# CSCI 6333 Data Mining & Warehousing

**Term Paper**

**For this term paper assignment, we have two topics, and you can choose one of the two to complete the work. Moreover, you can work individually or with a group of at most 3 members.**

1. **Research Topic One: Finding association rules with the help of inverted indexing**

The key to finding association rules is to find frequent itemsets. The classical apriori algorithm uses a level by level approach: First find all frequent 1-itemsets . For find all candidate frequent -itemsets with , then find the support of each candidate in and eliminate infrequent ones so that to obtain frequent -itemsets. Repeat this process until no more candidates are available.

There are two challenges involved in the above approach: One is how to generate with , and the other to how to find the support of each candidate in .

Inverted indexing is a classical technique in information retrieval and web search engines. This technique allows every index term (or keyword) to link to all the documents containing the keyword (along with frequency and locations in the documents). The inverted indexing can be built adaptively and in linear by scanning the document collection once.

We can view a transaction as a document and items in a transaction as keywords. For example, the following dataset has 10 transactions with items a, b, c, d, and e.

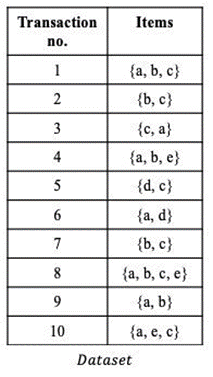


Figure 1. Dataset of 10 transactions

When viewing each transaction as a document and an item as a keyword, we can build an inverted index for the dataset. See Figure 2.

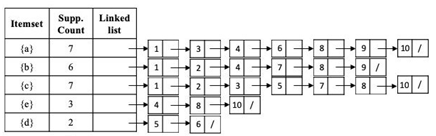


Figure 2. An inverted index for the Dataset.

Once, this inverted index is built, we can use it to help find frequent 1-itemsets .

for we can follow the classical approach to find all candidate frequent -itemsets with . Now, I ask you to research on the following: Can you use the inverted index to help find the support of each candidate in without scanning the dataset and then use the support to eliminate infrequent candidates so that to obtain frequent -itemsets.

1. **Research Topic Two:** Study one outlier/anomaly detection technique or an application of such a technique, give pros and cons of this technique/application, and then present your own technique with good improvement.
2. **Term paper format:**

* A total of five pages, single spaced, 11 pt., Times New Roman (or comparable) font, with 1-inch margin on four sides.
* First page includes a title, course number and semester, your name and SID.
* The content of your report shall be structured as follows
  + An overview of the existing technique(s). About 1/3 of length.
  + A comparative analysis: what are good? What are not good? About 1/3 of length.
  + Your new technique(s) along with justifications. About 1/3 of length.
  + References must be included and cited in the report.

1. **Notes:**

Plagiarism in any form is not allowed, and if found, the term paper will receive no credit, or the case will be reported to the Dean of Students.