Context Aware Recommendation for Data Visualization

W. A. D. Kanchana
kanchana.12@cse.mrt.ac.lk

G. D. L.
Madushankagdlmadushanka@
gmail.com

H. P. Maduranga
prasad.12@cse.mrt.ac.lk

M. D. M. Udayanga
udayanga.12@cse.mrt.ac.lk

D.A. Meedeniya
dulanim@cse.mrt.ac.lk

G. I. U.
S.Pereraindika@cse.mrt.ac.lk

Department of Computer Science and Engineering
University of Moratuwa
Moratuwa, Sri Lanka

ABSTRACT

Visualization is the final and most important step in a data mining process. The effort we put in mining information become useless if we fail to convey the findings properly to users. Therefore selecting the best visualization available for data with the right context is very important. Some of the data scientists may not have the expertise in data visualization. Often they have to process data that come from unknown domains. If we can take the service from a domain expert he/she can easily recommend commonly used best visualization types for us. Again availability of a domain expert in a data analysis project teamcannot be guaranteed. This paper proposesan automated system for recommending the most suitable visualization method for a given dataset using state of the art techniques. Our system is capable of identifying and matching the context of the data to chart types, which will enable the data scientists to take visualization decisions without the help from a domain expert.

CCS Concepts

• Human-centered computing → Visualization → Visualization application domains → Information visualization • Information systems → Information systems applications → Data mining.

Keywords

Recommender systems; Data visualization; Context awareness; Machine learning; Rule base engine

1. INTRODUCTION

Information technology plays a major role in modern day to day life. There is a tendency to collect user details by many e-commerce companies such as Amazon, eBay and professional networks such as LinkedIn, ResearchGate in order to provide suggestions for users, identify different design patterns, etc. Manufactures and sellers are increasingly interest in collecting information, which is related to their customers. YouTube, Amazon and eBay are examples for that such involvement. In interactive machine learning, users interact with the learning algorithm to solve problems. Training and recommendations of such systems require expert knowledge and complex mechanisms. Thus, it is important to have interactive systems with continuous feedback on the learning process and performance of the model. The decision making process and the visual analysis of such systems are supported by data visualization. A visual representation facilitates better accessibility, understandability and usability of data and supports to communicate information clearly and efficiently.

As the field of information visualization matures, there are thousands of information visualization tools. But most of them do not focus on the context the data when visualizing it [1,2]. In this research paper we introduce a framework that focuses on the context of the data and recommend the best visualization technique for the given data set automatically.

In this paper we provide a mechanism to characterize user data and map into a group of matching visualization techniques. Context awareness component use to extract the context/ domain specific metadata from a given dataset and send to rule based recommendation component. Each domain has its own way of depicting data. Therefore, the visualization is highly dependent on the targeted domain [3]. Recommendation component usesmachine learning techniques heavily in order to reveal hidden features when we are selecting visualization technique for a data set. Rule based engineis used to refine recommendations provided by the ML component. After that we visualize user data using selected visualization technique. At that point we use front end visualization libraries such as chart.js, Google chart libraries and D3.js.

The paper is structured as follows. Section 2 describes related work exist in the literature. Section 3 explains the features of the proposed system and Section 4 details the implementation methodologies. Proposed User Interfaces are shown in Section 5 and finally section 6 concludes the paper.

2. BACKGROUND

Different tools and techniques are available for data management and visualization. Rapidminor[2] is a more powerful and comprehensive tool than Weka[4] that can be used to machine learning, data mining, text mining, predictive analytics and business analytics. As the name implies it enables rapid prototyping and application development. It allows easy implementation of ETL process (Extract, transform, load)[5] and workflow architecture of Rapidminor nearly eliminate the need of writing codes. It has “operators” performing specific tasks. Each operator’s inputs and outputs are well defined. We can create a process by linking operators in a workflow. However, automated data context identification is still not there. Google Fusion Table[6] is an online application for the data management which is specially target for the users who are new to the computing environment. It supports for the new features such as data acquisition, collaboration, visualization and web publishing.

This application mainly fulfill following facts.

- Data management tool need to support collaborative among multiple users and multiple organizations

- Data management tool can be easily used by a person who does not have good technical knowledge
- Data management tool should be immediately compatible with the web

Fusion Table supports user to visualize data with combination of data visualization and SQL like querying. If dataset contains data about geographic data it offers several map view options. If it contains data and time then it shows timeline and motion charts. If data contain numeric columns it shows bar charts and pie charts. Fusion Table mostly focusing on geographical data visualization. Further user can add those charts into a blog or website by given HTML snippet.

