# Context-aware Recommendation for Data Visualization

L. A. D. P. Athukorala

(178053R)

Degree of Master of Science (Research)

Department of Computer Science And Engineering

University of Moratuwa

Sri Lanka

June 2019

# Context-aware Recommendation for Data Visualization

L. A. D. P. Athukorala

(178053R)

Thesis submitted in partial fulfillment of the requirements for the

degree of Master of Science (Research) in Computer Science and Engineering

Department of Computer Science And Engineering

University of Moratuwa

Sri Lanka

June 2019

## Declaration

I declare that this is my own research Thesis and this Thesis does not incorporate without acknowledgement any material previously published submitted for a Degree or Diploma in any other university or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Signature: …… Date:

L.A.D.P.Athukorala

I have read the Thesis and it is in accordance with the approved university Thesis outline.

Signature of the supervisor:…….. Date:

Dr. D. Meedeniya

Signature of the supervisor:…….. Date:

Dr. G. I. U. S. Perera

**Abstract**

Today projects with data analysis play a significant role to give us suggestions to our daily problems. While understanding those analysis data user needs to get the meaning and the nature of the data. Data visualization is the best option to observe the data. The human eye can easily analyze those data with the help of visualization. Moreover when visualizing a dataset better to have an understanding of data types and user intention or preferences. Recommendation systems are the best approach to address the above problem. In this ,study we discuss recommendation application which gets the help of machine leaning and mapping algorithm. Context awareness is a help while giving recommendations to chart types. Even though from users perspective suggestions can be changed. Therefore the proposed solution improves with the help of user's feedbacks. For each test-run system is collecting user feedbacks and use them to improve the training dataset. At the initial stage, there are only a few training data. Users can interact with the system and based on their feedbacks the outcome of the system will get more accurate. Based on user feedbacks recommendation will get more reliable in the long-run. In this study, we are looking at how much accuracy it has in the initial stage and how it varies with the number of test runs in the system. Therefore user interaction plays a significant role to help recommendations. Feedbacks from users help when improving the recommendations. The System recommendations are provided using a combined method of machine learning and rule based components and the evaluation has shown an accuracy over 80%. As this is a trending research area, contribution made through this study can be useful for the industry and the research community.

**Keywords:**

Human-centered computing, Recommender systems, Content awareness, Data Visualization, Information systems applications.

## *Acknowledgements*

I would never have been able to finish my dissertation without the guidance, support and encouragement of numerous people including my mentors, my friends, colleagues and support from my family. At the end of my thesis I would like to thank all those people who made this thesis possible and an unforgettable experience for me.

First and foremost, I would like to express my sincere gratitude to my supervisors Dr. Dulani Meedeniya and Dr. Indika Perera, for the continuous support given for the success of this research both in unseen and unconcealed ways. This would not have been a success without your tremendous mentorship and advice from the beginning. Your wide knowledge and logical way of thinking have been of great source of inspiration for me. You have always extended his helping hands in solving research problems. The in-depth discussions, scholarly supervision and constructive suggestions received from you have broadened my knowledge. I strongly believe that without your guidance, the present work could have not reached this stage.

**Thank you!**

# TABLE OF CONTENTS

# Chapter 1

# INTRODUCTION

## 1.1 OVERVIEW

Today In our daily activities, data plays a significant role when getting decisions in an accurate way. Think of taking the fastest rout when going on a journey. Data will be come to your help in those situations to get it. It will avoid even road traffic data when suggesting the path. Additionally when we use online shopping data will help with suggestions. These are the situations which use recommender systems. [1] Recommendation systems are used to filter the information and predict suggestions according to the user expectation. [2] Main reason is there is information overload. Such situations recommendation system will be the most reliable option. While creating RS, there are many techniques that used to succeed in the process. Like above said with the increase of data amount it becomes hard to analyze them. In that kind of situation will be needed the professional help to examine them. If using an RS, we may not need the help of data science professionals to help us on our decisions. Therefore we can say Recommender systems play a significant role in our day-to-day activities.

Getting the recommendation is not going to enough with the time. The reason behind that is day by day the amount of data elements get increased. Therefore the naked human eye may not able to analyze them by directly looking at those recommendations. In this case, the most proper option will be using a visualization technique. [3] [4] Data visualization is the most suitable approach to distinguish holistic view of data and discovery of data values from big data. In data science, information visualization plays a central role when exploring and analyzing data. Current usage of visualization makes analysis and discovery task easier for viewers.

There are various kinds of visualizations and all of them my not perfect for every data set. In those situations, recommender systems come in handy. But when we look from the users' perspective, the experts may know what that visualization says. Still, the common user may not have any idea about it. Therefore have to select the best one according to the user therefore if there is a way to gather users' feedback and give suggestions according to them that will be more accurate. [5] Here it's better to have an interactive method.

## 1.2 MOTIVATION

When we consider today's improvement of data content, getting a recommendation is an important part. Today we see almost all tasks which needed to human will involve data. They may not be just pure data. Most of them are big data. [6] Therefore need to further process to extract required data. Therefore the mainly the recommendation system will do minimize the human workload while getting an idea related to the data set. Basically, the system will give us the suggestions, and it will save mostly the time of users and also with the time cost also will be deducted. When time waste comes less cost also will be low [7]. These recommendations will help scientist when they analyze data and coming in to conclusions. Not only scientist ordinary people also will have help to continue their day-to-day life.

Those recommended data need to present to the user in a user friendly manor. Otherwise come up with conclusions according to the data will not be a much easy task. Therefore visualization comes to this part to help organize data in to more understandable way. The reason is people can grab more data out of visualization than reading just raw data set. [8] Scientist will come up with their conclusions by looking at suitable visualizations. That's why above recommendation systems are essential. RS can give suggestions for a most suitable recommendation for a given data set according to the data behaviors and patterns [9] [10]. By looking at those data cannot get a complete idea but with the help of visualization, we can minimize the decision making time also. As we see it will be easy to understand visualization, therefore, time to analyze will get deducted.

## 1.3 PROBLEM STATEMENT

Today recommendation systems play a significant part in every activity. Most situations we may need visualization because data itself is harder to understand by only looking at them. Above both parts needed to be combined to get a visualization for a given recommendation. When looking at recommendation systems, there are few problems that need to address. [1] Like cold start problem and also a spatial problem because unrated situations can be found. Additionally, with the growing information, there should be a way to handle even them also.

Contemporary data scope is very high and diverse, making it harder to analyze data, to take data-based decisions for ordinary folk. Thus, professional supports has to be sought to filter noise and analyze essential information. Therefore, data visualization is the most appropriate approach to distinguish a holistic view of data and discovery of data values from big data. But still, visualization may not perfect. Most visualization will only understand by the professionals related to the selected area [4] [11]. Meanwhile, common folk does not realize recommendations about visualization since

mainstream studies are mainly limited to case studies while visualizing is only focused by experts on the area.

When we look in to the above situation, have to consider every user perspective, and to that, there should be a way to gather feedbacks from the users of the system [12]. All of these reasons this area has become a more challenging research area.

## 1.4 OBJECTIVES

A major problem when it comes to this kind of recommendation systems is they won't give out the quality output visualization. We know that most tools have limited their visualization for a specific method. All data types cannot be visualized using the same way when considering that there should be a verity. Current tools provide visualization only for the specially selected method. The output should be according to the context provided. According to provided dataset should match those visualizations accurately. Furthermore should include a way to give recommendations to the user. The method should consist of an interactive way to get user feedbacks to improve recommendation.

The objective of this study is to develop a generalized platform to:

- Implement a prototype that recommends visualizations for datasets
    1. Recommend visualizations according to the context of provided data
    2. Recommend algorithms for visualization recommendation of different datasets
    3. Increase the quality measures of recommendations for data visualization with user feedbacks.

## 1.5 RESEARCH QUESTION

Core Research Question: How to get chart based visualization for a given data set that satisfied the needs of the current user?

Research hypothesis: Current trends of emergent and changing requirements for recommender systems and data visualization can be better supported by:

- Mapping the visualization types with provided data set according to patterns in the data.
- Improve the mapping according to user feedbacks.

## 1.6 EXPECTED OUTCOME

Successful completion of this research will lead towards a tool which will provide visualization recommendation for a context of a given dataset according to input data type and also according to behaviors and data patterns. Therefore the analysis party can analyze the data by looking at those visualizations rather than looking raw data itself. Plus point is user can provide feedbacks for a given visualization recommendation. With the help of user feedbacks also the recommendation quality will improve more.

Therefore outcome of the research will be increasing the recommendation quality of the visualizations. Thus not only experts but also common folk can get an idea or extract information from those visualizations. We know that they are very understandable.

## 1.7 SCOPE

This research manly focused on visualizing different type of datasets. As we know visualization makes easier to understand data rather than looking at raw data. Here we have consider different visualization chart types to visualize different data sets. While giving visualization recommendations to given data sets we have used machine learning and rule-base combined hybrid method. With user feeds we improve our training data set. There are limitation when in the initial stage because there are less training data but rule-base help to overcome that.

As outcome of the research we have built a system to give visualization to given data sets. Here we have only consider the csv data types. In future we can improve it to more data types. Moreover we have built a admin component which able to add new chart types and more rules to the rule-base.

In this study we have used Decision tree algorithm as classification method. We can use neural network deep learning method to improve the system. But it may effect to the processing time because it needs more processing power than our method. This thesis will discuss about the research outcome.

# Chapter 2

# LITERATURE REVIEW

## 2.1 INTRODUCTION

The Literature review will summarize the important parts which have done by the others that related to this research. This section will describe existing research solutions related to Context aware recommendation and data visualization. Mainly in this section will address about Recommendation systems, Interactive machine learning, and data visualization related researches. Here will discuss more currently using techniques and tool related to each sub area. Furthermore will focus on limitations in those tools and techniques and ways to prevent them.

## 2.2 DATA SETS

Data set is a collection of data which can access entities individually, by groups or whole at once. Data sets can be ordered or organized according to a specific structure like you can see in databases. The database itself also can be taken as a data set. Within the database, it can be containing varies kinds of information in each table. For example, one may contain employee data one may have sales data, etc.[13].

If we talked about point datasets, it contains continuous data collections and data which can be represented as point wise. Areal data will be related to a special pattern according to a given region.

When we consider about recommender systems, there have taken the entertainment and e-commerce industries data in there processing. Amazon, Netflix, and Spotify are few good examples for this situation [14].

### 2.2.1 Point Data Sets

Point data sets are the data which can be represented using points. It can be plotted in a 3D or 2D graphs which axes are x, y, and z.

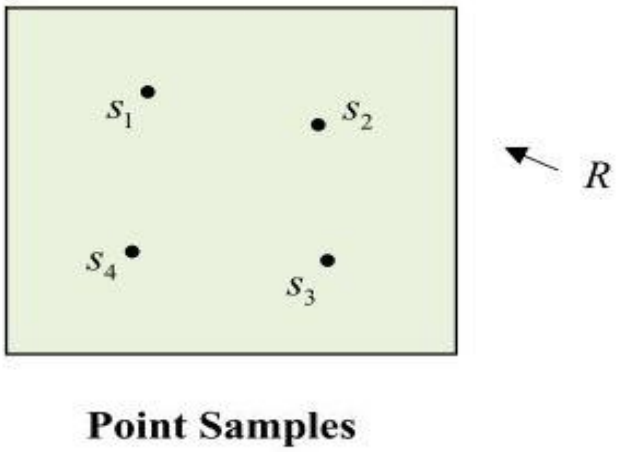


Figure 2.1: Point Samples

Here Figure 2.1 you can see four sample points, are, in the region, R [15].

### 2.2.2 Aerial Data Set

Areal data consider the qualities of each sub areas in the given full area of map.[15].

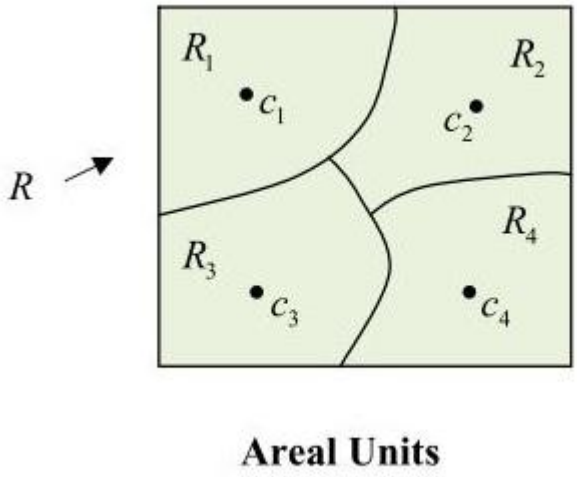


Figure 2.2: Areal Units

Here the major area divided in to for sub units {R1, R2, R3, R4}. Such areal units, Ri, will show by the center of each unit, ci is belongs to Ri , and they cnd be city names etc.[15].

### 2.2.3 Discrete Data

Discrete meaning is distinct or separate. Therefor discrete data means quantitative data which rely on the count. It only contains finite values, and sub-division values are not possible. It includes only those values that can only be counted in whole numbers or integers and are separate which means the data cannot be broken down into fraction or decimal [16].

**For example,** the Number of computers in the lab, Number of students in the school

### 2.2.4 Continuous Data

Continuous data is described as an unbroken set of observations; that can be measured on a scale. It can take any numeric value, within a finite or infinite range of possible value. Statistically, range refers to the difference between highest and lowest observation. The continuous data can be broken down into fractions and decimal, i.e., it can be meaningfully subdivided into smaller parts according to the measurement precision [16].

**For Example,** Age, height or weight of a person, the time is taken to complete a task, temperature, time, money

### 2.2.5 Example Data sets

Following will be selected example data sets related to RS [14].

- MovieLens – MovieLens web site provide us their datasets related to ratings. They are good for a start.
- Jester – Is good for creating a collabarative filter and it will be a simple one. Also it has 4.1 Million reatings which are continuous.
- Million Song Dataset – This data set is related to music recommendation. Here it is good to begin with collobarative filtering and later you can join with content based filtering also.

## 2.3 DATA PRE-PROCESSING

Data pre-processing is a necessary and critical step in data mining. Through pre-processing, it can convert the dataset in to accurate, reliable and quality data. When creating data warehouses and WWW also used preprocessing not only in data mining. [17] Data preprocessing is a base of association rules, classification, prediction, Pattern evaluation, Knowledge presentation for the step of data mining. Moreover data pre-processing contains data cleaning. Data preprocessing is a base of association rules, classification, prediction, Pattern evaluation, and Knowledge presentation for the step of data mining [18]. There are imputation and embedded methods to fill out the missing

values. Filtering will be used to eliminate those missing values prior to the imputation and integrated methods. Sampling or algorithm modification method will be used to handle those imbalanced data.

### 2.3.1 Data Cleaning

In data pre-processing one of the main processes is data cleaning. Clearing is a must because when creating a warehouse and in data mining real world data is not clean. Data can be very dirty with noisy data, missing values in a tuple and also inconsistency. If we apply mining techniques without any pre-processing outputs will not be reliable and also will be poor outcomes.

- Filling missing values
- Smooth noisy Data
- Duplicate removal

**Filling missing values**

If any missing values in the given data set there are few techniques which can be used in that situation [17]**.** The first technique is ignoring the tuple. If any tuple in the given dataset has blank values, then that tuple will be ignored.

Another way is filling the values using a manual method. Or can be used a constant, mean or a most repeated value to fill those blanks.

**Smooth noisy Data**

In a given data set it contains noisy data [17]. With those data, the further process won't be easy to do and will have a poor outcome. That kind of situation we can use binning, regression and clustering as techniques to overcome the problem.

In the binning method, data will be distributed in to bins. Then values related to the bins will be replaced with the each bin mean. This method called as smooth by bin means. Another method will be a regression. In this method the best line will find out using two relevant attributes and using that line will be able to predict the others.

Binning and regression will be only used to numeric data. But clustering can use not only for numeric. This method consided the data nature to group them. A here a similar type of data will be grouped together. Non related data will be organized outside those given clusters. Quality of the method will depend on heights distance of two objects in the cluster. Any two objects can be take.

**Duplicate removal**

If we consider the dataset with the duplicate values, it also will effect to the final recommendations. When considering that for an accurate recommendation we should have to remove those duplicates. First, we need to identify exact duplicates. Even some times the values look duplicate with the timestamp it can be a different record. While removing duplicates, we have to aware of those things also.

Then if we discuss further more [19] have talked about cleaning large datasets. In the document, they talked about how data comes with error and reasons for that. Data entry errors, measurement errors, distillation errors and data integration errors are few of them. Following will be data cleaning teachings broken down by data type.

*Quantitative data:* Quantitative data are integers or floating point numbers which measure quantities of interests. It can be multi-dimensional complex arrays which gathered data respect to the time. Basically, the outlier detection is in this data type.

*Categorical data:* It has names or codes which had assigned to those categories. Here we don't need to look for outliers. They are in the distance to those categories. Here mainly do map those category names with the uniform namespaces. Another thing is looking for miscategorization and add to the correct category. Furthermore misspelled ones need to handle if not another different category will consider for those data.

*Postal address:* When considering postal addresses we are mainly looking for the structure of the data and redundancy or ambiguity. Additionally, need to identify duplicates which may have different spellings, etc. Moreover, there can be two persons from the same address; those things need to be identified.

*Identifiers:* Identifiers or keys are special cases which gives unique identity for each property. Basically, those structured keys give out specific data related to the property. That need to identify. For example our Identity card, Passport numbers or telephone numbers. There are many things we can grab. Like from a telephone number we can tell country which province, district, etc. From Identity card numbers in Sri Lanka can say owner's birthday.

### 2.3.2 Data Transformation

Data inputs to the system can be different types or different format or even in different languages. These data need to be transformed into a common schema. It will be easier to perform actions to a specific schema rather than different formats. If this process applied in a data mining or later process the operations might get slower. The reason is many calculations will happen in those stages.

Aggregation, generalization, and normalization will be used as techniques for data transformation [17].

In Aggregation will calculate a summery related to the given attributes. In generalization, all low level data will be replaced with the high level data. If we moved to normalization, it would consider about specific range. Min-max normalization, z-score normalization, and decimal scaling are used in normalization technique.

### 2.3.3 Data Reduction

Data input can become larger because of every day collection. If data mining applied for this larger data set, it probably takes a long time to process. Therefore the outputs of the data mining will be more complex and before further operations will use data reduction. Following are the techniques used [18].

- Supervised Discretization – This method using class information and their split points with calculations.
- Unsupervised Discretization – Not using class information. Attributed value is divided into a fixed partition.
- Splitting – This use recursive way to find the best neighbour to the given points.
- Cluster Analysis - This technique applied to discretize a numerical attribute. Concept hierarchy can be define using this.

## 2.4 RECOMMENDER SYSTEMS

Recommender systems are a combination of software tools and techniques to provide suggestions according to user expectations [20]. Currently, recommender systems have become a more popular tool in daily life. When users are searching books, movies, news, and articles, recommendation systems will give suggestions to them according to their need. Those suggestions should satisfy the user. Suggestions will be related to various domains such as which book to read which music to listen or which item to buy.

Recommendations will do according to selected algorithms such as collaborative filtering and content-based filtering. Furthermore consider when we plan a journey, according to road traffic recommender systems will suggest us different routes which will be faster. Suggestions by RS should be useful to the user to take his decisions easily, and they need to satisfy the user.

### 2.4.1 Importance of Recommender Systems

When giving recommendations, RS needs to overcome challenges. That's why RS is an ongoing research area. First of all, when giving recommendations [20], RS will have to handle and manipulate large data sets. Therefore need good techniques to process them. Furthermore while handling larger data sets system can be getting slower. That also a challenging area to look at. To improve the processing, we can enable ways to processes offline, but still recommendation quality is relatively poor. If there are fewer data to give recommendations, still the RS outcome won't be accurate. Less data means suggestions also get in limited range. Therefore the recommendations that give with fewer data set will be poor [21]. Furthermore, privacy concerns need to address. Some user information provided by the user can contain privet information. That is also an issue.

### 2.4.2 Content awareness

Context awareness is the ability to understand the data in the specific file, folder data store or application. [22] That means should be able to understand those data in rest, use or transit. In the current day with almost all operations related to data, this content awareness is important fact. It's important to identify the exact datathe before starts the processing. Today to achieve and improve this contents awareness there are many techniques used in applications. For that most times will be used one or more techniques or mechanisms from following.

