Project 3
Advanced Database Systems

Amandeep Singh
as3947@columbia.edu

Evangelia Sitaridi
es2996@cs.columbia.edu

May 4, 2011

1 Dataset

This dataset [1] describes the data of interviews of World Trade Center 911 workers. We found it interesting since it combined demographic attributes and attributes describing a variety of consequences from the events of 911. We were interested to analyze the physical and mental conditions of people after this event. The columns of the dataset are described in table 1 and their name in their initial dataset is in parentheses.

2 Data Preprocessing

The initial dataset [1] contained 71427 rows and 52 columns. Thus the initial size was too large both in terms of rows' and columns' number. To reduce the execution time of the program and facilitate the testing of our code we did the following modifications:

1. keep the first 1000 rows of dataset
2. select 17 columns we found most interesting, and are the ones described later. We first removed the columns that had mostly NULL values. Some of the columns were removed after experimentation since they did not add to the results.
3. due to a problem in the format of the initial dataset the value of the participants' income was represented using comma, e.g. 25,000 instead of 25.000. This confused our program since it regarded as an extra column. That is why we replaced each *,000* ocurrence with string *000*.
4. attibutes were renamed to more intuitive ones

Finally, the quotes around the value of each column were added by OpenOffice SpreadSheet program but we let them since we did not feel they affected the legibility of the results.

3 Implementation

The dataset is stored in the DataSet class object, each row in the dataset is defined as a Transaction class object. If a dataset has a title in the first line, we set the item's value as the name of the attribute concatenated with the value of the attribute. Also, if a Dataset does not have a title then the second parameter of Dataset's constructor in the main class should be **set to FALSE**.

Table 1: 911 Dataset Selected Attributes

VARIABLE_NAME	VARIABLE_LABEL	QUESTION ASKED?
GENDER (gender)	Gender	What is the enrollee's sex?
AGE_GROUP(age_911grp)	Age Group on 9/11/01	What was the enrollee's age on 9/11/01?
CENSUS_RACE(census_race)	Race/Ethnicity	What is the enrollee's combined race/ethnicity?
MARITAL_STATUS(mar_status)	Marital Status	What was the enrollee's marital status at the time of interview?
EMPLOYED_911 (employed_911)	Employment Status on 9/11/01	Was the enrollee employed on 9/11/01?
SMOKER_ST (smkstatus_b)	Smoking Status	What was the smoking status of the enrollee at the time of interview?
COUGH (cough_nw)	New or Worsening Cough	Did the enrollee report new onset or worsening of a persistent cough since 9/11/01?
DEPRESSION (depression_nw)	New or Worsening Depression	Did the enrollee report new onset or worsening depression, anxiety, or other emotional problems since 9/11/01?
HEADACHE (headache_nw)	New or Worsening Headache	Did the enrollee report new onset or worsening frequent, severe headaches since 9/11/01?
BREATH (breath_nw)	New or Worsening Shortness of Breath	Did the enrollee report new onset or worsening of shortness of breath since 9/11/01?
SKIN (skin_nw)	New or Worsening Skin Problems	Did the enrollee report new onset or worsening of skin rash/irritation since 9/11/01?
THROAT (throat_nw)	New or Worsening Throat Irritation	Did the enrollee report new onset or worsening of throat irritation since 9/11/01?
DUST (dust)	Dust	Did the enrollee indicate that he/she was outdoors on 9/11/01 within the dust or debris cloud resulting from the collapse of the World Trade Center?
EDUCATION(educ)	Education Level	What was the highest grade or year of school completed by the enrollee at the time of interview?
INCOME_SLAB(income_interview)	Household Income	What was the enrollee's total household income in 2002 before taxes?

```
Preprocessing Text Algorithm:

DataSet(fileName, hasTitle)
    if(hasTitle)
       title_attrs=title.split(",")
    foreach (line)
        line_attrs = line.split(",")
        if(hasTitle)
            // adds item: title_attrs[i] = line_attr[i]  in the list
            transactions.add(line_attrs,title_attrs)
        else
            transactions.add(line_attrs)
```

After each iteration of the Apriori algorithm we generate a new instance of ItemSet class. So, for the first round of the algorithm we use.

```
GenerateL1()
    singles // stores the frequency of each item.
    itemsL1 // stores the items which have support > min_support

    transactions = dataset.getTransactions();
    foreach (transactions)
         foreach (item in transactions)
              calculate the frequency for each item.

    foreach (singles)
        supp=singles.get(s) / (dataset.size() * 1.0)
        if (supp >= minSupp)
            itemsL1 // add a new ItemSet(singles)
```

