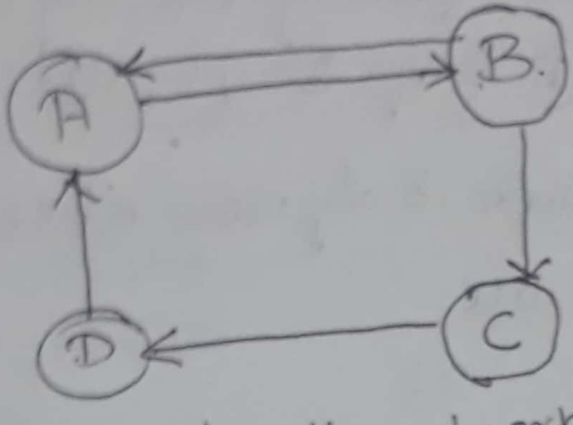
Assignment - 2

To identify the important node in the network by determining the rank of each node present in the network.



→ **Step-1:-** Consider the network with 4 nodes [A, B, C, D] and it has 6 directed links b/w the nodes

i.e,

* Node A links to node B
* Node B links to node C
* Node C links to node D
* Node D links to node A
* Node A links to node C
* Node B links to node A

Step 2:- Assign the rank "r" values for each node i.e,

Ra, Rb, Rc, Rd

* Node A has 2 outbound links i.e, B & C
* Node B has 2 outbound links i.e, A & C
* Node C has 2 outbound links i.e, D
* Node D has 2 outbound links i.e, A

Hence, we can write 4 equations:

$R_a = 0.5 * R_b + R_d \rightarrow (1)$

$R_b = 0.5 * R_a \rightarrow \textcircled{2}$

$R_c = 0.5 \times R_a + 0.5 * R_b \rightarrow \textcircled{3}$

$R_d = R_c \rightarrow \textcircled{4}$

Hence, there are 2 outbounds links from node A and node B, Both will be sharing its half of the influence respectively ∴ eqn ① & eqn ② & eqn ③

Step-3:- Solving mathematically now represent the equation in the form of matrix

$$\begin{array}{c|cccc} & R_a & R_b & R_c & R_d \\ \hline R_a & 0 & 0.50 & 0 & 1. \\ R_b & 0.50 & 0 & 0 & 0 \\ R_c & 0.50 & 0.50 & 0 & 0 \\ R_d & 0 & 0. & 1 & 0 \end{array}$$

Step-4:- Simplification of the problem of the using the initial rank values & then compute new rank values till they stabilize

Initial rank = $\frac{1}{n}$ = $1/4$ = 0.250

no of nodes

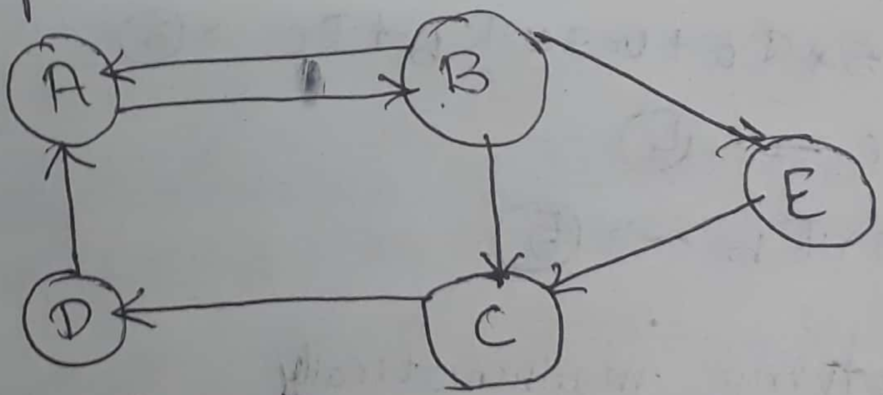
variable	Initial Value	Iteration-I
Ra	0.250	$R_a = 0.5 * R_b + R_d = 0.375$
Rb	0.250	$R_b = 0.5 * 0.250 = 0.125$
Rc	0.250	$R_c = 0.5 * 0.250 = 0.250$
Rd	0.250	$R_d = R_c = 0.250$.