VizRec [8] is a system that learns user profiles and feedbacks to give recommendations. They had gathered data by a crowd sourcing approach. They use a hybrid approach by combining a Rule based RS and collaborative filtering RS. This is the most similar approach we found up to now. But this differs from our system by the second part. Instead of collaborative filtering of user profiles we use content based filtering of charts.

3. SYSTEM FEATURES

Main intention of this project is to recommend the most suitable visualization technique for agiven data set. User does not need to worry about his or her knowledge on data analytic and visualization. The system will fulfill the knowledge gap between user and the data analytic expertise.

Onceuser uploads the data into the system and he/she can choose columns for visualization. Thereafter user receives a set of recommendations for the given data set. Users can select best matching technique out of them and rate the recommendations. Those ratings will be stored in order to refine the recommendation process. This GUI is provided for normal users;Apart from that API has been provided for developers to interact with the system.

Furthermore, administrators can add new chart types by defining the characterization of particular chart type. Further, depending on the user feedbacks to chart recommendations rule adjustments are auto generated by the system. Administrators can decide to accept or reject those adjustments to improve system. If some adjustment is recommended more frequently is means that change is important thing to do. After user accepting a rule adjustment those changes must be added to the rule engine.

4. SYSTEM METHODOLOGY

4.1 Data Flow

Our solution is an API for developers to use. Using that either a web or standalone application can be built. Here we describe the functionality of a web application created on top of our API.

Consider the data flow diagram shown in Figure 1.First user must upload their data set to the site. Then it is moved to the context awareness component, which recognizes data types, continuous and discrete nature and some special context features such as date time, percentage values, and geo location data and the inter column dependencies of data. Then we show the user the recognized context and ask for alterations and confirmations. This user feedbackis recorded and used for the improvement of context awareness component.

After user's confirmation recognized context object is sent to machine learning and rule engine components. Those components will use that to recommend matching chart types. In initial stages results from the model may be noisy and may include false positives. Therefore rule engine is there to maintain the accuracy while model become consistent gradually. The systemdoes not just use the rule engine as a helping module, but it also gets improved by the usage through its own learning. Necessary rule adjustments are automatically generated by the system and developer can accept or reject them to improve rules.

