People Clues – Business Intelligence Tool to dynamically select the best team for a given project

Rohan Samarasinghe, Gayathma Perera, Nadeesha Liyanage, Pathum Senaratna and Lakshan Samarasingha
Research Group, Sri Lanka Institute of Information Technology, Sri Lanka
*Contact: rohan.s@sliit.lk, phone +94-71-861 0664
gayathma3@gmail.com, phone +94-75-602 0115
niklakshaya@gmail.com, phone +94-71-074 9024

ABSTRACT— Today in Sri Lanka, companies with a huge number of employees, workforce management is a critical factor who is working as teams in Projects. Managing individuals and groups in an effective way has a big impact on decision making process and productivity of the company. But with the churning of the employees knowledge transferring process has been really complex and also training a new candidate takes great amount of time and effort. Therefore selecting the right person for the job by satisfying time, scope, budget and required quality can change the outcome of these problems. People Clues is a business intelligence tool which selects the best team for a given project by analysing their past experience, Educational Qualifications and Past Performance. People Clues has been built on the concept of Team Dynamics. In this paper we present a desktop and a web application that facilitates the task of automating dynamic team generation depending on the optimality or feasibility based on different knowledge areas such as team dynamics, predictive modelling, business intelligence, data mining and team characteristics. An ETL (Extract, Transform and Load) tool will analyse the data and four prediction models will be designed in order to generate the optimal and feasible solutions for both apparel and IT industry. Depending on the relationship of the attributes People Clues will dynamically visualize the most optimal or feasible team on the Dashboard.

Keywords: Business Intelligence, Team Dynamics, Data Mining, Predictive Modelling, ETL, Hadoop

I. INTRODUCTION

People Clues - Business intelligence tool syndicates operational data with analytical tools to present complex and competitive information to decision makers, in order to dynamically select the best team for a given project.

Data in businesses can be very useful only if data is analysed properly which will aid in strategic business decision making. Therefore, we must find solutions to automatically analyse data, classify and summarize it along with discovering and characterizing trends and flagging anomalies in order to ease the decision making process effectively in a company. People Clues targets at developing a set of tools, technologies and programmed products that are used to collect, integrate, aggregate and make data available for better, faster decision making.

Today in Sri Lanka, in IT, Apparel and BPO sectors workforce management is a critical factor since there is a huge workforce working under these industries. Managing individuals and groups effectively has a big impact on decision making processes and productivity of a company. This research involves cutting-edge technical skills and domain knowledge in order to achieve what it is actually meant to be achieved as goals and it will be based on Team dynamics. "Team dynamics is the term used to define how people work and interact together in teams. Team dynamics are the hidden strengths and weaknesses that operate in a team between different people or groups and they affect how a team reacts, behaves or performs [1]." The effects of team dynamics are often very complex. There are various forces that could influence team dynamics, these include nature of the task, the organizational context, and team composition. Not only limited to ETL (Extract, Transform, Load) tool, this project involves the predictive model building, analysis and implementation in regards to the data set available.

Finally, after processing all data set based on predefined predictive models, it generates comprehensive, self-descriptive and exploratory Dashboards which provide fact – driven visualization of the data to decision makers of a company. A heterogeneous or homogeneous team will be given as a solution to the provided task in a representative manner. "The terms "homogeneous team" and "heterogeneous team" used in the current literature cover many different aspects. It is important to note that while all agents in genetically homogeneous teams share the same genes, agents can nevertheless be behaviourally heterogeneous. Heterogeneous teams are those in which agents are, on average, not genetically more similar to team members than to agents in the rest of the population [2]." Apart from that many software engineering principles and information technology knowledge areas are applied to this project. Especially desktop based application development skills, web development skills, data mining tools [3] and techniques are essential. In addition, some basic knowledge on business process and decision making process are also required throughout the development of the system.

II. OBJECTIVES

For a project to sustain its way towards success it is a common understanding that the outcome of a team operation

depends on the team members and their composition .a proper team should be selected and team combination should be perfect. The Team Dynamics Concepts can be used to achieving this goal. The Proposed People Clues is a research project focused on several objectives. With the completion of the project we are supposed to fulfil these research objectives. The main objectives of the research project are mentioned as follows.

Introduce an intelligent system to decide the team composition based on the team's historical outcomes and apply this system to compose project teams in Apparel and IT industries. The system relies on historical data of the procedures performed in the past. Depending on the project characteristics given an optimal or feasible team will be generated. The optimal team composition is the one with the lowest probability of unfavourable outcomes an optimal solution is the theoretically proven solution. But it might not be the logically suitable solution and we might have to come up with the feasible team. Hence the tool has the option of providing the most feasible (possible and practical) solution as well.

