NoSQL Model Comparator

By : Salma KERTIT - Marouane MAATOUK

March 2019

1 Context

The main reason of this paper is to show what we both have done so far. The first steps were about knowing more about the NoSQL structure storage, and about Hackolade as a data modeling tool. The type of NoSQL data that our group is going to work on is the Document-oriented one, so one of the tasks for this first week was to get familiar with both the data type and the ML algorithms, before finding more sample schema corresponding to NoSQL type, and scoring them.

2 Hackolade and DB type

After downloading and installing the last version of Hackolade file, we got to actually get to know about what the whole tool is about. Hackolade is a data modeling tool that makes it possible for us to design our tables, collections and graphs, model our data and visualize it.

The next step was to know more about the type of data, which is the document-oriented DBs. A document database, also called as a document store, is a type of NoSQL database which is used for storing and managing semi-structured data. Documents in this database are addressed via a key, which represents the document. Encoding in use include XML , JSON and BSON.

3 Data Software

Getting familiar with these concepts is also about knowing the data we're going to use. For this project, ones we're going to use are mainly MongoDB, Couchbase and JSON.

MongoDB is platform-independent document-oriented database program which is written in C++. Its main objective is to handle these data never needing a predefined schema.

CouchBase Server is a NoSQL document-oriented database software package, offering services for interactive applications.

One of the other formats is the JSON one, which is an open-standard file format which is very used for asynchronous browser-server communication. Its basic data types are : Number, String ,Boolean, Object and Null.

4 Machine learning algorithms

The first one is K-Means. K-Means is an unsupervised learning algorithm that solve the well known clustering problem. Clustering is a technique for finding similarity in data. It attempts to group individuals in a population together by similarity. The whole algorithm is about classifying given data through a certain number of clusters (K ones for example).

Next, Boosted Decision trees are a decision support tool that uses a tree-like model of decisions and their possible consequences to help identify a strategy, to take a decision or to basically reach a specific goal. It's about displaying an algorithm that only contains conditional control statements.

Last but not least, Random forest classifier is an ensemble algorithm, which combines more than one algorithm for classifying objects. It creates a set of decision trees from randomly selected subset of training set. It then aggregates the votes from different decision trees to decide the final class of the test object.

5 Conclusion

As said before, the whole objective of this week was to get familiar with the concepts of NoSQL databases, with the specified algorithms, and with the software and the NoSQl data type we are going to use, which we hope was reached, as shown in this paper.