IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 19, Issue 2, Ver. IV (Mar.-Apr. 2017), PP 78-82
www.iosrjournals.org

Analysis of Development Factors for Asian Countries using DWM on Big Data

Pallavi Varandani, Sharvari Jalit, Mrunmayee Mujumdar, Pradeep Lalwani, Prof. Arthi C I, Prof. Priya R L

Department of Computer Science & Engineering Vivekananda Education Society's Institute of Technology Mumbai, India.

Abstract: *In today's world, most of the developing countries are rising to become a developed country. There have been analysis of countries that experienced banking crisis in the past. However, the analysis included only data preparation process and the data mining server application for subgroup discovery induction. This paper proposes a data analytical system to perform the analysis on World Bank Indicators for Asian Countries from 1960 to 2015. The past dataset from the World Bank and other sources can be a source to predict the duration required for a country to be called a developed country. The purpose of the paper is to help the government of a nation to collect information and work on the path for the development more accurately. The analysis can be done using various methodology such as MapReduce, Canopy Clustering and Kmeans. Clustering and Reduction techniques are applied in parallel to enhance the existing technology. The outcome will be the prediction of development factors through analysis of various parameters such as Population, Gross domestic product, Trade and Employment, which affects the growth of developing countries.*

Keywords: *Asian Development, World Bank Indicators, Kmeans, clustering, canopy, big data.*

I. Introduction

Development of a country has become an essential factor to meet the requirements for healthy lifestyle in a country. There are many factors which affect the growth of a country. Thus, many countries are still in the developing phase from decades. There are countries that look similar but provide varied living standards to their citizens. For example, in 2009 a citizen in Burkina Faso earned on average 510 USD[1] in comparison to Japanese citizen that earned 37,870 USD. However, in Burkina Faso 29 percent of the adult population was literate and a new-born baby are expecting to live 53 years but all adults in Japan were literate and a Japanese newborn baby could expect to live 83 years. This scenario concludes that it is not just the economy which plays major role in the development of the nation. Other factors such as Life expectancy, birth rate, mortality rate, population, trade, etc. also have a part in the development of a country.

To get a better understanding of the development, the countries can be clustered based on its region and factors of development. Here, only Asian countries are taken into consideration for analyzing the development features. The main Goal of this paper is to propose a data analytical system for developing countries demonstrating the duration required and the main factors hindering their development. The data required for the analysis are obtained from World Bank. The World Development indicators is the open source database provided by the World Bank [14]. The analyzing techniques will include Map Reduce and Clustering algorithms. The Map Reduce method will first reduce the data by grouping the countries as per their region. From these group, clustering algorithm will be applied to the Asian countries only. The clustering algorithm will be performed in two stages: Pre Clustering and clustering.

This proposal will help the government of the nation to understand the hindrance for their respective countries. In the following, section I will describe the literature survey of the topic and extended and section II describes the proposed system design. Finally, section III will explain the future scope of the system.

II. Literature Survey

There have been studies on the world development indicators with respect to banking crises in 2013[2]. However, the analysis has been conducted using the data mining server application for subgroup discovery induction. The result of this analysis concludes five subsets of countries with only banking crises. Among the five subsets, three are known as financial driven types, while the rest two are of socioeconomic problems. Also, there are comparative studies of China and the world's Information Development. The data set has been compared with the data of all continents instead of individual countries. The evaluated result does not indicate the comparison of China with respect to the Asian Countries [3]. The research conducted states that China has made a steady growth in informatization.

The proposed paper is based on data mining and big data. It helps to understand the overall development of the country. It is always essential to implement indicator selection stage before performing the analysis. The Laplacian score methodology is used by Mariam, Ali and Denis in [1] to select the relevant income indicators. This method is based on a similarity matrix. Laplacian score is mainly used for the feature selection. In RStudio, this selection can be processed with package, graph.laplacian. The given input is a graph which will generate the output in the form of a symmetric graph. In this technique, the dataset is divided into training and testing sets.