	II	III	IV	V	VI
Ra	0.3125	0.34375	0.328125	0.3359375	0.33203125
Rb	0.1875	0.15625	0.171875	0.1640625	0.167968
Rc	0.25	0.25	0.25	0.25	0.25
Rd	0.250	0.25	0.25	0.25	0.25

e	VII	VIII
Ra	0.333984375	0.333007812
Rb	0.166015625	0.1669928187
Rc	0.25	0.2499999
Rd	0.25	0.25

∴ The strongest node is Ra

2b To identify the important node in the network by determining the rank of each node present in the network



→Step -1:- Consider the network with 5 nodes [A, B, C, D, E] & it has 8 directed links b/w the nodes i

* Node A links to node B

* Node B links to node C

* Node C links to node D

* Node D links to node A

* Node A links to node C

* Node B links to node A
* Node B links to node E
* Node C links ~~to~~ to node C

Step-2:- Assign the rank "R" values for each node i.e
Ra, Rb, Rc, Rd, Re

* Node A has 2 outbound links i.e, B & C
* Node B has 3 outbound links i.e, A, C & E
* Node C has 1 outbound links i.e, D
* Node D has 1 outbound links i.e, A
* Node E has 1 outbound links i.e, C

Hence, we can write 5 equations

$$R_a = 0.3 * R_b + R_d \rightarrow ①$$

$$R_b = 0.5 * R_a \rightarrow ②$$

$$R_c = 0.5 * R_a + 0.3 * R_b + R_c \rightarrow ③$$

$$R_d = R_c \Rightarrow ④$$

$$R_e = 0.3 * R_b \rightarrow ⑤$$

Step-3:- Solving mathematically,

	Ra	Rb	Rc	Rd	Re
Ra	0	0.3	0	1	0
Rb	0.5	0	0	0	0
Rc	0.5	0.3	0	0	1
Rd	0	0	1	0	0
Re	0	0.3	0	0	0

Step 4: Simplified of the problem of using the initial rank value & then compute new rank value till they stabilize.

Initial Rank = $1/n = 1/5 = 0.2$

n ↓ no of nodes.

Variable	Initial value	Iteration.
Ra	0.2	$0.3 \times R_b + R_d = 0.3 \times 0.2 + 0.2 = 0.26$
Rb	0.2	$0.5 \times R_a = 0.3 \times 0.2 = 0.1$
Rc	0.2	$0.5 \times R_a + 0.3 \times R_b + R_c = 0.36$
Rd	0.2	$R_d = R_c = 0.36$
Re	0.2	$R_e = 0.3 \times R_b = 0.3 \times 0.2 = 0.06$

	II	III	IV	V	VI	VII
Ra	0.28	0.399	0.2545	0.24385	0.31175	0.2582
Rb	0.13	0.115	0.1995	0.12725	0.121925	0.1556
Rc	0.22	0.184	0.273	0.2216	0.21995	0.2304
Rd	0.36	0.22	0.184	0.273	0.2216	0.2200
Re	0.03	0.039	0.0345	0.05985	0.038175	0.0366.

VIII	IX	X	XI	XII
Ra= 0.2667	0.2691	0.2524	0.2592	0.2512
Rb= 0.1291	0.1334	0.1346	0.1262	0.1292
Rc= 0.2124	0.2188	0.2133	0.2066	0.2079
Rd = 0.2304	0.2124	0.2188	0.2133	0.2066
Re = 0.0467	0.0384	0.0400.	0.0404	0.0579

XIII	XIV	XV	XVI	XVII	XVIII
Ra = 0.2454	0.2456	0.2391	0.2360	0.2332	0.2286
Rb = 1256	0.1227	0.1228	0.1196	0.1180	0.1166
Rc = 0.2023	0.1992	0.1973	0.1932	0.1907	0.1879
Rd =0.2079	0.2023	0.1992	0.1973	0.1932	0.1907
Re = 0.0388	0.0377	0.0368	0.0368	0.0359	0.0354

