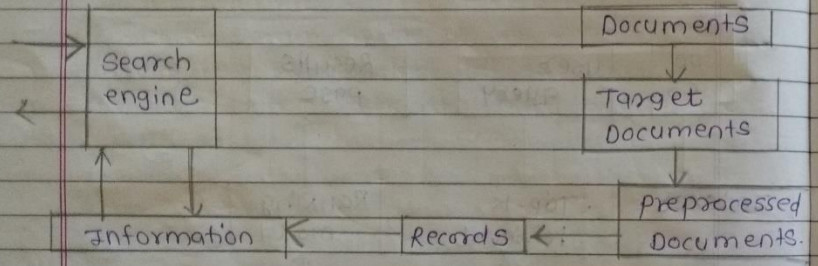


Fig:- IR processes : Ingestion.

Process → (I)

1) An information retrieval process begins when a user enters a query into the system.
2) queries are formal statements of information needs, ex:- search string in web search engines.
3) In information retrieval a query does not uniquely identify a single object in the collection. instead several objects match queries.
4) object is entity that represented by information in a content collection or database.
5) User queries are matched against the database information.
6) In information retrieval results returned may or not may matched the query, so results are typically ranked.
7) Depending on application the data objects may be, for ex:- text documents, images, audio, videos.
8) Documents themselves are not kept or stored directly in IR system. but instead represented in the system by metadata.

b) Explain precision & recall.

→ Precision :- (I)

1) In the field of information retrieval predision is the fraction of retrieved documents that are relevent to query.

2) precision = $\frac{|\{\text{relevent documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$

3) for example :- for a text search on a set of documents precision is the no. of correct results divided in by no of all returned results.

Recall :- (I)

1) Recall in information retrieval is the fraction of the documents that are relevent to the query that are successfully retrieved

2) Recall = $\frac{|\{\text{relevent docs}\} \cap \{\text{retrieved docs}\}|}{|\{\text{relevent documents}\}|}$

3) for example :- for text search on a set of documents recall is the no. of correct results divided by the no of results ~~of~~ that should have been returned.

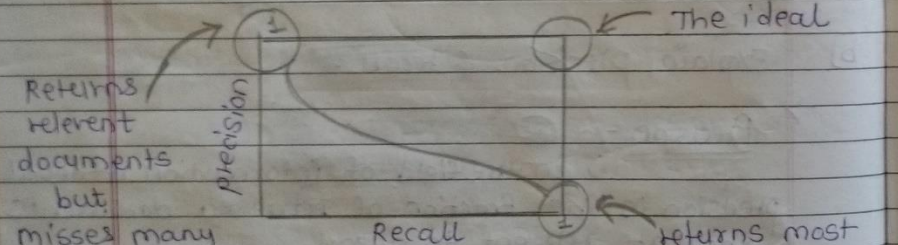


Fig :- trade-off bet^n Recall & precision

c) What is web crawler? explain working of basic crawler

→ Web crawler :-

1) A web crawler is an automated program that scans or crawls through the internet pages to create an index of the data.

2) A web crawler is also known as web spider, web robot, bot & automatic indexer.

3) Search engines makes use of web crawler to collect information about the data on public web pages, their primary purpose is to collect data so that when a user enters a search term on their site, they can be quickly be provided with relevent web sites.

4) The search engines web crawler visits a web page it collects information like visible text, hyperlinks & various tags like keyword rich meta tag

5) This information may be used by the search engine to determine what the site is about & index the information

6) web crawling is considered to be an important method for collecting data & keeping up with the expanding internet.

Different types of crawler :-

i) Traditional crawler :- visits entire web & replaces index

ii) periodic crawler :- ~~selectively~~ selectively searches visits portions of web & update

the web subset of index

iii) incement crawler :- selectively searches the web & incrementally modifies index

iv) focused crawler :- visits pages related to particular subject

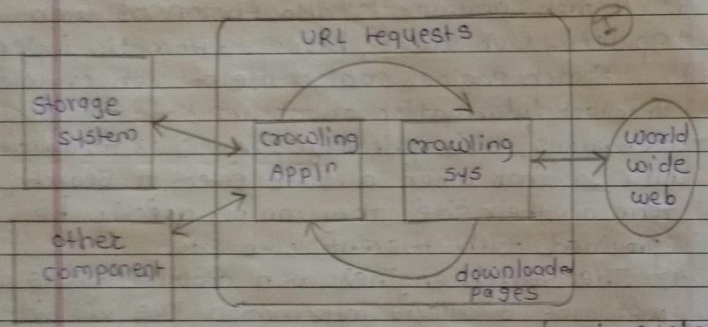
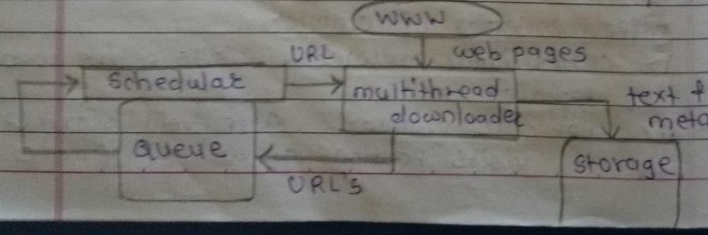


Fig:- Crawler basic system Architecture.

Basic crawler operation :-

1) Begin with known "seed" pages.
2) fetch & parse them
3) Extract URL they points
4) place the external URL on queue.
5) fetch each URL on queue & repeat.



Q:7

a] Define Big data & state few challanges of big data.

→

Big data :-

1) Big data is an evolving term that describes any voluminous amount of structured, semistructured, unstructured data that has potential to be mined for information.