Subsequent rounds of Apriori use this base itemsL1, new string is added into the item set using nextString(ItemSet) method.

```
nextString(ItemSet i1, ItemSet i2)
    // as the item set is generated iteratively, length of i1 & i2 should match
    if len(i1) != len(i2)
        return null
    else
        len_x = length of i1
    // check both the strings are equal except the last element
    while (len_x > 1)
         if i1.item() != i2.item()
              return null
         // check the last element
         x = last element of i1
         y = last element of i2
         if (x < y)
              return y
    return null
```

The support of the itemsets are cached so it is not computed multiple times.

4 Results

During our data analysis we faced the problem of many complex rules so apart from saving the frequent item-sets and rules we create an extra file: top_output.txt containing only top five-rules for each possible conclusion. For the second setting we only add the file containing only the top rules since due to the low support we had set it included many itemsets. Below we analyze the most interesting associations present in the dataset.

Minimum Support:10% Minimum Confidence: 60%

Rule 1

["COUGH"="No", "EMPLOYED_911"="Yes", "INCOME_SLAB"="$150000 or More", "SKIN"="No", "THROAT"="No"]
=> ["EDUCATION"="College +"] Conf: 95.24% , Supp: 12.01%

This rules associates education of the participants to income, smoking and hearing condition after 911 events. It seems that people with high income, employed and relatively not physically affected 911 events are educated.

Rule 2

["COUGH"="No", "GENDER"="Male", "INCOME_SLAB"="$150000 or More", "SKIN"="No"]
=> ["CENSUS_RACE"="White (non Hispanic)"] Conf: 92.66 % , Supp: 10.11 %

Again, there is another demographic relation that wealthy males were white.

Rule 3

["AGE_GROUP"="25-44 years", "CENSUS_RACE"="White (non Hispanic)", "DEPRESSION"="No", "DUST"="No"]
=> ["EMPLOYED_911"="Yes"] Conf: 98.26 % , Supp: 11.31 %

Young volunteers who were white and not exposed to dust, not suffered from depression were employed during 911 events.

Rule 4

["COUGH"="Yes", "THROAT"="Yes"] => ["BREATH"="Yes"] Conf: 64.50% , Supp: 12.91%

This rules correlates different conditions caused by 911 events. It seems that participants experiencing cough and throat irritation also experience shortness of breath.

Rule 5

["AGE_GROUP"="25-44 years", "COUGH"="No", "DEPRESSION"="No", "EDUCATION"="College +", "HEADACHE"="No", "THROAT"="No"]
=> ["SMOKER_ST"="Never smoked"] Conf: 72.60% , Supp: 10.61%

Young people, who are educated and not seriously affected from the attack were most likely non-smokers.

Rule 6

["BREATH"="Yes", "THROAT"="Yes"] =>["DEPRESSION"="Yes"] Conf: 69.41% , Supp: 11.81%

This rules relates a mental symptom to physical symptoms. People physically affected from 911 events likely experience symptoms of depression.

Rule 7

["DUST"="Yes", "THROAT"="Yes"] =>["DEPRESSION"="Yes"] Conf: 74.64% , Supp: 10.31%

Rule 8
["DEPRESSION"="Yes"] =>["GENDER"="Female"] Conf: 44.62% , Supp: 20.32%
["DEPRESSION"="Yes"] =>["GENDER"="Male"] Conf: 55.38% , Supp: 25.23%
This rule implies that depression is more prevalent to male participants.

Rule 9
["SMOKER_ST"="Current smoker"] =>["EDUCATION"="College +"] Conf: 63.92% , Supp: 10.11%

Most current smokers seem to be educated.

Minimum Support: 5% Minimum Confidence: 35%
We decided to focus on underpresented groups of participants which might not be covered by the previous setting of the algorithm.
Rule 10
["CENSUS_RACE"="White (non Hispanic)", "GENDER"="Female", "THROAT"="Yes"]
=>["DEPRESSION"="Yes"] Conf: 81.48% , Supp: 6.61%

["CENSUS_RACE"="White (non Hispanic)", "EMPLOYED_911"="Yes", "GENDER"="Female", "THROAT"="Yes"]
=>["DEPRESSION"="Yes"] Conf: 82.86% , Supp: 5.81%

The second of the two rules above sets an extra condition, that a woman is also employed and has higher confidence. We can conclude for that depression is slightly more common among employed women.

