Data Mining with Declarative Programming

CS240A: Databases and Knowledge Bases

Tianyu Zhang Jiapeng Wan

UID: 805332081 UID:305353896

zhangtianyu@g.ucla.edu wanjp15@g.ucla.edu

**Introduction**

The objective of this project:

(i) to explore the new power of Datalog and SQL made possible by aggregates in recursive queries,

(ii) to become familiar with the Datalog/RaSQL system that support them on Apache Spark.

We also seek to understand better the challenges and di\_culties facing users. Thus, in your report you should also comment on the following:

1. di\_culties you faced in implementing these algorithms and

2. ease-of-programming with these declarative languages.

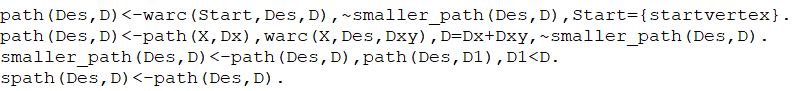
**1. Graph Algorithms and Sorting**

**1A Shortest Distance in a Directed Graph**

Our program computes the shortest distance from one starting node to every other node in a directed graph.

**Datalog**

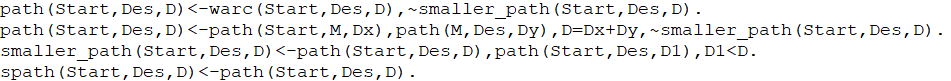
We found the shortest distance by applying a constraint to the predicate path(Des, D). With formal non-monotonic semantics, our program uses a Pre-Mappable constraint by pre-mapping the min computation into the recursion rules.



Then we verify the PreM condition by using the classical theory of Functional Dependencies (FDs) and Multivalued Dependencies (MVDs) in relational database schema design.

First, we obtain the relation R by joining path(Des, D) and warc(Start, Des, D). In R, the MVD holds: X->>Dx and X->>Des, Dxy. Therefore, if a tuple satisfies the constraint X, DesDx+Dxy, then X, DesDx and X, DesDxy also holds. Thus we can augment the previous result into X, Des, DxyDx, XDx. The latter one is a drop-in constraint and does not change the mapping.

We also explored the a la Floyd rules for non-linear shortest paths problems. The form is similar to our original rules.



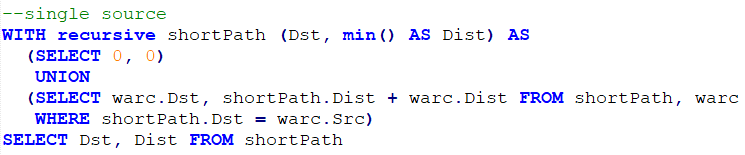
We can also verify the PreM condition.

First, we obtain the relation R by joining path(M, Des, Dy) and path(Start, M, Dx). In R, the MVD holds: Start, M, Des->>Dx and Start, M, Des->> Dy. Therefore, if a tuple satisfies the constraint Start, M, DesDx+Dy, then Start, M, DesDx and Start, M, DesDy also holds. Thus we can infer from previous results that Start, MDx, M, DesDy, which are both drop-in constraints and don’t change the mapping.

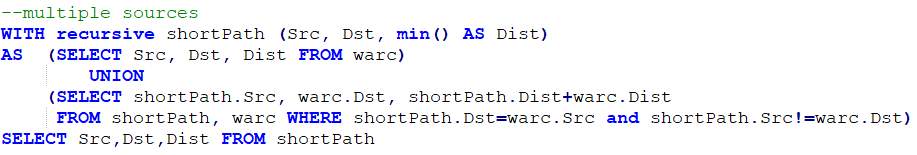
However, the program didn’t work well because the predicate path is initially undefined.

**RaSQL**

We wrote recursive rules for RaSQL. The semantics is guaranteed by the PreM property. First, we initialized shortPath(Dst, Dist) by using the base relation warc(Src, Dst, Dist) where Src is 0 and Dist is 0. Then we recursively computed the distance to other nodes by combining the Src of warc and the Dst of shortPath. Finally, we used the min aggregation select the shortest paths.



We also explored the case where the shortest paths of multiple starting points or sources are required. The form of recursive rules is similar to the case of single source except that we need to ignore the records where shortPath.Src equals warc.Dst.

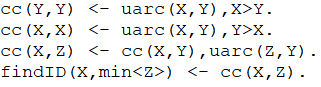


**1B Strongly connect component in an undirected graph**

Our program uses the smallest node id to represent the component id of every node.

