Report of Frequent Item-Set Mining

Wei Wang 40041116
futurename@gmail.com

Jiahui Wang 40070981
wjhmars@gmail.com

Chongwen Li 40042828
lichongwen1994@gmail.com

Bin Xue 40059809
binxue6@gmail.com

Algorithm

Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. Following is the details of process of our project. After open the input file. First is raw data handling, we record the amount of lines and support points. Then, we put each of every read numbers into a treeset data structure, the main purpose of using a treeset data structure is due to we want to find a most efficient data structure and we do data handling process in this structure. Meanwhile, we record the numbers which's frequencies surplus the support point in a arraylist. After handling the whole of data. We use another data structure to record amount of appearance of each combination which the number is from frequency item arraylist and do comparing with data in treeset. After recorded, the print function and writing file function directly implemented combinations from this data structure. The most important and the most significant process in Apriori algorithm and even in our whole project is build combination from frequent items which meet requirement. Most of implementation of relevant programs used recursion to obtain all of combinations. We considered recursion would reduce the speed, so we applied iteration. When we pass one numbers list to our creating combinations algorithm. Firstly, we assume every digit of these list is 1, then we use first digit to make combinations, after creating all of combinations include the first digit, transfer first digit from 1 to 0, and do the second digit until every digit become 0.

Optimization

As for optimization of our programming. We implemented two approaches to try to find an optimal and best situation. First of all, after we read the input file in

the beginning of program, we obtain the whole of data from raw to processed in our individually designed data structure. The main purpose of this optimization is we want to do the following all of operations in the memory which is the fastest component, another purpose is we want do just one times in reading file operation. Reading file is a particular high consumption operation, therefore we want decrease these consumptions. Secondly, we implemented iteration in the process of creating combinations. After our testing, iteration can reduce more than 20% running time than recursion. In future works, we want implement more optimization tactics in this program.

Contribution

In this project, we divided into four parts. Data structure design & Data processing, Apriori algorithm implementation, Function of testing combinations and Process of generating data & programming test. Jiahui Wang finished data structure design and data processing. Wei Wang focused on Apriori algorithm implementation. Bin Xue worked in function of testing combinations. Chongwen Li was responsible for the process of generating data and programming test.