The selection process is conducted on training set and the accuracy measure on the testing set. Further the next step is the selection of the algorithm for predictive analysis. As the system implements predictive analysis, it can use unsupervised learning algorithm. The selected indicators from the Laplacian Score methodology is further clustered. The clustering algorithms such as K-means is considered. After research and survey in 2006, K-means clustering algorithm was found to be second among the ten top popular data mining techniques [7]. Moreover, when it comes to the selection between k-means and k-medoids it totally depends on the type of the data set being used [4].

However, the major issue with K-means algorithm is computation complexity during analysis on very huge dataset. The solution for this are Canopy Clustering [5] and initial centroids based on weighted average [8]. Canopy clustering [5] is an unsupervised pre-clustering algorithm it processes huge data sets, but the resulting clusters are rough pre partitioned data set. This pre-partitioned data set helps to reduce the time complexity of the slow k-means algorithm. The selection of initial centroid has a great impact on the final cluster sets [8]. Initial centroid based on weighted average methodology overcomes the complexity of K-means by calculating the initial centroid by taking the average of the weights for each cluster instead selecting randomly in the first phase of the algorithm.

di=x1,x2,x3,…,xn
di(average) = (w1*x1+w2*x2+w3*x3+….+wm*xm)/m
where, x= the attribute value, m = number of attributes and
w = weight to multiply to ensure fair distribution of cluster.

In addition, Map Reduce technique can be applied in parallel to K-means algorithm [7]. Preprocessing of the data set is necessary before the application of the clustering algorithm. The preprocessing of data set includes understanding data, reducing the data by omitting the redundant entries, pattern creation etc. These preprocessing of the data set can be performed by MapReduce. Preprocessing the big data will lead to inefficiency [7]. Thus, Veronica S. Moertini and Liptia Venica in [7] shared a proposal that instead of preprocessing the attribute selection, data cleaning, etc. can be performed in parallel with clustering algorithm.

III. Proposed System

The proposed system does the analysis on World Bank indicators of various developing countries of Asian continents. This analysis will help government of a nation to work towards necessary development factors to achieve their goals.