- Exact data matching
- Structured "fingerprinting" of data
- Statistical analysis
- Rules and regular expression matching
- Conceptual definitions
- Keyword and file-tagging look-up
- Watermark recognition

### 2.4.3 Recommendation Algorithms

Today we can see that many processes use big data in their operations. Therefore there is a need for getting recommendations from those data. Otherwise, need to go through all data to get a single outcome. With the help of a recommender system, it comes very easy to analyze the big data. To get filtered data from a big dataset, there are few algorithms which use. Mainly most RS use collaborative filtering and content-based filtering as their approach. Still using the same algorithm for each and every data set is not accurate. Primarily with different datasets, the algorithm should be changed. To overcome drawbacks in these algorithms mainly considers about hybrid solutions of

those algorithms. They are the areas which needed to be addressed in the future related to this area. Mostly in this chapter will discuss what are the available algorithms and their limitations.

#### 2.4.3.1 Content-based filtering

With help of past user likings recommend simila items vis content-based filtering. For example, one user liked to a song by the specific singer. Then the system will provide recommendations from that singer. [23] But when considering a large amount of data had to get summarized data. It may reduce the quality of the data. Recommendation quality will be relatively poor. It's too general. Recommendations should help user to discover new items. [24] Have presented a set of collaborative and content-based approaches for recommending items on social book-marking websites. Here ranked lists using both content-based and collaborative filtering and merge these results to get the final list by using adaptively weighted averages.

Even though most users not rating all the items, Therefore that becomes a problem when rating. When added new items there are no ratings for them as well. Therefore still there are some limitations.

Collaborative method data can be input even explicitly [7]. It will be like suggestion provide by a friend that already knows you well. Here in Content-based systems there are lesser ratings to address user needs. For that here they have used hybrid approaches. Using content-based recommendations can solve the problems of the unseen items by others. When using collaborative method we can even give recommendations to users who haven rated. [25] There are many suggestions based on both content-based and collaborative methods to implement hybrid methods. They are weighted based on both method, Mixed Hybrid, and Cross-Source Hybrid. Weighted method there are different weights for both content-based and collaborative methods. In Cross-Source Hybrid first applied, then each recommendation's weight and sources appeared will consider when giving recommendations.

#### 2.4.3.2 Collaborative filtering

This filtering method will consider the similar type of users to give recommendations to the active user. For example,If the active user has a taste for comedy movies. There is another user who has a taste in comedy films in the past. That user's other likings will provide as recommendations to the active user.

If we looked into further, we would be able to find several types of collaborative filtering. User-user collaborative filtering and item-item collaborative filtering will be them.

User-user collaborative filtering: Here mainly lookalike users to the live user and according to them the recommendations will be provided. This type is accurate, but it needs to compute every customer pair information. Therefore this method takes a lot of time.

Item-item collaborative filtering: In this approach will look for items look alike. Will not look customers look alike. When find a look alike item using the matrix we can recommend that item to the customer. This method needs fewer resources than the previous method. Moreover, it takes less time for a new user than in user-user collaborative filtering.

Pure collaborative filtering recommenders only give suggestions based on user rating matrix. All the users and items will consider in an atomic level in those methods. [24] [23] Uses a combination of collaborative filtering, content based and cluster models. [26] Hulu's recommendation system is primarily based on ItemCF. They have added many improvements to the system on top of the ItemCF algorithm in order to make it generate better recommendations. It is suitable for sites where there are a lot more users than items. Easily explain recommendations given users' historical behaviors. Not only those can every user behavior reflect user preferences and more than implicit data, explicit data is important. Furthermore, recent behaviors are much more important than old behaviors. All the items are not rated and need to have a way to rate newly added. [7] Collaborative systems help to improve user data with help of other user's interaction data. Content-based systems solve all of its suggestions based on pure collaborative. Using content-based recommendations can solve the problems of the unseen items by others. By combining both can overcome limitations in both techniques. [27] Have used Automated Collaborative Filtering to their recommendations. It will use neighboring friends to give recommendations. But when friend counts getting low, the accuracy also will be low. [28] Furthermore discuss recommendations of collaborative filtering is manly focus nearest neighbor. Such methods provide predictions based on calculations.

#### 2.4.3.3 Demographic

This type of system recommends items based on the demographic profile of the user. For example, take Sri Lankan persons may like to visit and see cricket related websites. According to the demography when a new user comes from the specific niche will get recommended which related to that niche. [7] Similar people tend to behave in a similar way. Demographic filtering systems use the general features of a cluster of similar people or a stereotype of a person to infer the interests of a particular user. [20] Provide recommendations on social bookmarking websites. While providing recommendation demography niches will be helpful even though it's the more generalized approach. All the users in the niche may not know alike.

[7] Demographic filtering will learn from the description related to each individual and will learn relationships between the items respect to those descriptions.

#### 2.4.3.4 Knowledge-based

As Knowledge-based systems will suggest recommendations based on pre collected knowledge and it will consider the usefulness for the corresponding user with those recommendations [20]. According to user requirements, these systems will select the perfect match. First, need to create the knowledge base according to the selected area. Before started need to gather information about the selected area and need to research about knowledge related to that area. Otherwise, the final output will be wrong. [29] Then need to define the minimum knowledge that you are going to need while creating the knowledge-base. Then need to describe how contents are going to be grouped, how to rate them and also how they will change with respect to the time. Then need to plan which will be the framework going to use and etc. With the knowledge, you gather to create the knowledge base and have to take steps while creating to decrease waiting time and improve the quality of the outcome. Finally, need to look on continuous improvement related to knowledge-base. There will be stepping while building knowledge-base.

Recommendations will be taken according to the user profile only. Other users' preferences or profile won't be the effect on the final result. [30] When using the knowledge-based technique it can be done even with lesser data also. Even with the cold start, it will be work fine. Not only with that, but has a greedy sheep problem also handled in the knowledge-based recommendation. The limitation will while considering this method before we do process we have to gather knowledge and need to create a knowledge base. It is a usually complicated task.

#### 2.4.3.5 Community-based

This method mainly consider abut perspective of user's friends [20]. These systems will look for users' friends and their preferences. According to those preferences recommendations will be provided. The community can be defined as a social network which characteristics of people in the same community basically same. Therefore we can use matrixes to calculate similarities. [31] Paper they have used community based methods for scholar recommendation in academic, social network sites. It will address even the cold start problem. [2] Additionally have used community based methods in their recommendation implementation to give support to sales staff with registered documents and included articles. Here they consider community as the same department of the same company. But the user can be disliked to those his community like that can be a problem.

#### 2.4.3.6 Cluster Models

Cluster Models will find the similar customers related to the live user [23]. The model automatically categorize the user in to a specific cluster based on its features. Purchacing and rating based on users will be use when giving the specific suggestuions.

#### 2.4.3.7 Rating method

Two sampling methods using here to came up with solutions [21]. They are prize and quiz datasets. Test data will be used to evaluate performance. Randomly selected users and there ratings will be taken as a qualifying dataset. To assign a data set users most recent data and prize data will be selected. There are situations that ratings are lesser than 18 then the most recent one-half of their ratins will be selected as assign data set.

#### 2.4.3.8 Hybrid systems

Currently many types of research we see that hybrid approaches. Mainly and content-based filtering methods will combine to approach hybrid methods. In hybrid systems, recommendations for both types will provide separately. Then combine those two recommendations to get the final recommendation.

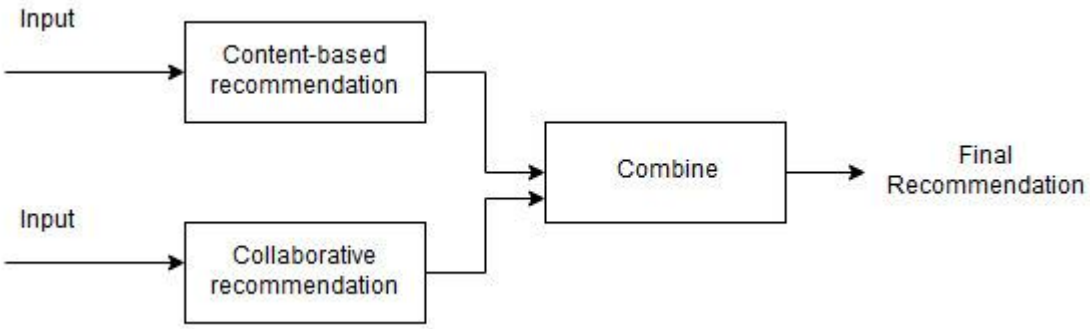


Figure 2.3: Hybrid recommendation model

Have suggested Hybrid methods with content-based filtering and collaborative filtering methods [25]. They are Weighted Hybrid, Mixed Hybrid, and Cross-Source Hybrid. Weighted hybrid is using different weights of these content-based and a collaborative methods. In Cross-Source Hybrid first applied, then each recommendation's weight is multiplied by the number of sources in which it appeared.

But still, there are common problems like systems such as cold start and the data paucity problem.

### 2.4.4 Existing recommender systems

- Amazon.com recommendations [23]
- Hulu's recommendation system [26]
- Netflix Prize [21]
- A Taxonomy of Personalized Agents on the Internet [7]
- SmallWorlds visual interactive system [27]
- Interactive hybrid recommendation system [25]
- VizRec Recommender [32]

**Amazon.com's recommendations**

In Amazon.com's recommendations they have used item-to-item based collaborative filtering [23]. It won't be matching similar customers but will match each of these rated items by the user to similar items. Combination of those similar items with recommendation list will be next step.

Based on customers purchasing item algorithm will provide item matches. They didn't include a product-to-product algorithm because many products didn't have any common customers. Additionally, the used algorithm is efficient in both processing time and memory usage. When looking at processing time, its intensive which O ($N^2M$) will be the worst case and always will be close to O (NM). Sample customers with best-selling also reduce the time.

After algorithm finds items similar to the customer according to ratings and purchasing, then aggregate those and give the most popular or correlated item as a recommendation. This calculation will be fast. Ratings or purchasing number will effect on calculations.

Even though there are few drawbacks, When it performing off-line recommendations qulity will get low. Then will fail to provide recommendations with interesting, targeted titles.

**Netflix Prize**

Netflix provided there large dataset for computer science community to improve the data mining, machine learning to beat the Cinematch tools performance by certain amount. For that Netflix gave over 100 million ratings from over 480 thousand randomly-chosens, Subscriptions by the anonimus users on eighteen thousond movies. The company is to provide a prize to the team which develops the tool with 10% more accurate than Cinematch. But most of them were only able to do 5% [2].

**Hulu's recommendation system**

It was difficult for Hulu's users to find the best matching videos according to their historical interests. The goal of the Hulu's Recommender system is to find the best interest videos for their users. Here they have used collaborative filtering as the technique to give recommendations [26].

The system has two architectures related to offline and online. If we consider about online architecture first, it will build a profile for a user when the user comes for the first time. With the time this user profile will contain users' historical behaviors of the users. To generate preferences on topics, those behaviors will be considered. Then pull all similar to raw recommendations and then filtering will do because recommended results cannot give directly to users. Then the re-ranking will do. After users find out shows (videos) which they like but they haven't seen before. Finally, the explanation part will create rules according to given recommendations to future references.

And the offline method will have different architecture. In data center contain all behavior data and it will contain in side relational database and Hadoop clusters. With the help of collaborative filtering and based on content related tables will generate. Then will group in to similar content. Additionally, use user feedbacks to improve the recommendation quality. Finally, the report generator will give reports according to the given recommendations.

**A Taxonomy of Personalized Agents on the Internet**

First in the system will do the profile generation. When a new user comes initial profile will be generated [7]. This profiles it the one that responsible for tracking user activities, such as user's feedbacks and interests. Not only profile generation, but maintenance is also done by the system. To this process need information from the user. While the user is interacting with the system, he or she will provide interests. If we monitor the behaviors, we can gather habits and knowledge of those users.

Next, need to do the filer the data. In the system, the first data will filter using the demographic method. We can see that demographically similar person behave in similar ways on most occasions. It uses a general feature of the cluster to filter the users. Then the content-based filtering will use to get recommendations according to the past likings of the user. Then collaborative filtering will use to match people with similar interests to give recommendations.

Still there some drawbacks like still system need much effort from the user's end. People can not specify their interests because sometimes unconsciously takes some decisions.

**SmallWorlds visual interactive system**

SmallWorlds is an interactive visual tool which deployed with limited data set from facebook [27] API. The system will give the recommendations according to the profile data. Tis system also uses a Hybrid approach of both content-based filtering and collaborative filtering. Moreover,

recommendations will be improved by using data of the friends. It will be an effect on the decisions made by the system. From the system, they have found that users with more than 200 users have more satisfaction than users with less than 50 friends. With the help of friends' data also the recommendation can make accurate.

Problems of the system will occur with a lower number of Friends. If friends are lower the accuracy of the output will get low. Furthermore, the facebook API only limits access to the user profile. That also will be a draw back.

**Interactive hybrid recommendation system**

First in the system will initialize a profile music preferences using the facebook API [25]. If using the Wikipedia model will Google search APIs, and those music preferences will map. For each search, the top item will be selected. Then the content-matching algorithms will use to generate recommendations. Overall weight will be calculated with the current search and with compared to the previous searching.

If the facebook model used, it will look into the preferences of the friends of the user also. It will use the way that similar to collaborative filtering. It usually gives predictions according to similar users are liked previously. Furthermore will consider the friends of the user in the process.

There are few interaction methods also. They will be profile interaction, source interaction and full interaction. Profile interaction only can view and fine-tune the weights. In source interaction, addition to profile interaction will able to change the weights of the sources. In full interaction, the user could see the turning actions on the recommendations.

By using those modes will generate weights for each search item. Then by combining all this will provide the final output. Here the user can change his or her profile according to preferences.

Fully interaction method will be helpful to the highest utility source.

In this evaluation demonstrate that by providing a user with an interactive control mechanism and by combining traditional overlap in preference profiles with preexisting social connections make recommendation very rich.

The following point will be very helpful when designing a recommender system.

- Via user interface we can explain easily the hybrid recommendation system.
- When improving user experience and accuracy recommendation time plays major role.
- It will be better than traditional CF when using different API to provide recommendations.

### 2.4.5 Related work in Recommender systems

Related to recommender systems, there are many systems and tools have created. They have been very useful while taking decisions in those applied areas. There is a tool for social bookmarking websites for their recommendations. It uses Collaborative and also the content-based filtering for item recommendation [20]. Amazon.com also uses RS in their recommendations. Their recommendations use Item-to-item collaborative filtering [23]. Another related work is Hulu's RS. Hulu's recommendation system helps content owners promote their video [26]. Netflix also uses RS to give their recommendations to users according to ratings that they have received [21].

Table 2.1 will show a few of the related work and methods related to those work.

Table 2.1: Methods related recommender systems

| | **Content-based** | **Collaborative Filtering** | **Demographic** | **Knowledge-based** | **Community-based** | **Cluster Models** | **Rating Method** |
|---|---|---|---|---|---|---|---|
| [24] | ✓ | ✓ | | | | | |
| [20] | ✓ | ✓ | ✓ | ✓ | ✓ | | |
| [23] | ✓ | ✓ | | | | ✓ | |
| [26] | ✓ | | | | | | |
| [21] | | | | | | | ✓ |
| [7] | ✓ | ✓ | ✓ | | | | |
| [27] | | ✓ | | | | | |
| [25] | ✓ | ✓ | | | | | |
| [33] | | | | | ✓ | | |
| [34] | | ✓ | | | | | |
| [35] | ✓ | ✓ | | | | | |
| [36] | | ✓ | | | | ✓ | |
| [37] | | ✓ | | | | | ✓ |
| [38] | ✓ | ✓ | ✓ | | | | |
| [30] | | | | ✓ | | | |

## 2.5 MACHINE LEARNING FOR INTERACTIVE SYSTEMS

When we consider a human there are lots of was to understand each other. Like they can speak to each other, or they can understand by looking at their facial expressions, etc. Actually, it is a complex task. Even the same thing can be defined in different ways according to time, or the way express things. If looked in to machines they also need a way of understanding data. Need to address the data and identify special behaviors of them.

Today machine learning is a growing technical area which will be connected with data science and knowledge discovery. [39] Have talked about health interactive machine learning. By monitoring human behaviors can be come up with health information. Furthermore by collecting those data after further processing can give accurate answers.

Machine Learning has introduced in Human computer Interaction also. The first approach related to HCI in machine learning will be [40] Hidden Markov Models in Speech Recognition application. But when it comes to Interaction, it will be sequential, Therefore it may be challenging. The user always has a non-deterministic behavior.

There will be two methods which are related to machine learning. One will be supervised learning, and the other is unsupervised learning. Supervised learning will give output for a given input according to the training dataset. A supervised algorithm is learning from the training dataset. Learning will stop when the algorithm comes to an acceptable level of performance.

[41]Un-supervised are not like supervised Algorithms left their alone to discover and present patterns and interesting data. Additionally, there are semi-supervised methods. Here un-supervised can use to discover knowledge and supervised method to get the best guess for those unlabeled data.

But still there challenges and this is an ongoing research area.

### 2.5.1 Supervised learning

As Supervised learning is a machine learning task which used to create functions for new data inputs by using a training dataset. In those training data sets, there will be outputs for every combination provided. After analyzing that supervised learning will derive an algorithm which can be used in a new dataset to map with the outputs. Furthermore, this algorithm should identify those unseen possibilities also.

[42] First step in the process will be collecting the datasets. If experts available get their suggestions related to those data. If not use to measure everything and informative features can be gathered. If the data were pre-processed data mining algorithms will operate faster and also very effectively.

Here can create new features by using the existing features. These new features will lead to a more accurate algorithm.

There are two sub categories in supervised learning.

- **Classification**

  Classification will be used when output is a category. For example colors like “Blue” or “Green.”

- **Regression**

  Regression will be used if the expecting data is a real world value. For example: “Height”.

**Classification**

Classification is a process which is a function which will assign each item in a data collection in to a specific class or category.[43] The task of classification is accurately defined a class for each data provided. Classifications do not have an order and here mainly use categorical data. The simplest way of classification is binary classification. It will only have two classes to classify — for example, higher income people with more than 100 000 rupees. Less than that will be lower income people.

**Regression**

Regression is a process which used to predict the numbers.[44] Sales, distance, height, and weight will predict according to this. This will begin with known target values. Linear regression can be taken as an example of this. In this function, there will be a value for each input is given. This was tested with various statistical methods with actual and the predicted values.

Few supervised machine learning algorithms:

- In regression problems mostly use linear regression.
- Mainly for classification and regression Random forest will be used.
- In classification problems SVM (Support vector machine) will be used.

### 2.5.2 Unsupervised learning

In unsupervised learning is to derive functions using unlabeled data. If the provided data is not labeled cannot do analyze or pattern prediction. Moreover cannot evaluate the accuracy of those structures. That’s how unsupervised different from another method. For that kind of situation unsupervised learning will be used.

In unsupervised learning will follow the following steps [45]. First will get random patches from the unlabeled data. Then if the data cleared the next step will be easier. If data is not cleared need to clear first. Then will learn features related to data using the unsupervised learning.

Unsupervised learning can more divide in to clustering and association.

- **Clustering**: In clustering will identify the inherit groups of data with respect to the behaviors.
- **Association**: Will use to discover the rules related to large part of the data. For example, people buy tea also tend to buy milk.

**Clustering**

In clustering will find data elements similar to each other. [46] Unlike in classification here clustering the data which haven't previously defined. Clustering is very useful while exploring data.

Clustering will be helpful when detecting anomalies when most data have similarities if there can be found data with differences that also will be detected. This can be taken as anomalies or outliers.

**Associative**

Association is predicting the probability of co-occurrences. [47] For example it will monitor customers buying another item with some specified item. For example, the customer always tends to buy shampoo with soap.

This can be applied for many domains as well. In e-commerce, this has been used vastly. While presenting data on a web page also use this. For people visit a specific page also visit another page. Likewise, we can come up with many predictions.

Following algorithms are for unsupervised learning:

- Clustering problems used k-means.
- In association apriori algorithm used to rule learning.

### 2.5.3 Semi-supervised learning

In semi-supervised learning use, both supervised unsupervised learning methods. For the labeled data supervised learning will be applied and for those unlabeled data unsupervised method will apply. In a way, this will be a hybrid process.