2) By Gartners defination :-

" Big data is high volume, high velocity, & high variety information assests that require ~~few~~ new forms of processing to enable enhanced decision making, & discovery & process optimization "

Challanges →

1) Hetrogenity & Incompleteness :-

machine analysis algorithms expect homogeneous data & cannot understand nuance

2) Scale :-

Managing large & rapidly increasing volume of data has been a challanging issue for many decades

3) Timeliness :-

The larger the data set to be proceed, the longer it will take to analyze.

4) Privacy :-
The privacy of data is another huge concern & one that increases in the context of big data.

5) Human collaboration :-
A big data analysis system must support input from multiple human experts & shared exploration of results The data system has to accept this distributed expert input & support their collaboration.

b] compare business Intelligence & Big data.

Big data	Business Intelligence
1) Big data helps to find the questions you don't know you want to ask.	1) Business Intelligence helps to find answers to questions you know.
2) Big data is about technology	2) BI is always about informing business.
3) It includes gather & analyze data	3) It includes generation, aggregation, analysis of data.
4) This is technology that stores & processes the data from sources both internal & external to your company.	4) BI is data-driven decision making. It includes the generation, aggregation, analysis & visualization of data to inform & facilitate business management & strategizing

c] Write a note on Reinforcement learning.

→ Reinforcement learning :-

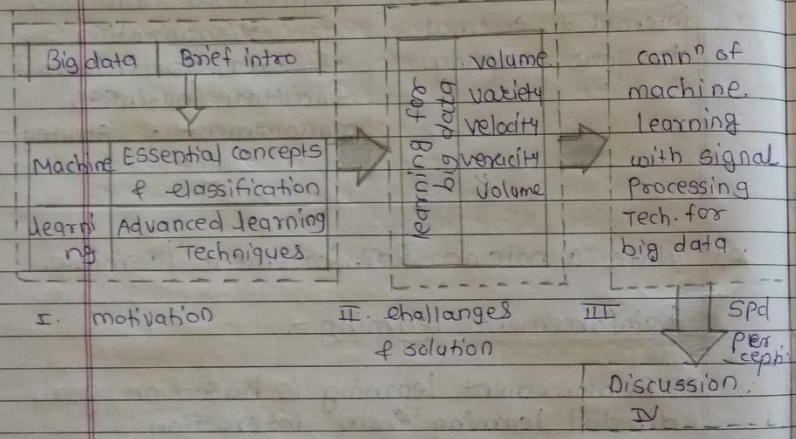
1) Reinforcement learning is based on goal directed learning from interaction.
2) Reinforcement learning maximizes a numerical reward signal by mapping situations in the actions
3) Reinforcement learning is by characterizing a learning problem not by method
4) All reinforcement learning agents have explicite goals & are intelligence to find the aspects of their environment.

Elements of Reinforcement learning :-

1) A policy
2) A Reward function
3) A value function
4) A model.

Q.1.8 a] Draw a diagram showing generalized systematic machine learning framework & explain

→ Machine learning →



1) Machine learning investigates how computers can learn (or improve their performance) based on data.
2) A main research area for computer programs to automatically learn to recognize complex patterns & make intelligent decisions based on data.
3) machine learning is a field of research that formally focuses on theory, performance & properties of learning systems & algorithms.
4) It is a highly interdisciplinary field building upon ideas from many different kinds of fields such as artificial intelligence, optimization theory, info. theory, statistics.
5) Most traditional machine-learning based systems are designed with the assumption that all the collected data would be completely loaded into memory for centralized processing.
6) However the data keeps getting bigger & bigger the existing machine learning tech encounter great difficulties when they are required to handle the unprecendented volume of data.

Machine learning Technologies :-

1) Supervised learning
2) Unsupervised learning
3) Semi-supervised learning
4) Active learning

b) Write a short note on multi perspective learning.

→ Multiperspective learning :-

1) multiperspective learning is needed for multiperspective decision making.
2) multiperspective learning refers to learning from knowledge & information collected from different perspectives.
3) multiperspective learning builds knowledge from various perspectives so that it can be used for decision making process.

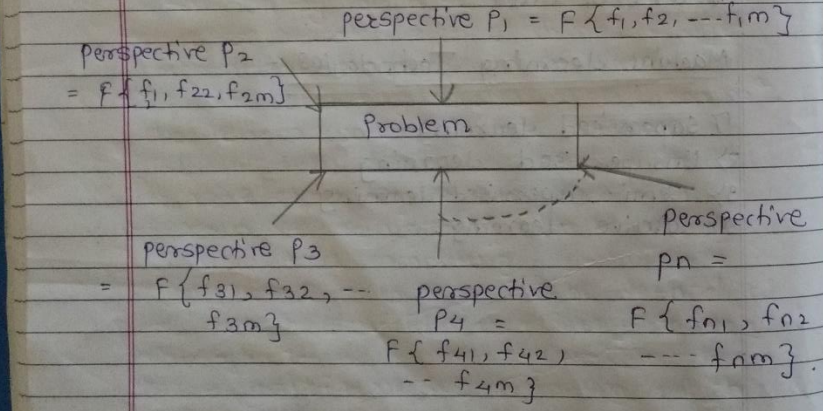


Fig:- multiperspective learning.

4) The perspective includes context, scenario & situation the way we look at a particular decision problem.
5) fig:- P1, P2, P3, -- Pn refers to different perspective in learning process.
6) Each of this perspective is represented as a function of features.
7) There may be an overlap among the perspective.
8) feature difference may be there as some feature which possibly visible from from one perspective may not be visible from the other perspective.