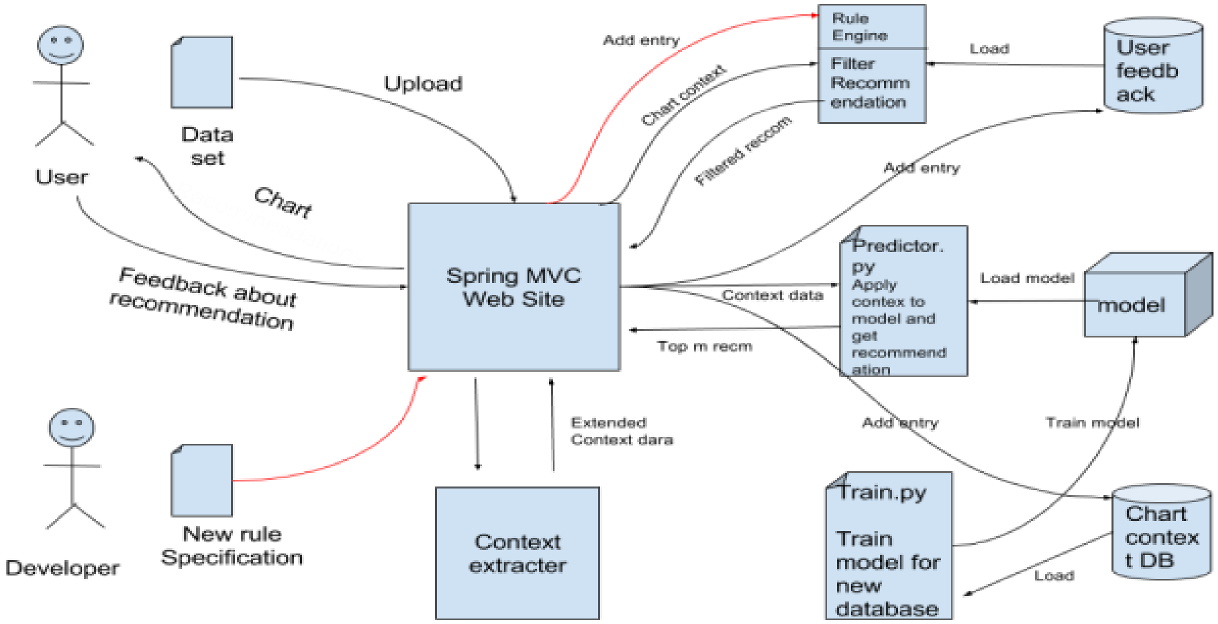


Figure1: Data Flow View of the System

Finally a recommending list of charts in the order of importance is given to the user. User can apply them, see the visualization and finally select best match for the data set. Then their feedback is also recorded in appropriate databases used in next iteration of training the model.

Apart from this usual process, developers can submit new charts types and new rule definitions to the system. Also they can receive auto generated rule adjustments and accept or reject them to improve the system. These developer tasks are not mandatory to the functionality of the system and these will not occur frequently.

4.2 Graph Characterization

In order to provide chart recommendations we have to somehow encode charts and user input data sets using common mechanism. So we have derived a mechanism for that characterization.Chart or data set is characterized using number of dimensions it has, intension(comparison, composition, relationship, distribution) of the visualization and the properties of each dimension [9]. Those dimensional properties include cardinality, continuous/ discrete nature, context of the data (numerical, nominal or ordinal) and the partial orderings exist between columns [10].

4.3 Basic Components

4.3.1 Context Aware Component

4.3.1.1 Schema Extractor

Purpose of this module is to identify the schema of given table. Table can be on CSV, MYSQL, and XML etc. This component is capable of recognizing all input data format. For some of the databases, schema is already given with the dataset (such as sql file). The dataset schema can be easily identified by analyzing metadata of the given dataset.

Schema extractor module goes through all of the records and checks it with possible matching string which is given in a lookup table. Lookup table contain possible string and each string mapped to relevant ontology. Since there can be string misspelling, Levenshtein algorithm is used to identify most relevant string from given lookup table. Based on the user feedback, lookup table update with new string and mapped ontology.

Further, based on the user feedback new entities are added to the lookup table and improve the functionality of schema extracting.

4.3.1.2 Entity Extractor

Purpose of this module is to identify the type of entity given in a column. This module goes through all selected cells in each column and separates it into different ontologies.

First data should be sampled using sampling technique to reduce execution time to the process. Random sampling is used for this task. Then each cell load into the regular expression filter. This filter puts string into to regular expression entity recognition and extract entities such as date, time, and currency etc. remainingunrecognized string passed into DBpedia SPARQL NER module. This module creates SPARQL statement and sends it to DBpedia server by using HTTP request. DBpedia dictionary looks up for the string and returns possible ontology. This interface can return query result as HTML, XML, JSON, CSV etc. JSON used for this application since it is lightweight and easy to implement. SPARQL module analyzes the returned JSON object and identifies related ontology. This SPARQL interface can be also used to get other related information for the dataset such as geo coordinate for given location or city.

4.3.2 Recommendation Component

Once context of each column is identified with the above component, this information is fed into recommendation engine in order to relevant visualization suggestions. Recommendation is handled by basically two components.

- Model developed using Machine Learning
- Rule based recommendation engine

Given set of rules sometimes may not work withsome cases. Most of the times it is not possible to define a clear cut separation between chart types. Often chart type may depend on the cardinality of dimensions. We cannot define an exact threshold for switching between chart types. Hence, it is better to have a model, which is trained using practical instances of visualization recommendations.

But in cold start situation our model might not be matured enough in certain cases to give recommendations. Sometimes In those cases its recommendations may give irrelevant and incorrect results. Thus in such a case there is rule based component, which dominates the recommendation list. This rule based component can also be considered as a filter for results produced by the previous model. With that, system becomes more robust under anomalies and faulty user feedbacks (Since system uses user feedback to improve the model, incorrect user feedback may lead to incorrect recommendations). With this component wecan ensure the robustness of the framework where system will not provide under any circumstances.

The other fact is the sustainability of the framework. It should be fairly easy to add new chart types to the framework. With the usage and usefulness those new types will be appeared at the top in the recommendations.

4.3.2.1 Machine Learning Component

Fundamental purpose of this component is to reveal hidden features when we are selecting visualization technique for a data set. And also this component makes it easy to add new chart types to framework. This component helps users to get most trending and most relevant chart types for their data set.