A. Conceptual System Design

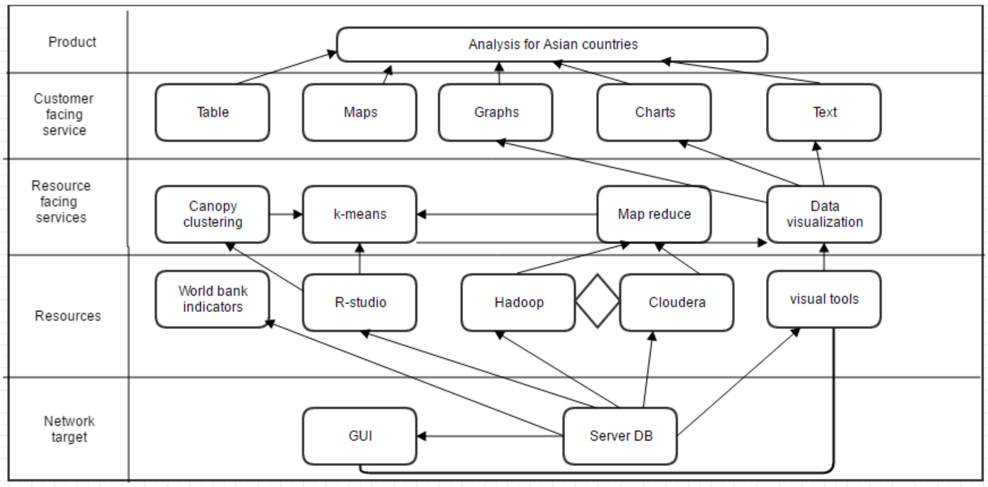

Figure 3.1.1

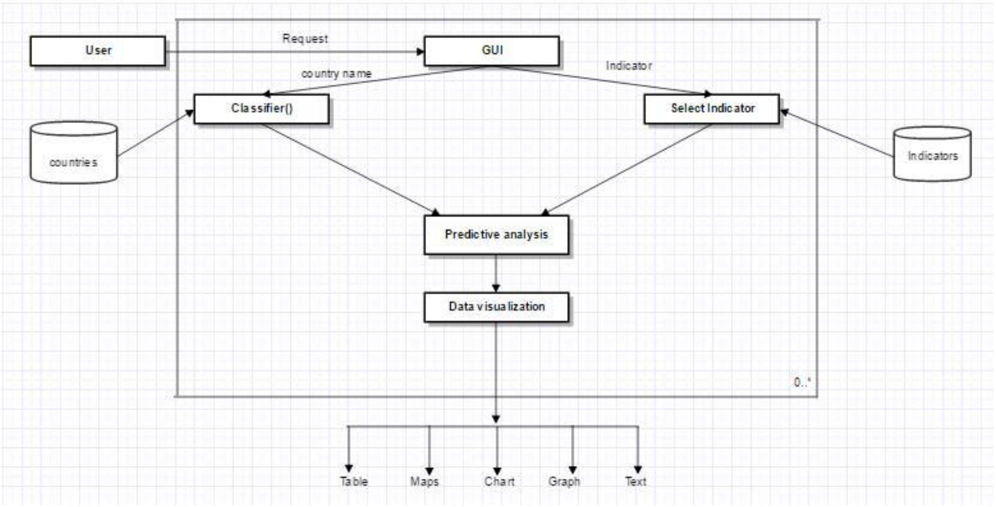

Figure 3.1.2

The figure 3.1.1 describes the conceptual model of the proposed system. It demonstrates various layers such as:
(a) Product: It shows the result of the system.
(b)Customer facing service: This module demonstrates the output visual format of the result which will help the government and non-government organization for better and simple understanding of the output.
(c)Resource Facing Service: Here, the model describes the techniques which can be used for the process of analysis and clustering.
(d)Resources: This phase states the tools to use and the source from where the data set is used.
(e)Network target: This module gives two main part of the system: GUI and Database.
The figure 3.1.2 describes the abstract view of the system. It consists of User, GUI, Database – countries and Indicators, classifier, Select Indicator, Predictive analysis and Data visualization.
User – The user can be Government or the Non-Government Organization.
GUI – The GUI accepts three input components from the user: Country, Indicator and Duration/Year.
Database – The database has two tables, Country: which provides the basic details, and Indicators: which provides the indicator code along with its value and year for all countries.
Classifier – This function will divide the countries into Asian and Non-Asian countries using MapReduce. Thus, creating Hadoop Clusters.
Select Indicators – This module will help in the selection of indicator from the list of all indicators.
Predictive analysis – Here the function will get input of the country and its indicator along with the duration/year. This module will give the output in the data visual format such as Table, Maps, charts, Graphs, and Text.
B. Data Flow Design