XIX	XX	XXI
Ra = 0.2257	0.2222	0.2186
Rb = 0.1143	0.1129	0.1111
Rc =0.1847	0.1822	0.1793
Rd = 0.1879	0.1847	0.1822
Re = 0.0350	0.0343	0.0339

∴ The Strongest node is Ra

3) Define Text mining ? Brief out its applications and Explains and Explain the best practices of text mining.

→ Text mining is the art and science of discovering Knowledge, insights and patterns from an organized.

Collection of textual databases.

Text Mining applications:-

* Marketing:- The voice of the customer can be captured in its native and raw format and then analyzed for customer preferences and complaints.

a) Social personas are a clustering techniques to develop customer segments of interest. Consumer input from the social media sources, such as reviews, blogs, and tweets, contain numerous leading indicators that can be used towards anticipating and predicting consumer behaviour.

b) A 'listening platform' is a text mining application, that in real time, gathers social media, blogs and other textual feedback and filters out the chatter to extract true consumer sentiments.

c) The customer call center conversation and records can be analyzed for patterns of customer complaints.

* Business operation:- Many aspects of business functio-ning can be accurately guauged from analyzing text.

a) Social network analysis and text mining can be applied to emails, blogs, social media and other data to measure the emotional states and the mood of employee populations.

b) Studying people as emotional investors and using text analysis of the Social internet to measure mass psychology can help in obtaining superior investment returns.

* **Legal** :- In legal applications, lawyers and paralegals can more easily search case histories and laws for relevant documents in a particular case to improve their chances of winning.

a) Text mining is also embedded in e-discovery platforms that help in minimizing risk in the process of sharing legally mandated documents

b) Case histories, testimonies, and client meeting notes can reveal additional information, such as morbidities in healthcare situations that can help better predict high-cost injuries and prevent costs.

* **Governance and Politics** :- Government can be overturned based on a tweet originating from a self-immolating fruitvendor in Tunisia.

a) Social network analysis and text mining of large-scale social media data can be used for measuring the emotional states and the mood of constituent populations.

b) In geopolitical security, internet chatter can be processed for real-time information and to connect the dots.

c) In academics, research streams could be meta-analyzed for underlying research trends.

Text Mining Best practices:-

* The first and most important practice is to ask the right question. A good question is the one which gives an answer and would lead to large payoffs for the organization.

* A second important practice is to be creative and open in proposing imaginative hypotheses for the solution.

* Another important element is to pursue the problem iteratively. Too much data can overwhelm the infrastructure and also befuddle the mind.

* A variety of data mining tools should be used to test the relationships in the TDM. Different decision tree-algorithms could be run alongside cluster analysis and other techniques.

4) Explain SVM Model and kernel method.