Rule 11
["BREATH"="Yes", "HEADACHE"="Yes", "THROAT"="Yes"] => ["DEPRESSION"="Yes"] Conf: 83.87% , Supp: 5.21%

Remembering Rule 6, adding headache condition the chance for a person to be depressed increases significantly from 69.41% to 83.87%.

Rule 12
["AGE_GROUP"="25-44 years", "INCOME_SLAB"="$25000 to less than $75000"]
=>["MARITAL_STATUS"="Not Married"] Conf: 52.73% , Supp: 5.81%
This rules gives us again information on the demographic profile of the participants. Participants aged 25-44, having low income are more likely not to be married.

Rule 13
["DEPRESSION"="No", "INCOME_SLAB"="$150000 or More", "MARITAL_STATUS"="Married", "SMOKER_ST"="Never smoked"]
=>["EDUCATION"="College +"] Conf: 100.00% , Supp: 5.11%

Rule 14
Top Reasons for depression:
["BREATH"="Yes", "HEADACHE"="Yes", "THROAT"="Yes"]
=> ["DEPRESSION"="Yes"] Conf: 83.87% , Supp: 5.21%

["CENSUS_RACE"="White (non Hispanic)", "EMPLOYED_911"="Yes", "GENDER"="Female", "THROAT"="Yes"]
=> ["DEPRESSION"="Yes"] Conf: 82.86% , Supp: 5.81%

["SKIN"="Yes", "THROAT"="Yes"]
=> ["DEPRESSION"="Yes"] Conf: 82.35% , Supp: 5.61%

["BREATH"="Yes", "EMPLOYED_911"="Yes", "HEADACHE"="Yes"]
=> ["DEPRESSION"="Yes"] Conf: 82.05% , Supp: 6.41%

["CENSUS_RACE"="White (non Hispanic)", "GENDER"="Female", "THROAT"="Yes"]
=> ["DEPRESSION"="Yes"] Conf: 81.48% , Supp: 6.61%

We observe that THROAT and Gender="Female" increased the risk for a participant to experience depression.

Rule 15
["AGE_GROUP"="45-64 years", "BREATH"="Yes"]
=>["INCOME_SLAB"="$25000 to less than $75000"] Conf: 39.39% , Supp: 5.21%
We observe that middle-aged participants that suffer from shortness of breath is most likely to have low income.

Rule 16
["EDUCATION"="Some College", "SKIN"="No"]
=>["INCOME_SLAB"="$25000 to less than $75000"] Conf: 38.13% , Supp: 5.31%
Minimum Support: 3% Minimum Confidence: 60%
Rule 17
["AGE_GROUP"="45-64 years", "COUGH"="Yes", "DEPRESSION"="Yes", "GENDER"="Female"]
=> ["THROAT"="Yes"] Conf: 89.19% , Supp: 3.30%

Rule 18
["BREATH"="No", "CENSUS_RACE"="White (non Hispanic)", "DEPRESSION"="No", "EMPLOYED_911"="
"HEADACHE"="No", "INCOME_SLAB"="$150000 or More", "SKIN"="No", "SMOKER_ST"="Never smoked", "THROAT"="No"] => ["EDUCATION"="College +"] Conf: 100.00% , Supp: 3.70%

Minimum Support: 70% Minimum Confidence: 80%
Rule 19
["HEADACHE"="No"] =>["SKIN"="No"] Conf: 90.77% , Supp: 76.78%

Rule 20
["EMPLOYED_911"="Yes"] => ["HEADACHE"="No"] Conf: 84.62% , Supp: 75.98%

Appendix

File Listing

We list below the files included in our submission. The compressed file contains the following sub-folders: [dataset, src, output]. The name of our INTEGRATED-DATASET is **911_1000.csv**. The files are:

- README.pdf

- src/{APriori.java, ItemSet.java, DataSet.java, Transaction.java}
- src/Makefile
- src/dataset/911_1000.csv
- For high support values we are giving the output.txt file and for lower support values, where the result size is large, we provide top_output.txt file containing the top rules for each conclusion. output/{output_s10_c60.txt , output_s70_c80.txt, top_output_s10_c60.txt, top_output_s5_c35.txt, top_output_s3_c60. }

Compilation

- **Compile**: make
- **Clean-up**: make clean (removes .class files and output files)
- **Execute**: java APriori *min_supp min_conf* (from src directory)

References

[1] *911 Dataset*, `http://www.nyc.gov/html/datamine/html/data/terms.html?dataSetJs=raw.js&theIndex=22`