### 2.5.4 Related Work on interactive machine learning

Table 2.2: Related Work on interactive machine learning

| System Name | Method | Benefits/ Conclusion | Limitations | Reference |
|---|---|---|---|---|
| JAABA: The interactive machine learning for automatic annotation of animal behavior | Labeling behaviors by looking at videos. Then based on labeled automatically classify behaviors in other videos. The interactive user interface to visualize to non-machine learning user. | Softwares freely available for windows, mac, and Linux.<br>JAABA obtained less error rate.<br>Wide variety of data can be consider because of interactive method.<br>Convert high dimensional data in to understandable data. | The complex output will be provided. Therefore only will useful to experts. | [48] |
| ReGroup: The interactive machine learning for on-demand group creation in social networks | When new person added to group learn probabilistic modal related to group membership in that group. Then will suggest additional members and group characteristic for filtering. | Gropes with unlearnable will addressed.<br>From end user mainly need to create groups more than training classifier. | Still, have to be improved when addressing missing values. | [5] |
| Deploying an Interactive Machine Learning System in an Evidence-Based Practice Center: abstrackr | Use active learning and dual supervision to this semi-automated citation screening. Keep track of labeled. | Useful as a collaborative annotation tool, independent of its machine learning components.<br>Modal is interactively trained with the help of experts. | Will deliver to domain experts. Therefore output may be some messy. | [49] |
| Interactive Machine Learning System for Automated Annotation of information in text | Start with not annotated data which what is to be learned.<br>Iterative, interactive user sessions will train annotators.<br>Trained annotators will use to discover more annotations in text. | Annotations that receive a high confidence level can be automatically accepted, and those With low confidence levels can be automatically | There are cost and time limitations to the amount of text data people can annotate. | [50] |

| | | | | |
|---|---|---|---|---|
| | Provided a convenient and efficient interface Therefore that the context of use can be verified if necessary in order to evaluate the annotations and correct them. | rejected. Provided a convenient and efficient interface Therefore that the context of use can be verified. | | |
| SmallWorlds: Generates accurate and useful item recommendations based on a combination of the Facebook user profile. | Predictions will be based on automated collaborative filtering. Neighborhood user based approach. Will look for user similarities.<br>Based on user interest will define ratings. | Via visual componants user will understand the macanism of ACF algorithm. | Facebook API limits the access to the user profiles.<br>With a lower number of friends, richness recommendations will also go low.<br>When measuring Algorithm accuracy can't adupt the larg scale data. | [27] |
| EnsembleMatrix: Allows users to interactively manipulate visualizations of the output of a multiclass classifier system | Combine multiple component classifiers. Confusion<br>Human observers identify patterns of misclassification. For exploration will use Partitioning and linear combination operations. | People frequently can find good solutions without doing an exhaustive search. Enhance machine learning with insights gleaned from human problem solving. | With fewer participants, accuracy will be low.<br>The participant should provide correct answers. | [51] |
| TasteWeights: Hybrid music recommendation system with an interactive interface | Users can understand the ongoing hybrid algorithm very easily. | The interface can increase user satisfaction.<br>Interaction improves recommendation and user experience.<br>Because of the hybrid method, it's better than CF. | Because real-time processing will take time.<br>Social media APIs limits data. | [25] |
| Interactive Machine Learning for Health Informatics | System learns from the data and imrove with more interactions. When the doctor in the loop which provides feedbacks and also | Doctor in the loop improves the accuracy of output.<br>Interactive learning improves satisfaction. | The performance will be low in larger datasets.<br>Multi-tasking like learning problem together with | [39] |

| | | | | |
|---|---|---|---|---|
| | will help to improve more. | | multiple related sources not possible. | |

## 2.6 INFORMATION VISUALIZATION

If we consider our daily events, they give us information. Our bank statements, newspapers, emails that we receive and convocations we do with others give us information [10]. Now computation gives us the luxury to store that information and retrieve them. This information always leads us when we take most of our decisions.

Visualization is the way that gives that information to the user in an understandable way [52]. Here it will be easier to understand and also to manipulate. [32] Day by day volume of the data will get increased. [53] [3] Visualization is a way to deal with data overload problem and presenting it to the user in an understandable way.

When considering those reasons this area has become a research area which needed to dig more [11]. Therefore people always try to build new techniques or tools to simplify this area. Many techniques have built in reason past, Therefore it becomes a vast area which needed to grow more.

When considering these systems related to information visualization still have challenges. Example user friendliness, address users problem and also finding universal technique [8]. If we consider those reasons, still this area becomes more and more necessity of researching.

### 2.6.1 Information structures

Information structures are the way of giving data in a sentence in a structural way. According to the variations of the sentences it will be having different structures [54]. There are different types; then there should be ways to identify them as well [55]. There will be specific patterns according to each structure Therefore when extracting data from it have to consider according to them. There are a lot different in each structure and still some types hard to extract data.

#### 2.6.1.1 Tabular structures

People use many ways when they are communicating like printed documents, spreadsheets and hand wrote documents, etc. [54]. Tabular structure or tables can be taken as a good method when using the above methods because we can categorize and easily look at data that we needed. [56] Tables have both logical and physical views. Physical structure will define headers, rows, columns, and cells, etc. The logical structure will say how cells in the table are going to connect with each other.

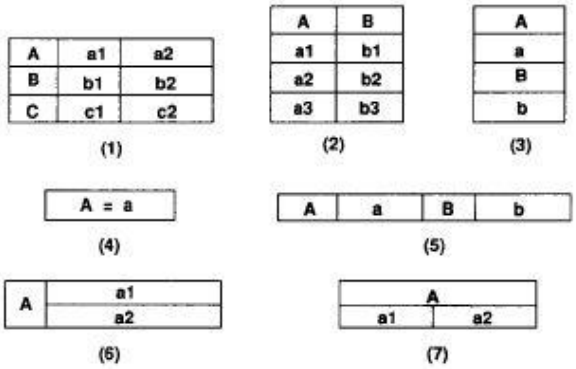


Figure 2.4: Tabular structures

From the above figure we can see those recognizable tabular structures [57]. According to structure information retrieval can be changed.

**Spatial structure**

Spatial structures mainly consider the data which have geometric coordinates. [58] Space/spatial structures refer to a structure made of an assemblage of linear members interconnected to each other in space [59]. There are many areas related to computer science that we can find these data. Geographic Information Systems is one area which can be taken in to those areas [60]. Specific ways to identify and process those data have been implemented. But still, we can see researches are ongoing.

**Temporal structure**

Temporal data is the data which collected respect to the time. It will represent point data in time sampled at a specific period of times. [61] These data will be stored in a single attribute data which can use the timeline to visualize them. Respect to the time need to identify variations in those temporal data, and we can come into conclusions according to them.

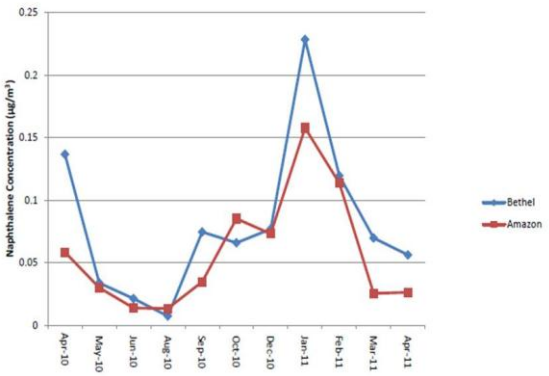


Figure 2.5: Temporal structure for Naptholine consantraction

Figure 2.5 you can see example related to temporal data. It gives naphthalene concentration in bethel and Amazon [62].

**Tree structure**

The tree structure is a method of placing data in a hierarchical manner. Division points will call as nodes. From nodes, there will be spreading branches. [63] [10] With more and more child nodes tree will be gigantic. The starting point will call as root, and a maximum number of children will call as the order of the tree. A number of levels of the tree will call as depth.

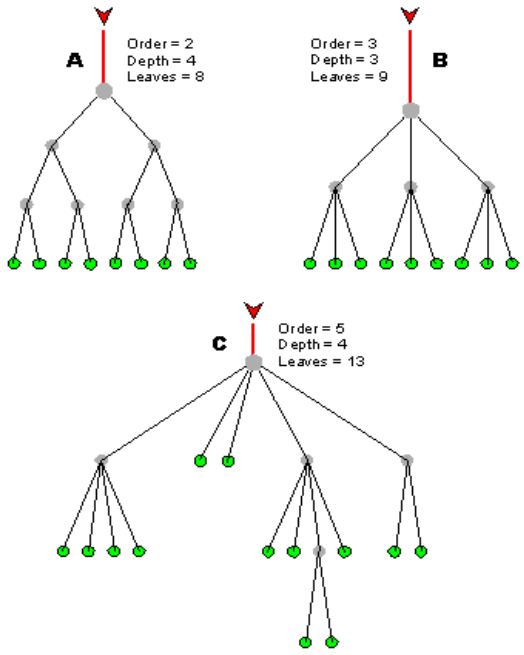


Figure 2.6: A Tree structure Example

Figure 2.6 shows us a few examples related to the tree structure.

**Network structure**

Above tree structure have a hierarchical structure. But in network structure, it is an interconnected node. There will be no hierarchical way it can be connected to any node. Network graphs can be shown as in Figure 2.7 [64].

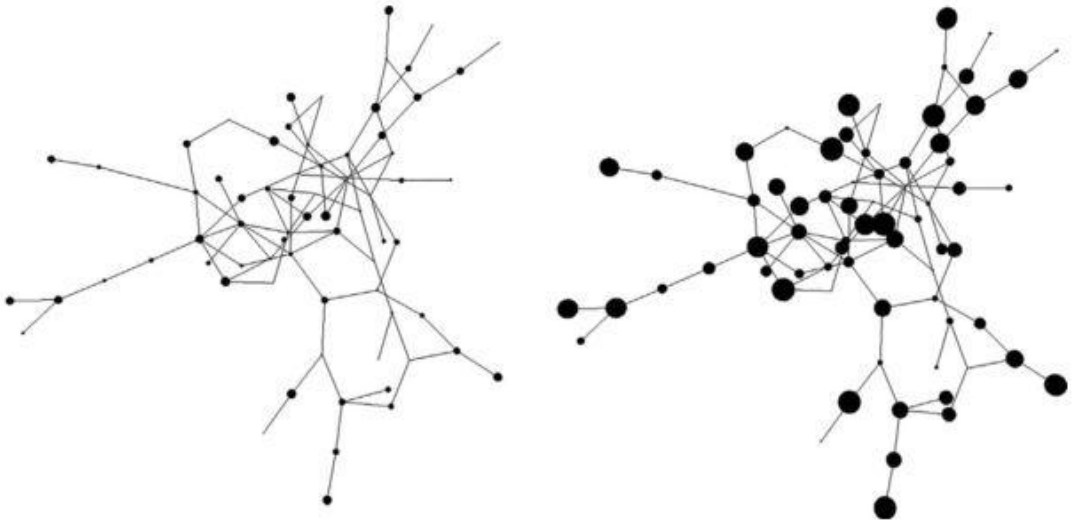

Figure 2.7: Network structure Sample

This will be the visualization of a network data structure.

### 2.6.2 Data Visualization strategies

With an increment of big data and storage capacity task of data analyzing had become more and more difficult and challenging. Not only had the amount of data, but complexity also increased. Therefore our naked eye can identify and analyze those data easily because of that visualization is developed. By using visualization there are two advantages [65].

- Get huge amount of data into a graph.
- Provide an easy way to analyze the data.

They will be in an easily understandable format. This also a trending area therefore many types of research going on this area to find the most effective visualization for a given dataset [32] [53] [27].

**Scatter plots**

Scatter plot is using to show the relationship between two sets of data. This will be similar to a line graph. But rather than a line, it plots data points in x and y coordinates. In here the diagram will show how much one variable effect by another variable. Those relationships call as correlation.

There are positive and negative correlations. If the plots are making lesser angle than a right angle with the x axis that relationship is a positively correlated relationship. If the plots making lager angle than a right angle with x axis, it will be a negative correlation. Figure 2.8 shows perfect negative and positive correlated graphs [66].

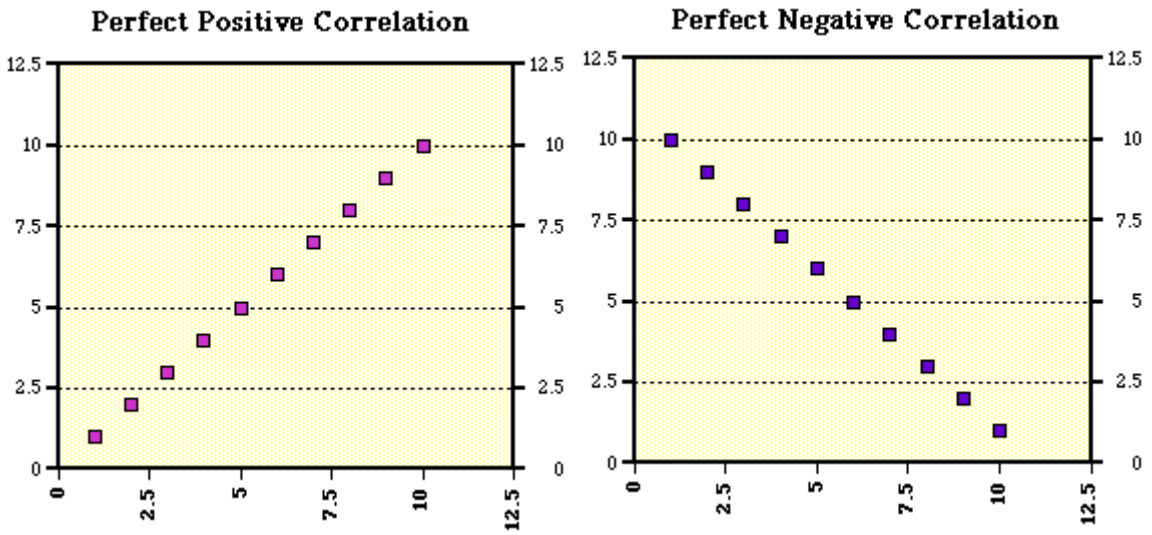


Figure 2.8: Perfectly correlated graphs

But the plots always will not be in a straight line. Most times it will be not perfect.

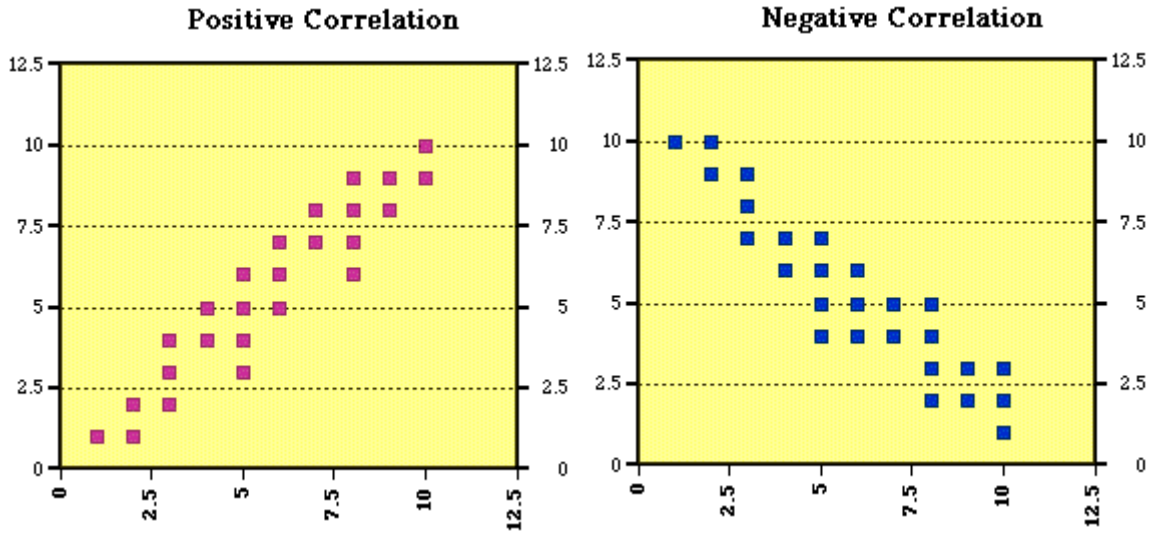


Figure 2.9: Not Perfectly correlated graphs

There will be no correlated graphs also.

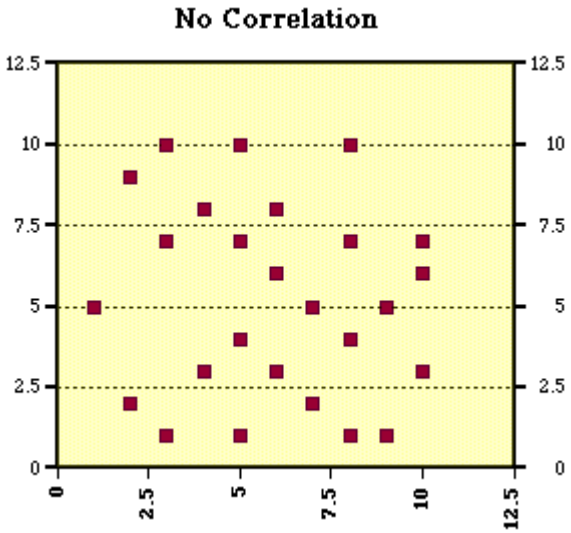


Figure 2.10: No correlated graphs

**Line chart**

Line chart commonly uses to visualize some value changing over time. There are x and y axis to represent all the data. X axis is independent because most times represent the time. In most cases graph deal only with positive numbers [67]. Data function will go with a line from dot another dot.

Here each x value only will have only one y value. Easy to understand without any expert knowledge. Figure 2.11 can take as an example graph for a line chart.

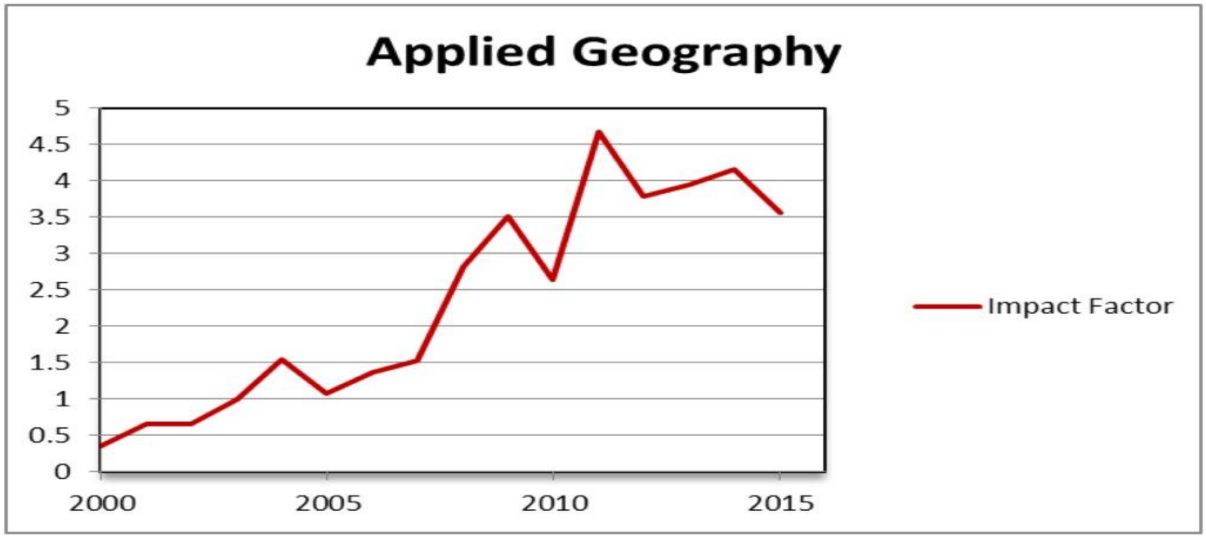


Figure 2.11: Line chart for Applied Geography Impact factor

**Bar chart**

Bar chart also is known as bar graph represent categorical data in a rectangle manner according to a scale. Heights and length of those rectangles will represent those values in scale. Bar chat can be

drawn in both vertical and horizontal ways. Bar charts can show a comparison between discreet categories. Bar charts are asy to understand without any expert knowledge.

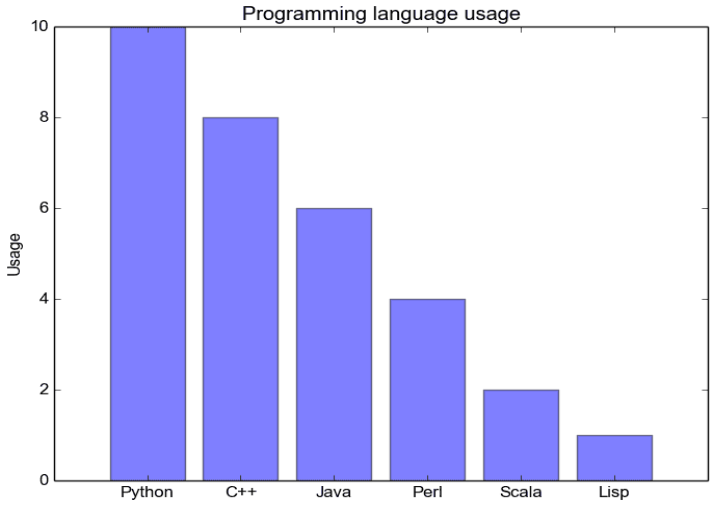


Figure 2.12: Bar chart of programming useages

**Pie chart**

The pie chart is a circle that represents different types according to their quantity as a part of its full quantity as slices. The arc of each slice is proportional to quantity it represents. Not only arc but the angle of each slice also proportional to the value represent.