→

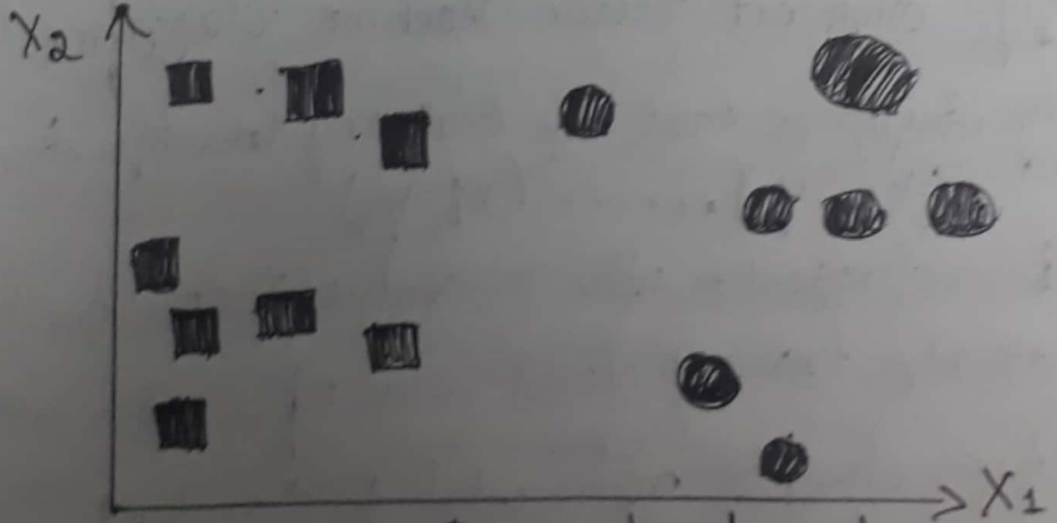


fig:- Data Points for classification

An SVM is a classifier function in a high-dimensional space that defines the decision boundary between two classes. The support vectors are the data points that defines the 'gutters' (or) the boundary condition either side of the hyperplane; for each of the two classes.

SVM takes the widest street approach to demarcate the 2 classes and thus finds the hyperplane that has the widest margin, i.e, largest distance to the nearest training data points of either class.

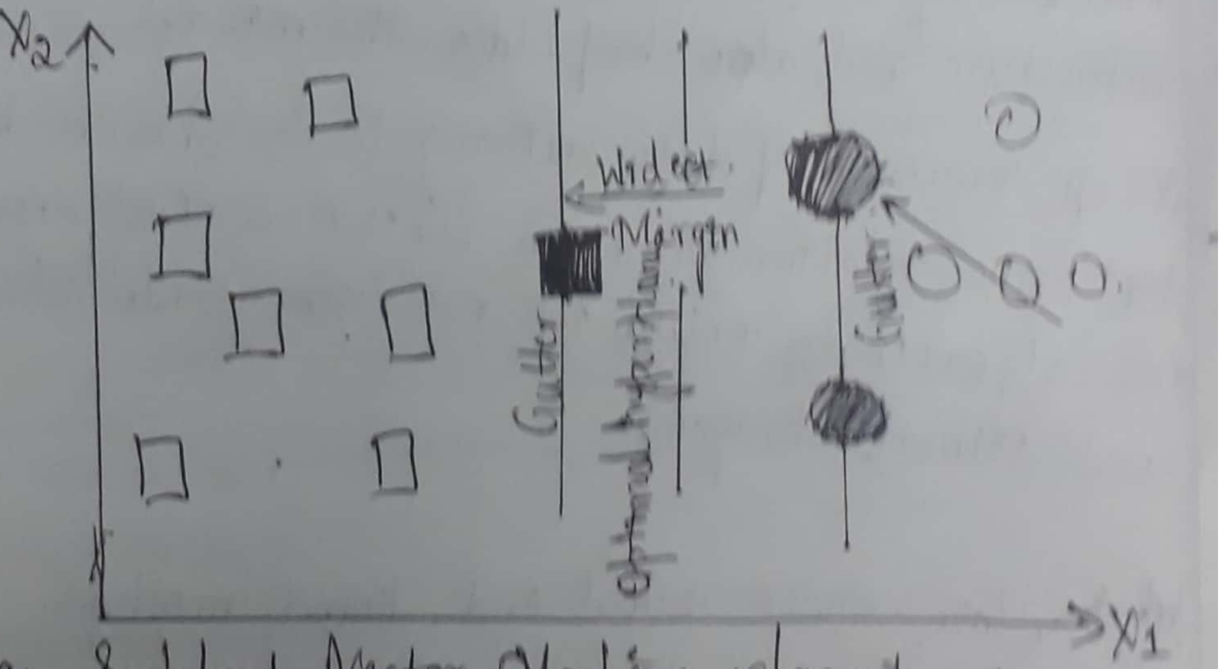


fig:- Support Vector Machine classifier

Abstractly, suppose that the training data of n points is,

$$(X_1, y_1) \ldots\ldots (x_i, y_i)$$

where, X_i represents the p-value vector for point i and y_i is its binary class value of 1 (or) -1. Thus, there are two classes represented as 1 and -1

Assuming that the data is indeed linearly separable,

$$W \cdot x + b = 0.$$

where, w = normal vector to the hyperplane.