**Datalog**

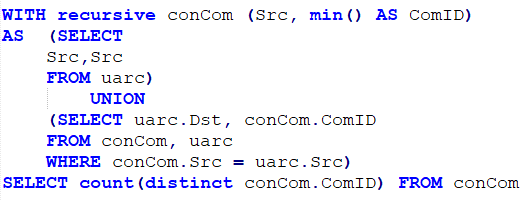
Our program generates the label of every node in an undirected graph. Because there is no distinct count method in datalog, our result lists the id of every node along with its ComID. Our program lists all the ComIDs, but not every node is listed. We can count the distinct number of ComID manually to ensure the result is correct.



Then we verify the PreM condition. First, we obtain the relation R(X,Y,Z) by joining cc(X,Y) and uarc(Z,Y). In R, the MVD holds: Y->>Z. Therefore, if a tuple satisfies the constraint ZX, then YX also holds. Thus it is a drop-in constraint and does not change the mapping.

**RaSQL**

Our program generates the label of every node in an undirected graph. The semantics is guaranteed by the PreM property. First, the ComID of every node is initialized as its own id. Then in the recursive rule, the ComID of every node is updated to its neighbors’ ComID. The min aggregation is applied to ensure the ComID of every node to be the minimal possible id. Our final result is the count of distinct ComIDs.



**2. Naive Bayes Classifier (NBC)**

The Naive Bayes Classifier is one of the fast, simplest, yet popular machine learning model due to its interpretability. It is also very easy to build and is particularly useful for large datasets. It is a simple "probabilistic classifier" based on applying Bayes' theorem with strong (naive) independence assumptions between the features. In other words, it assumes that presence of a particular feature in a class is unrelated to presence of any other feature. Even if the features depend on each other, all of them contribute independently to the probability.

In order to do so, we created the following scripts where each denotes a step in the data mining process:

**1. Load data and verticalize the training table** (NBCpreprocess.py)**:** This script is used to download the two datasets in csv file. Then we need to verticalize the data by reading and write them into a separate file. The final format of dataset will be readable by our RaSQL system. In order to handle the numeric attributes, like in the case of Bank Marketing Dataset, we have applied categorization and binning to some attributes, and used Gaussian distribution for others, depending on the nature of the attributes. We treated discrete numeric attributes with less number of distinct values as categorical. For the attributes that made sense to be categorized, we performed binning and divided it into categories. The continuous numeric attributes were handled using Gaussian distribution using the equation given below: We compute the above probability for each attribute value-class pair. Here, mean and standard deviation are used to compute the probability that a particular attribute belongs to a class. In order to perform the task of mining, this script partitions the data into training and testing set (in the ratio of 4:1), with each set containing of verticalized records of the original dataset. We select random rows and place it into training and test set to avoid any kind of bias.

The process of load data, verticalize data, handle numeric, partitioning are all built in this python file.

**2. NBC Model implementation (NBC.SQL):**

**Result**

* 1. We obtained the following accuracies for the provided datasets:
  2. **Mushroom Dataset: 95.75%**

We faced various problems while building this classifier:

**.**

* 1. **b. Handling missing values:** We did so by considering them as an attribute value. We also tried ignoring them, but that reduced the accuracy of the classifier. Also, replacing them by the mean/median is not possible, since the missing attributes are categorical.
  2. **c. Handling the frequency zero problem**: Initially, when the classifier encounters unseen combination of attribute value and class label, it used to assign zero probability. But this is not correct as we are ignoring the effect of other features. To remedy this, we used the Laplace smoothing technique, which adds 1 to the count for every attribute value-class combination (Laplace estimator), due to which the probabilities will never be zero

**3. K-Nearest Neighbors Classifier (KNN)**

A KNN classifier is a non-parametric machine learning classification model that predicts the class of a test instance based on the class labels k training instances closest to it.

**3.1. Datalog Model (DeAL)**

**1. The problem of syntax in RaSQL system**

At first we want to implement the the KNN in RaSQL system, however, during the process of implementation. We found out that there are still many syntax problems in RaSQL system to write datalog program.

The problem is the result of a sequence of operation. In one of the operation, we need to store the result of square of the difference of two variables:

*squares(Id,Tid,Col,C) <- test(Id,Col,V1),train(Tid,Col,V2),C=(V1-V2)\*(V1-V2).*

*sumsq(Id,Tid,sum<(C)>) <- squares(Id,Tid,Col,C).*

The first query generates a table with Id, Tid, and column with the square of difference of two values of two tables. The second query will sum up the values with same Id and Tid and create a table. The problem is that, when we want to create another table from *sumsq*, such as:

*minsq(Id,Tid) <- sumsq(Id,Tid,C).*

it will throw an error which tells us that:

*cannot resolve '`Id1`' given input columns: [Id, Tid, Aggr\_1];*

This problem happens if and only if both queries are applied. The table can be generated with column which resulted from variable operation and can be generated with column which sums up column with corresponding Id. The error will occur if and only if the column is generated by operation and then sums up. Since no more value can be extracted from the table, the further operation can not be proceeded.