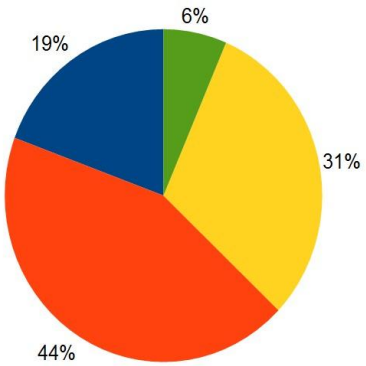


Figure 2.13: Pie chart example

**Parallel coordinates**

Parallel coordinates are the way that uses to visualize data across many dimensions. Data will be display with a connected line with each dimension.

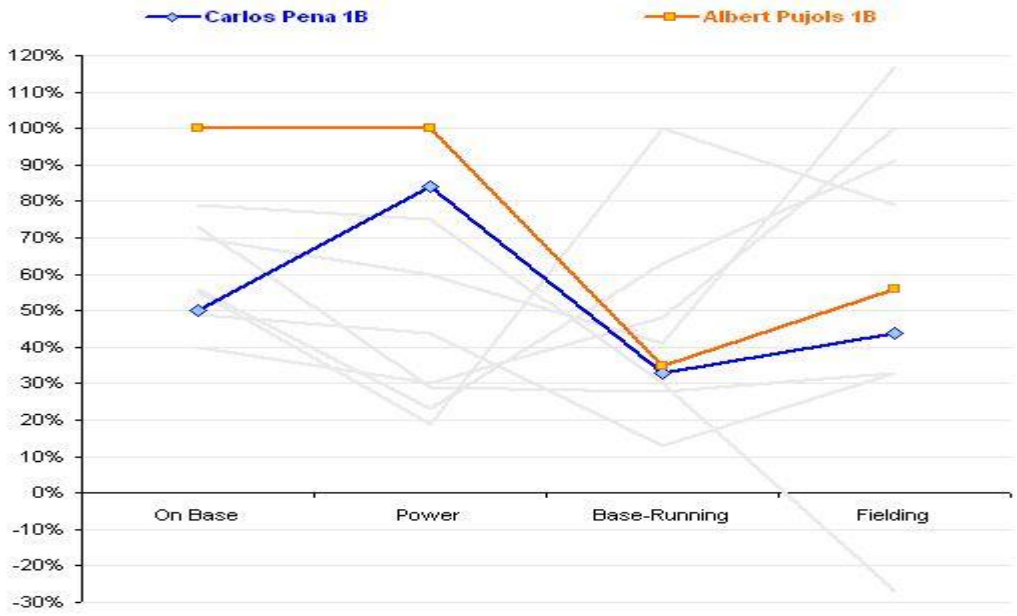


Figure 2.14: Parallel coordinates

Figure 2.14 is representing a player with performance four characteristics. Two players have been selected to compare those values.

**3D surface plots**

This strategy uses to visualize three dimensional data. It will show diagrams in a 3D space of X, Y and Z. These diagrams will be helpful when analyzing the relationship between one dependent and two independent variables [68]. As data to these diagrams, there are 3 variables for each point for x, y and z. Figure 2.15 is an example of the 3D surface plot.

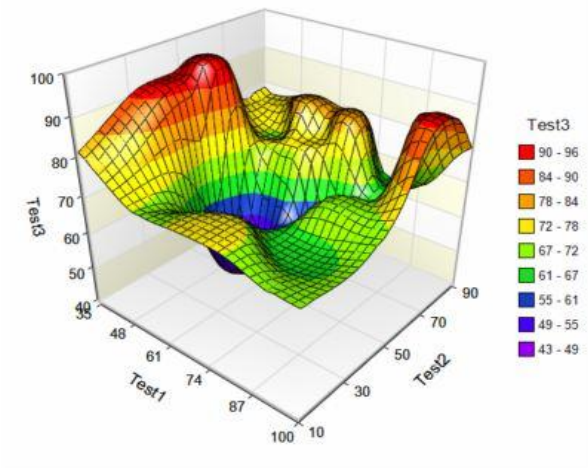


Figure 2.15: 3D surface plots

**Radial layouts with hierarchical edge bundling**

Mainly this method will be based on visually bundling the adjacency edges, non-hierarchical edges, together. The first hierarchy will define using a tree hierarchy method. Next bend each adjacency edges, toward the polyline. Visualization will implicitly adjacency edges between parent nodes will occur because of adjacency edge between their child nodes [69] [70].

Figure 2.16 will be example graph.

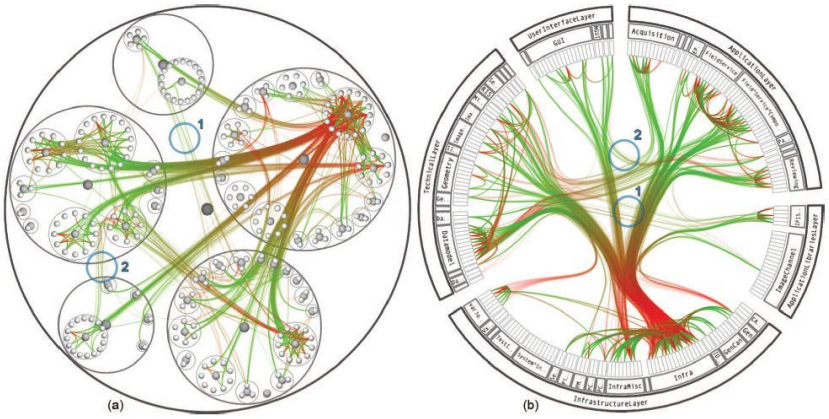


Figure 2.16: Hierarchical edge bundling Example

**Histogram**

A histogram is a plot that mainly shows continuous data and its frequency distribution. And also help to discover. This will be very helpful when analysing the data. (e.g., normal distribution), outliers, skewness, etc [71].

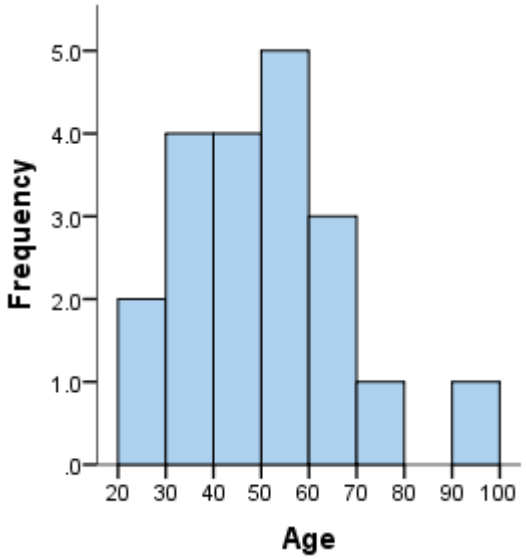


Figure 2.17:Histogram Example with scoure frequencies

**Bubble chart**

Bubble charts are use to presen 3 dimentional numeric data. Here X and Y axis will represent two datas and size of the bubble will represent the 3rd data. Its like scatter chart but scatters are different bubble sizes [72].

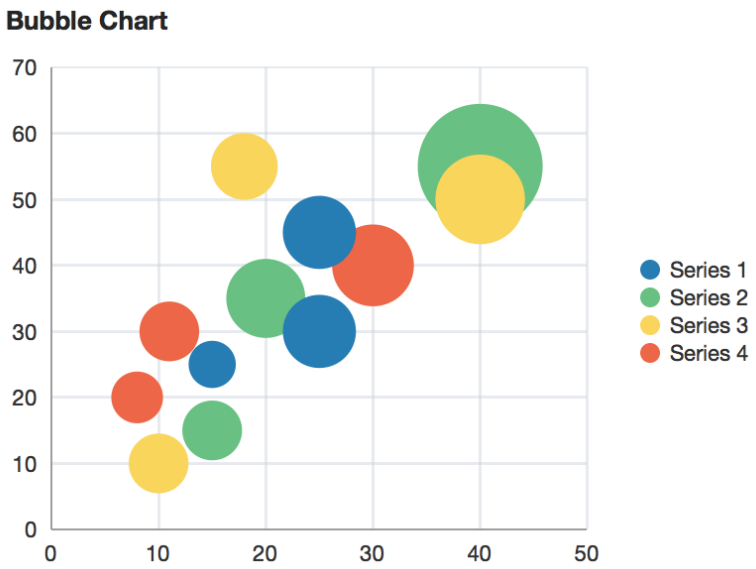


Figure 2.18: Bubble chart for four serises

**Area chart**

An area chart is used to represent data which change with the time like line charts. Here the data are shown as ares respect to the X axis. Area charts are different from line graphs, however, because the area between the x axis and the line is filled in with color or shading [73].

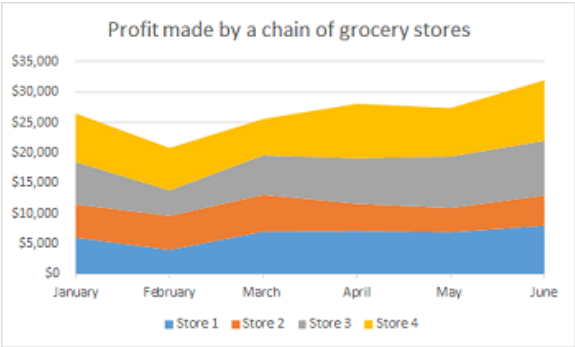


Figure 2.19: Area chart for profit made by grocery stories

**Column chart**

Column chart is mainly focus on group of data to track their diferences [74]. Similar to bar charts and easy to understand without any expert knowledge.

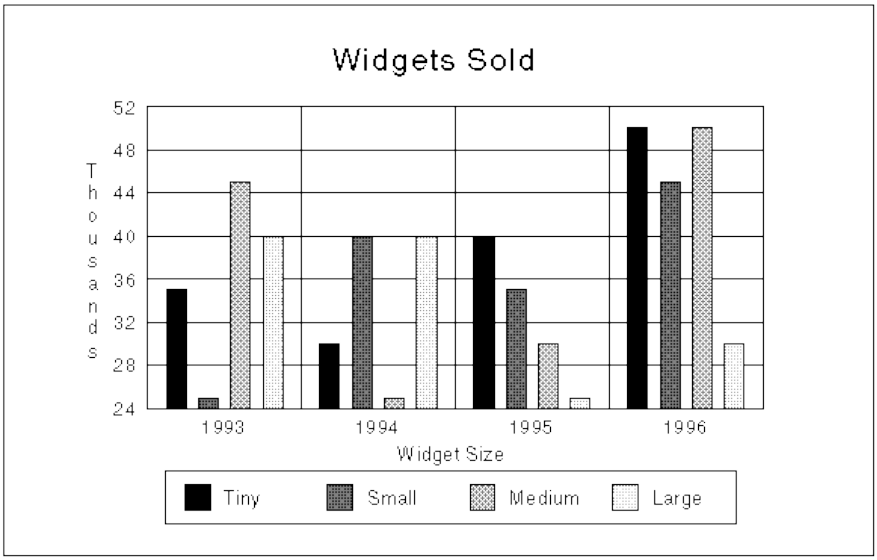


Figure 2.20: Column chart Example for Widgets Sold

**Stacked column chart**

Column charts are good in showing total and their portion. It will be good for specific datasets only. For example, to compare the number of opportunities created each month by campaign source in a report. Grouping values are mainly consider. The chart displays a single bar for each month, broken down by source, with each source shown in a different colour [75] [76].

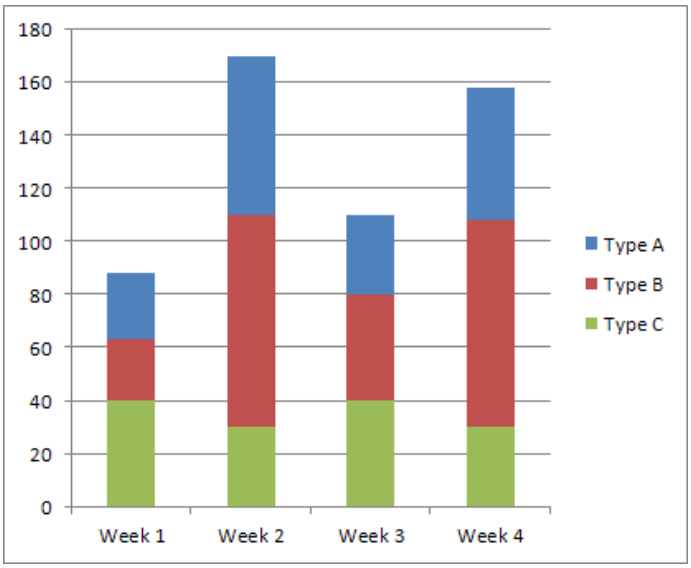


Figure 2.21: Stacked column chart Sample

**Geo chart**

Geo charts show data on a map by location. They are geo area, geo bubble, and geo heatmap charts [77]. And these geo charts can display six types of geographical data, which are:

- Country
- State
- County
- Zipcode
- Point (latitude/longitude)
- Other sub-nation regions (for international countries)

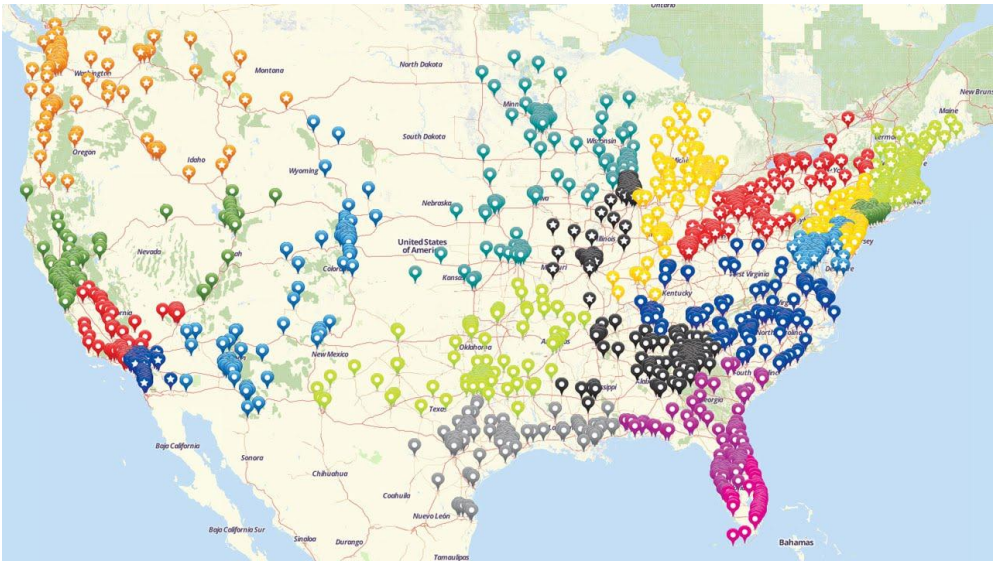


Figure 2.22: Geo chart Pinned location example

### 2.6.3 Related work on data visualization

1. Central limit theorem

A central limit theorem which states that the sum of a large number of independent random variables is distributed as a Gaussian distribution. [78] Get a weighted sum of the random dimension variables. We can see that central limit theorem will grverned their sum and visualize in one dimension by "histogramming." In this figure, we clearly recognize the characteristic "Bell-shape" of the Gaussian distribution of projected and histogrammed data. But Gaussian distribution is also the most uninformative distribution. Will Give a fixed mean and variance, the Gaussian density represents the least amount of information among all densities with the same mean and variance.

2. Levenshtein distance algorithm

[79] This visualisation and algorithm is used to visualize misinformation spread. Here we need significant human involvement so hard to scale.

3. Plot bursts (Twitinfo)

[79] Tweet bursts will automatically identified and calculate the top tweeted URLs. Based on different colors they will be plot in a map. Google also have done a similar approach. But still, it's only for limited users which only willing to participate. Because of collaborative approach it will be easier to the non-technical users.

4. Social network analysis

[64]Social network analysis is used to identify features and properties of the people. Visualisation of those data will provide interactions to explore the data and it can visualize temporal data. Data range will be limited and images will use to visualize.

5. Temporal graph drawing

[64] These techniques use to display real time data with help of animations so the user can extract data easier. Here the GRIP algorithm's modified version will used to generate animations.

6. Quantitative visualization techniques

As 2D scatter-plots cover the main properties of many approaches such as 2D bar-charts, and 2D pie charts, throughout this paper they have use scatter-plot as a general term to refer 2D quantitative visualization techniques. Generally, scatter plots use two axes of Cartesian coordinating to form spatial mappings of the data where each axis corresponds to one data attribute. The main idea behind scatter-plots is to represent the dependency between different data attributes. Scatter-plots do not represent relational information to the user; analysts cannot trace intercommunication patterns. Our method, however, [64]provides an animated representation of both relational and statistical information for a social network.

## 2.7 EXISTING TOOLS ON DATA VISUALIZATION AND RECOMMENDER SYSTEMS

### 2.7.1 Data recommender tools

**In-tag**

In-tag is an innovative content discovery tool which provided by "CodeFuel." This tool will provide a recommendation for each visitor. The algorithm used in the tool will enable to show the right content to the right user at the right time [80].

It works different way than other systems because it will place recommendation automatically on the page rather than going in to another page. We can use In-tag to recommend our own content for more page views.

**LensKit**

LensKit is using collaborative filtering methods in implementing the tool and also provide benchmarking tools. I will provide common API for recommendation algorithms. Additionally will provide a framework for the offline evaluation. The aim of this tool is to provide common recommender framework and will helpful when reusable applications [81].

This tool still under development. Programmers are currently trying to develop rating based and history based methods to this tool.

**Duine Toolkit**

The duine toolkit is an open source product which allows users to develop recommendation engines. It provides many recommendation algorithms. Additionally, it provides methods to join a few algorithms and implement the new combined algorithm. But not only that it will enable the user to create new algorithms [82]. Duine has been developed by the Telematica Instituut/Novay. Basically it was based on scientific research on personalisation and specifically into recommender systems.

Duine Recommender offers 7 recommendation techniques

- Collaborative Filtering
- Information Filtering
- Case-based Reasoning (items which similar have been rated by the user in the past)
- Genre LMS (reasoning on genres)
- Top N Deviation (popularity)
- Already Known
- User Average

**CoFE (the COllaborative Filtering Engine)**

CoFE is a collaborative filtering engine use when applying collaborative filtering. The tool was developed by the IIS research group at Oregon State University. This is a free tool which any one can use and can be able to setup in a java platform.

Features of engine:

- Recommendations for individual items
- Top-N recommendations for all the items and one item

### 2.7.2 Data visualization tools

**Plotly**

Plotly is an open source tool which enables web based interactive data visualization. It also contains composing, editing, and sharing [83]. Users can show their data without the use of coding. That will be a plus point when considering users. There are online chart creation libraries. They will enable creating those interactive graphs.

**Polymaps**

This provides a speedy display of multi-zoom data on maps. Additionally, support a variety of visualizations fo titled vector data. Polymaps can load data in a scales at full range. It shows information from the country level to down like states, cities, neighborhoods, and individual streets [84].

Polymaps uses SVG when displaying data. Furthermore, you can use CSS to modify the layout of the outcome as you wish.

**Zingchart**

Zingchart is a JavaScript based charting library which has rich API to create various types of charts. It helps to build interactive charts. Not only that it will allow HTML charts and also Flash charts. It gives us more than 100 types of chart types to display our data.

It has fast rendering techniques. It can render 100,000 data points in under 1 second. It will enable to add CSS designs to the charts. Furthermore will give real-time processing chart types.

**Dygraphs**

This tool enables to create a visualization with the help of JavaScript. It is a fast and flexible javascript chart library that enable users to explore and visualize data sets. This tool is highly customizable and works in all the major browsers. Not only personal computers also will work fine on mobile and tablet devices. Following are the main features:

- Handles huge data sets
- Interactive out of the box
- Strong support for error bars
- Highly customizable

### 2.7.3 Machine learning support tools

**RapidMiner**

This was formerly known as YALE. The tool was written using java language and provide advanced analytics. Template based framework used to provide advanced analytics. Rather than a software, it provides a service. Following will be the functionality provides.

- Data pre-processing
- Visualization
- Predictive analytics
- Statistical modelling
- Evaluation
- Deployment

The major thing is it provides learning schemes and also can use scripts in R programming language and in weka. It has various applications like visualization and algorithms to analyze and predict the modeling. It has a plus point related to RapidMiner because it can customize as a user like.

**WEKA**

Originally developed a non-Java version. It was used to analysing data from the agricultural domain. With the help of java, it was very advanced. There are lots of functions that can be done using it. Tasks that support by weka listed below.