First it reads data from a data set which consists of mapping between different contexts and corresponding visualizations. Then it fits those data to a model. We have implemented probability based decision tree[7] approach in this case. Since this is a classification problem those algorithms provided good results. With this methodology we got the most relevant chart type for particular data set.

Performance of this component is highly dependent on the output coming from the context awareness component. Since the internal process information is dependent on the metadata collected on the columns. Data set is also a critical aspect of this component. Maturity of the data set is very important because the coverage of model directly depends on the data set. User selections and ratings are used to improve the data set in turn improves the whole system of recommendations.

4.3.2.2 Rule Based Component

Purpose of this component is to refine recommendations provided by the Machine Learning (ML) component. In a cold start situation ML component may provide incorrect results. In such case, this component should provide better results based on specified set of rules. In a case where ML component providing totally irrelevant result this component should filter outthem. This type of failure is possible since ML component depends on user feedback and users may provide misleading feedbacks intentionally.

First it reads context/metadata of the data set given by the user. Then with the defined rules it funnels the possible chart types for visualization. We can guarantee that this list of charts is capable of visualizing given data. Then it considers the list provided by the ML component and its confidence of recommendation. Using the list generated by rule based engine, it can filter and refine the list generated by the ML component. If the confidence of the list generated by the ML component is low, then the new list will dominate the final recommendation list.

4.4 Implementation View

Using this proposed API developers can implement both web and standalone applications. Figure 2 describes the implementation with respect to a web application. Presentation layer consists of a set of User Interfaces (UI) to interact users with the system easily. Data upload UI to upload their files, Column selection UI to select set of columns to be visualized and to apply changes to the automatically recognized context, Recommendation and chart view UI to show the ranked order of chart recommendation and apply those chart types to data. Also, there is a feedback UI to collect user feedbacks. From the perspective of developers there must be a UI to add new charts and accept and reject rule adjustments.

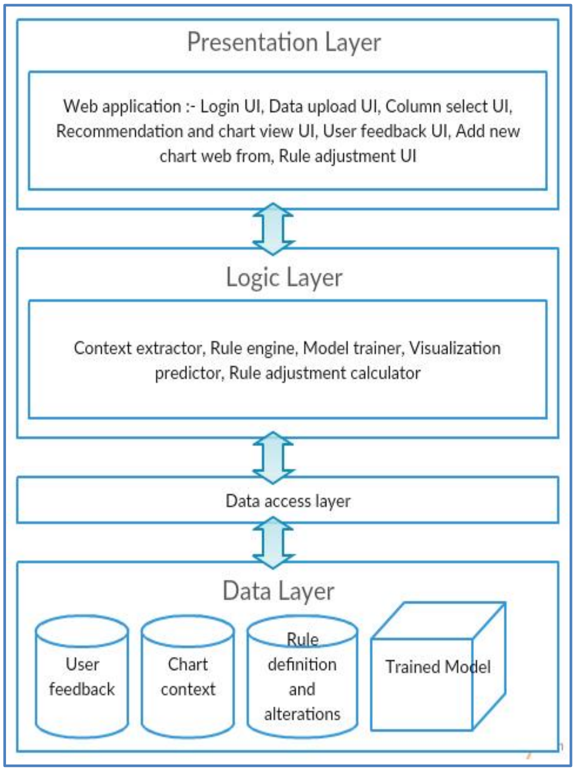
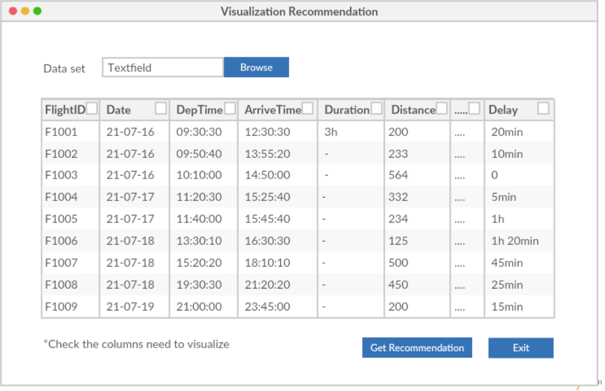

Presentation Layer
Web application :- Login UI, Data upload UI, Column select UI, Recommendation and chart view UI, User feedback UI, Add new chart web from, Rule adjustment UI
Logic Layer
Context extractor, Rule engine, Model trainer, Visualization predictor, Rule adjustment calculator
Data access layer
Data Layer
User feedback
Chart context
Rule definition and alterations
Trained Model

Figure 2: Layered Architectural View of the System

Logic layer provides the core functionality of the system including context aware rule engine, model trainer, rule adjustment generator, and visualization prediction components. Moreover, non-volatile data and the API separated by a data access layer which provides access to user feedback, chart context, rule definition databases and finally to the train models.

