Ricard Meyerhofer Parra

Joel Cantero Priego

2016-2017 Q1

CAIM Lab, Session 6: MongoDB and MapReduce

Introducción

En esta sesión utilizaremos **MongoDB** para utilizar una base de datos **NoSQL** y aprenderemos a utilizar el algoritmo **map-reduce** utilizando **pymongo**.

La parte principal de esta sesión es la implementación en Phyton que nos ha conllevado más problemas.

Desarrollo

La primera parte de la práctica consistía en leer el documento que se nos proporcionaba de datos (groceries.csv) y guardar los datos correspondientes en nuestra base de datos. Esta parte ha sido relativamente fácil pues los ficheros que tenemos como ejemplo, nos han servido como referencia (de hecho, hemos tenido más problemas en instalarnos pymongo y jen leer toda la documentación que se nos proporcionaba que en realizar el código).

Una vez creada la base de datos con los datos introducidos, hemos procedido a cómo guardar estos datos de forma que los cálculos que hiciéramos posteriormente fueran lo más eficientes posible.

Al pensar en hacerlo eficiente, nos hemos dado cuenta de la necesidad de guardar todos los datos para hacer el cálculo y que no podíamos ahorrar en almacenamiento de datos.

Por otra parte debido a que los cálculos de support i confidence sobre (a,b) son unidireccionales (a ->b o b -> a), hemos añadido a la base de datos “a#b” y “b#a” (# nos sirve para definir la implicación de izquierda a derecha).

Finalmente tal y como dice el enunciado también hemos tenido que añadir cada término para poder contar su número de apariciones.

El reduce a diferencia de los mappings, puede ser el mismo para tanto los pairs como los términos pues se cuentan de la misma manera. En el caso de los mappings hemos tenido

que diferenciar en dos modelos de MapReduce los cuales cuentan el número de apariciones de cada pair y término (pairCounts y termCounts).

Por último hemos rellenado la tabla con los valores de support y confidence (**task1**) para cada implicación donde los thresholds se han calculado sobre la misma BD (al ser un proceso costoso).

Task 1: With these programs, fill out the missing values of the following table:

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

Task 2: Give the list of association rules found corresponding to rows 4, 5 and 6 of the table of Task 1

Row 4:

0 Associations.

Row 5:

Association	Support	Confidence
other vegetables ⇒ whole milk	7.483477376715811	38.67577509196006

rolls/buns ⇒ whole milk	5.663446873411286	30.790491984521836
whole milk ⇒ other vegetables	7.483477376715811	29.287703939514525
yogurt ⇒ whole milk	5.602440264361973	40.160349854227405

Row 6:

Association	Support	Confidence
other vegetables ⇒ whole milk	7.483477376715811	38.67577509196006
whole milk ⇒ other vegetables	7.483477376715811	29.287703939514525