The People Clues enables users to generate a team for a given project by selecting relevant attributes for team formation depending on their company or industry. Most of the BI tools in the current market have a very complex interface where professional knowledge is required to perform the tasks. People Clues will not require technical knowledge or professional expertise to interact and it has been designed in a simple way with less number of controls in order to increase the adaptability and user friendliness.

People clues will generate the most optimal and feasible team for apparel industry considering attributes related to the IT and apparel sectors. Depending on the project characteristics and the relationships between them most optimal and most feasible team will be selected. The prediction is expected to be highly accurate and final decision is represented as a simulation.

III. SYSTEM OVERVIEW

This section describes the system architecture and the design of the proposed system. It mainly focused on few specific areas.
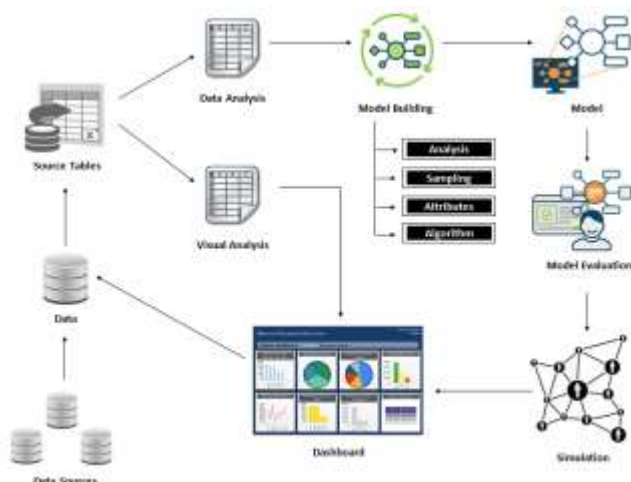
ETL (Extract Transform Load) Tool

Data in a business can be very useful if and only if analysed properly which leads to take strategic business decisions. It is acceptable that the data can be in various formats and locations. Thus People Clues should be able to extract data automatically in whatever form it is available.

The user can select only required attributes to analyses the data hence People Clues will only consume the selected attributes and precede for predictive analysis. Furthermore People Clues will automatically detects odd data and remove it from processing. Hence it will result in high- accuracy and best quality output. We only focused on HR systems, Finance systems, Project tracking systems as data sources. Basically ETL consists of 3 parts – Extract, Transform, and Load.

Extract includes 2 sub-parts, "identification of the data sources" and "extraction" of selected data automatically. In most of the organizations, there can be different data sources related to the system, stored in different locations in different formats such as xls, xlsx, csv, text, xml etc. So the system should be able to identify all these data sources and process them. After identification, necessary attributes has to be extracted and stored in a suitable way for local processing.

After that, data needs to be transformed to a more organized format. This is done in two steps, "data cleansing" and "data transformation". Data cleansing is performed by detecting and removing and/or correcting a database's dirty data (i.e., data that is incorrect, out-of-date, redundant, incomplete, or formatted incorrectly). Data imputation techniques will be applied in order to cope up with missing data values and remove duplicate data so as to fit it to the models [4]. Then in the transformation step, data may need to be merged, aggregated, enriched, summarized, or filtered depending on the nature of the integration scenario.

Finally the cleansed and transformed data are loaded and stored in data mart for further processing. The data mart is a subset of the data warehouse that is usually oriented to a specific business line or team. "A data mart is a simple form of a data warehouse that is focused on a single subject (or functional area), such as Sales, Finance, or Marketing. Data marts are often built and controlled by a single department within an organization. Given their single-subject focus, data marts usually draw data from only a few sources. The sources could be internal operational systems, a central data warehouse, or external data [4]."

Predictive Model Building

People Clues mainly works with data mining [3], trend and forecasting. Once the data is collected and required, relevant data is selected, People Clues will make the statistical model which will generate predictions.

So in building the predictive model People Clues will follow the following steps;

• Hypothesis Testing [9].

- Data Sampling [5]-[7].
- Algorithm building; Classification, Association, Factor Analysis and Regression [4].

This will generate a team for a given project depending on whether it is optimal or feasible. This optimality or feasibility will be selected during the prediction of the predictive model. With the help of this predictive model, the business performance can be shown as how it has been in the past, present and what will happen in the future, thus business predictions can be made wisely and easily.

The predictive model should be made up of a number of predictors, which are variable factors that are likely to influence future behaviour or results. First step is to sample the main dataset to obtain a representative, and statistically valid sample of the whole using a suitable sampling technique. A proper sampling technique such as randomized statistical sampling or probability sampling like cluster sampling and stratified sampling must be used for this purpose [6]. Using those techniques the system has to produce Test Set, Training Set and Holdout set in order to bench mark the models and find out the most accurate model.

Next, a suitable algorithm needs to be selected, based on the type of prediction required. This can be classification, association, factor analysis or regression. As each of these algorithms have different behaviours and produce different results, selection of proper algorithm has to be done with proper justifications.