- Data preprocessing
- Clustering
- Classification
- Regression
- Visualization
- Feature selection

**Orange**

Orange is becoming more popular because it is a python based tool. Moreover, Python is simple and easy to learn yet powerful. As a python developer you look for a tool for data mining it should be Orange because it will be easier for both novices and experts.

Include functions:

- Machine learning
- Add-ons for bioinformatics
- Text mining

**KNIME**

KNIME mainly uses to data preprocessing task. There are three tasks that needed while data preprocessing. They are extraction, transformation, and loading. KMINE does all these three tasks. KMINE developed using JAVA language. It makes easy to extend and add more plugins. More functions also can add while going. It is an open source platform which enables data analytics, reporting, and integration. It has a machine learning and data mining techniques through its modular data pipeline.

## 2.8 EVALUATION OF VISUALIZING DATA IN RECOMMENDER SYSTEMS

### 2.8.1 Comparison of related work on visualizing and recommender data

Table 2.3: comparison of related work on visualizing and recommender data

| Related work | Description | Advantages | Limitations | Main features(RS algorithm/visualization used) |
|---|---|---|---|---|
| [27] | Give the recommendations according to the Facebook profile data. | Recommendations improve according to friends.<br>Will help the system to make decisions. | If friends are lower the accuracy of the output will get low.<br>Facebook API has limited profile access. | A hybrid approach of both content-based filtering and collaborative filtering.<br>Web based networked graph visualization. |
| [32] | Uses either the user preferences or the user and item profiles or a combined version of both to get a recommendation. | User comfortable with using a particular representation for data analysis. | Did not concentrate on whether users are willing to provide information for visualization. | User-based collaborative filtering.<br>Four typess of visualizations integrated into the system (bar chart, timeline, line chart, geo chart). |
| [20] | Give recommendation for bookmaking | The system improves with the provided data. | Given a set of recommendations may not be applicable forever | Collaborative and content-based filtering for item recommendation |

| | | | | |
|---|---|---|---|---|
| | site according to user preferences. | | | |
| [26] | The goal of the Hulu's Recommender system is to find the best interest videos for their users. | Users have suggestions they like but haven't seen before.<br>Feedbacks improve the recommendation quality. | At the start, quality may be low. | Used collaborative filtering as the technique to give recommendations |
| [25] | Related to music preferences overall weight is calculated with the current search and with compared to the previous searching with the help of context matching. | Through a user interface increase user satisfaction.<br>Hybrid strategies improve recommendation than traditional CF. | API only give limited access according to the user. | Recommend related to users friends.<br>A hybrid method of content based and collaborative filtering. |
| [78] | Use Gaussian distribution for visualizing collected data. | Many dimensions can be represented | Gaussian distribution is also the most uninformative | Central limit theorem |
| [79] | Use the Levenshtein distance algorithm, Plot bursts (Twitinfo) for visualizing. | automatically identifies the bursts | Have outlined the structure of an information visualization platform | Levenshtein distance algorithm |
| [64] | Social network analysis, Temporal graph drawing, Quantitative visualization techniques, Focus-context techniques | HTLM represents edges between social actors, decrease the visual complexity caused, Users have the ability to control the visibility of data items | Visual clutter and scalability are weaknesses of the HTLM. | HTLM as visualization method. |

### 2.8.2 Comparison of recommender algorithms

Table 2.4: Comparison of recommender algorithms

| **Algorithm** | **Method** | **Advantages** | **Limitations** | **Reference work** |
|---|---|---|---|---|
| Content-based | The system learns and will help to recommend items | Always provide suggestions based on user preferences. | Recommendations will make only with seen items. | [24],[23],[26] |

| | | | | |
|---|---|---|---|---|
| | that are similar to the ones that the user liked in the past. | | | |
| Collaborative Filtering | Recommends to the active user the items that other users with similar tastes liked in the past. | Unseen Items also will have as suggestions. | All the items are not rated, and there isn't a way to rate newly added. | [27],[24],[23] |
| Demographic | This type of system recommends items based on the demographic profile of the user. based on language, country or age. | Recommendations will be according to a specific category of people. | All the people in the same category may not like the same. | [20], [7] |
| Knowledge-based | Knowledge-based systems recommend items based on specific domain knowledge about how certain item features meet users' needs and preferences. | Can use where other techniques can't use. Like content-based or collaborative filtering.<br>Cold start problem is addressed. | Explicitly need to define the recommendation knowledge. | [20] |
| Community-based | This type of system recommends items based on the preferences of the user's friends. | Can find new interest related to the user community. | Those interest find may not match with all users. | [25] |
| Cluster Models | The algorithm's goal is to assign the user to the segment containing the most similar customers. It then uses the purchases and ratings of the customers in the segment to generate recommendations. | Cluster models can perform much of the computation offline. | Recommendation quality is relatively poor. | [23] |
| Rating Method | These ratings were randomly assigned, with equal probability to three subsets quiz, test, and probe. Selecting the most recent ratings reflects predicting | Can be applied to mass data set. | If a user has fewer ratings (less than 18 in this document) only the most recent one-half of their ratings will select to assign to the subsets. | [21] |

| | | | | |
|---|---|---|---|---|
| | future ratings based on past ratings. | | | |

### 2.8.3 Comparison of data visualization techniques

Table 2.5: Comparison of data visualization techniques

| Technique | Method | Advantages | Limitations | Reference work |
|---|---|---|---|---|
| Central limit theorem | A central limit theorem which states that the sum of a large number of independent random variables is distributed as a Gaussian distribution. | Many dimensions can be represented using this theorem. | The Gaussian distribution is also the most uninformative distribution. | [78], [53], [19] |
| Plot bursts (Twitinfo) | Automatically identified "bursts" of tweets. Twitinfo also calculates the top tweeted URLs in each burst, and plots each tweet on a map, colored according to sentiment. | Automatically identifies the bursts. | Limited users which only willing to participate. | [79] |
| Levenshtein distance algorithm | These visualizations are helpful in the study of the spread of misinformation. | Identify misinformation. | The approach requires significant human labor and thus is more difficult to scale | [79], [40] |
| Social network analysis | Social network analysis is the study of social actors and their relations through formal methods to reveal statistical properties. Exploration systems generate visualizations of net- work datasets and provide interaction mechanisms such as zooming and | Generate visualizations of net- work datasets and provide interaction mechanisms | It has limited scalability | [64] |

| | | | | |
|---|---|---|---|---|
| | filtering to convey details to the user. | | | |
| Temporal graph drawing | Temporal graph drawing techniques utilize animations and use variations of force directed methods to represent temporal, social network datasets as node-link graphs. In a modified GRIP algorithm is used to produce animations of evolving graphs. | Circular orientation in hyperbolic space provides us for generating readable overviews of large temporal, social network datasets. | Only will concern about temporal data. | [64] |
| Quantitative visualization techniques | As 2D scatter-plots cover the main properties of many approaches such as 2D bar-charts, and 2D pie charts. | Easy charts to understand by the user. | Only represent 2D data. | [10] |

## 2.9 EVALUATION OF RELATED WORK

When we consider about the related work, there are many limitations that we can see. First, let us look on the JAABA (interactive machine learning for automatic annotation of animal behavior) [48] system. It is labeling behaviors by looking at videos. Then based on labeled automatically classify behaviors in other videos. There is an interactive user interface to visualize to non-machine learning user. Even though provided outputs are complex Therefore only will useful to experts in the area.

Next, let see ReGroup (interactive machine learning for on-demand group creation in social networks) [5] system. In this system when new person added to group learn probabilistic modal related to group membership in that group. Then will suggest additional members and group characteristic for filtering. But still, there is a problem of addressing missing values.

Interactive Machine Learning System for Automated Annotation of information in the text [50] is another related work. Provide a convenient and efficient interface Therefore that the context of use can be verified if necessary in order to evaluate the annotations and correct them. Here also cost and time limitations were applied.

SmallWorlds [27] generates accurate and useful item recommendations based on a combination of the Facebook user profile. Still, Facebook API limits the access to the user profiles. With a lower number of friends, richness recommendations will also go low. Cannot adopt the standard large scale automated tests to evaluate the quality of a recommendation algorithm in terms of accuracy.

TasteWeights [25] Hybrid music recommendation system with an interactive interface. Here because of real-time processing will take time. This situation also we can see social media APIs limits data. Deploying an Interactive Machine Learning System in an Evidence-Based Practice Center is the abstrackr [49]. In this system, final output needs to deliver to domain experts because output is messy and without any expert knowledge hard to understand.

## 2.10 QUALITY ATTRIBUTES

When dealing with computer systems failures can happen at any time. when it uses a way to minimize those failures by providing solutions to them, the outcome will be perfect. Most times to get the quality outcome we depend on assumptions [85]. But there should be a proper way to check those assumptions are correct or not. Based on them we can achieve the objectives in high standards.

### 2.10.1 Accuracy

Accuracy is the closeness of the measured value to standard or known value. For example, we measured the weight of the given substance. The measured value becomes much lesser or higher than the expected value. That means the Accuracy is very low [86]s. To be accurate value should be near to the expected value. For accuracy Precision, recall and F1 are widely used. Precision is a number of correct outcomes from all outcomes received. The recall is a number of correct ones among the total number of relevant instances. F1 is the measure that joins both precision and recall. It will be harmonic mean of precision and recall.

Precision = |{relevant information} ∩ {retrieved information}| / |{ retrieved information} |

Recall = |{relevant information} ∩ {retrieved information}| / |{relevant information} |

F1 = 2 * (Precision * Recall)/(Precision + Recall)

### 2.10.2 Performance

When we create an application, we must check the code for its performance. Therefore we need to check how our code is running. By evaluating the performance, we can able to minimize the running time. That will become more cost effective when the system started to run. For this performance evaluation, kappa statistic has used. All disagreements between those observers will be treated as total disagreements event if those neighboring values of ordinal scale performance indicators are used. Nominal values will use to represent playing surface areas.

### 2.10.3 Reliability

Reliability is mainly the quality of the measurement. Reliability analysis will provide the ability to study the measurement scales and items that compose in those scales. Reliability analysis will look in to commonly used measurement and scales to check reliability. Additionally, it will provide a relationship between individual items.

Few models of reliability:

**Alpha -** Model of internal consistency, it will be based on average of inter-item correlation.

**Split-half –** the scale will split into two parts and examines the correlation between those two parts.

**Guttman -** computes Guttman's lower bounds to get the true reliability.

### 2.10.4 Usability

Simply usability is the quality of the user experience when the user interacts with those softwares or system. To check the usability should give it to users to a test run. Otherwise will not be able to find errors from users' perspective. Mainly the interfaces should be understandable to the user. Then he can perform the actions according to the planed flow [87]. To check you may use the following methods.

- Provide surveys or interviews to get user feedbacks to achieve goals.
- Check that the user follows the expected path

### 2.10.5 Scalability

The ability of software programs to continue the function after its context or volume changed in order to satisfy the user. For example, if you perform a task from the data set which contains fewer data. Then you need be able to perform the same task with a large amount of data also. Scalability also can categorize as following areas.

Load scalability: Check how different data loads can be handled.

Administrative scalability: Check how many users can handle by the system at one time.

Functional scalability: Check enhancement of the system by adding a new function.

# Chapter 3

# RESEARCH METHODOLOGY

## 3.1 INTRODUCTION

This section mainly discusses methodology in the study. The main component of this research is a recommendation component. That component consists of two sub components. Those components are Machine learning component and rule-base component. Both sub components process separately to come up with suggestions, and in the final step, both will be combined to give a most efficient final recommendation. Moreover, there is a pre-processing step that helps to clean and smooth those data. Context identifier in the system helps to grab the context in provided data. In the final stage visualization module gives chart output. This section covers those components and how they perform as one system to succeed the ultimate goal.

## 3.2 SYSTEM DESIGN

### 3.2.1 The process of the system

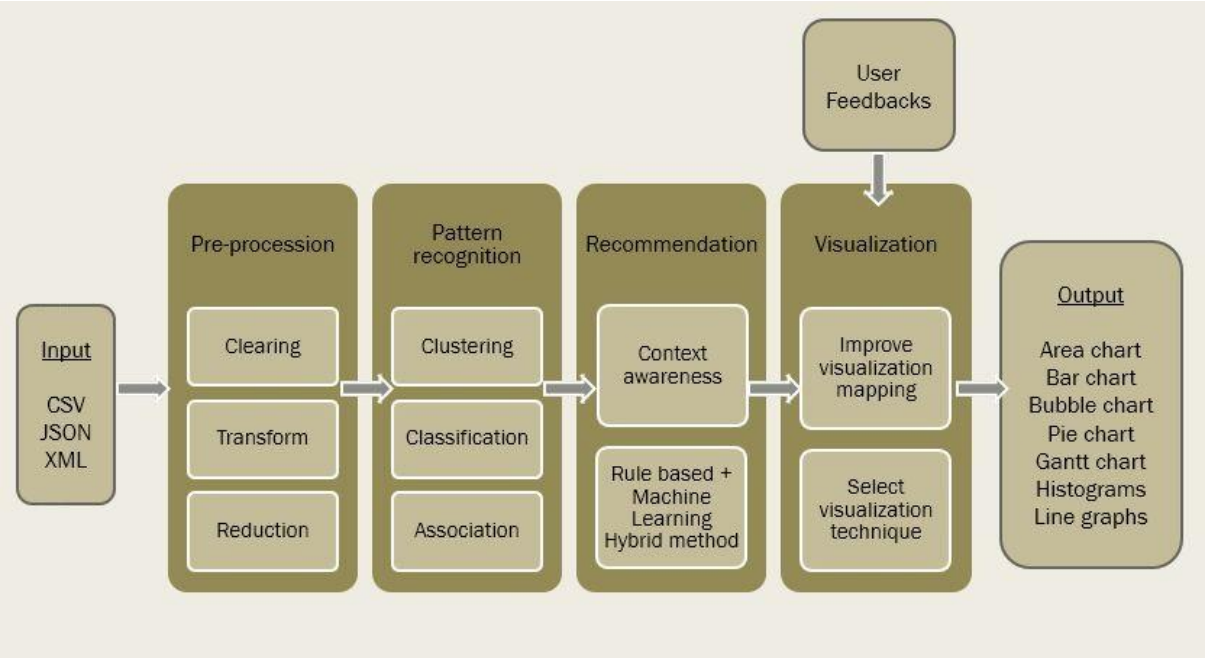


Figure 3.1: Design

In the first step, data will be preprocessed. With those provided data there can be many behaviors that can effect further processing. By preprocessing can convert data into more reliable and accurate data. Here missing values will be filled; duplicate data will be removed. Then data will arrange into

common schema Therefore it will be easy on future processing tasks. Then need to find behaviors of the data to give recommendations. For that data mining will be used. With the help of decision tree based machine learning algorithm which will be map data context and variables with output visualization charts provide us recommendations. Then rule-based mapping also will provide recommendations. Finally, both will combine to come up with a final recommendation. With the help of user feedbacks training, data will improve.

### 3.2.2 Architecture design

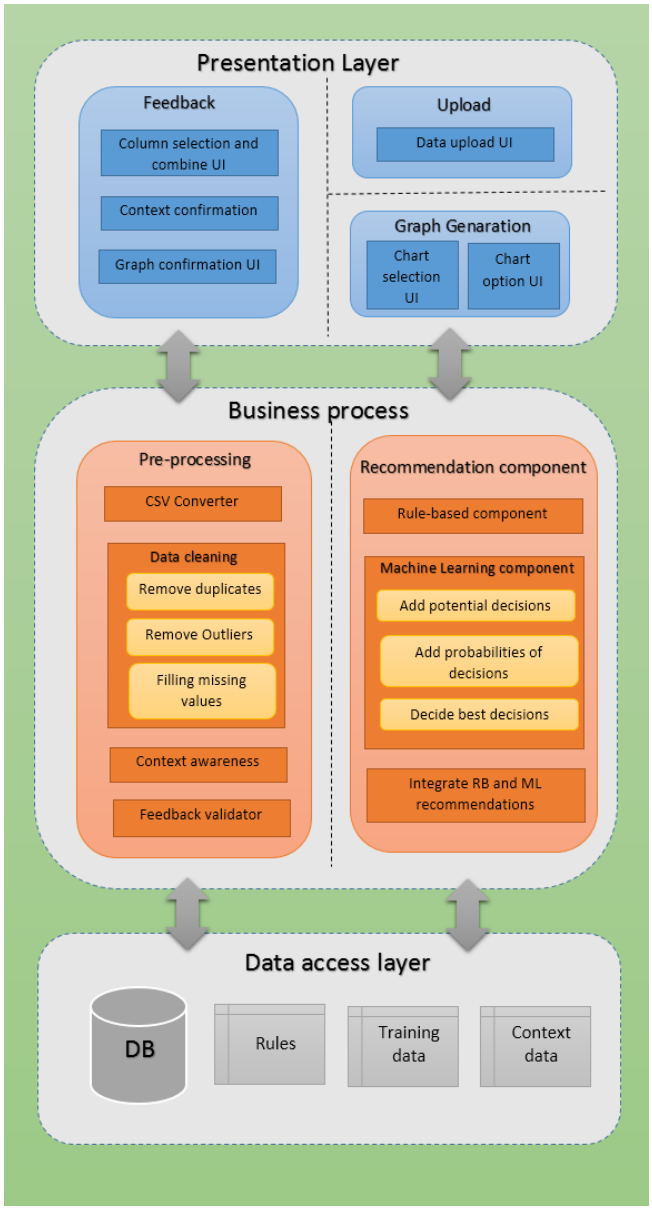


Figure 3.2: Architectural Design

Data flow will happen among layers in Figure 3.2. In the presentation layer, there are UIs to upload the data, to provide suggestions to the user and to get a feed from the user. In the logical layer, it will define the operation that needs to take based on the user inputs such as input data, selected

columns, confirmation of recommendations. Then the data access layer will access the database data according to the system needs. When identifying the context, it will look previous context identifications. While in the recommendation process it will consider both training data and feedback data. Finally, the data layer will hold the data which will impotent to the system such as user feedbacks, context related data and also the training data.

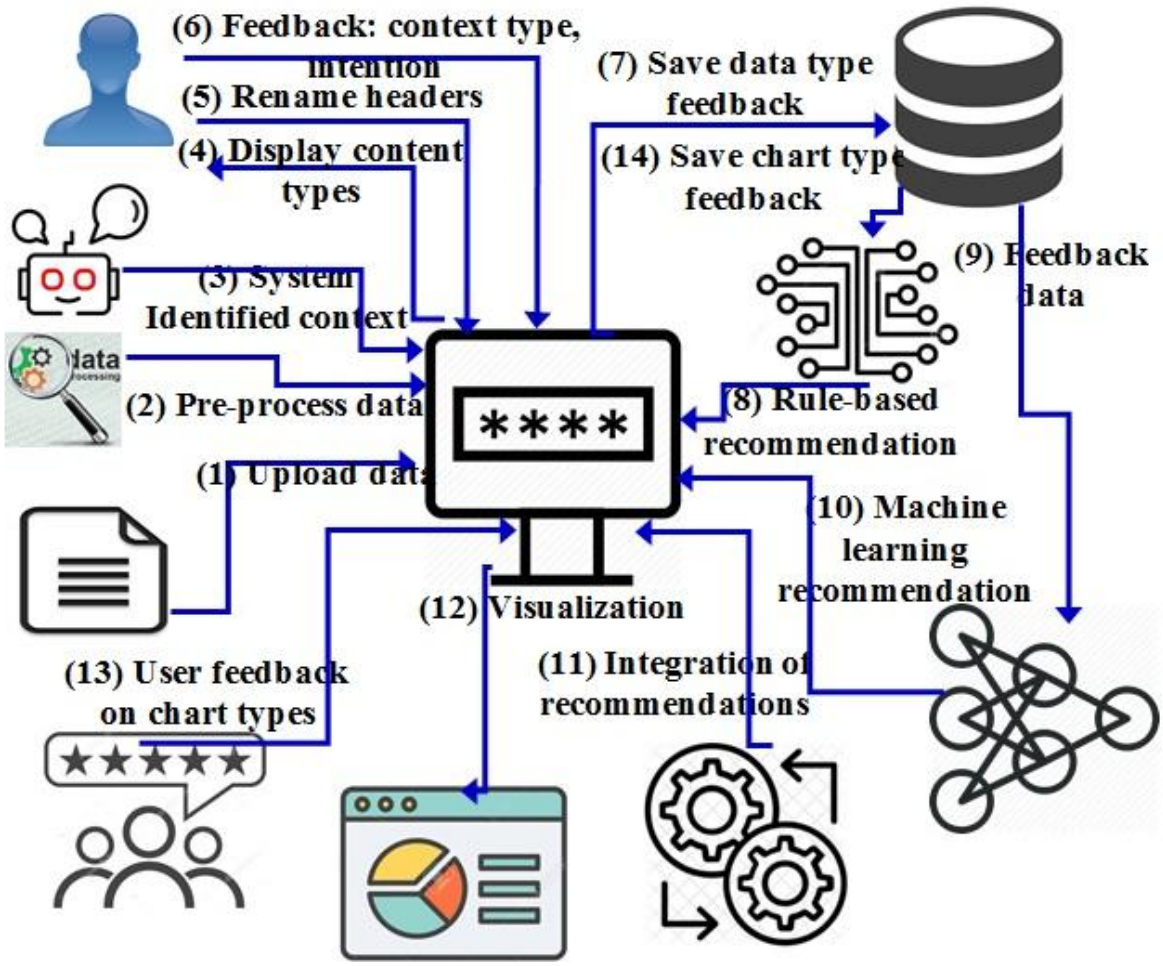


Figure 3.3: Data Flow Diagram

## 3.3 PRE-PROCESSING

### 3.3.1 Duplicate tuple removal

While inputting the data, there can be mistakes happen such as apply same data more than one time. That kind of situations needs to detect. Otherwise, conclusions come up based on that visualization can be wrong.

