JOURNAL OF COMPUTATION IN BIOSCIENCES AND ENGINEERING

Journal homepage: http://scienceq.org/Journals/JCLS.php

Research Article Open Access

Unstructured Data: Qualitative Analysis

R S M Lakshmi Patibandla1, SanthiSri Kurra[2], PrasadAnde[3], N.Veeranjaneyulu[4]

[1]Assistant Professor, Department of IT, VFSTR University, India.

[2, 4] AssociateProfessor,Department of IT,VFSTR University, India.

[3]AssociateProfessor,Department of CS,VS University, India.

. *Corresponding author : R S M Lakshmi Patibandla, Mail Id: patibandla.lakshmi@gmail.com

Received: July 17, 2015, Accepted: August 27, 2015, Published: September 11, 2015

ABSTRACT

Major research is taking place in analyzing and processing of massive amounts of data produced by different organizations. The lumps of data received must be stored and different techniques are needed to visualize the numerical statistics. We have to focus on factors related to performance levels. This paper provides big data disaster management and different admission policy techniques to analyze the performance based on socio-economic factors and educational achievements along with comparison of admission policies.

Keywords: visualize, policy, socio-economic, performance, admission, management

INTRODUCTION

The research system "Big Data Analysis of University Data" is a profound analysis of large datasets of universities and students data, by using this datasets we are measuring, some interesting patterns and correlations with respect to universities performance.

Big Data can help in four phases of disaster management such as prevention, preparedness, response, and recovery. Two major sources of big data, dedicated sensor networks (e.g., earthquake detection) and multipurpose sensor networks (e.g., social media such as Twitter using smart phones). However, significant big data research challenges arise because of disaster management requirements for quality of service (e.g., highly available real time response) and quality of information (e.g., reliable communications on resource availability for the victim). Two of the major big data challenges are: Variety (integration of many data sources including dedicated sensors and multipurpose sensors), and Veracity (filtering of Big Noise in Big Data

to achieve high quality information). To fulfill the potential benefits of applying big data to disaster management, we need to bring together the best minds from around the world. From the big data view, we need the application pull of disaster management researchers to apply big data techniques and tools to solve real world problems.

Therefore, different methods and techniques from the fields of exploratory, quantitative and visual data analysis as well as statistics are deployed. The research is assisted by the University Guide which provides access to the datasets, and the business intelligence tool. Qlik View which is partly used to conduct the data analysis. Here the first part of this report describes the framework of the research system in further detail, including an introduction of the system resources as well as the applied process model. The following part is the disquisition concentrates on the actual data analysis.

System Design

Initial Situation and System Goal :The initial step of the system is data collected from the University, that has already conducted a brief analysis of the data, investigating on common patterns that might be interesting for parents to help with their universities choice. This will help the universities with respect to choosing the best one in all aspects.

However, this paper presents additional data records originated from reliable statistic offices may be integrated to draw a wider picture of national factors influencing the educational sector.

Finally the end result of the system will be helpful to educational as well as socio-economic factors and educational achievement.

System Resources

Data :The data provided by University is specified in the table below,

Description	Years
University performance	2004-2014
University Examinations tables	2004-2014
University Census	2004-2014
A-level people Examinations table	2004-2014
Students census	2004-2014

Figure 1: Data provided by the university

All the above data in the table are collected from the University.

Tools to be used

QlikView

Datasets from the university processes and analyses the data

with the business intelligence tool QlikView. QlikView consolidates data from multiple files and supports different data analysis and visualization[1] methods (QlikTech International AB, 2011). In this system, QlikView is used to combine, the search and discover the different datasets as it provides an intuitive interface for the analysis of bigdata. Further a simple statistical calculations and the first general visualization of the data are conducted.

