SHRUTI SINGH, TE3 (D-BATCH), 67. Date:

ASSIGNMENT No. 1

Q1] Data warehouse for weather bureau.

→ IRD:

Data-dimension	region-dimension	temp-dim	pereciption-dim
Time	probe-location	temperature	precipitation
month	region_name	temp-range	prec_range
Quarter	district	temp-desc	prec_desc
Year	city	temp-unit	
season	country		
	state		

Fact: region-map(), count(), area(), Pressure()

1. STAR SCHEMA

region_dim
- region_location
- region_name
- district
- country
- city
- state

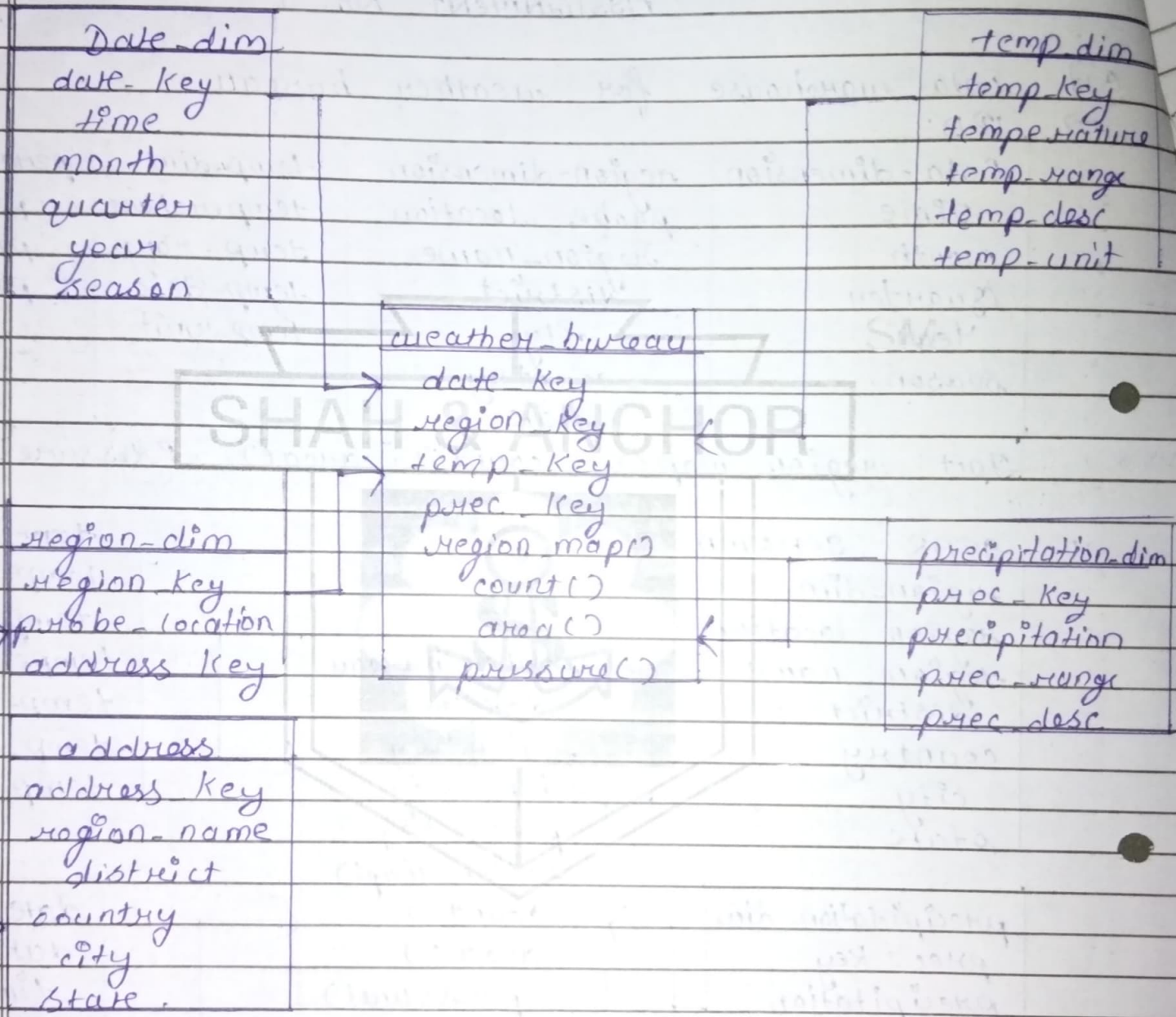
weather_bureau
- date key
- region_key
- temp_key
- prec_key
- region_map()
- count()
- area()
- pressure()