Following are the code steps to remove duplicate tuples.

**Pseudocode 1** Duplicate tupple remove

```
Require: Input Dataset
Ensure: Remove duplicate
Return: duplicate removed list
1. Start
2. while (Read each line of data) {
3.  // build a "line" from the parsed data
4.  line = data_raw;
5.   // if the line has been seen, skip it
6.   if (isset(lines[line])) continue;
7.   // save the line
8.   lines[line] = true;
9. }
10.  End
```

Here each line will be saved in an array. If the line in the current loop can find in the newly created array it will be skipped.

### 3.3.2 Normalization

Normalization also can be done using the PHP-ML library. Basically, Normalization is the process of scaling individual samples to have unit norm.

Following is the PHP-ML code based on normalization

$samples – provided dataset.

**Pseudocode 2** Normalization

```
Require: Input Dataset
Ensure: Normalization
Return: Normalize data
1.  Start
2.
3.  use Phpml\Preprocessing\Normalizer;
4.
5.  $normalizer = new Normalizer();
6.
7.  $normalizer->preprocess($samples);
8.
9.   End
```

### 3.3.3 Filling missing values

When collecting or recording the data there can be missing values. If data need further processing, missing values need to be addressed. When filling missing values usually replace all missing values with the same constant value. Here also we can use PHP-ML library to this operation. It will address missing values, often encoded as blanks, NaNs or other placeholders. To solve this problem, PHP-ML uses the Imputer class.

Here we have to consider both numerical and nominal data. Therefor its better use the replace missing using the most frequent value along the axis.

$data – input dataset.

**Pseudocode 3** Filling missing values

```
Require: Input Dataset
Ensure: Fill missing values
Return: Data with no missing values
1.  Start
2.
3.  use Phpml\Preprocessing\Imputer;
4.  use Phpml\Preprocessing\Imputer\Strategy\MostFrequentStrategy;
5.
6.  $imputer = new Imputer(null, new MeanStrategy(), Imputer::AXIS_COLUMN);
7.  $imputer->transform($data);
8.
9.   End
```

### 3.3.4 Outlier removal

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population [88]. When plotting data outlier data needs to be removed. Otherwise, it will take larger space to represent just that data. The following algorithm will address those data.

Let A denote an n×n symmetric matrix of non-negative real numbers whose diagonal entries are all equal to 0 and whose off-diagonal entries are all unique. An example of such a matrix, which is used for diameter-based outlier removal, is the matrix in which Ai,j denotes the distance from point i to point j in an n point set in which all interpoint distances are unique.

**Pseudocode 4** Outlier remove

```
Require: Input Dataset
Ensure: Outliers removed
Return: Data with no outliers
1.  Start
2.
3.  for g = 1 to c + 1 do
4.  Ai,j ← maximum entry of A
5.  x ← (c + 1)-st largest entry in A i,*
6.  T ←  T̃U{ k : Ai,k≥x} {* mark c + 1 largest entries in row/column i*}
7.  x ← (c + 1)-st largest entry in A*,j
8.  T̃ ← ˜TU{k :Ak,j ≥x} {* mark c + 1 largest entries in row/column j*}
9.  A←A \{ i, j} {* delete i and j*}
10. return  T̃s
11.
12.  End
```

## 3.4 CONTEXT IDENTIFIER

If we look for context awareness tools, there are many tools that can identify those context of the data [89] [90]. The first user selects the columns which needed to visualize. The system considers both header and the actual data to identify the context. The solution is able to understand the context of the data such as location, date-time, percentage, numeric and nominal. Some occasions data types may not be able to identify by the system. For example data relevant to 'Year' can be tag as a numeric

value, where it should be categorized as date-time type. Therefore we have used a confirmation method.

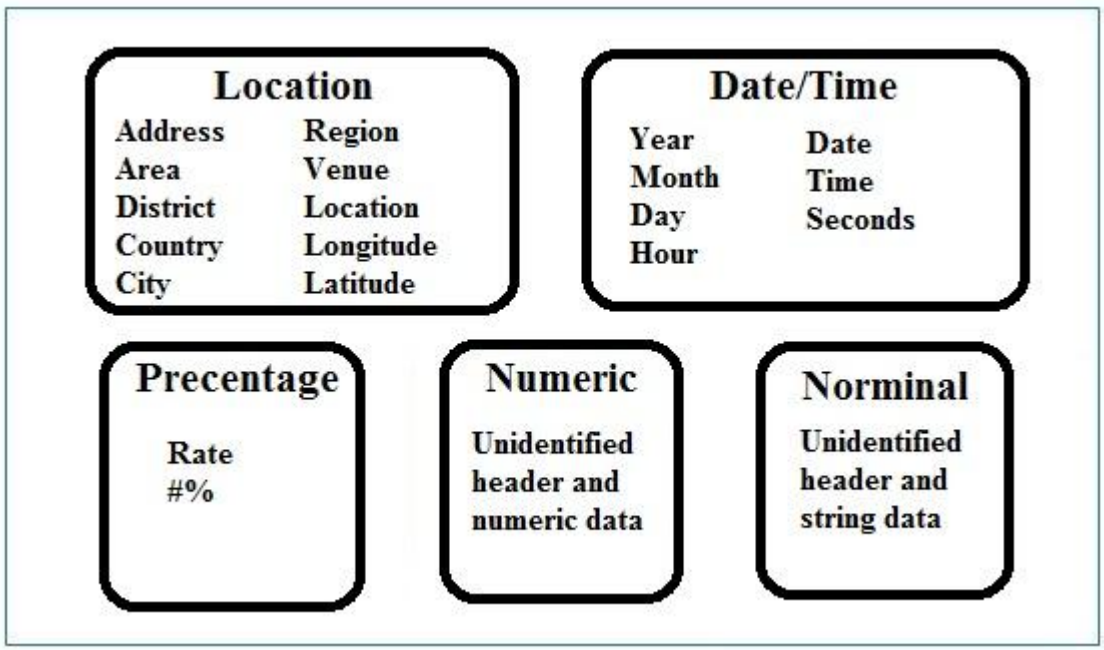


Figure 3.4: Context Identification Phrases

This method displays identified context data to the user and asked the user to confirm the identification or can select the correct context from the given dropdown. This method will mitigate the unsatisfactory of the user. Those user feedbacks related to context data will be used while identifying the context of the same dataset in later occurrences. When user feedbacks increase the accuracy of context identification gets higher. Moreover with more test cases above context identification phrases will increase.

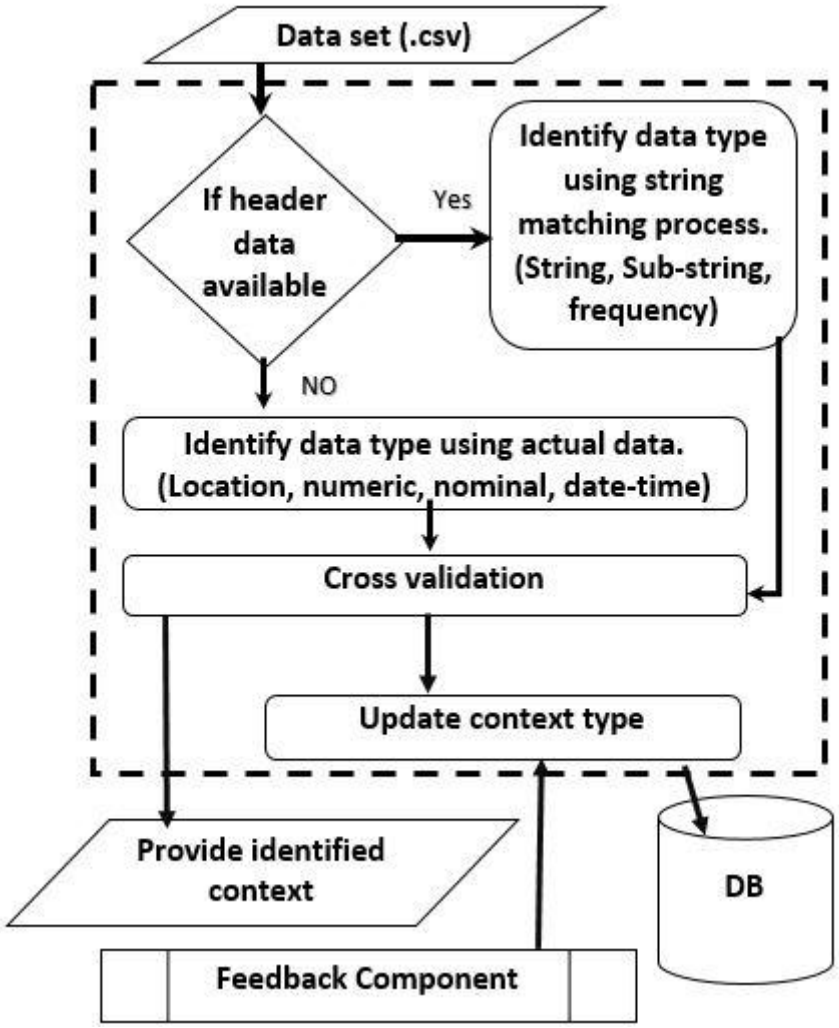


Figure 3.5: Flow of Context Identifier

Figure 3.6 shows the design of the context identifier component.

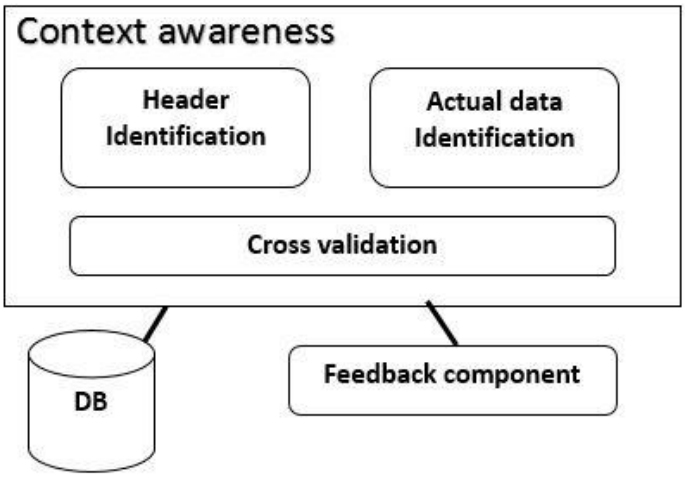


Figure 3.6 Content Awareness Component

In the context awareness module, there are few sub-modules which help to come up with the decisions in the module. First is the header data identifier. If the header data is available this module responsible for identifying the context data type using header name strings. Then the actual data identifier will use actual data and its ranges to identify the data types. Finally, the cross validator considers both identifications and provides the final context data.

Following are the terminology used in the pseudocode for context awareness.

- header_data : all the header related data
- actual_data: actual data set

**Pseudocode:** Context awareness

```
Require: Input Dataset
Ensure: Find suitable data types
Start
If(header_data){
    SELECT type, COUNT(*)
    FROM header_tags table
    WHERE ' header_data'
    LIKE CONCAT( \'%\', name, \'%\' )
    GROUP BY type
    ORDER BY COUNT(*) DESC
    Limit 1

Give header_datatypes
    }
  else{
if(is_numeric(actual_data)){

if(date-time format){ Actual_datatypes='date-time'}
elseif(location){ Actual_datatypes='location' }
else{ Actual_datatypes='numeric'}

}else{
Actual_datatypes='norminal'
}
  }
context = cross_validate(header_datatypes,actual_datatypes)
return context
End
```

## 3.5 RULE-BASED COMPONENT

Rule-based reasoning is a logic-based reasoning technique where rules are used to define the production of new statements and conclusions from initially known statements of the model (describing the application domain) or from the ones derived earlier [91].

Basically when considering context awareness most occasions, rule-based approach come very usefully. Based on context taking decisions based on given rules make easier when taking correct decisions. The rule-based system represents knowledge in terms of a set of rules that tells what to do or what to conclude in different situations [92].

Any rule-based system consists of a few basic and simple elements as follows:

- A set of facts. These facts are actually the assertions and should be anything relevant to the beginning state of the system.
- A set of rules. This contains all actions that should be taken within the scope of a problem specify how to act on the assertion set. A rule relates the facts in the IF part to some action in the THEN part. The system should contain only relevant rules and avoid the irrelevant ones because the number of rules in the system will affect its performance.
- A termination criterion. This is a condition that determines that a solution has been found or that none exists. This is necessary to terminate some rule-based systems that find themselves in infinite loops otherwise.

In most studies, we collect data which important to do further studies. Therefore there should be an appropriate presentation method of those data to get more understanding about those data. There is major problem when it comes to selecting the correct chart types, styles, and methods of presenting those data. it can be confusing and difficult to pick the right one [93].

In this component mainly consider mapping between data types and chart types based on pre-defined knowledge. There are few studies which follow how to select correct chart type for a dataset based on their data types. We followed those studies to collect data related to map those chart types with the data types we identified.

When we consider about cold start related to our system the machine leaning process may not be accurate. In those situations, this mapping between data types and chart types are helpful to come up with recommendations.

### 3.5.1 Why rule-based

There are many outcomes of selecting a rule-based approach to the suggestion system.

- Availability: Availability of the system for the user is not an issue
- Cost efficient: This system is cost efficient and accurate in terms of its end result
- Speed: You can optimize the system as you know all the parts of the system. Therefore to provide output in a few seconds is not a big issue
- Accuracy and less error rate: Although coverage for different scenarios is less, whatever scenarios are covered by the RB system will provide high accuracy. Because of these predefined rules, the error rate is also less
- Reducing risk: We are reducing the amount of risk in terms of system accuracy
- Steady response: Output which has been generated by the system is dependent on rules, Therefore the output responses are stable, which means it cannot be vague

- The same cognitive process as a human: This system provides you with the same result as a human, as it has been handcrafted by humans
- Modularity: The modularity and good architecture of the RB system can help the technical team to maintain it easily. This decreases human efforts and time

Some drawbacks also there

- Lot of manual work: The RB system demands deep knowledge of the domain as well as a lot of manual work
- Time consuming: Generating rules for a complex system is quite challenging and time consuming
- Less learning capacity: Here, the system will generate the result as per the rules, Therefore the learning capacity of the system by itself is much less
- Complex domains: If an application that you want to build is too complex, building the RB system can take a lot of time and analysis. Complex pattern identification is a challenging task in the RB approach
- It is not easy to mimic the behavior of a human.
- Time consumption of human effort is too high.

### 3.5.2 User Intention

When selecting a chart type, it is important to identify the intention of the user to come up with the best chart suggestion. Therefor we have to consider it is to show a relationship between data points, a comparison of data points, a composition of data, or a distribution of data.

**Relationship**

A relationship tries to show a connection or correlation between two or more variables through the data presented, like the market cap of a given stock over time versus the overall market trend.

**Comparison**

A comparison tries to set one set of variables apart from another, and display how those two variables interact, like the number of visitors to five competing web sites in a single month.

**Composition**

A composition tries to collect different types of information that make up a whole and display them together, like the search terms that those visitors used to land on your site, or how many of them came from links, search engines, or direct traffic.

**Distribution**

A distribution tries to lay out a collection of related or unrelated information simple to see how it correlates, if at all, and to understand if there's any interaction between the variables, like the number of bugs reported during each month of beta.

### 3.5.3 Select a chart based on user's intention

After you identified intention of the user next need to select the best method to display those data. Different chart types can be categorized based on user need. For example, scatter plots are best used to show distributions, while line charts (scatterplots with a defined trend) are better suited for relationships. Pie charts do well when you're trying to communicate a composition, but make for poor comparisons or distributions.

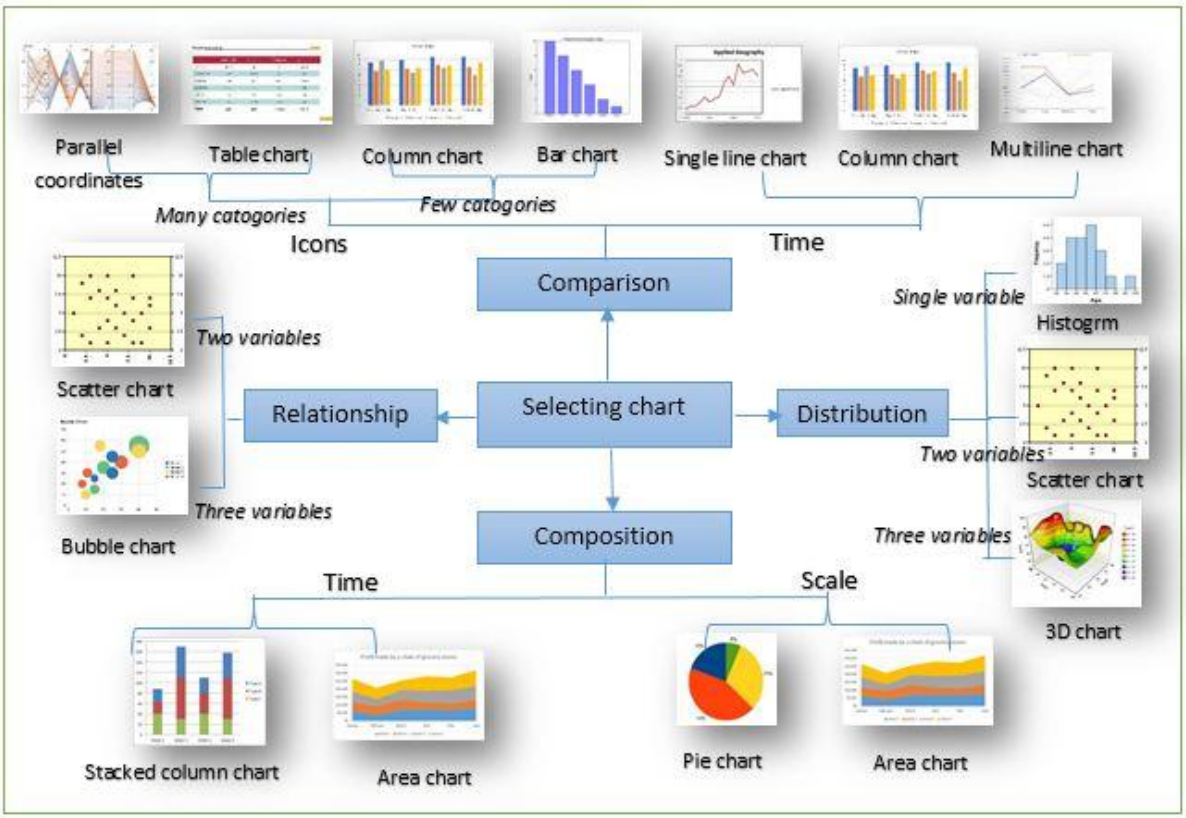


Figure 3.7: Chart Suggestions

### 3.5.4 Steps of process

By using the above chart, we can easily understand the chart type that needed for he given data types. Here in our system, we have selected some of charts to develop the system. Based on given rules it will provide suggestions to given data sets. Even though currently using a few charts we can develop it to visualize more charts with more observation. Here in our system, we have mainly considered

the data types or the context of the given data. Then the user intention to make easier to suggestion process. Then look if there any time related variables. If time related variables included there will be specific chart types. If not there will be different chart suggestions. Finally look for the number of variables included in the data set. Based on the variable count also the outcome can be changed.

Selected rules were more simplified to process easier and get the suggestions. Main involvement of this module will occur when in the cold start of the system with lesser data in the training data to a machine learning component.

There are major steps including while identifying chart types using this component.

