Chapter 1
Fuzzy-Based Querying Approach for Multidimensional Big Data Quality Assessment

Pradheep Kumar K.
BITS Pilani, India

Venkata Subramanian D.
Hindustan Institute of Technology & Science, India

ABSTRACT

This paper is intended to design a fuzzy based approach to assess standards and quality of big data. It also serves as a platform to organizations that intend to migrate their existing database environment to big data environment. Data is assessed using a multidimensional approach based on quality factors like accuracy, completeness, reliability, usability, etc. These factors are analysed by constructing decision trees to identify the quality aspects which need to be improved. In this work fuzzy queries have been designed. The queries are grouped as sets namely Excellent, Optimal, Fair and Hybrid. Based on the fuzzy data sets formed and the query compatibility index, a query set is chosen. A data set that has a very high degree of membership is assigned a fair query set. A data set with a medium degree of membership is assigned a optimal query set. A data set that has a lesser degree of membership is assigned a Excellent query set. A data set which needs a combination of queries of all the above is assigned a hybrid query set. The fuzzy query based approach reduces the query compatibility index by 36%, compared to a normal query set approach.

INTRODUCTION

In today's world with an increase in the amount of data processing and information requirement it is essential to develop strategies to effectively manage and assess the data for essential quality checks. The database forms the basis of day to day decisions taken by the organization. Data obtained from employees need to be periodically updated for effective utilization. In this work, an attempt has been made to assess data quality based on certain measures or parameters like Accuracy, Completeness, Reliability, Usability, etc as discussed by Pradheep et al in (2014). Based on these parameters the data set is queried to assess

DOI: 10.4018/978-1-5225-1008-6.ch001

the effectiveness of attributes like accuracy, usability, reliability, timeliness, etc. The parameters or quality factors such as Accuracy, completeness, etc are further subdivided into minor factors. Accuracy is sub divided into Syntactic and Semantic accuracy as explained by Pradheep et al (2014). The sub factors in turn have a parameter which is a measure and this parameter has an acceptable set of values.

To assess the effectiveness of the big data, a model is constructed for a Knowledge Management System which is a Multi-Dimensional Framework for quality checks. The different attributes are each modeled as a decision tree. The combination of all these form a Decision forest tree. A model for the decision forest tree was proposed by Criminisi et al in (2011). The decision forest tree model was a probabilistic model based on classification and regression analysis. The data under consideration could be textual, video, photographs to form a random forest of decision trees. To analyse this data effectively for information several data mining techniques were proposed. Berendt and Preibusch (2014) have proposed several techniques to extract data from databases based on the choice of attributes. Another technique which is map and reduce technique for large data sets had been proposed by Doulkeridis and Norvag in (2014). A large number of data visualizing techniques have been explained in this regard by Doulkeridis and Norvag (2014), Venkat et al (2011), Gorodov et al (2013), Serban et al (2013), Shamsi et al (2013), Jennex and Olfman (2003), Evans et al (2013) and Banerjee et al (2014).

A data dictionary needs to be available which acts as a repository for storing the data. The data dictionary would contain metadata of the data related to the nature, type, volume, etc. Data integrity is another feature which decides on the reliability of data. Data access should be provided according to the role based privileges. This is done based on access privileges and functional aspects. The access privileges may vary from time to time based on the effectiveness of the queries which are also assessed to ensure minimal processing time and memory. Based on this approach the queries are classified into sets. Based on the decision tree analysis carried out the entire data is partitioned into smaller datasets. The size of the dataset may vary arbitrarily based on data volume and processing speed of the database.

The datasets are discretised based on the degree of membership. The queries are classified based on the query processing time and memory. A data set that has a very high degree of membership would have data points very closely spaced. This type of data set has a scattering distance which is negligible. The distance is measured by Euclidean method. In other words the data points would be clustered and would belong to only a single fuzzy set. Hence it would be ideal to use a query set which has a low query compatible index. This would not have a very high impact on the query cost, thereby optimizing the same. When the degree of membership is partial, we would need a slightly more efficient query set to work on this type of data set. Here the data set is spaced a bit and may belong to two different fuzzy sets. To accomplish this, the query would need to work on both fuzzy sets to return precise information. In such cases the data points are separated by a finite scattering distance. Here there is no clustering. For these data sets the query set should have an average or optimal query compatibility index.

When the degree of membership of the data is very low, the data points are widely scattered with a very large scattering distance. Here there is no question of clustering and the data points are distributed among multiple fuzzy sets. This would imply a very large scattering distance. To extract precise information, it is essential to assign highly efficient query set with a very high query compatibility index. A Mamdani's fuzzy inference engine is used as the inputs are non-linear and are provided as fuzzy sets. In this work a triangular membership function has been used to determine the degree of membership. The entire approach aims at ensuring no compromise in performance and optimal assignment of datasets with query sets.