PART- A DMT DA - D1 SLOT PRIYAL BHARDWAJ 18BIT0272

Q.3 The Apriori algorithm makes use of prior knowledge of subset support properties.

(a) Prove that the all nonempty subsets of a frequent itemset must also be frequent.

(b) Prove that the support of any nonempty subset s0 of itemset s must be at least as great as the support of s.

(C) Given frequent itemset l & subset s of l, prove that the confidence of the rule "s0 ⇒ (l−s0)" cannot be more than the confidence of "s ⇒ (l−s)", where s0 is a subset of s.

(d) A partitioning variation of Apriori subdivides the transactions of a database D into n non overlapping partitions. Prove that any itemset that is frequent in D must be frequent in at least one partition of D.

A.3 (a) Let s be a frequent itemset.

Let min-sup be the minimum support.

Let D be the task-relevant data, a set of database transactions.

Let |D| be the number of transactions in D.

Since s is a frequent itemset support_count (s) = min-sup × |D| .

Let s0 be any non-empty subset of s .

Then any transaction containing itemset s will also contain itemset s0.

Therefore, support_count (s0) ≥ support_count (s) = min-sup × |D|.

Thus, s0 is also a frequent itemset.

(b) Let D be the task-relevant data, a set of database transactions.

Let |D| be the number of transactions in D.

By definition, support (s) = $\frac{\text{support\_count}(s)}{|D|}$

Let s' be any non-empty subset of s.

By definition, support (s0) = $\frac{\text{support\_count}(s0)}{|D|}$

From (a) support(s0) ≥ support(s). This proves that the support of any non-empty subset s0 of itemset s must be as great as the support of s.

(c) Let s be a subset of l. Then confidence $(s \Rightarrow (l-s)) = \frac{support(l)}{support(s)}$

Let $s0$ be any nonempty subset of s. Then confidence $(s0 \Rightarrow (l-s0)) = \frac{support(l)}{support(s0)}$

From (b) support $(s0) \geq$ support (s), therefore, confidence $(s0 \Rightarrow (l-s0)) \leq$ confidence $(s \Rightarrow (l-s))$.

i.e The confidence of the rule "$s0 \Rightarrow (l-s0)$" cannot be more than the confidence of the rule "$s \Rightarrow (l-s)$".

(d) Any itemset that is frequent in D must be frequent in at least one partition of D. Let us prove this by contradiction.

Assume that the itemset is not frequent in any of the partitions of D.

Let F be any frequent itemset.

Let D be the task-relevant data, a set of database transactions.

Let C be the total number of transactions in D.

Let A be the total number of transactions in D containing the itemset F.

Let min_sup be the minimum support.

F is a frequent itemset which means that $A = C \times min\_sup$.

Let us partition D into n non-overlapping partitions, $d_1, d_2, \ldots d_n$.

Then $D = d_1 d_2 d_3 \ldots d_n$.

Let $c_1, c_2, c_3 \ldots c_n$ be the total number of transactions in partitions $d_1, d_2, d_3 \ldots d_n$, respectively. The $C = c_1 + c_2 + c_3 + \ldots + c_n$.

Let $a_1, a_2, a_3 \ldots a_n$ be the total number of transactions in partitions $d_1, d_2, d_3 \ldots d_n$ containing the itemset F respectively. Then $A = a_1 + a_2 + a_3 + \ldots + a_n$

We can rewrite $A = C \times min\_sup$ as $(a_1 + a_2 + a_3 + \ldots a_n) = (c_1 + c_2 + \ldots c_n) \times min\_sup$.

Because of the assumption that we made at the beginning, we know that F is not frequent in any of the partitions $d_1, d_2, d_3 \ldots d_n$ of D.

This means that $a_1 < c_1 \times min\_sup$; $a_2 < c_2 \times min\_sup$; ....., $a_n < c_n \times min\_sup$.

Adding up all the inequalities we get $(a_1 + a_2 + \ldots + a_n) < (c_1 + c_2 + \ldots + c_n) \times \cancel{min\_sup}$ S or simply $A < C \times min\_sup$, meaning that F is not a frequent itemset.

But this is a contradiction since F was defined as a frequent itemset at the start of the proof. This proves that any itemset that is frequent in D must be frequent in at least one partition of D.

PART-B

Q.1 Radio-frequency identification is commonly used to trace object movement & perform inventory control. An RFID reader can successfully read an RFID tag from a limited distance at any scheduled time. Suppose a company wants to design a data warehouse to facilitate the analysis of objects with RFID tags in an online analytical processing manner. The company registers huge amounts of RFID data in format of (RFID, at location, time), & also has some information about the objects carrying the RFID tag, for example (RFID, product name, product category, producer, date produced, ~~time~~ price).

(a) Design a data warehouse to facilitate effective registration & online analytical processing of such data.

(b) The RFID data may contain lots of redundant information. Discuss a method that maximally reduces redundancy during data registration in the RFID data warehouse.