However, to apply QlikView in order to do complex statistical calculations, elaborate skills and thorough knowledge is required. Due to the limited time scope of this research system, these skills could not be acquired. File format used by the business intelligence tool QlikView

Microsoft Excel 2010 The detailed data visualization as well as the correlation coefficient calculations, shown in this paper are conducted in Microsoft Excel 2010. This is used is due to the prior knowledge and experience of the researcher.

System Scope

In this system time factor plays an important role. Because of large amount of data it take more time to generate output through the interesting Patterns [2] statistical approaches so the time factor is the main limiting factor in this system, as it does not allow an in depth analysis of all provided data.

The data analysis is not conducted by a sociologist or expert but by a computer scientist. Therefore the investigation is exclusively based on statistical calculations and aims to be without bias concerning socioeconomic assumptions and theories

System Methodology

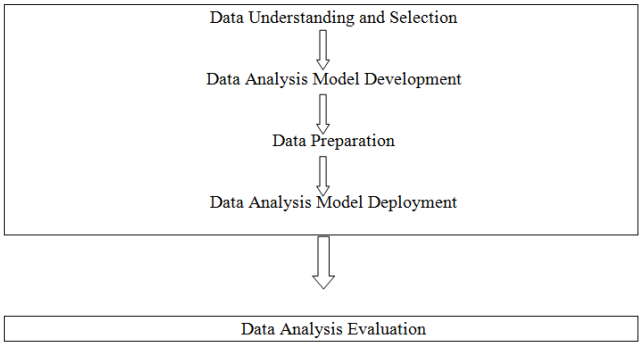

Figure 2: System Methodology.

Figure 2: System Methodology.

During the initial stage of the data analysis process, a general overview and understanding of the data is attained. Therefore, different techniques for visualization and statistical measures are applied, to investigate on single data attributes and discover correlations between different data attributes. Furthermore the quality of the datasets and its attributes are evaluated. After that vision and detailing of the research question, the system stage is concluded with the selection of data for the analysis.

The second step of the process concentrates on the development of an appropriate data analysis model, to approach the research question. Under the examination of the desired structure of the analysis result, diverse statistical methods are selected and combined. After the data analysis model is achieve, the data preparation stage is set off for the proper result. In respect of the data analysis methods chosen in the previous stage, the selected data is optimized in a cleaning and integration process. Finally, the model is deployed and the results of the data analysis are summarized and visualized.

In the final process stage, the data analysis model and the consequent results are evaluated. The validity and coherence in regard to the research questions is assessed, [4], reviewing each stage of the data analysis process. Subsequently, the informative value of the new results is rated.

SYSTEM IMPLEMENTATION

This stage deals with two components that are

Understanding the Data

Data Attributes

First, to gain a general overview of the data, the attributes of the University and students data are roughly explored. As summarization of this investigations [3]. This data model is the basis for the specification of the research area, as it reveals the relations between the data sets and the feasibility of their integration. Due to the time constraint, an in depth analysis of all provided data is not feasible. In order to exclude data from further interpretation, a rough specification of the research area is required at this system stage. However, an investigation on national socio-economic factors influencing school performance has not been conducted yet. For further dissolution on this research topic, regional datasets comprising appropriate information have to be included and set in relation to the existent datasets. For this purpose, the address characteristics of the university league tables are capable of serving as foreign keys. As the data attributes comprised in the university tables also give information about the university's performance, they provide adequate information for further research. Hence, the student's data, examination tables and reports are excluded from further analysis.

Data Visualization

Data visualization is the most significant technique It is not only used to visualize results but is data analysis tool by itself, as it provides a summary of the data.

At this stage of the system, the A-Level university datasets and their Competitive Index [5] of datasets are visualized to explore patterns and evolved interesting research questions. Furthermore, distinct visualization techniques are used, to discover zero values and outliers.