**2.The implementation of KNN on DeAL.**

To implement the program and run it successfully, we tried to implement it on DeAL.

**(a) Preprocess the data(KNN.fac).**

In our program, we download the file from the website in data format and then load the data on csv file, then we verticalize the data to make sure the format is readable by DeAL. Then we manually put the data to DeAL. Since the DeAl is not able to deal with large amount of data in file, we will only put a small portion of data to it. The dataset we used are:

Hill\_Valley\_without\_noise\_Testing.data

Hill\_Valley\_without\_noise\_Training.data

**(b) Build the KNN classifier (KNN.deal)**

This file builds the KNN classifier from the training dataset and can be used to predict labels of test dataset. At first, we compute squares of difference in attribute values for all rows in test dataset. Then, we sum the squares by grouping using test Id. Though DeALS does not have a function to get the square root, we will work with squares since it will not cause any change in answers. Then we sort the square values in ascending order. To sort the table, we need to manually define a sorting function. After sorting and extracting top K values from the table, we will count the labels for each class for each test Id. In that case, we will be able to get the maximum value which represents the majority class out of K closest training records. This gives us the predicted label of the test record.

Though we are trying to build the classifier without recursion, it is not applicable. Because to compute the top K values, we have to manually write the sort function which cannot be accomplished with out recursion.

Moreover, the recursion in sort function can be improved if we use PreM(pre-mappability) during the calculation of top k closest neighbors. When the queries contains aggregations such as sum<>, min<> or max<>, PreM can be used to optimize the model semantics for exoprograms. To formulate PreM in DeAL, we just need to add drop-in goals. By using the drop-in goal in the body of rule, we can prove the PreM because it does not change the mapping defined by queries. Moreover, we can also prove the PreM by mentioning that the join in the body of recursive rules. By deriving the functional dependency using mixed transitivity, PreM is satisified.

**3.2 SQL model (RaSQL)**

We have encountered a similar problem as stated in datalog model. The created table can not be used further. But we still write SQL queries based on RaSQL system. In order to build the classifier in SQL, we created the following scripts, each doing a particular task in the mining process.

**1. Preprocess Data (KNNpreprocess.py)**

The script is used to first download the data by reading .data file and writing .csv file. Since the converted data file does not have the format that we will need in our classifier. We need to verticalize the data and write them in new file which will be used by our program.

**2. Build the KNN Classifier(knn.sql)**

This script creates a KNN Classifier by employing the training data-set, and predicting the classes for the testing dataset. We used Euclidean distance to calculate the distance between two instances. We follow the steps given below:

(a). We compute the square of difference in values for each of corresponding attributes between each set Id and all the training records.

(b). We calculate the distances by summing up distances and take square root of them for each attributes grouped by train ids.

(c). Then we calculate top K records having lowest values for Euclidean distance by using the Rank function in SQL. This function helps us to calculate top K values.

(d). Find the majority class among the top K records. This majority class gives the predicted label.

(e). Calculate the accuracy of classifier by calculating the ratio of correct predictions to total number of predictions, similar to what was done for NB classifier.

**Results**

we trained and tested the classifier on only one dataset, but the same can run on different datasets, just by changing the load script in terms of dataset names. The resulting accuracies for the dataset (Hill\_Valley\_without\_noise\_Testing.data & Training.data ) is:

For K = 5, 61.2% for 40 sets; 50.57% for 100 sets; 55.49% for 300 sets.

For K = 7, 59.3% for 40 sets; 56.64% for 100 sets; 62.2% for 300 sets.

As K grows larger, the program takes longer to execute.

**Conclusion**

We have faced many challenges and problems in terms of RaSQL system both in datalog languages and RaSQL languages. Since such systems have their own syntax, modifying them is much harder, as compared to other systems. Moreover, even a slight mistake in syntax might cause the entire program to fail. The declarative version of the procedural language is much more complex to implement and understand. The error message generated by declarative languages is difficult to understand and debug i.e. some error messages are not intuitive.

(a). The RaSQL can directly load the file with correct format. However, we need to preprocess the file manually to make it readable by programs. In that case, we will need to write a preprocess program and read, write, and verticalize the data in the file.

(b). The main problem of RaSQL system is that though it supports datalog, there are many operations and functions are not supported by RaSQL system. In that case, if want to take the sum of a column after the column did multiplication, there will be errors. These kind of problems are hard to be shown on documents. To make the file compile successfully, we need try by changing syntax and see if that works.