temp
- temp-dim
- temp_key
- temperature
- temp-range
- temp_desc
- temp_unit

precipitation_dim
- prec_key
- precipitation
- prec_range
- prec_desc

date_dim
- date_key
- time
- month
- quater
- year
- season

2. SNOWFLAKE SCHEMA



Date:

Q2 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70

a) Mean = $\frac{809}{27}$, 29.96 years .

median = $\frac{(27+1)}{2}$, 14th term = 25 years

b) Mode = 25, 35
Data's modality is Bimodal

c) mid Range = $\frac{(max + min)}{2} = \frac{(70+13)}{2} = \frac{83}{2}$, 41.5 year

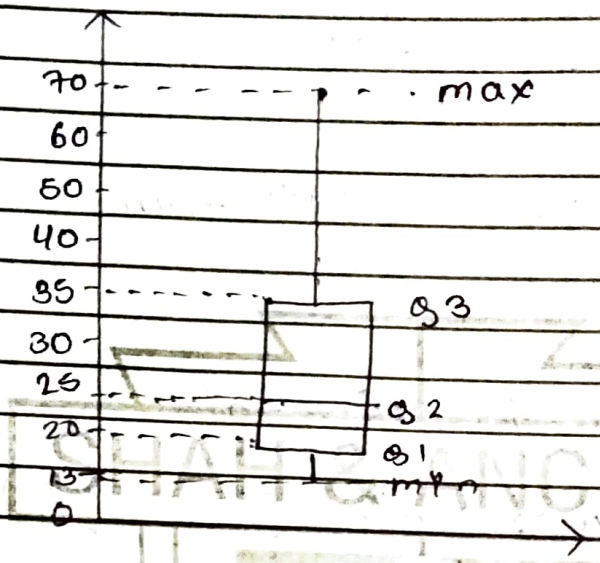
d) Q1 will be median of first half of dataset i.e median of
13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25
∴ Q1 = 20 (first quartile)

Q3 will be median of second half of dataset i.e median of
30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70
∴ Q3 = 35 (third quartile)

e) 5 number Summary :
min = 13
1st quartile Q1 = 20
median Q2 = 25
3rd quartile Q3 = 35
max = 70

Date:

f) Box Plot :



Q3]

DEPARTMENT	AGE	SACARY	COUNT	STATUS
Sales	31...35	46K...50K	30	Senior
Sales	26...30	26K...30K	40	Junior
Sales	31...35	31K...35K	40	Junior
Systems	21...25	46K...50K	20	Junior
Systems	26...30	66K...70K	5	Senior
Systems	41...45	46K...50K	3	Junior
Systems	36...40	66K...70K	8	Senior
marketing	31...35	46K...50K	10	Senior
marketing	41...45	41K...45K	4	Junior
Secretary	46...50	36K...40K	4	Senior
Secretary	26...30	26K...30K	6	Junior

use Naive Bayesian to find status of Tuple :
("System", "26.30", "46K...50K")