A-Level university data

To begin with, the university datasets of the different years are merged together and interpreted disregarding the year of their origin to conform to the unbiased approach of exploratory data interpretation. The following diagrams summarize the data included in the A-Level university league tables [6], focusing on the characteristics A-Level points per peoples or students as indicator of the university's performance. First, the scatter plot in Figure is created, to illustrate the value distribution of the attribute and highlight missing values and outliers. As a second numeric dimension is necessary for this visualization technique, the A-Level points per pupil are opposed to the corresponding value of the attribute p

Peoples or students. The value dissemination of the A-Level points clearly exhibits two aggregations, one between 100 and 500 points per people and the other between 400 and 1000. Furthermore a not negligible amount of null values is revealed.

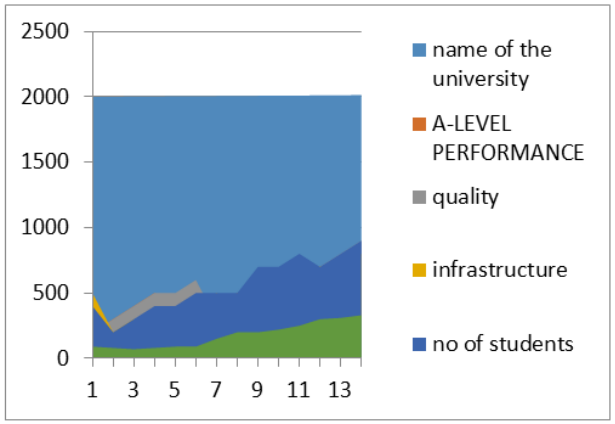

Figure 3: University comparative study (2000-2014)

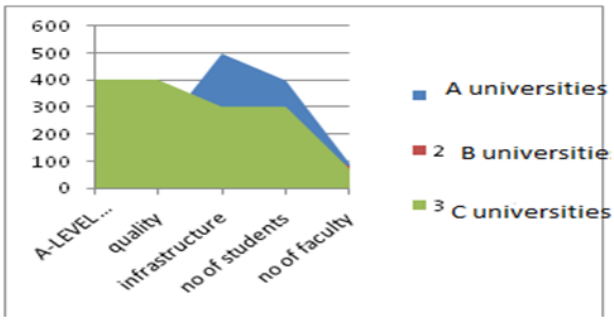

Figure 4: Comparative study on A-level universities data

Data Selection

The university data relevant to the research question is comprised in the A-Level performance tables. Due to the different A-Level point tariff system in the datasets 2003 to 2005 and the incomparability of the universities type attribute in the dataset 2011, the sample dataset for this analysis is composed of the datasets 2006 to 2010. In addition, to investigate on the second research question, all accessible datasets comprising the years 2005, 2006 and 2008 to 2010.

Data Analysis Model Development

The investigation on the dependencies between universities performance and different universities attributes as. In contrast to the university properties, the impact of the socio-economic [5] factors on the performance of a university is likely to be time-delayed. Therefore, two models are developed, each adapted to one of the two research questions.

Correlation between A-Level performance and university properties

The development of a data analysis model starts with the selection of the model class that determines the structure of the data analysis result correlation coefficients. Function between two quantitative attributes [5]. In contrast to correlation coefficients, the correlation ratio measures the correlation between nominal and quantitative attributes. The table below indicates the methods used for the correlation analysis

Data preparation

Before the data analysis models are deployed, the data sample that was selected in Chapter3.2 is prepared, to optimize its quality and applicability. For the data visualization stage, the A-Level performance tables for each year are already merged to one dataset that undergoes the cleaning process, documented in the next section.

Data Cleaning

First the A-Level data is cleaned from missing values and outliers that were partly identified during the data visualization stage. All records with missing values for the A-Level points per people attribute are excluded from the sample data set, as this attribute is essential for both analysis models. The remaining amount of records for the university type Special is no representative for further analysis and excluded as well. Beside this comprehensive dataset[13], several sub datasets for the breakdown analysis are prepared, excluding data records with missing values for the clustering criteria.