Finally, selected algorithm needs to be customized to suit current scenario based on available data and produce final prediction model.

Predictive Model Evaluation

Predictive Model Evaluation is carried out by using 3 techniques - Classification Table (Confusion Matrix), ROC (Receiver Operating Characteristic) and Model Accuracy [8].

Classification Table shows the number of correct and incorrect predictions made by the classification model compared to the actual outcomes (target value) in the data.

ROC curve is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. The curve is created by plotting the true positive rate against the false positive rate at various threshold settings.

Model Accuracy is measured by the ratio of correct predictions to the total number of cases evaluated.

Dashboard Simulation

Once the data is analysed the Dashboard of People Clues will present it in a format which will make the user understand the distinction between raw data and predictive data in a meaningful way that would be able to depict current status of the entity. This data representation will help top management to take decisions wisely. People Clues dashboard will present data in an interactive way using graphs, charts, hierarchies and tables. Dashboard will provide summary of the results and key points regarding analysed data. This approach enables user to generate an optimal or feasible team with their specific roles for a given characteristics of a project. This is not just a static dashboard where user is presented with a set of predefined datasets every time it loads. The system is expected to provide better interaction with the user through interactive dashboard. While user can view high level result of the prediction, user may also trace selected data points and explore much deeper to find more detailed information of them.

III. METHODOLOGY

This section includes detailed descriptions about the techniques and mechanism employed to make this project a reality. The descriptions include how software implementation of our project is carried out, what are the materials and data needed, and how they will be collected. It also includes time frames and schedules that are required in achieving its objectives. In addition to them, the research areas that we have identified in order to carry out this project are explained rationally.

Our main objective is to develop a Business intelligence tool to select best team for a given project by analysing Experience, Qualifications and past performance. In order to achieve the goals of the proposed project there are 4 major tasks to be completed. Those are,

1. Traditional ETL Tool
2. Hadoop Integrated ETL Tool
3. Predictive Model for ICT Industry
4. Predictive Model for Apparel Industry

Traditional ETL Tool

Implementation of Hadoop Integrated ETL Tool

Apart from using traditional ETL we have decided to Integrate Hadoop to the ETL in both IT and Apparel Industries. This change is necessary to handle large, fast and unstructured data since traditional ETL can become and overhead and bottleneck because it is difficult to implement while taking a long time to execute and costing a huge amount for maintenance[22].

Apache Hadoop provides a cost-effective and massively scalable platform for ingesting big data and preparing it for analysis. Using Hadoop to offload the traditional ETL processes can reduce time to analysis by hours or even days. Running the Hadoop cluster efficiently means selecting an optimal infrastructure of servers, storage, networking, and software [23].

Hadoop is more recommended if the user is processing a data set with at least more than 20000 records. Talend open studio for big data has been used in order to design the cleansing mechanisms with different workflows for each and every rules. Therefore user get a chance to select whatever the preferred rules for the cleansing and transformation of the data.

Data Extraction – Extracting different data formats has been done using python and all the attributes has been mapped automatically by referring to a dictionary which has been built by the research team.

Data Transformation and Cleansing – Talend open studio for big data has been used in order to Cleanse and

transform data into a meaningful order. A separate workflow has been implemented for each and every rule and then that workflows has been called within the ETL java application by considering the user preferences in rules.

Data Loading – A data warehouse has been built in MySql with snowflake technique in order to load the cleansed and transformed data.

IV. FUTURE WORK

By improving the proposed research we are planning to take People Clues in to the international by localizing it to major languages over the world. A knowledgebase will be added to the data ware house with a neural network in order to increase the model accuracy. It will be evaluated to the point where user can customized it according to their company without any prior knowledge.

V. CONCLUSIONS

REFERENCES

[1] Shuib Basri, and Rory V. O'Connor, "Software Development Team Dynamics in SPI: A VSE Context," In Proc. Asia-Pacific Software Engineering Conference 19th'04, 2012, pp.1-2.

[2] Markus Waibel, Member, IEEE, Laurent Keller, and Dario Floreano, Senior Member IEEE, "Genetic Team Composition and Level of Selection in the Evolution of Cooperation," IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, VOL. 13, NO. 3, pp. 649, JUNE 2009.

[3] [Online].Available:http://www.bisoftwareinsight.com/business-intelligence-stats/ [Accessed: March 01, 2016].

[4] Elena M. Hernández-Pereira, Diego Álvarez-Estévez and Vicente Moret-Bonillo, "Automatic Classification of Respiratory Patterns Involving Missing Data Imputation Techniques", Biosystems Engineering, 138.March 2016 ,2015, 65–76 .

[5] Oracle Documentation, "Oracle® Business Intelligence Standard Edition One Tutorial," Oracle, E10312-01. [Online].
Available: http://docs.oracle.com/html/E10312_01/dm_concepts.htm. [Accessed: March. 02, 2016].

