AUTOMATED DATA GATHERING FOR DESTINATION DRIVEN ‘RENEWABLES AND EFFICIENCY INCENTIVES PROGRAMS’ DATA WAREHOUSE USING LINEAR REGRESSION

K.R.Nishanth[1] | Dr.S.Karthik[2]

[1](Computer science and Engineering, SNSCT, Coimbatore, India, itsmenishanthkr@gmail.com)
[2](Computer science and Engineering, SNSCT, Coimbatore, India, kkarthikraja@yahoo.com)

Abstract— *In this proposed method the data warehousing schedule is automatically calculated with the help of a machine learning algorithm called ‘linear regression’ and will be used to predict when to request for new data from operational data sources. In this proposed system the frequency and the time of the data being delivered to the data warehouse are initially recorded for a particular period of time and is trained with linear regression algorithm. With the trained model, the time for the data delivery can be predicted and a request is made to the ‘Renewables and efficiency incentives programs’ database. According to this prediction, the data can be requested at the time and be stored in the data warehouse. The stored information can further be used for other purposes like data mining and analytics.*

Keywords— Data Warehouse; Machine Learning; Linear Regression; Scheduling

1. INTRODUCTION

A data warehouse is a large storage system that stores a wide range of data from various sources. The data that is accumulated in a data warehouse is used to make management decisions. Data warehouses are classified into two types based on their architecture. They are source driven architecture and destination-driven architecture [1]. In source driven architecture the data sources are periodically sent to the warehouse. In destination driven architecture the data warehouse requests for new data from various sources [1]. Traditionally the request for new data is either scheduled or is handled by a person when the data is ready to be delivered. In this case, the proposed system works with destination driven architecture. It is not always certain that the data for the request will be delivered to the warehouse immediately or regularly at a particular interval of time. So there is a need to wait for the data to be received. In such cases, one has to check for database updates for changes. The proposed system uses linear regression to predict the time at which a data request can be made to the sources.

2. ORGANIZATION OF DATA

A. *Structured data*

A structured data is a data that is well organized in the form of rows and columns that will fit in a database table. These data are said to be structured because they are arranged in a way that is universally accepted for data storage and processing. In general, the SQL databases store data in a structured format.

B. *Unstructured data*

Unstructured data are data that do not have any predefined structure to store and organize the data. Instead every time a new data is generated, the structure will vary in different forms. In real time images, graphics, pdf files, videos, etc. are different from each other and hence called as unstructured data.

C. *Semi-structured data*

Semi-structured data is basically a structured data that is not organized. A semi-structured data is a data that is formatted in tags and other markers. Though it has a structure to represent a data, the structure is not same in all the case. The attributes might vary at different times for a semi-structured data.

In this scenario, the data that will be used for warehousing will be obtained in a semi-structured format called JSON.

3. WAREHOUSE DATA SOURCES

The data from sources will be sent to data warehouses in various formats. They can either be structured data or semi-structured. The data received will be extracted and transformed to the required structure before it is loaded into the warehouse [3]. In this proposed system, a semi-structured data that has information about incentives provided for the programs that make use of renewables and efficiency resources are used as source data. This data is fetched from an operational database in a semi-structured format called JSON. The JSON data is parsed and transformed according to the needs and is stored in data warehouse [4]. While the destination-driven data warehouse is operational, the time and the frequency of the data request are recorded for a sufficient period of time. The recorded dataset are used to train a supervised machine learning algorithm called ‘Linear regression’.

4. MACHINE LEARNING

Machine learning is a field that combines both Artificial intelligence and data mining areas. Machine learning algorithms find hidden patterns in data to infer a function that helps make better decisions.

A. Types of Learning

Machine learning algorithms are programs that are capable of learning from experience concerning some task. The performance measure of the learning algorithm improves with experience. The types of learning are

1. Unsupervised Learning
2. Supervised Learning

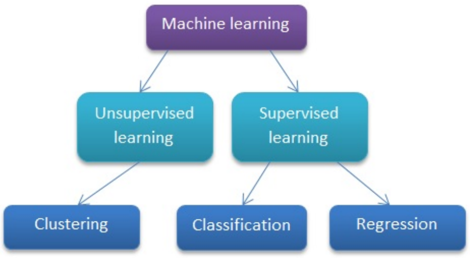

Fig. 1. Types of learning

B. Unsupervised Learning

Unsupervised learning finds hidden patterns within the data. The Data used will contain a subset of data which may show similarity within them, such that there might exist several subsets that show similarity among them which are grouped. The groups so formed are called as clusters. Unsupervised learning uses algorithms to identify clusters within data. They do not know any target values. They identify clusters based on the relation between every single instance in instance space. Algorithms like K-Means clustering are used in unsupervised learning.

C. Supervised Learning

Supervised learning infers a function that maps the input to the output by feeding the algorithm with the required number of sample data. Supervised learning deals with classification and regression problem. Algorithms that can perform binary class classification and multiclass classification like SVM, Logistic Regression, K-nearest neighbors, Naïve Bayes are used in classification problems. Algorithms like linear regression, ANN are used in regression problems.