Finding Prior Probability

P(C1: status : "junior") = 6/11 = 0.55

P(C2: status : "senior") = 5/11 = 0.45

Finding conditional probability

P(department = "systems" / status = "senior") = 2/5 = 0.4

P(department = "systems" / status = "junior") = 2/6 = 0.33

P(age = "26... 30" / status = "senior") = 0/5 = 0

P(age = "26... 30" / status = "junior") = 3/6 = 0.5

P(salary = "46K... 50K" / status = "senior") = 2/5 = 0.4

P(salary = "46K... 50K" / status = "junior") = 2/6 = 0.33

using above probabilities, we obtain

P(x / status = 'junior') = 0.33 × 0.5 × 0.33
= 0.05

P(x / status = 'senior') = 0.4 × 0 × 0.4 = 0

P(x / status = 'junior') · P(status = 'junior') = 0.05 × 0.55
= 0.0275
= 0.03

P(x / status = 'senior') · P(status = 'senior') = 0 × 0.45
= 0

∴ 0.03 > 0 naive bayesian classifier predicts status of given tuple as "junior"

Q4

Point	x	y
A1	2	10
A2	2	5
A3	8	4
B1	5	8
B2	7	5
B3	6	4
C1	1	2
C2	4	9

Initial cluster centres : A (2,10) B(5,8) C1 (1,2)

calculate the distance of each point from each cluster center

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

for A1:

d (A1, A1) = 0
d (A1, B1) = 3.60
d (A1, C1) = 8.06

d (A2, A1) = 5
d (A2, B1) = 4.24
d (A2, C1) = 3.16

d (A3, A1) = 8.49
d (A3, B1) = 5
d (A3, C1) = 7.28

d (B1, B1) = 0
d (B1, A1) = 3.6
d (B1, C1) = 7.21

d (B2, A1) = 7.07
d (B2, B1) = 3.60
d (B2, C1) = 6.70

d (B3, A1) = 7.21
d (B3, B1) = 4.12
d (B3, C1) = 5.38

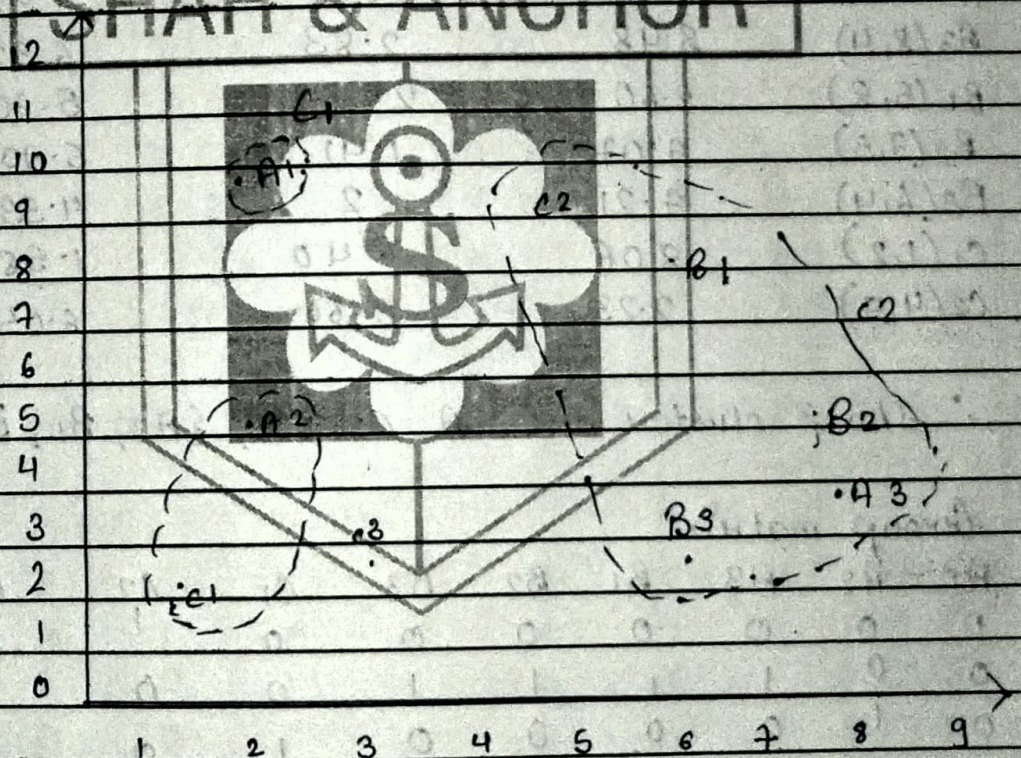
d (C1, A1) = 8.06
d (C1, B1) = 7.21
d (C1, C1) = 0

d (C2, A1) = 2.23
d (C2, B1) = 1.41
d (C3, C1) = 7.61

Date:

POINT	DISTANCE from A_1	DIST from B_1	Dist from C_1	Cluster
$A_1(2,10)$	0	3.6	8.06	A_1
$A_2(2,5)$	5	4.24	3.16	C_1
$A_3(8,4)$	8.49	5	7.28	B_1
$B_1(5,8)$	3.6	0	7.21	B_1
$B_2(7,5)$	7.07	3.6	6.7	B_1
$B_3(6,4)$	7.21	4.12	5.38	B_1
$C_1(1,2)$	8.06	7.21	0	C_1
$C_2(4,9)$	2.23	1.41	7.61	B_1

∴ The 3 cluster are : {A4} {A3, B1, B2, B3, C2}, {A2, C1}



Group Matrix a :

A_1	A_2	A_3	B_1	B_2	B_3	C_1	C_2	Centroid
1	0	0	0	0	0	0	0	$A_1(2,10)$
0	0	1	1	1	1	0	1	$B_1(5,8)$
0	1	0	0	0	0	1	0	$C_1(1,2)$

Date:

∴ new centroids after first iteration

$C1 = (2, 10)$

$C2 = \left(\frac{8+6+7+6+4}{5}, \frac{4+8+5+4+9}{5}\right) = (6, 6)$

$C3 = \left(\frac{2+1}{2}, \frac{5+2}{2}\right) = (1.5, 3.5)$

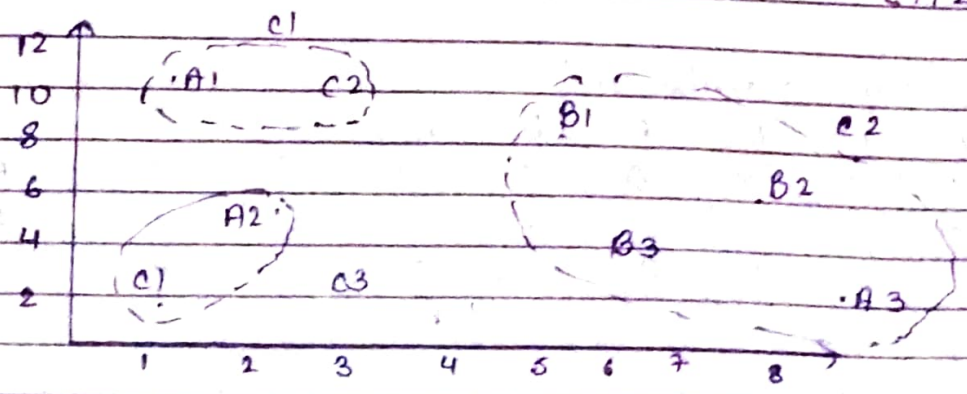
Iteration - 2:

Point	Dist from C1	Dist from C2	Dist from C3	Cluster
A1 (2,10)	0	5.65	6.52	C1
A2 (2,5)	5	4.12	1.58	C3
A3 (8,4)	8.48	2.83	6.52	C2
B1 (5,8)	3.60	2.23	5.70	C2
B2 (7,5)	7.07	1.41	5.70	C2
B3 (6,4)	7.21	2	4.53	C2
C1 (1,2)	8.06	6.40	1.58	C3
C2 (4,9)	2.23	8.66	6.04	C1

∴ The 3 cluster are {A, C2} , {A3, B1, B2, B3} {A2, C1}

Group matrix:

A1	A2	A3	B1	B2	B3	C1	C2	Centroid
1	0	0	0	0	0	0	1	A1 (2,10)
0	0	1	1	1	1	0	0	B1 (5, 8)
0	1	0	0	0	0	1	0	C1 (1,2)