5. USER INTERFACES

The proposed system consists of set of Graphical User Interfaces that facilitate users to interact with the system. Data uploading interface of this system is shown in Figure 3. This interface is used to upload the data set to the system. Then the system will preview the content of the dataset as a table. Then the user can select the columns (attributes) which need to be visualized.

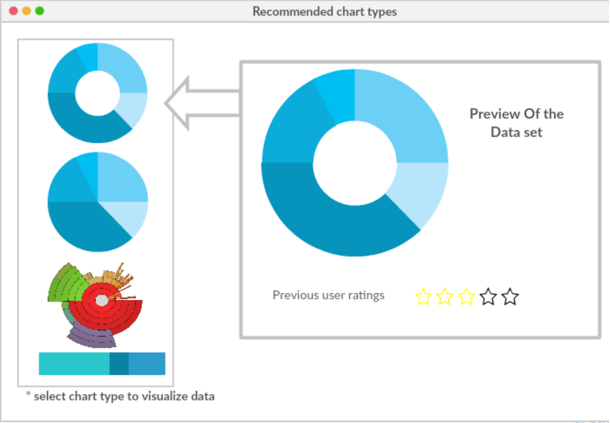


Data set Textfield

FlightID □	Date □	DepTime □	ArriveTime □	Duration □	Distance □	..... □	Delay □
F1001	21-07-16	09:30:30	12:30:30	3h	200	....	20min
F1002	21-07-16	09:50:40	13:55:20	-	233	....	10min
F1003	21-07-16	10:10:00	14:50:00	-	564	....	0
F1004	21-07-17	11:20:30	15:25:40	-	332	....	5min
F1005	21-07-17	11:40:00	15:45:40	-	234	....	1h
F1006	21-07-18	13:30:10	16:30:30	-	125	....	1h 20min
F1007	21-07-18	15:20:20	18:10:10	-	500	....	45min
F1008	21-07-18	19:30:30	21:20:20	-	450	....	25min
F1009	21-07-19	21:00:00	23:45:00	-	200	....	15min

*Check the columns need to visualize

Figure 3: DataUploadingWindow



* select chart type to visualize data

Figure 4: RecommendationWindow

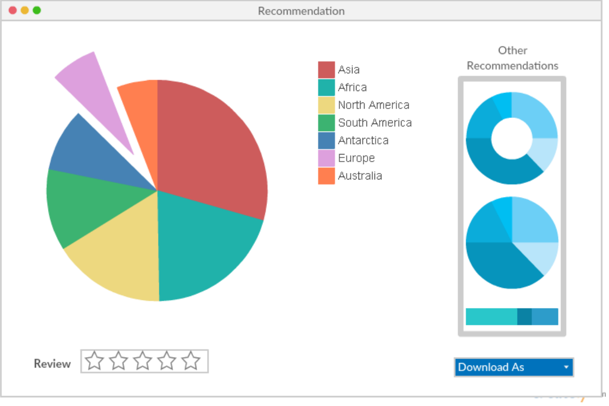


Figure 5: UserFeedBackWindow

Figure 4 shows the Recommendation user interface of the system that shows the recommended chart types that are generated through the system. By hovering the mouse over each recommendation, there will be a preview of the data set by the chart type user point the mouse. Also that preview include the ratings of the previous users (how much of the users have use this chart type to visualize similar kind of data set).

The User feedback window shown in Figure 5 visualizes the data set with the selected chart type by the user. After applying recommended visualizations user select the best visualization method for him/her. Then user can rate that visualization as a measure of to which extentthat chart type represent the required information and to which extent it correctly represents the context of dataset. Here, a user can download the chart in a given type such as PNG, JPEG, etc.