[6] Jeffrey D. Wells and others, "An Evaluation of Sampling Methods Used to Produce Insect Growth Models for Postmortem Interval Estimation", International Journal of Legal Medicine, 129.2 ,2015, 405–10

[7] [Online].Available:http://web-assets.domo.com/blog/wp-content/uploads/2013/11/2013-Data-Driven-Marketing-Survey-Report-Final.pdf[Accessed: March 7, 2015]

[8] Keith W Buffinton, Kathryn W Jablokow and Kathleen a Martin, "Project Team Dynamics and Cognitive Style", Operations Management, 14.3 ,2002, 570–77

[9] [Online].Available:http://www.312analytics.com/analysis-types-reporting-hypothesis-testing-data-mining/[Accessed: March 01, 2016]

[10] Leandro Soriano Marcolino, Haifeng Xu, Albert Xin Jiang, Milind Tambe, Emma Bowring, "Give a Hard Problem to a Diverse Team: Exploring Large Action Spaces,"

[11] Nahid Saberi, Mohsen Mahvash, and Marco Zenati, Member, IEEE, "An artificial system for selecting the optimal surgical team,"

[12] Zhongcheng Zhang, Xuefan Yang and Weikai Liu, "Optimization Model for Team and Member Selection in Mathematical Contest in Modeling," IEEE International Conference on Wireless Communications' 2010, pp.60 - 63.

[13] Waibel, Markus, Keller, Laurent, Floreano, Dario, "Genetic Team {C}omposition and {L}evel of {S}election in the {E}volution of {M}ulti- {A}gent {S}ystems," {IEEE} {T}ransactions on {E}volutionary {C}omputation, vol. 13, no. 12, July, 2005. [Online serial].Available:http://infoscience.epfl.ch/record/111745/files/Waibel09-IEEE.pdf. [Accessed Mar. 2, 2016].

[14] Agustín-Blas, Luis E.,Salcedo-Sanz, Sancho, Ortiz-García, Emilio G.,Portilla-Figueras, Antonio,Pérez-Bellido, Ángel M.,Jiménez-Fernández, Silvia, "Team formation based on group technology: A hybrid grouping genetic algorithm approach Computers & Operations Research, vol. 38, issue. 2, p. 484-495, 2011. [Online serial]. Available: sciencedirect,
http://www.sciencedirect.com/science/article/pii/S0305054810001371. [Accessed March 06, 2016].

[15] E. Semsar-Kazerooni, K. Khorasani, "A game theory approach to multi-agent team cooperation" American Control Conference, 2009. ACC '09.' 2010, pp. 4512 - 4518.

[16] Chen, Shi Jie Gary, Lin, Li, "Modeling team member characteristics for the formation of a multifunctional team in concurrent engineering", vol. 51, issue. 2, p. 111-124, 2004.[Online serial]. Available: researchgate.net,
https://www.researchgate.net/profile/Shi_Jiegary_Chen/publication/3076819_Modeling_Team_Member_Characteristics_for_the_Formation_of_a_Multifunctional_Team_in_Concurrent_Engineering/links/53f227f40cf2f2c3e7fcc723.pdf. [Accessed March 06, 2016].

[17] Xu, Y, Ma, J, Guo, X, "Modeling researchers' characteristics for the formation of research team", 2012. [Online serial]. Available: pacis-net.org, http://www.pacis-net.org/file/2012/PACIS2012-182.pdf. [Accessed March 06, 2016].

[18] Becchetti, Luca,Castillo, Carlos, "Online Team Formation in Social Networks", p. 839-848, 2012. [Online serial]. Available: research gate, https://www.researchgate.net/profile/Carlos_Castillo/publication/262240293_Online_Team_Formation_in_Social_Networks/links/548aeb130cf214269f1d7243.pdf. [Accessed March 06, 2016].

[19] [Online].Available: http://www.softwareadvice.com/bi/ [Accessed: March 07, 2016]

[20] [Online].Available:https://www.oracle.com/solutions/business-analytics/business-intelligence/index.html [Accessed: March 07, 2016]

[21] Elena del Val, Juan Miguel Alberola, Victor Sanchez-Anguix, Alberto Palomares,Ma Dolores Teruel, "A Team Formation Tool for Educational Environments" ", vol. 293, p. 173-181, 2014.[Online serial]. Available: http://link.springer.com/chapter/10.1007/978-3-319-07476-4_21#page-1

[22] [Online]. Available: https://tdwi.org/Articles/2015/02/24/Can-Hadoop-ReplaceMy-ETL-Tool.aspx?Page=2 [Accessed: March 31, 2016]

[23] [Online].Available:https://software.intel.com/sites/default/files/article/402274/etlbig-data-with-hadoop.pdf [Accessed: March 31, 2016]