∴ new centroids after 2nd iteration

$C_1 = \left(\frac{2+4}{2}, \frac{10+9}{2}\right) = (3, 9.5)$

$C_2 = \left(\frac{8+5+7+6}{4}, \frac{4+8+5+4}{4}\right) = (6.5, 5.25)$

$C_3 = \left(\frac{2+1}{2}, \frac{5+2}{2}\right) = (1.5, 3.5)$

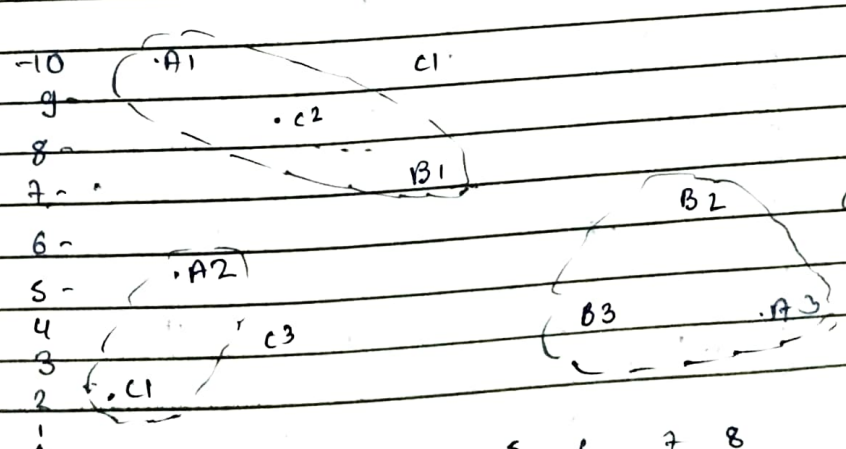
Iteration 3:

POINT	Dist from C1	Dist from C2	Dist from C3	cluster
$A_1(2,10)$	1.12	6.54	6.52	C1
$A_2(2,5)$	4.61	4.51	1.58	C3
$A_3(8,4)$	7.43	1.95	6.52	C2
$B_1(5,8)$	2.5	3.13	5.70	C1
$B_2(7,5)$	6.02	0.56	5.70	C2
$B_3(6,4)$	6.26	1.35	4.53	C2
$C_1(1,2)$	7.76	6.39	1.58	C3
$C_2(4,9)$	1.12	4.61	6.04	C1

∴ The cluster are - $\{A_1, B_1, C_2\}$ $\{A_3, B_2, B_3\}$ $\{A_2, C_1\}$

Group matrix

A1	A2	A3	B1	B2	B3	C1	C2	Centroid
1	0	0	1	0	0	0	1	$A_1(2,10)$
0	0	1	0	1	1	0	0	$B_1(5,8)$
0	1	0	0	0	0	1	0	$C_1(1,2)$



∴ new centroids after 3rd iteration are:

$C_1 = \left(\frac{2+5+4}{3}, \frac{9+8+10}{3}\right) = (3.67, 9)$

$C_2 = \left(\frac{8+7+6}{3}, \frac{4+5+4}{3}\right) = (7, 4.33)$

$C_3 = (1.5, 3.5)$

Point	Dist from C1	Dist from C2	Dist from C3	Cluster
A1(2,10)	1.95	7.56	6.52	C1
A2(2,5)	4.33	5.04	1.58	C3
A3(8,4)	6.61	1.05	6.52	C2
B1(5,8)	1.66	4.18	5.70	C1
B2(7,5)	5.20	0.67	5.70	C2
B3(6,4)	5.51	1.05	4.53	C2
C1(1,2)	7.49	6.44	1.58	C3
C2(4,9)	0.67	5.56	6.04	C1

