

Hand-out

Advanced database technology

April 3, 2006

1 ADBT exam, June 2005: Data mining (20 %)

Consider the following 8 “market baskets”:

shampoo	beer	beer	shampoo	chips	milk	beer	diapers
milk	diapers	chips	diapers	coke	eggs	milk	milk
diapers	chips		eggs	beer	diapers	coke	
eggs			flour				

a) Suppose we run the Apriori algorithm on the above data to find pairs of items with support at least 3.

- Which items are found by the first pass?
- Which pairs of items are found by the second pass?

b) For each possible association rule with support at least 3, i.e., involving the pairs found in a), state the confidence and interest of the association.

The Apriori algorithm runs in two passes under the assumption that there is sufficient internal memory to hold all occurring pairs of high support items in internal memory. Obviously, it would be nice to run the algorithm efficiently also in the case where the size, M , of internal memory is not sufficiently large.

c) Let N denote the total size of the data processed and stored by the Apriori algorithm, i.e., the market baskets, the item counts, and the pair counts. Show that both the first and the second pass of Apriori can be implemented to run in $O(\frac{N}{B} \log_{M/B}(N/M))$ I/Os without any assumptions about M . (**Hint:** Use an algorithm discussed in the course as a subroutine in your solution.)

2 Last hand-in (due April 24)

ADBT exam, June 2003, problems 3 and 5. **Hint:** Problem 3 is easier than it might look.