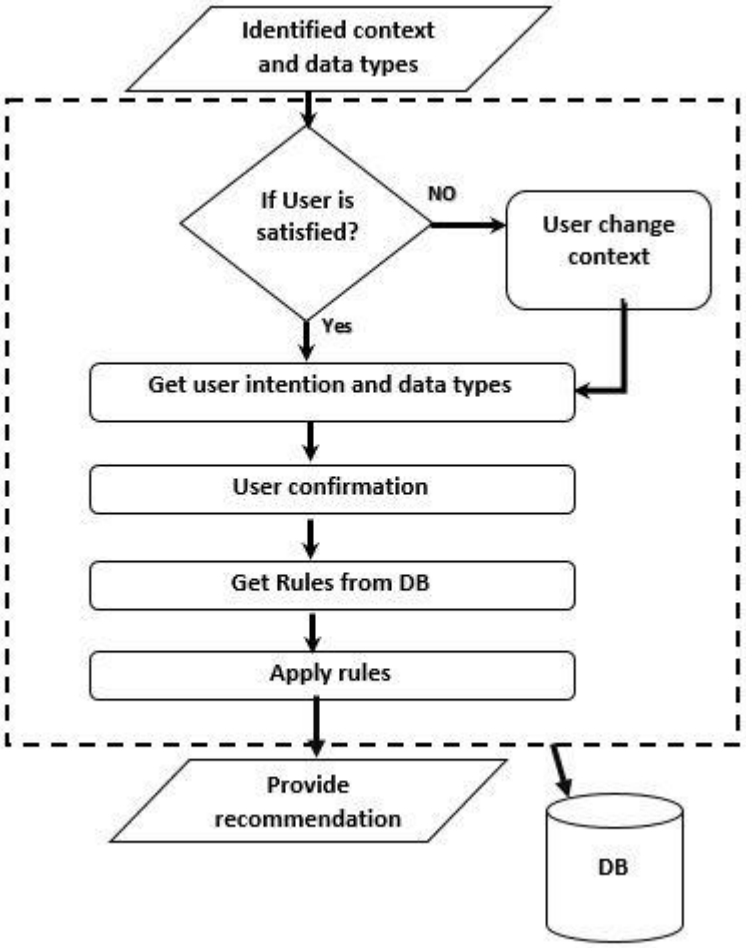


Figure 3.8: Chart Identification Steps

Before coming to rule-based identification, there is a pre-identification process of the context. In this process, the system has a mapping between data column headers and the data context. Those mapped data are stored in the database.

First, each column header match with the pre identified header names. If a system is able to find any matching context, it will be given as the output. Still, there can be multiple context identification for some headers. Therefor system will look for most occurred. If the system enables to find any match

to the given header, then look for the data to find about context type. Then the user needs to confirm the suggested context of if they are mis-identified user can change to the correct one. This confirmation will be recorded in the system and will be helpful in the next text cases.

Then that identified context will forward to the next step with the user intention. The user needs to provide the intention. Then based on user intention, identified context, number of variables and the existence of date-time variables rules will give the suggestion to the given data set.

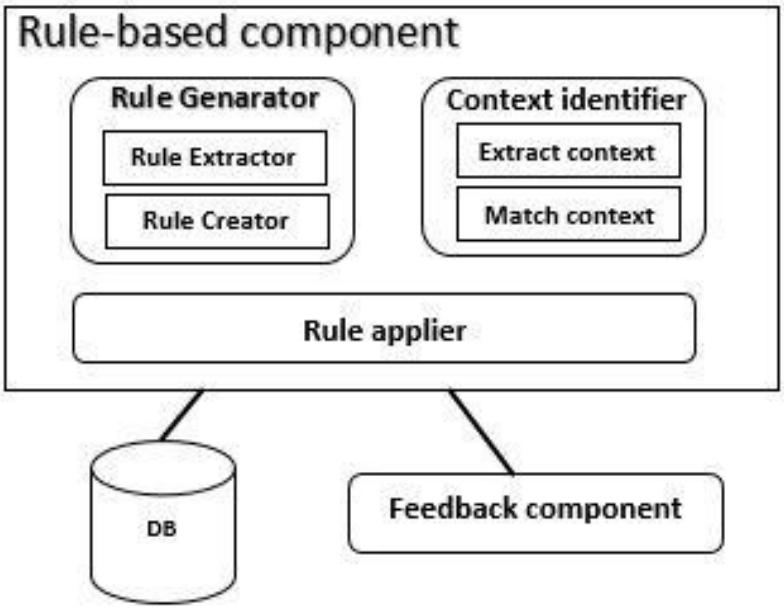


Figure 3.9 Rule-Based Component

After charts suggested based rules these constrains are using to order the chart recommendations.

Constrains and conditions

C1: Bigger datasets more than 50 rows

C2: Non-linear data

C3: Datasets less than 50 rows

C4: Datasets less than 20 rows

C5: Bigger datasets more than 20 rows

C6: Date-time variable should include

C7: Nodes connected with only two levels (parent and child)

C8: Nodes connected with more than two levels

C9: At least one numeric or percentage variable

C10: Two or more data columns

C11: One variable should be Numeric

C12: At least one numeric or percentage variable

C13: Any type of variable can add

C14: Two or more data series

C15: Datasets less than 10 rows

C16: All 3 variables should be Numeric

C17: Represent percentage data

C18: Covered area used to conclusions

C19: Linear length used to conclusions

C20: Data with different levels

C21: When smaller changes exist

Table 3.1: Rules for deciding chart types

| | intention | Possible Data types | Number of variables | Chart type | Data set size & Special constrain | References |
|---|---|---|---|---|---|---|
| 1 | Relationship | Numeric | 2 | Scatter plot | C1, C2 | [94],[95] |
| 2 | Relationship | Numeric | 3 | Bubble chart | C3, C2 | [94] |
| 3 | Relationship | Date-time, Nominal | 2 | Gantt chart | C4, C6 | [95] |
| 4 | Relationship | Nominal | 2 | Tree diagram | C4, C7 | [96] |
| 5 | Relationship | Nominal | 2,3 | Network diagram | C5, C8 | [97] |
| 6 | Relationship | Nominal | 2 | Hierarchical edge binding | C1 | [69] |
| 7 | Comparison | Numeric, Nominal, Location, Percentage | 1+ | Parallel coordinates | C4, C10 | [98] |
| 8 | Comparison | Nominal, Numeric | 2 | Bar chart | C4, C11 | [99] ,[94] |
| 9 | Comparison | Numeric, Nominal, | 2+ | Radar chart | C4, C10 | [99] |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | Location, Percentage | | | | |
| 10 | Comparison | Nominal, Numeric, Percentage | 2+ | Column chart | C4, C12 | [99] |
| 11 | Comparison | Numeric, Nominal, Location, Percentage | 2+ | Table | C5, C13 | [99] |
| 12 | Comparison | Date-time, Numeric, Nominal | 2 | Single line chart | C1, C6 ,C21 | [99], [94] |
| 13 | Comparison | Date-time, Nominal, Numeric, Percentage | 2+ | Column chart | C4, C12 | [99], [94] |
| 14 | Comparison | Date-time, Nominal, Numeric, Percentage | 2+ | Multi-line chart | C12, C14,C21 | [99], [94] |
| 15 | Comparison | Date-time, Numeric | 2 | Radial Bar Chart | C15, C6 | [98] |
| 16 | Distribution | Numeric | 3 | 3D chart | C1, C16 | [100],[95] |
| 17 | Distribution | Numeric | 1 | Histogram | C4 | [99], [94] |
| 18 | Distribution | Numeric | 2 | Scatter plot | C1, C2 | [94],[95] |
| 19 | Composition | Date-time, Numeric | 2 | Area chart | C1, C2, C6 | [99] |
| 20 | Composition | Date-time, Numeric | 2+ | Stacked column chart | C4 | [101], [95] |
| 21 | Composition | Date-time, Numeric | 2+ | Stacked area chart | C1, C2, C6 | [95] |

| | | | | | | |
|---|---|---|---|---|---|---|
| 22 | Composition | Numeric, percentage | 2 | Area chart | C1, C2 | [99] |
| 23 | Composition | Numeric, percentage | 1+ | Pie chart | C4, C17, C18 | [99],[95] |
| 24 | Composition | Nominal, Numeric | 3 | Sunburst chart | C4, C20 | [102] |
| 25 | Composition | Nominal, Numeric | 2+ | Doughnut chart | C4, C19 | [103] |

## 3.6 MACHINE LEARNING COMPONENT

This is one of the main components in this system. This component will process according to the identified context types and number of variables in each provided dataset. There is a trained dataset with a number of variables and identified context data. Those data are using while giving visualization suggestions to provided data sets. After datasets uploaded in to the system user can provide the visualization columns and combine those columns if necessary. Then the machine learning component will give suitable chart types as suggestions based on provided training data. The classification process of suggesting chart types is running based on decision tree based approach. While giving recommendations system will consider context and variable count as features. In the training data, chart types are the labeled data.

Moreover, the system will improve the training dataset with final user confirmation of the given visualizations. At the early stages, the accuracy of the machine learning component is lesser accurate. But with more user interactions system learns to give better recommendations with the help of those user interactions. For each user interaction, a new record will be added to the training dataset. Those data will be used to classify the data uploaded by the next user onwards. More user interactions mean the accuracy of the machine learning module is gets higher.

### 3.6.1 Why machine learning

Machine learning is a method of analyzing data to build models. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention. In this research first, we need to get trained data set to give visualization suggestions. Basically, machine learning was born from pattern recognition and the theory that computers can learn without being programmed to perform specific tasks; most cases

researchers use Machine learning in their tasks because computer programs are learning from the data to give suggestions related to input data [104]. It makes it easier for the user when interacting with systems. The user may not need to think deeply to come up with decisions because the system helps the user by suggesting related to past behaviors.

Improving the ability to apply complex mathematical calculations to big data automatically, iteratively and quickly is one of major tasks that machine learning can do [105]. In this study with more user interactions, training data increases and calculations get more and more complex. Therefore using machine learning methods makes it easier for the system.

### 3.6.2 Decision tree based approach

First, we have collected the data and trained them. Those trained data are stored in the database. For the process features and the labels need to consider separately. Those trained data will use when performing the operations to give suggestions. Here we have used decision tree algorithm as our prediction algorithm.

Classification or regression models can be built using decision tress in the form of a tree structure. Association decision tree incrementally build by breaking the dataset in to smaller subsets. The final result of the tree will consist with decision nodes and leaf nodes. A decision node has two or more branches. A leaf node is the one that represents the final result. Starting from top most decision node call as the root node. A decision tree can handle both categorical and numerical data [106].

The core algorithm of building the decision trees called ID3. It was proposed by J. R. Quinlan. It works as a top-down greedy search through the space of possible branches with no backtracking. ID3 uses Entropy and Information Gain to construct a decision tree. In ZeroR model there is no predictor, in OneR model we try to find the single best predictor, naive Bayesian includes all predictors using Bayes' rule and the independence assumptions between predictors, but decision tree includes all predictors with the dependence assumptions between predictors [106].

**Entropy**

We know that the decision tree is built using top to bottom approach. It starts from the root node and partitioning the data in to subsets which contains instances with similar values (homogenous). ID3 algorithm is using entropy to check the homogeneity of the sample. Entropy will be zero if the sample is completely homogeneous. If the sample divided equally, it might have an entropy of one.

p,q are probabilities of given all possible outcomes. (p,q means there are only 2 possibilities in this example equation)

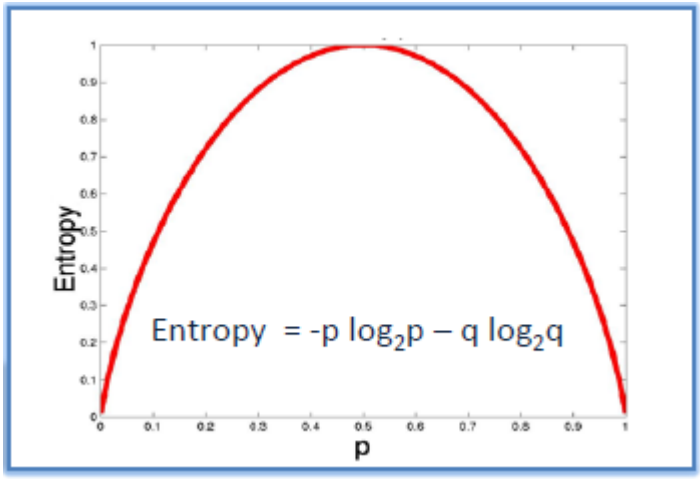


Figure 3.10: Entropy curve

When we are building a decision tree, we need to calculate two types of entropy using frequency tables as follows:

- Entropy using the frequency table of one attribute:

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

- Entropy using the frequency table of two attributes:

$$E(T, X) = \sum_{c \in X} P(c) E(c)$$

The main idea behind a decision tree is to divide the data set into smaller data sets based on the descriptive features. Then it will go until you reach a small enough set that contains data points that fall under one label.

Each feature of the data set becomes a root[parent] node, and the leaf[child] nodes represent the outcomes. The decision on which feature to split on is made based on resultant entropy reduction or information gain from the split.

**Important Terminology related to Decision Trees** [107]

Let's look at the basic terminology used with Decision trees:

- Root Node: It represents the entire population or sample, and this further gets divided into two or more homogeneous sets.
- Splitting: It is a process of dividing a node into two or more sub-nodes.
- Decision Node: When a sub-node splits into further sub-nodes, then it is called a decision node.
- Leaf/Terminal Node: Nodes do not split is called Leaf or Terminal node.

Advantages of Decision Trees [108]

- Decision trees are easy to interpret.
- To build a decision tree requires little data preparation from the user- there is no need to normalize data

Disadvantages of Decision Trees [108]

- Decision trees are likely to overfit noisy data. The probability of overfitting on noise increases as a tree gets deeper.

PHP-ML makes most operations easier because most of the steps were done by the library itself. In our system when the classification process starts first division will happen based on a number of variables. Then tree will expand based on the context type. The depth of the tree will be defined based on the number of context types identified related to those variables. A number of levels will go until the maximum number of context identified in the training dataset.

Based on a searching algorithm correct path can identify. The leaf node of the path will provide the predicted chart type.

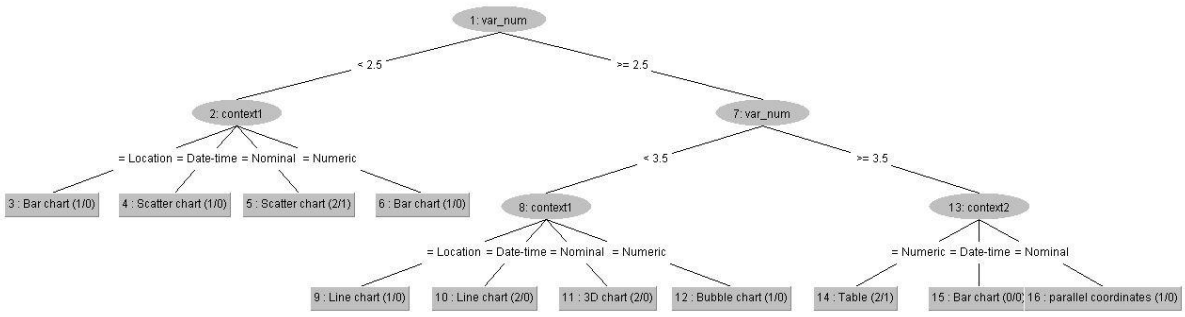


Figure 3.11: Decision Tree with Initial Data

### 3.6.3 Flow of module

First of all, module trains the collected training data to predict from them. Then right after test data provided based on decision tree algorithm PHP-ML will provide the prediction to the data. Currently, prediction algorithm will go for two iterations to get two recommendation predictions. For the second iteration training data set will be changed. Because we need two different suggestions therefore given chart suggestion need to remove before train the data according to the second iteration.

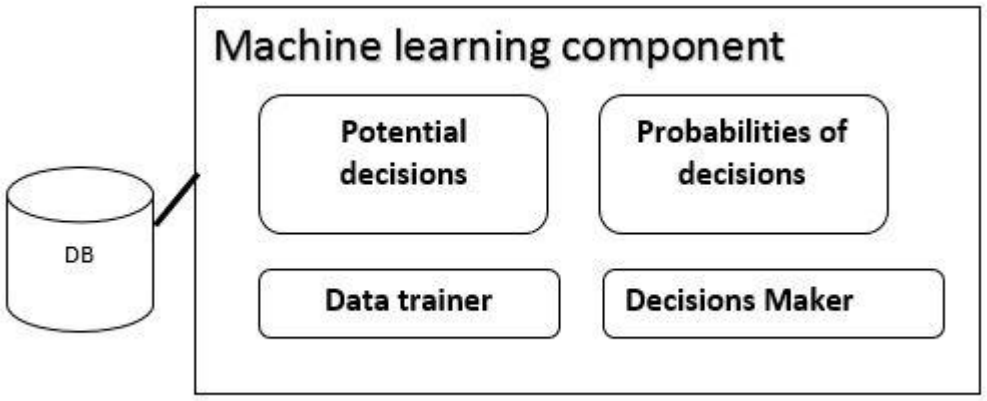


Figure 3.12: ML Module

In the code first, need to set the classifier model. Classifier modal will set by using the PHP-ML libraries. Then need to divide the training features and the labels from the training data. Then using both features and labels train the training dataset using decision tree classifier. Then using the input features predict the suggestion. Then remove the rows which contain label of provided recommendation. Then train the new dataset and apply those input features to get the second recommendation.

Here in this system, we have only gone for two iterations, but if we need more predictions, we can run the same operation for more and more times to get the predictions.

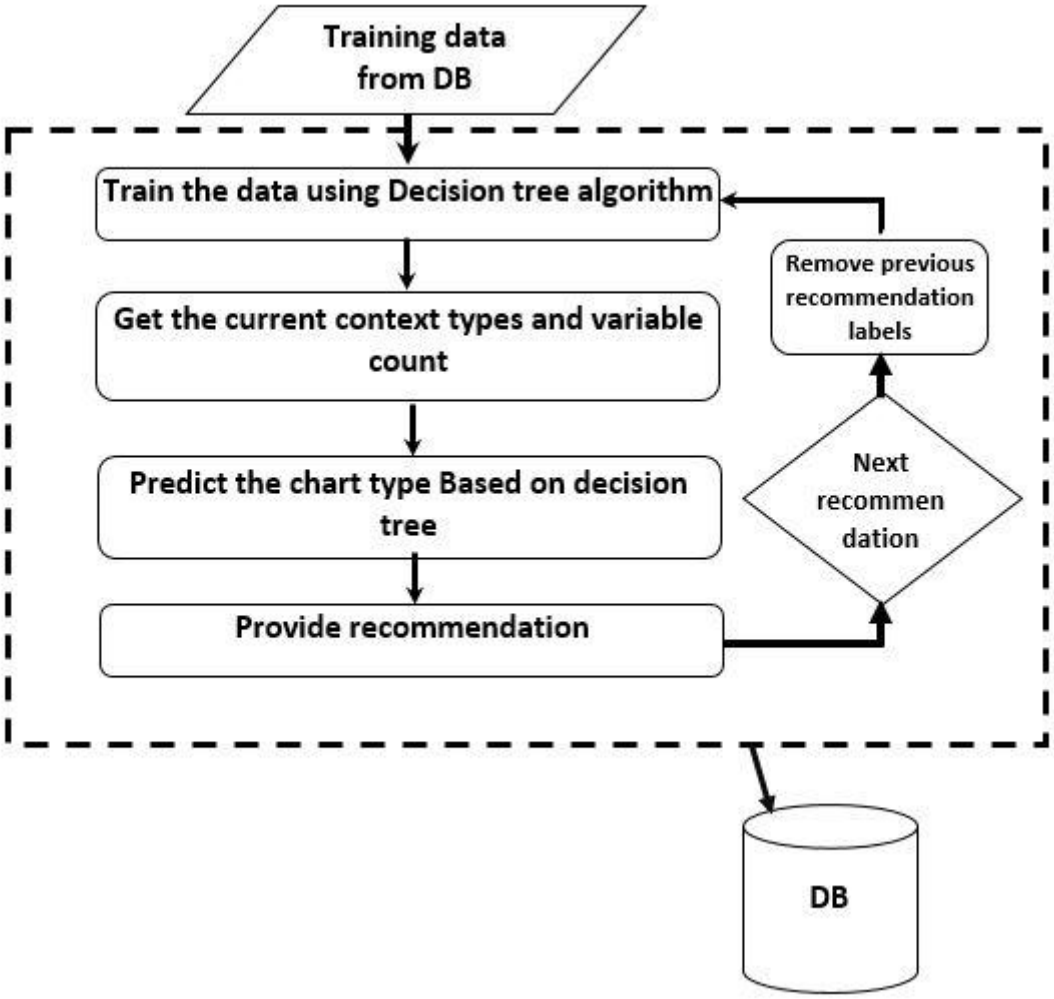


Figure 3.13: ML Recommendation Process

Later both recommendations using this component and the rule-based component combine to get final recommendations. Then users are confirming the suitable recommendation according to their satisfaction. Those confirmations will be used to improve the training dataset. Then feedback data also can use in next test runs to improve the accuracy of the machine learning module.