6. CONCLUSION

As the interest on data science is increasing rapidly there is a need for tools which are user friendly, accurate and powerful enough to support for data visualization. In this paper we introduced a new dimension of the data visualization using the machine learning based context identification. This framework will recommend the most appropriate chart types to a given dataset by considering the context of each dimension of the data. This framework will become a prominent tool between the data analyzers and the open source community.

Up to now this framework supports 22 basic chart types. This framework can be improved to support more chart types by adding new constraints and context information to the rule base system. At this stage user is able to upload dataset which is in CSV and XML format. The system is expected to be improved to support JSON, SQL, and Excel data types in future.As of now the system is able to identify partial ordering dependencies.

At the initial stage system supports up to 5 columns of data visualization. In the next iteration the system will be improved to support any number of dimensions that are given at the data set. Since this application use DBpedia SPARQL interfaceto identify geodata from a given string, this application can be extended to update machine learning model by using DBpedia ontologies. This model can be given more accurate data by analyzing the context of the strings given in each cell of the table. Undiscovered relation between given data set and graphs can be identified by analyzing context of the table.

7. REFERENCES

[1] Minitab, I., 2000. MINITAB statistical software.Minitab Release, 13.

[2] R. Platform, "RapidMiner: Open Source Predictive Analytics Platform", RapidMiner, 2016. [Online]. Available: https://rapidminer.com/. [Accessed: 16- Jun- 2016].

[3] van Wijk, J.J., 2006. Views on visualization.IEEE transactions on visualization and computer graphics, 12(4), 421-432. DOI=DOI:10.1109/TVCG.2006.80

[4] Hall, M., Frank, E., Holmes, G.,Pfahringer, B., Reutemann, P. and Witten, I. H.,The WEKA Data Mining Software: An Update; SIGKDD Explorations, 2009, Volume 11, Issue 1

[5] Vassiliadis, P., Simitsis, A. and Skiadopoulos, S., 2002, November. Conceptual modeling for ETL processes. In Proceedings of the 5th ACM international workshop on Data Warehousing and OLAP.ACM, New York, USA, 14-21. DOI=http://dx.doi.org/10.1145/583890.583893

[6] Gonzalez, H., Halevy, A.Y., Jensen, C.S., Langen, A., Madhavan, J., Shapley, R., Shen, W. and Goldberg-Kidon, J., 2010, June. Google fusion tables: web-centered data management and collaboration. In Proceedings of the 2010 ACM SIGMOD International Conference on Management of data.ACM, New York, USA, 1061-1066. DOI=http://dx.doi.org/10.1145/1807167.1807286

[7] Safavian, S.R. and Landgrebe, D.,A survey of decision tree classifier methodology,1990

[8]Mutlu, B., Veas, E., Trattner, C. and Sabol, V., 2015, March. Vizrec: A two-stage recommender system for personalized visualizations. In Proceedings of the 20th International Conference on Intelligent User Interfaces Companion.ACM, New York, USA, 49-52. DOI=http://dx.doi.org/10.1145/2732158.2732190

[9] Herman, I., Melancon, G. and Marshall, M. S., “Graph visualization and navigation in information visualization: A survey,” IEEE on Visualization and Computer Graphics, vol. 6, no. 1, pp. 24-43, 2000. DOI= 10.1109/2945.841119

[10] Ware, C., Information Visualization: Perception for Design, Elsevier, 2013.

Authors’ background

Your Name	Title*	Research Field	Personal website
G.I.U.S. Perera	Senior Lecturer	Data analytics/ Software engineering/ Human Computer Interaction / Virtual Reality	www.indika.perera.lk
D. A. Meedeniya	Senior Lecturer	Data analytics/ Software engineering	www.dulani.meedeniya.lk
H. P. Maduranga	Final year undergraduate	Big data analytics/ Db administration/ Data visualization	https://www.linkedin.com/in/prasadmaduranga
G. D. L. Madushanka	Final year undergraduate	Big data analytics/ Embedded systems/ Data visualization	https://lk.linkedin.com/in/lahiru-madushanka-24604691
W. A. D. Kanchana	Final year undergraduate	Big data analytics/ Db administration/ Data visualization	https://lk.linkedin.com/in/dilshankanchana
M. D. M. Udayanga	Final year undergraduate	Big data analytics/ Embedded systems/ Data visualization	https://lk.linkedin.com/in/dhanushkamadushan

***This form helps us to understand your paper better,the form itself will not be published.**

***Title can be chosen from: master student, Phd candidate, assistant professor, lecture, senior lecture, associate professor, full professor**