The hard margins can be defined by the following

$$W \cdot x + b = 1 \text{ and } w \cdot x + b = -1$$

The width of the hard margin is $(2/|W|)$.

The SVM algorithm finds the weights vector (w) for the features such that there is widest margin b/w the 2 slopes.

The Kernel method:-

* The heart of an SVM algorithm is the Kernel method.

* Most kernel algorithms are based on optimization in a convex space and are statistically well-founded.

*. Kernel stands for the core of the germ in a fruit.

* Kernel methods operate using what is called the 'Kernel trick'. This Kernel trick involves computing and working with the inner products of only the relevant pairs of data in the feature space, they do not need to compute all the data in high-dimensional feature space.

* Kernel methods achieve this by learning from instances. They do not apply some standard computational logic to all the features of each input.

(C Briefly explain the web mining structure"

→.

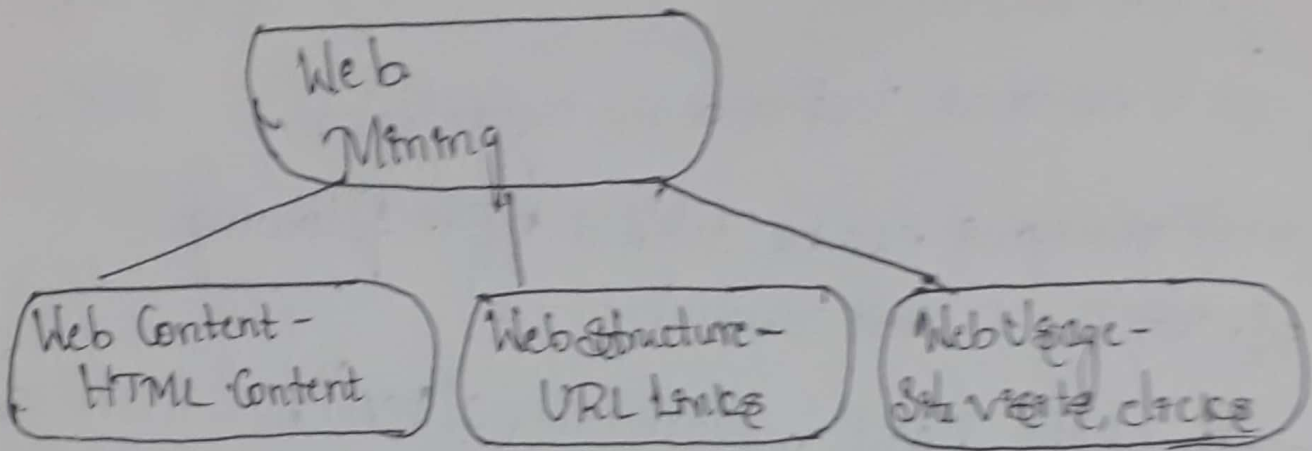

Web Mining
Web Content - HTML Content
Web Structure - URL Links
Web Usage - Site visits, clicks

fig:- Web mining Structure.

a) Web Content Mining:- A website is designed in the form of pages with a distinct URL [Universal Resource Locator]. A large website may contain thousands of pages. These pages are managed using specialized software systems called Content Management Systems.

b) Web Structure mining:- The web works through a system of hyperlinks using the hypertext protocol (http). There are 2 basic strategic models for successful website - Hubs and Authorities.

* Hubs:- These are pages with a large number of interesting links. They serve as hub or a gathering point

* Authorities:- Ultimately, people would gravitate towards pages that provide to most complete and authoritative information on a particular subject.

c) Web Usage mining:- As a user clicks anywhere on a webpage or application, the action is recorded by many entities in many locations. The browser at the client machine will record the click and the web server providing the content would also make a record of the pages served and user activity on those pages.

