MongoDB and MapReduce

Krishna Kalyan, Saul Garcia

12/18/2016

Objective

The objective of this Lab session is to get familiar with *MongoDB* and compute the *Association Rules* for Basket Analysis.

Process

The first step in this Assignment was to install *MongoDB* and get familiar with it. For this some sample code and a MapReduce algorithm was provided by the instructors.

We were provided a `groceries.csv` dataset, which contains 9835 rows, or different tickets, which contain each up to 32 products. If we were to handle this type of dataset with a `dataframe`, we will end up with a very sparse dataset, here is where *NoSql* comes in handy, which in this case, we are using *MongoDB*.

Note: It is important to consider that we are taking into account the full specific name of the item. For example, if there are two type of milks, there would not exist any relation between them for our analysis.

Once we have read the data in *Mongo*, in order to prepare our *Association rules**, two Map Reduce codes were built. The *first one* to map and compute the frequency for each individual item bought, and the *second one* in order to compute the total frequency of each possible pair of items from each different ticket.

Once we have both `documents` of data, we compute the `support` and `confidence` with the following formulas:

$$support(X \rightarrow Y) = \frac{count(X,Y)}{count(All)}$$

and

$$confidence(X \rightarrow Y) = \frac{count(X,Y)}{count(X)}$$

Results

Once we computed the *support* and *confidence* for each pair, this is how it looks (first 6 elements):

Product 1	Product 2	Support	Confidence
Instant food products	UHT-milk	0.0004067	0.0506329
Instant food products	baking powder	0.0003050	0.0379747
Instant food products	beef	0.0008134	0.1012658
Instant food products	berries	0.0003050	0.0379747
Instant food products	beverages	0.0004067	0.0506329
Instant food products	bottled beer	0.0006101	0.0759494

The support and confidence gives us a hint on which products are related, or most probably simultaneously bought. For instance, if we look for a product of high Confidence:

	Product 1	Product 2	Support	Confidence
483	baby food	brown bread	0.0001017	1

This means that both product `Baby food` and `brown bread` are always bought together. Looking at its support of 10^{-4}, it tells us that most probably this was the only time that *baby food* was bought in our data, so in this case it really gives us no information.

Considering that we have 19354 different rules, we follow the parammeters given in the assignment sheet to see how many rules we have according to a specific *Support* and *Confidence*.

Row	Support	Confidence	Nr. of association rules found
1	1%	1%	426
2	1%	25%	96
3	1%	50%	0
4	1%	75%	0
5	5%	25%	4
6	7%	25%	2
7	20%	25%	0
8	50%	25%	0