Apriori implementation using Matlab

Luting Chen 50133507, Yuze Liu 50207903, Vicky Zheng 50037709

Introduction

The apriori algorithm is used to determine association rules. In this homework, we implemented
the apriori algorithm by using Matlab. The dataset we used the algorithm on is gene data. The
applications of this are straight forward: with association rules we can see how genes relate to
each other and what the relation between genes can tell us about the data.

Data Preprocesing

The data provided is a .txt file that contains a 100 by 102 matrix. There are 100 samples each
corresponding to the genes of a patient with a disease. The diseases are: ALL, AML, Breast
Cancer and Colon Cancer. In order to optimize our program, we converted each of these values
into integers: 1,2,3,4 respectively. The data set also gives us a sequence of 100 genes labeled
as either up or down. Each gene at index i is treated as 1 if it is up and 0 if it is down. We did
this to optimize our program by having fewer string comparsions since integer comparisons are
much faster.

Implementation

For this homework, we choose to use Matlab because our input data is a matrix. Matlab has
helpful functions for matrix manipulation like determining whether a set of numbers is in a row

1

of the matrix. It also has helpful functions such as the union of two sets which helped us do the pruning necessary for an efficient apriori algorithm.

Conclusion

In conclusion, BioGalaxy can have reasonably timed queries since having multiple fact tables allows for fewer joins. Our schema also supports many to many relationships, temporal data, and hierarchal querying.

References and Notes

[1] Liangjiang Wang, and Aidong Zhang *BioStar Models of clinical and genomic data for biomedical data warehouse design*.

[2] Wide Skills
$http://www.wideskills.com/data-warehousing/dimensional-data-modelling$

[3] Wiki Pedia Website
$https://en.wikipedia.org/wiki/Fact_constellation$