Abstract: One of the challenges of frequent itemset mining is long running times of algorithms. Two major costs of long running times are due to the number of database scans and the number of candidates generated. Database scans take time and candidates require memory. In order to deal with these issues, we propose a new implementation of Apriori algorithm, NCLAT (Near Candidate-less Apriori with Tidlists), which scans the database only once and creates candidates only for level one. In addition, we show the results of choice of data structures used whether they are probabilistic or not, whether the datasets are horizontal or vertical, how counting is done, whether the algorithms are computed single or parallel way. Furthermore, we implement, explore and devise incremental algorithm UWEP with single as well as parallel computation (using multi-threading in a multi-core environment). We have also created a more efficient version UWEP2. UWEP2 fixes a bug that we found in UWEP and adds caching (keeping the counts of frequent itemsets), lazy evaluation (i.e. we do not count and update its count of a frequent itemset unless we need to).

While there has been a lot of work done on frequent itemset mining on structured data, very little work has been done on unstructured data. Combining the ideas from Boyer-Moore Horspool,

Rabin-Karp and Raita algorithms, we created a new hybrid pattern search algorithm, Double Hash, which can potentially be used in frequent itemset mining on unstructured data in the future. Double Hash was faster than all three algorithms aforementioned for any type of short or long patterns in all of our tests. We will present our work and test results on this as well.

Committee:

- Professor Xiaowen Zhang, Mentor, The Graduate Center
- Professor Ted Brown, Queens College
- Professor Subash Shankar, Hunter College

Outside member:

- Professor Professor Lixin Tao, Pace University