DMW

TT2

JUNAID GIRKAR
6000419005 57
TE COMPS A4

13/12/2021

Q1

Transaction ID	Items
100	I1, I2, I5
200	I2, I4
300	I2, I3
400	I1, I2, I4
500	I1, I3
600	I1, I3
700	I1, I3, I2, I5
800	I1, I3
900	I1, I2, I3

APRIORI ALGORITHM

Minimum support count = 2

Minimum confidence = 70%.

STEP 1: Frequency table of all items

ITEM	FREQUENCY
I1	7
I2	6
I3	6
I4	2
I5	2

STEP 2: Removing items with frequency < Min support count.
NO items removed.

6000419005 7

C_1

Item	SUPPORT COUNT
I1	7
I2	6
I3	6
I4	2
I5	2

L_1

Item	support count
I1	7
I2	6
I3	6
I4	2
I5	2

ITEMSET	support count
{I1, I2}	4
{I1, I3}	5
{I1, I4}	1
{I1, I5}	2
{I2, I3}	3
{I_2, I4}	2
{I3, I4}	0
{I3, I5}	1
{I2, I5}	2
{I4, I5}	0

6000419 0057

L	Item	support count
	$\{I_1, I_2\}$	4
	$\{I_1, I_3\}$	5
	$\{I_1, I_5\}$	2
	$\{I_2, I_3\}$	3
	$\{I_2, I_4\}$	2
	$\{I_2, I_5\}$	2

C	Item	support count
	$\{I_1, I_2, I_3\}$	2
	$\{I_1, I_2, I_5\}$	2
	$\{I_1, I_3, I_5\}$	1
	$\{I_2, I_3, I_4\}$	0
	$\{I_2, I_3, I_5\}$	1
	$\{I_2, I_4, I_5\}$	0

L	Items	support count
	$\{I_1, I_2, I_3\}$	2
	$\{I_1, I_2, I_5\}$	2

C	Items	support count
	$\{I_1, I_2, I_3, I_5\}$	1

So selecting above iteration frequent itemset

Association Rule :

$\{I_1, I_2, I_3\}$

$$\{I_1, I_2\} \rightarrow \{I_3\} = \frac{\text{supp}(I_1, I_2, I_3)}{\text{supp}(I_1, I_2)}$$

$$= \frac{2}{4} \times 100$$

confidence = 50%

$$\{I_1, I_3\} \rightarrow \{I_2\} = \frac{2}{5} \times 100$$

$$= 40\%$$

$$\{I_2, I_3\} \rightarrow \{I_1\} = \frac{2}{3} \times 100$$

$$= 66.67\ \%$$

$$\{I_3\} \rightarrow \{I_1, I_2\} = \frac{2}{6} \times 100$$

$$= 33.33\ \%$$

$$\{I_2\} \rightarrow \{I_1, I_3\} = \frac{2}{6} \times 100$$

$$= 33.33\ \%$$

$$\{I_1\} \rightarrow \{I_2, I_3\} = \frac{2}{7} \times 100$$

$$= 28.57\ \%$$

6000419005

$\{I_1, I_2\} \rightarrow \{I_5\} = \frac{2}{4} \times 100$

$= 50\%$

$\{I_1, I_5\} \rightarrow \{I_2\} = \frac{2}{2} \times 100$

$= 100\%$

$\{I_2, I_5\} \rightarrow \{I_1\} = \frac{2}{2} \times 100$

$= 100\%$

$\{I_5\} \rightarrow \{I_1, I_2\} = \frac{2}{2} \times 100$

$= 100\%$

$\{I_2\} \rightarrow \{I_1, I_5\} = \frac{2}{6} \times 100$

$= 33.33\%$

So strong association rules are

$\{I_1, I_5\} \rightarrow \{I_2\}$

$\{I_2, I_5\} \rightarrow \{I_1\}$

$\{I_5\} \rightarrow \{I_1, I_2\}$

G0004190057

Q2

ANS A web crawler is an automated program that scans or crawls through the internet pages to create an index of the data. It is also known as a web spider, web robot, bot, crawler and automatic indexer.

web crawling is considered to be an important method for collecting data and keeping up with the expanding internet and therefore search engines make use of web crawler to collect information about the data on public web pages. and their primary purpose is to collect data so that when a user enters a search term on their site, they can be quickly provided with relevent web sites

Types of crawlers :-

1) Traditional crawler
2) Periodic crawler
3) Incremental crawler
4) Focused crawler.

REGULAR	FOCUSSED
• Regular crawlers scrape every link	• It crawlers based on focussed keywords :: faster
• It has low precision	• It has higher precision.

G0000190057

Q3 b

ANS Data warehouse design is one of the key technique in building the data warehouse. Choosing a right data warehouse design can save the project time and cost. Basically there are two data warehouse design approaches that are popular :-

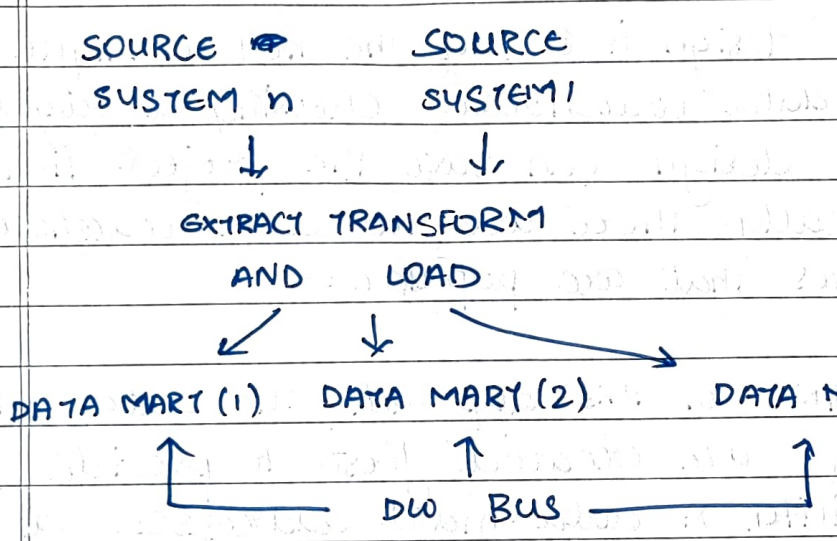
1) BOTTOM-UP DESIGN: In this approach, the data marts are created first to provide reporting capability. A data-mart addresses a single business area such as sales, finance etc. These data marts are then integrated to build a complete data warehouse.

ADVANTAGES :

- Documents can be quickly generated
- The data warehouse can be entended to accomodate new business units
- This model contains consistent data marts and these data marts can be integrated with other data marts and ~~devired~~ delivered quickly.

DISADVANTAGES

- The positions of the data warehouse and the data marts are reversed in the bottom-up approach.



2) TOP-DOWN DESIGN APPROACH: This is a data-driven approach as the information is gathered and integrated first and then business requirements by subjects for building data marts are formulated. Since, this method supports a single integrated data source, the data marts built from it will have consistency when they overlap

ADVANTAGES:
- Data marts are loaded from the data warehouses
- Developing new data mart from the data warehouse is very easy.

DISADVANTAGES:
- This technique is inflexible to changing department needs.
- The cost of implementing the project is high

6000190057