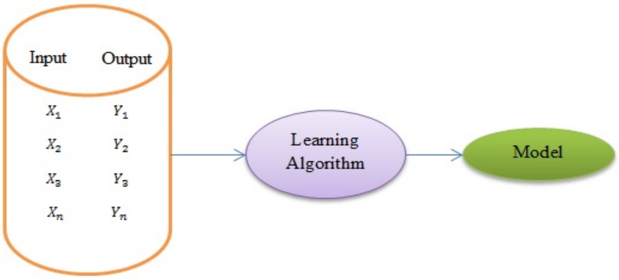

Fig. 2. Schematic diagram of a supervised learning

D. Linear Regression

Linear regression algorithm will be helpful in regression problems to predict and forecast the future depending upon the relationships identified between dependent and independent variables. Among a set of data points plotted in two-dimensional spaces, the best fit line is obtained by linear regression [2]. A linear regression line is of the form y=β_0+β_1 x_i1+⋯+β_p x_ip+ε_i

In this proposed system the time for the data request by the destination-driven data warehouse request is forecasted using linear regression algorithm.

Initially, for a few days, the data request is done manually and the request events are recorded. The recorded datasets will serve as training datasets for the linear regression algorithm. In this case, the attributes recorded are the day of the week and the hour for that day in which data request was made.

In this scenario, the training data used are tabulated below

TABLE I. TRAINING DATASETS

S.no	Day	Hour
1	1	3
2	2	3.4
3	3	3.5
4	4	3.6
5	5	3.4
6	1	3.5
7	1	3.6
8	2	4
9	5	3.4
10	7	3.3
11	6	3.4
12	6	3
13	2	3.2
14	3	3.4
15	4	3.5
16	5	3.3
17	1	3.2
18	1	3.2

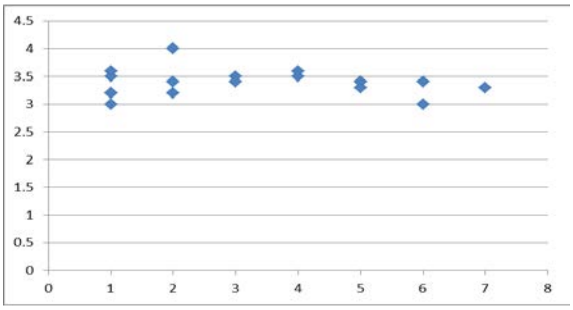

Fig. 3. Graphical plots of the training data points

The best fit line is calculated for the above dataset and is used as regression prediction model for future forecast.

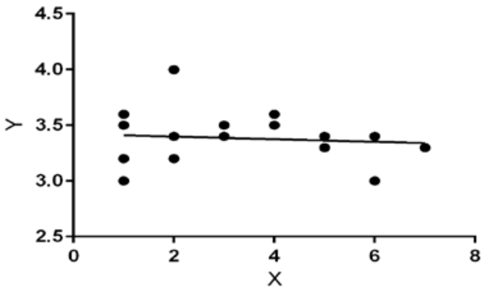

Fig. 4. Graphical representation of the best fit line calculated using linear regression

According to the developed regression equation, the future values are forecasted.

Thus giving the data warehouse, the ability to make a decision of when to make data request. After the regression equation is calculated the forecast of the value is made and the time to make the request is predicted and a new data request is passed to the data source

5. DATA TRANSFORMATION

Data transformation is the process of converting the data from one format to another to fit into destination system from a source system.

6. DATA SOURCE

In this proposed method the data source contains information about incentives provided by the government for renewables and efficiency resources used. The data retrieved from the data source is in JSON format with the following attributes.

S.no	Attribute
1	programid
2	code
3	state
4	implementation_sector
5	categories
6	type
7	name
8	fundingsource
9	budget
10	cities
11	zip
12	contact
13	email

These attributes are transformed to the required schema programmatically and are stored in the warehouse thus making it persistent.

The data is requested periodically at a particular time that was forecasted using the linear regression algorithm. Every time when the data is received it is transformed and is stored in the database.

7. CONCLUSION

In this system at last after the date has been requested the incentives information is received and is stored in the warehouse. Thus by making a machine learning algorithm like linear regression the data warehousing process which has destination driven architecture can be automated. This method reduces the interaction of humans with the data warehouse to operate. Even though there are automation tools for transformation phase, there is no automation process for taking human level decisions yet. This method can further be developed in future using various other machine learning algorithms to fully automate a warehouse from construction to operation.

REFERENCES

[1] Silberschatz A, Korth H F, Sudarshan S, (1997) "Database System Concepts", McGraw-Hill Series in Computer Science

[2] B. Ganapathy Subramaniam, T Rama Prabha (2017),"Linear Regression in Machine Learning", Rungta International Journal of Computer Science and Information Technology, Vol 2 Issue 1 & 2

[3] Vishal Gour et. al. (IJCSE) “Improve Performance of Extract, Transform and Load (ETL) in Data Warehouse”, International Journal on Computer Science and Engineering Vol. 02, No. 03, 2010, 786-789

[4] J.Anitha et al, “ETL Work Flow for Extract Transform Loading”, International Journal of Computer Science and Mobile Computing, Vol.3 Issue.6, June- 2014, pg. 610-617