Data Integration

The sub dataset for the correlation analysis between A-Level performance of universities and the UKCI is constructed, As the local authority areas were subject to minor changes between 2008 and 2009, incomparable regions are deleted from the dataset. Next the A-Level performance tables are extended by a data field for the local authority code, serving as primary key and foreign key, respectively. For this purpose a local authority code - postcode translation table, included in the university Guide metadata collection, is joined with the A-Level dataset on the university postcode attribute, using QlikView[10]. To indicate the A-Level performance of each local authority areas, the median of the A-Level performance of all local universities is calculated for the data set. Additionally, this horizontal integration is conducted for each subgroup clustered by university type. Regions that have missing values for the *A-Level points per people [10]* attribute are deleted from the datasets.

CONCLUSION

This system is completely based on the comparative study of the A-level university as well as people studied in the university Within the limited time frame, set out for this research, an insight in the large datasets of university and student data is gained, using an exploratory visual approach to data analysis. Here we are using Ms-excel to view the statistical results, for further improvement for research study better to use QlikView. During this process, interesting patterns[12] with respect to universities performance are discovered, resulting in two detailed research questions that require an in depth investigation. Additional data records are integrated, to analyze correlations between socio-economic factors and universities performance on local level and an appropriate data model is generated.

The results of the data analysis disclose interesting and in some cases significant correlations between university properties and university performance.

REFERENCES

1. Bergdahl, M.,Ehling, M.,Elvers,E., Földesi, E., Körner, T.,Kron, A.,et al.(2007).
2. Handbook on Data Quality Assessment Methods and Tools. *Handbook on Data Quality Assessment Methods and Tools* (pp. 9-10). Wiesbaden: European Commission.
3. Berthold, M., Borgelt, C., Hoeppner, F., & Klawoon, F. (2010). *Guide to Intelligent DataAnalysis: How to Intelligently Make Sense of Real Data.* London: Springer.
4. Department for Education. (2012). *Technical annex*. Retrieved July 05, 2012, from http://www.education.gov.uk/performancetables/pilot16_05/annex.shtml
5. Frigge, M., Hoaglin, D. C., & Iglewicz , B. (1989). Some Implementations of the Boxplot. *The American Statistician, 43*(1), 50-54. Google. (2012). *Welcome to Fusion Tables*. Retrieved June 09, 2012, from http://support.google.com/fusiontables/bin/answer.py?hl=en&answer=2571232
6. Gupta, S. (2009). Business statistics. Jaipur: Sultan Chand & Sons. Huggins, R. (2003). Creating a UK
7. Competitive Index: Regional and Local Benchmarking. *Regional Studies, 37*(1), 89-96.
8. Mirkin, B. (2011). *Core Concepts in Data Analysis: Summarization, Correlation and Visualization.* New York: Springer.
9. Office for National Statistics. (2012). *Datasets and reference tables*. Retrieved June 02, 2012, from http://www.ons.gov.uk/ons/datasets-and-tables/index.html .
10. Park, E., & Lee, Y. (2001). Estimates of Standard Deviation of Spearman's Rank Correlation
11. Coefficients with Dependent Observations. *Communications In Statistics: Simulation & Computation, 30*(1), 129–142.
12. QlikTech International AB. (2011). *QlikView Reference Manual.* Lund:QlikTech International AB.
13. Rodgers, J. L., & Nicewander, W. A. (1988). Thirteen Ways to Look at the Correlation Coefficient. *The American Statistician, 42*(1), 59-66. Seale, C. (2004). *Researching Society and Culture.* London: SAGE Publications Ltd.
14. Wang, R. Y., Storey, V. C., & Firth, C. P. (1995). A Framework of Analysis of Data Quality Research. *IEEE Transactions on knowledge and data engineering, 7*(4), 623 - 640.

Citation: R S M Lakshmi Patibandla et al. (2015). Unstructured Data: Qualitative Analysis. J. of Computation in Biosciences and Engineering. V2I3. DOI: 10.15297/JCLS.V2I3.02