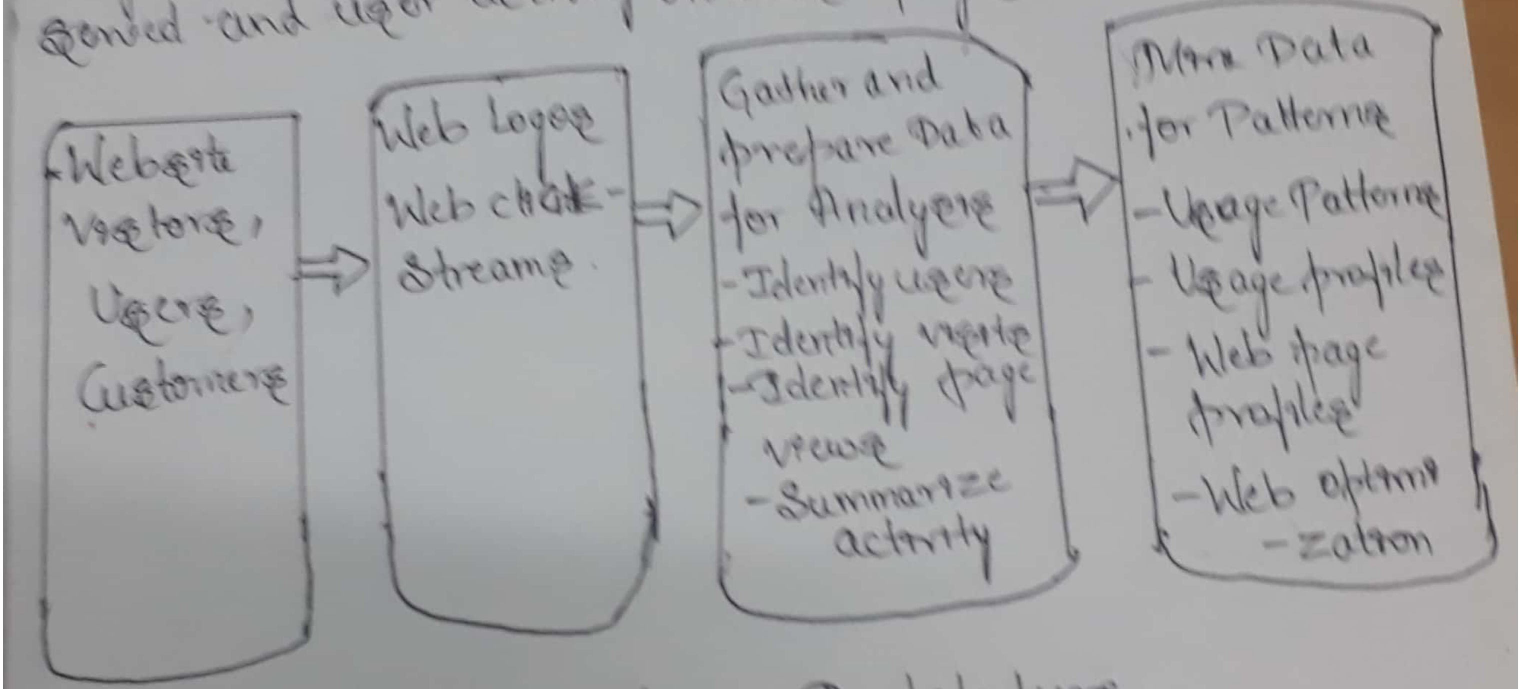


fig:- Web Usage Mining Architecture.

* Usage pattern could be analyzed using 'clickstream' analysis i.e, analyzing web activity for patterns of sequence of clicks and the location and duration of visits on websites.

* Textual information accessed on the pages retrieved by users could be analyzed using text mining techniques.

6. Differentiate between SNA and TDA

→