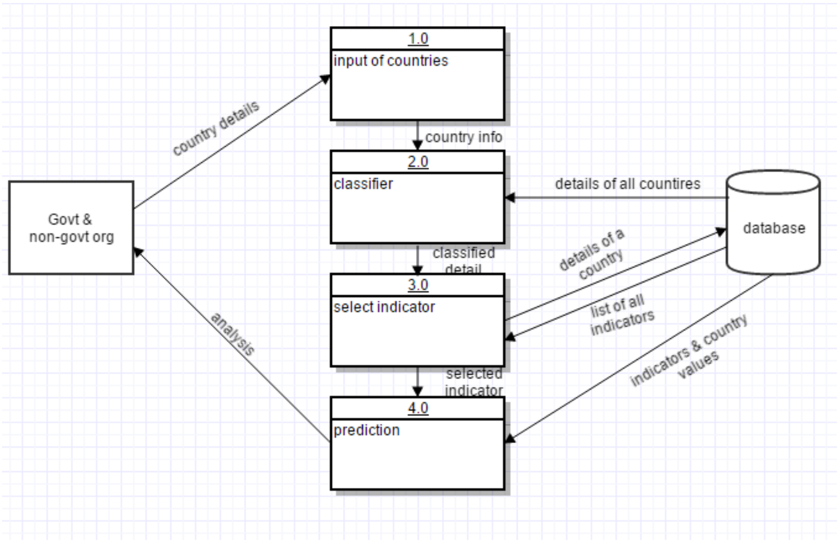

Figure 3.2.1 – Level 0 DFD

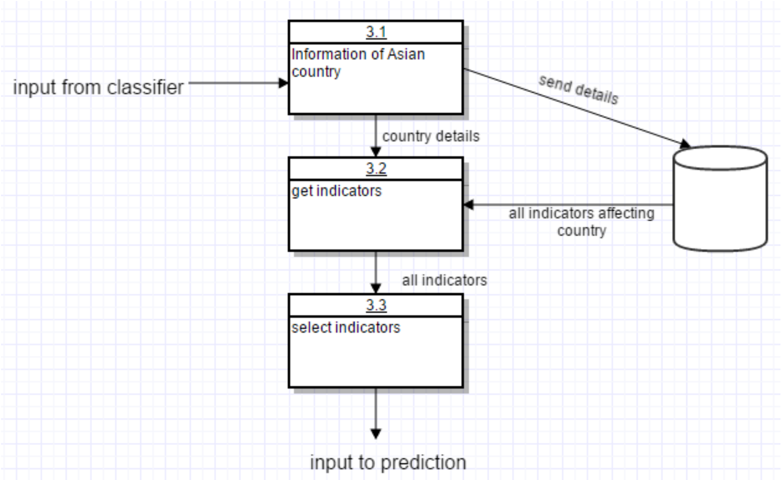

Figure 3.2.2 – Level 1(a)

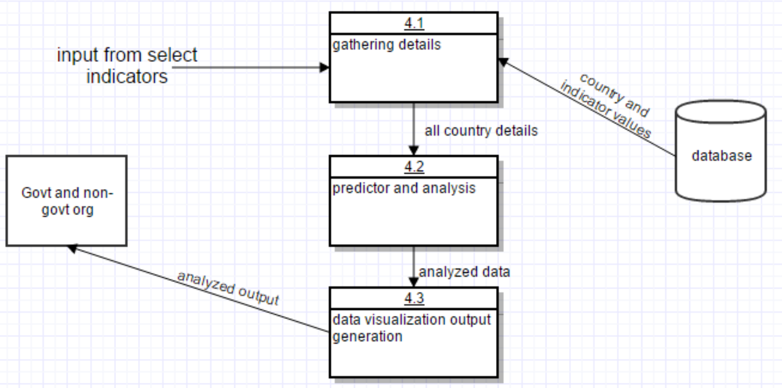

Figure 3.2.3 – Level 1(b)

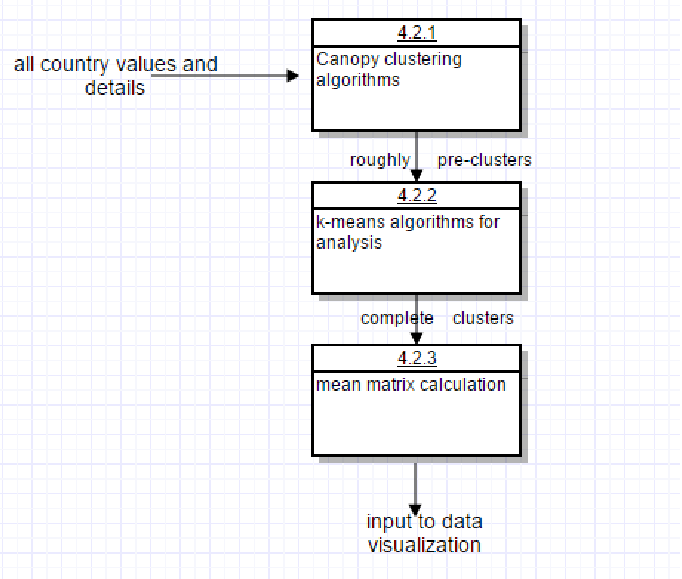
4.2.1
Canopy clustering algorithms
all country values and details
roughly pre-clusters
4.2.2
k-means algorithms for analysis
complete clusters
4.2.3
mean matrix calculation
input to data visualization

Figure 3.2.4 – Level 2

The above Design as depicts in figure 3.2.1 to 3.2.4 is the preliminary step to create an overview of the system, which can later be elaborated. It is basically the graphical representation describing the “flow” of the system. The stepwise flow of the system is:

1. The user of the system will give the input as indicator name, country name and time period.
2. The Raw Data will be mapped and reduced by using the MapReduce in Hadoop. This will help to reduce and clean the dataset by mapping them into groups.