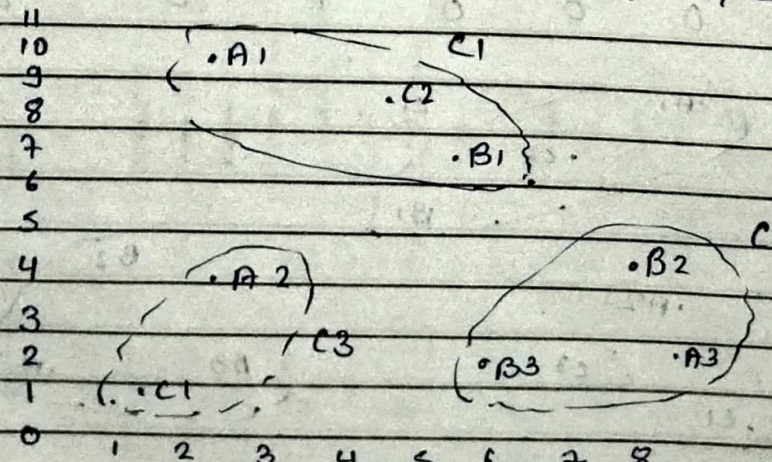
The 3 cluster are: {A1, B1, C2} {A3, B2, B3} {A2, C1}

Group matrix:

A1	A2	A3	B1	B2	B3	C1	C2	Centroid
1	0							
0	0							
0	1							

cluster in iteration 4 is same as in iteration 3

∴ Final cluster {A1, B1, C2} {A3, B2, B3} {A2, C1}



Q5 APRIORI

TID	items_bought
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y}
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, O, K, I, E}

minimum support = 60%, minimum confidence = 80%

Support = $\frac{\text{Tuples containing } (A+B)}{\text{no. of tuples}}$

$\frac{60}{100} = \frac{x}{5}$

$x = 3$ i.e. min support count = 3

C_1 =

item	Count
M	3
O	4
N	2
K	5
E	4
Y	3
O	1
A	1
U	1
C	2
I	1

After pruning, we get: (count >= 3)

	Item	Count
L1 =	m	3
	o	3
	K	5
	E	4
	Y	3

Generating C2 from L1:

	Item	Count
	mo	1
	mK	3
C2 =	mE	2
L1 ⋈ L1	mY	2
	oK	3
	oE	3
	oY	2
	KE	4
	KY	3
	EY	2

After pruning, we get (count >= 3)

	Item	Count
L2 →	mK	3
	oK	3
	oE	3
	kE	4
	KY	3

Date:

Generate C3 from L2:

C3 = L2 ⋈ L3

Item	Count
OKE	3
KEY	2
OEY	2
MKY	2
MOE	1
MKO	1
KOY	2

After pruning, we get (count >= 3)

Item	Count
OKE	3

∴ frequent item set = {O, K, E}

∴ we generate association rules for {O, K, E}

Confidence = Tuples containing A & B / Tuples containing only A

∴ Association rules are:

	Confidence
O ∧ K ⇒ E	3/3 = 100%
O ∧ E ⇒ K	3/3 = 100%
E ∧ K ⇒ O	3/4 = 75%
E ⇒ O ∧ K	3/4 = 75%
K ⇒ O ∧ E	3/5 = 60%
O ⇒ E ∧ K	3/3 = 100%

Now, we have to find strong association rule i.e confidence = 80%.

Date:

∴ The strong association rules are:

$O \wedge K \Rightarrow E$

$O \wedge E \Rightarrow K$

$O \Rightarrow E \wedge K$

Q6 Describe HITS algorithm

→ clever aimed at finding both authorative pages & hubs

→ Hubis a page that contains links to authoritative pages

→ The clever system identifies authorative pages & hubs by creating weights

→ HITS stand for Hyperlink-Induced topic Search

→ It is used to the web-link structures to discover & rank the webpage relevant for a particular search

1. Based on a given set of keywords, a set of relevant pages is found
2. Hub & authority measures are associated with these pages. Pages with highest values are returned.

→ w // www viewed as directed graph

q // query

s // support

OUTPUT:

A // set of authority pages

H // set of hub pages

HITS algorithm

R = SE (w, q)

B = R ∪ {pages linked to from R} ∪ {pages that link to page in R}

G = (B, L) = subgraph of w induced by B

G = (B, L) = Delete links in G within the same site

xp = q where (q, p) ∈ q ; // find authority of weights

yp = where (p, q) p2 ∈ q; // find hub weights

A = {p | p has one the height xp);

H = {p | p has one of the height yp);