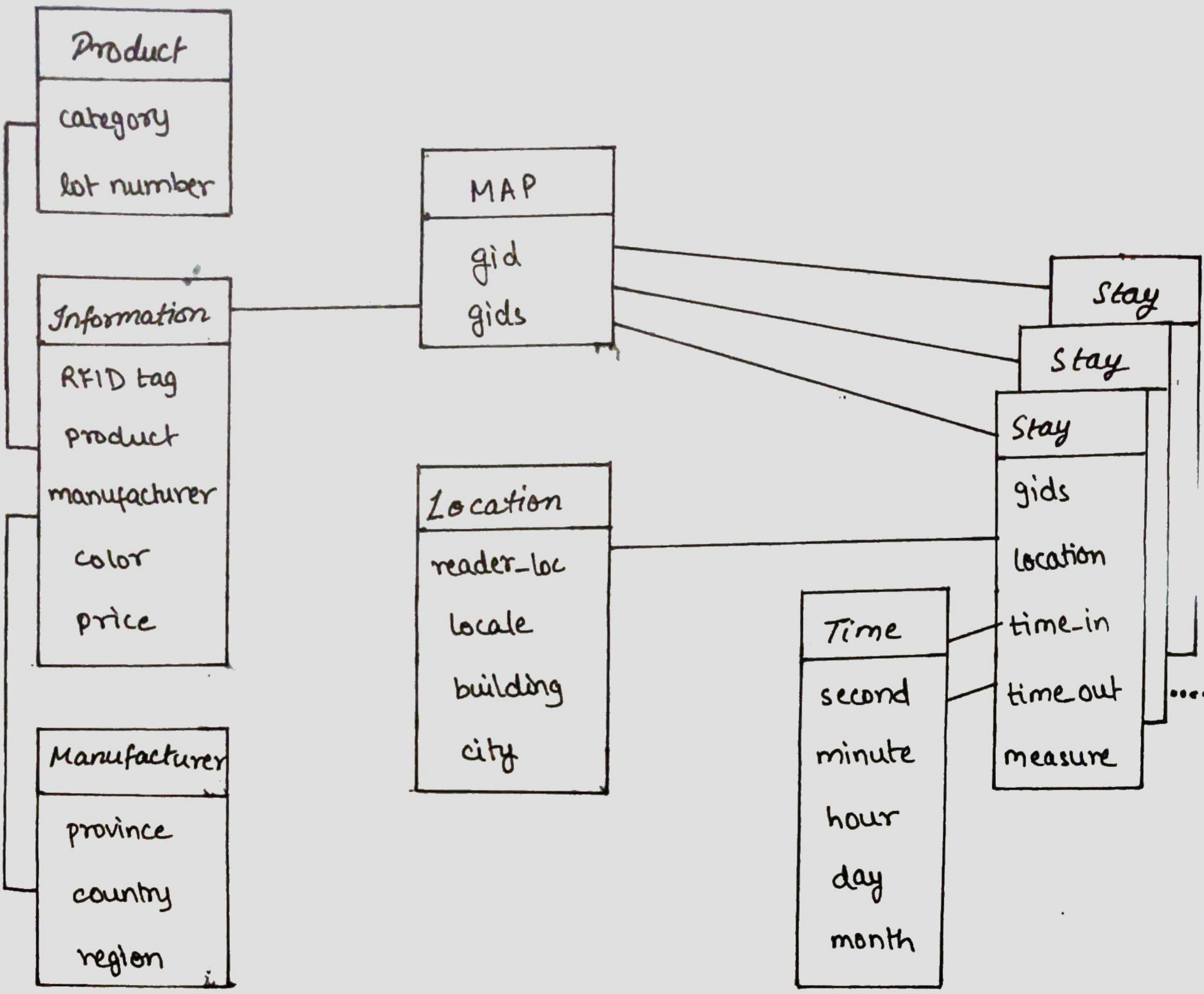
(c) The RFID data may contain lots of noise such as missing registration & misread IDs. Discuss a method that effectively cleans up the noisy data in the RFID data warehouse.

(d) You may want to perform an online analytical processing to determine how many TV sets were shipped from LA seaport to Best Buy in Champaign, IL, by month, brand, & price range. Outline how this could be done efficiently if you were to store such RFID data in the warehouse.

(e) If a customer returns a jug of milk & complains that it has spoiled before its expiration date, discuss how you can investigate such a case in the warehouse to find out what the problem is, either in shipping or in storage.

A.1 (a) A RFID warehouse needs to contain a fact table, stay, composed of cleansed RFID records; an information table, info, that stores path-independent information for each item; & a map table that links together different records in the fact table that form a path. The main difference between RFID warehouse & a traditional warehouse is the presence of the map table linking records from the fact table (stay) in order to preserve the original structure of the data.

• Star schema data warehouse:



(b) Each reader provides tuples of the form (RFID; location; time) at xed time intervals. When an item stays at the same location, for a period of time, multiple tuples will be generated. We can group these tuples into a single one of the form (RFID; location; time in; time out). For example, if a supermarket has readers on each shelf that scan the items every minute, & items stay on the shelf on average for 1 day, we get a 1440 to 1 reduction in size without loss of information.

(c) One can use the assumption that many RFID objects stay or move together, especially at the early stage of distribution, or use the historically most likely path for a given item, to infer or interpolate the miss & error reading.

(d) Compute an aggregate measure on the tags that travel through a set of locations & that match a selection criteria on path independent dimensions.

(e) For this case, we can obtain the RFID of the milk, we can directly use traditional OLAP operations to get the shipping & storage time efficiently.