3. The processed data from Hadoop can be then pre-clustered using canopy clustering in parallel with MapReduce technique.
4. Further, the pre-partitioned data can be used as an input for K-means Algorithm in R studio.
5. The resultant output will be generated in the form of graphical representation of the indicators affecting the development and growth of the countries.
6. Also, there will be a threshold which will define the stage at which the country will be a developed nation.

The drawback of this proposed system is, it is only restricted to the continent of Asian. Also, here number of indicators selected are restricted to four only, i.e. Population, Gross Domestic Product, Trade and Employment. This proposed system will generate the output of only one country at a time. Thus, the proposed system used to predict the duration required by a country to be a developed nation along with the factors that are hindrance to the development process of the nation.

IV. Future Scope

The proposed system will help government and non government organizations to analyze and predict the indicators which hinder the growth of a country. Thus, helping the country to work on the indicators in the optimal direction. With further enhancement in the data set the model can be used on global level for predictive analysis of all countries. As the predictive model give the duration required for the country to be a developed nation, it will create a threshold for every country.

V. Conclusion

The clustering process purposed in this paper will help in the predictive analysis of the indicators affecting the growth of the developing country. The main advantage of this clustering methodology is that the time complexity is reduced with the help of Map Reduce and Pre-Clustering algorithm using Canopy clustering technique. The result of this system will generate the duration required for a developing country in Asian continent to reach the threshold of development along with the indicators which can hinder the development. Also, the indicators act as a catalyst to reach the threshold of development. The methodologies provided in this work are more accurate and appropriate for predictive analysis of development factors of Asian countries.

Acknowledgments

The work presented in this paper has been supported by the institute “Vivekanand Education Society’s Institute of Technology” under the affiliation of University of Mumbai. This work is created under the supervision of Dr. Nupur Giri, Head of Computer Department of the institute.

References

[1]. Mariam Kalakech, Ali Kalakech and Denis Hamad, “Selection of World Development Indicators for Countries Classification”, International Conference on Digital Economy, 2016.

[2]. Dragan Gamberger, Drazen Lucanin and Tomislav Smuc, “Analysis of World Bank Indicators for Countries with Banking Crisis by Subgroup Discovery”, MIPRO, 2013.

[3]. Zhang Jianguang and Zhu Jianming, “A comparative Study of China and the World’s Information Development”, International Conference on Information Society, 2015.

[4]. Norazam Arbin, Nur Suhailayani Suhaimi, Nurul Zafirah Mokhtar, Zalinda Othman, “Comparative Analysis Between K-means and K-Medoids for Statistical Clustering”, Third International Conference onArtifical Intelligence, Modelling and Simulation, 2015.

[5]. Amresh Kumar, Yashwant S. Ingle, Abhijit Pande, Piyush Dhule, “Canopy Clustering: A Review on Pre-Clustering Approach to K-means Clustering”, International Journal of Innovation & Advancement inComputer Science, 2014.

[6]. Manuk Ghazanchyan, Janet G. Stotsky, and Qianqian Zhang, “A new Look at the Deminants of Growth in Asian Countries”, International Monetary Fund Working Paper, 2015.

[7]. Veronica S. Moertini and Liptia Venica, “Enhancing Parallel k-Means using Map Reduce for Discovering Knowledge from Big Data”, IEEE International Conference on Cloud Computing and Big Data Analysis, 2016.

[8]. Md. Sohrab Mahmud, Md. Mostafizer Rahman, and Md. Nasim Akhar, “Improvement of K-means Clustering algorithm with better initial centroids based on weighted average”, 7th International Conference on Electrical and Computer Engineering, 2012

[9]. Hans Rosling (Global Health Expert), “The best stats you’ve ever seen” and “Asia’s Rise- How and When”, TED talk, 2006.

[10]. Jiawei Han, Micheline Kamber and Jian Pei, “Data Mining Concepts and Techniques”, Third Edition and Radha Shankarmani, M Vijayalakshmi (Professor, VESIT), “Big Data Analytics”.

[11]. URL: link [www.worldbank.org]

[12]. URL: link [www.gapminder.org]

[13]. URL: link [www.kaggle.org]

[14]. URL: link [http://data.worldbank.org/indicator]