Here we can see how important the user feedback in these type of systems. This system includes a user interactive process and user Satisfaction should be mainly considered. In the beginning, the accuracy of the machine learning module may not be much accurate. Later with more feedbacks from the users, the training data set will be improved, and accuracy may get higher. More users and test cases mean the accuracy of the next test case is getting higher.

## 3.7 FEEDBACK COMPONENT

Feedback component uses to improve the system with user involvement. This component involves from the beginning of the recommendation process.

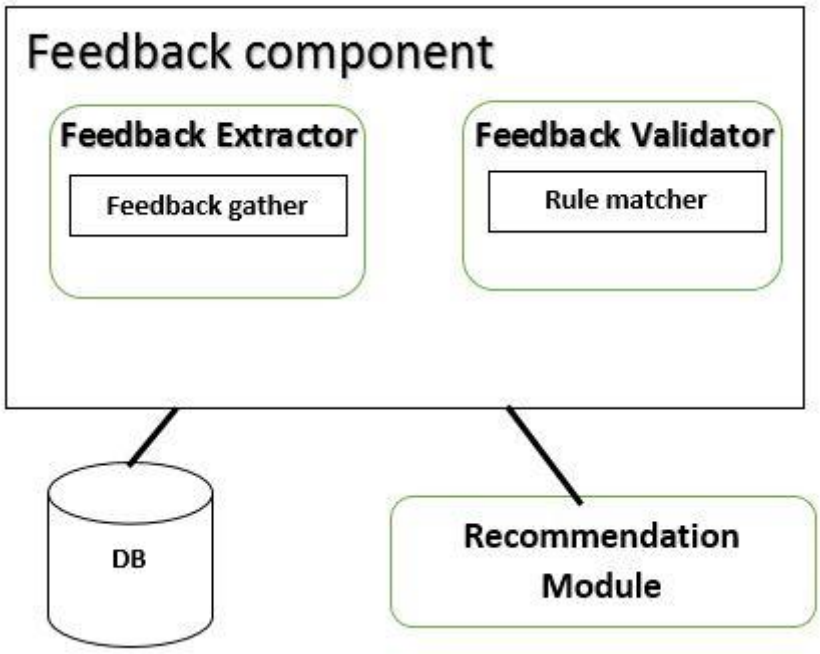


Figure 3.14 Feedback Component

In the feedback component, there are few sub-components which helpful while in the process. Feedback extractor helps to gather the user feedbacks. It provides user interfaces to gather the input data from the user. Then those gathered feedbacks need to go through a validation process. Feedback validator is the sub module which helps to validate those gathered feedbacks. Feedbacks will store in the database after the validation process Therefore they can be helpful to improve the system accuracy.

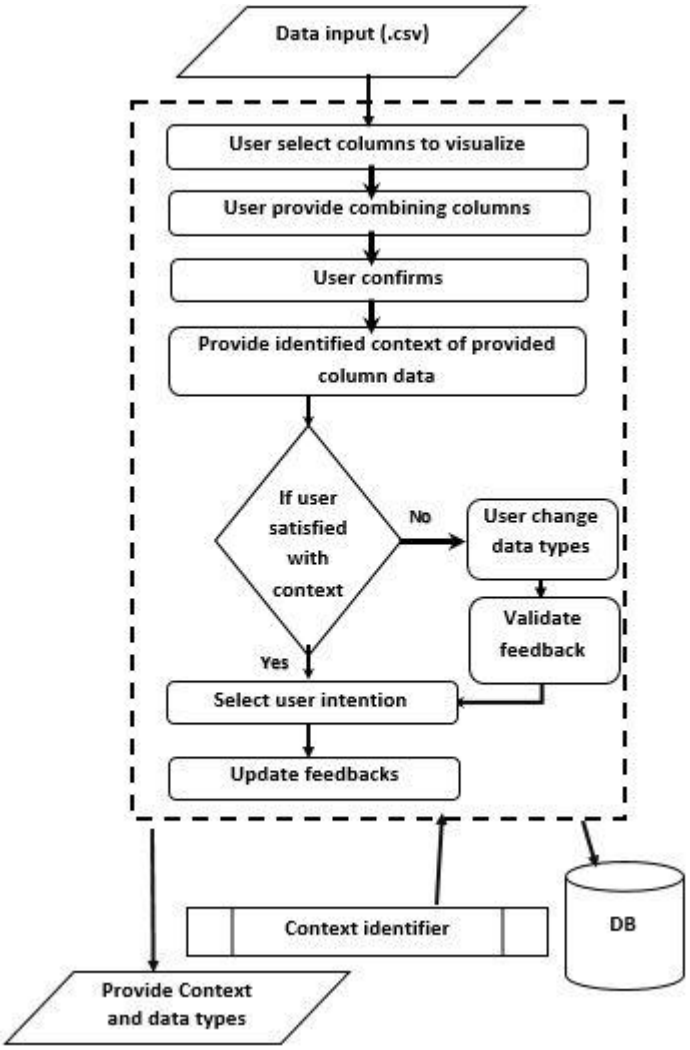


Figure 3.15 Feedback Component in Pre-processing Stages

From the beginning of the process, it depends on the user feedbacks. After uploading the data user will see the dataset in tabular format, and he or she needs to select the columns which going to visualize. Then the user can provide the columns which columns are going to combine. Then after user confirmation identified context data would provide to the user. Here the user can change if those context types are misidentified. With context confirmation user provide the intention also. I context type changed it will go through a validation process. Then database will update to improve the context data and provide final context identification to next stages.

Following are the terminology used in the feedback component used in pre-processing steps in Pseudocode.

- *columns*: user selected columns
- *context*: system identified context

**Pseudocode** Feedback process

**Require:** Input lables, Suggested charts
**Ensure:** Gather feedback to improve training data

1. Start
2. Get *columns*
3. Combine *colunms*

```
4. User Confirms
5.
6. Provide context
7. If(user agree with context) then
8.    confirms
9. else
10.    select NEW context
11.    confirms
12. validete context
13. update the database
14.
15. End
```

The first user provides columns for visualizing then the combining columns. Based on those feedbacks system will create a new csv file with selected and combined columns. Then after the user confirms the identified context after a feedback validation process, those data will be added to the database.

Later in recommendation stages feedback component plays a major role in improving recommendation quality.

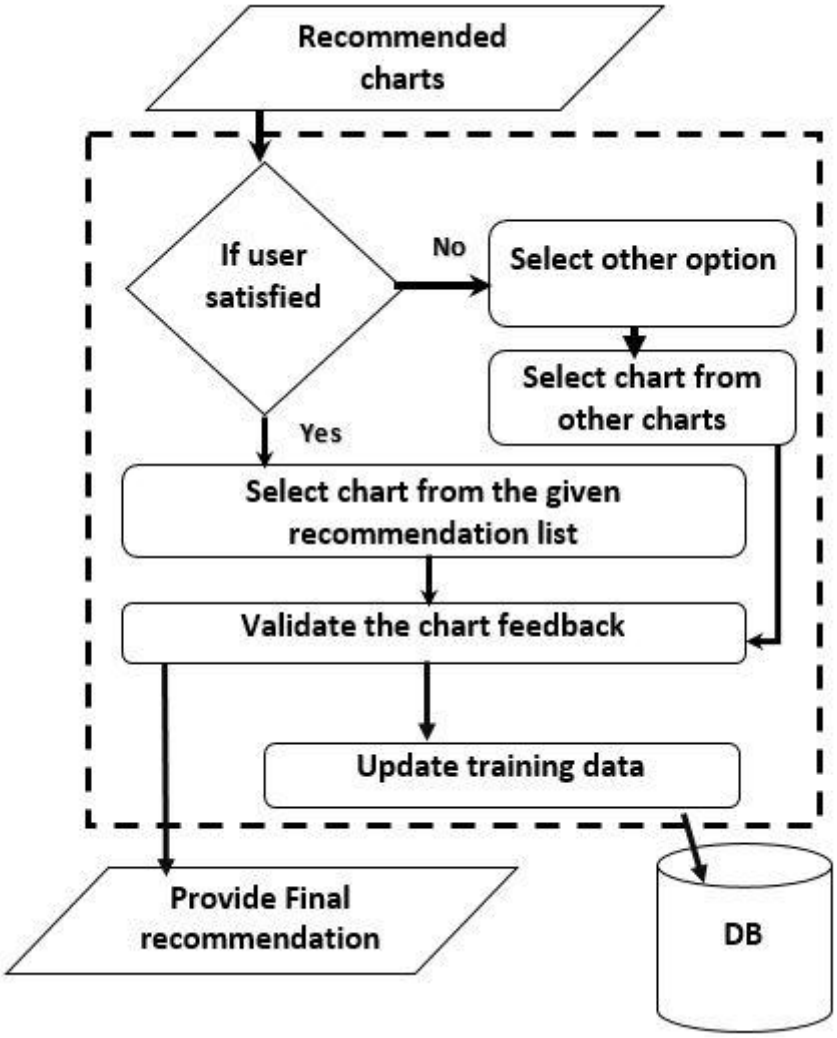


Figure 3.16: ML Improving Process with Feedbacks

After chart recommendations provided, the user can select a chart from the given list. IF the user satisfied with a given recommendation then select the most suitable selection to the user. Then the based on selected chart chart options will appear. After the user is given the correct options and confirms visualization will appear. Like wise user can select other chart types and compare each other. If the user did not satisfied with given recommendations, there is an option called "Other." Then the user can select other option to check other chart types and can visualize by changing the provided chart options. Based on user satisfaction user can confim the chart type. After valisate, those feedbacks data will be added to the training dataset to improve the ML algorithm.

Following are the terminology used in the feedback component used in the recommendation module in Pseudocode.

- *test_features*: features data set of the test case
- *final_suggestion*: feedback confirmation of the user
- *trained_data*: trained features and lables

**Pseudocode** Feedback process

**Require:** Input lables, Suggested charts

**Ensure:** Gather feedback to improve training data

```
1. Start
2. Get test_features and final_suggestion
3. If(agree) then
4.   confirms
5. else
6.  select NEW final_suggestion
7.  confirms
8. Get trained_data
9. add test_features and final_suggestion
10. update database
11.
12. Ends
```

In the first step get the features of the current test case and the given final suggestion. If the user agrees with suggested recommendations, the user confirms. If not set other selection as final suggestion. Then after validate, new suggestion get the training data and update with a new record to improve it.

## 3.8 INTEGRATION OF RECOMMENDATIONS

The final recommendation of the system is based on both machine learning and rule-based combination. Here in the system main priority is given to the machine learning component. First get recommendations from the ML component. Then get the suggestions from the rule-based component. After combining both recommendations, the system will generate least two and most four recommendations as final ones.

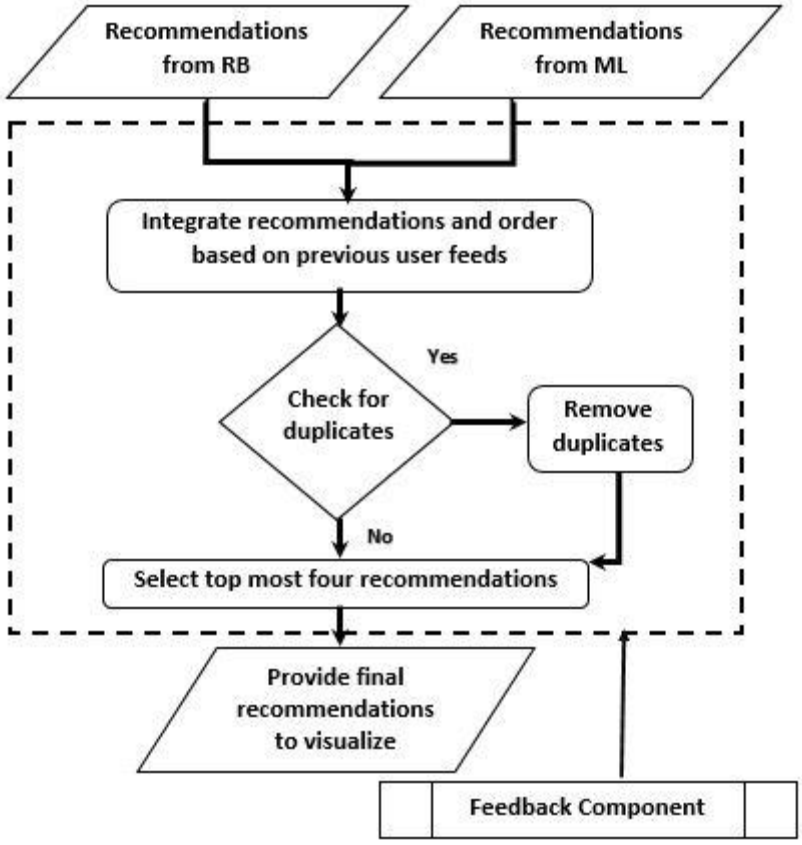


Figure 3.17: the Hybrid Process of Recommendation

Here you can see the design of the Integration module.

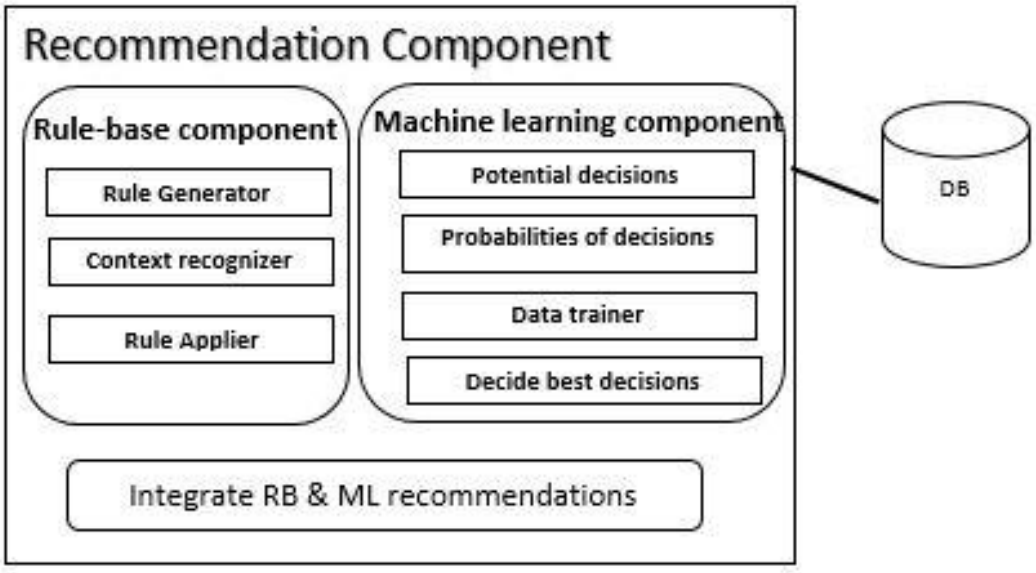


Figure 3.18 Recommendation Integrate Component

Recommendation component consists of both Rule-based and machine learning components with the integration component. The rule-based component provides recommendations based on the provided rule set, and the machine learning component provides recommendations based on the training dataset in the database. Finally using integration sub-component, it gives final ordered list based on previously provided feedbacks and select fist four from the given list as final recommendations.

## 3.9 VISUALIZATION COMPONENT

In data analytics, users are required to understand and analyze the collected data that are supported by data visualization. Charts give us an abstract view of the data, and it will be easier to understand the nature of the data. This visualization component is used D3 JavaScript, which is embedded in to a web page. This requires loading some D3 libraries to work with D3 [109]. The user can customize the charts. After data is well-formed D3 library can give visualization output. Chart is loaded in to specific  tag with specific id. The system will render the chart in to specific div tag. This component supports many chart types such as bar chart, histogram, line chart, scatter chart, pie chart, area chart, column chart, 3D Area chart and bubble chart. As the main outcome of the tool, this component provides a set of ordered recommendations for chart types. The order of the chart recommendations is based on the training data and the mapping function with pre-defined knowledge. This order can be changed based on the current user perspective.

The user needs to confirm the recommendation or select a suitable chart type from another chart list as the feedback to the system. Those are the feedbacks which will be used to improve the machine learning training date set.

### 3.9.1 Visualization process

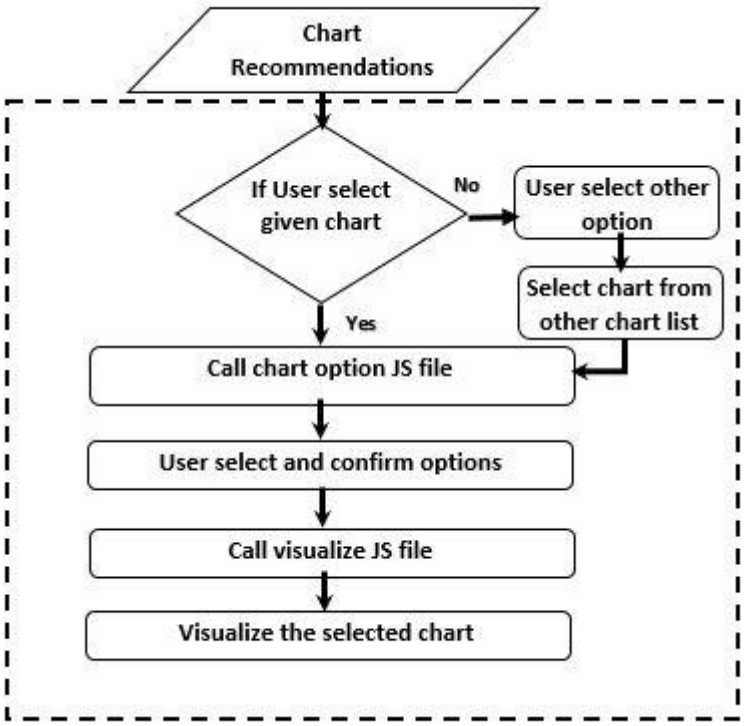


Figure 3.19: Visualization Process

In this module, there are different options for different charts. Some of the options are axis selection and series selection. If we take line chart as an example, there are x axis selection and y axis selection. Then need to select the series which are going to present using the line chart. The user can interactively change and check for the correct visualization.

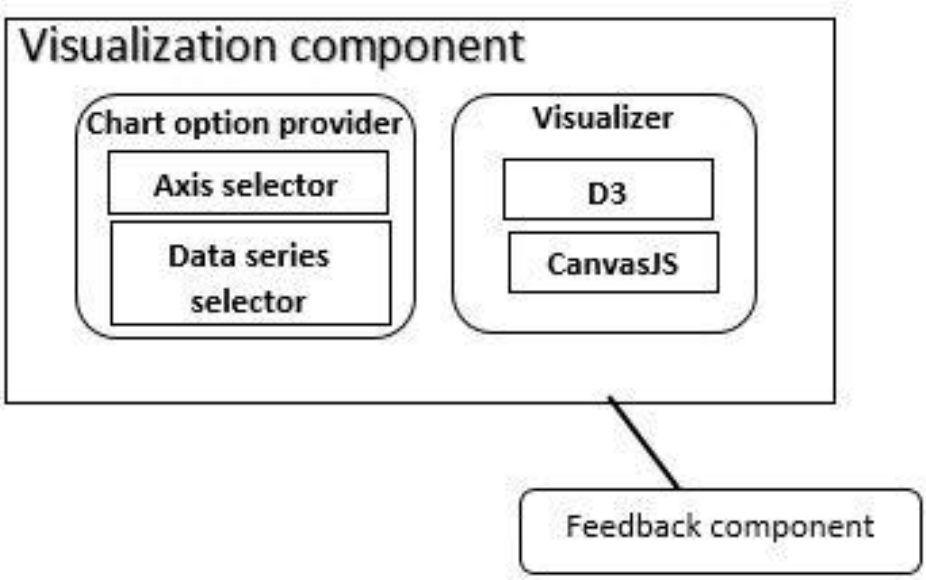


Figure 3.20 Visualization Component

In the visualization module, there are two sub-modules. First is chart option provider. This component works based on the chart_op.js file. There are different options for each chart type. Related to the provided chart type chart option provider responsible for giving the right options to the user. Then the Visualizer module is responsible for giving the visualization based on the selected chart and the given options. Through visualization js, it will load the correct chart in to a given .

### 3.9.2 Designing Visualizations

Data visualization plays an important part in data analysis. People need to understand the data to come up with conclusions from them. Data visualization is the main thing that helps people to understand the significance of the data in its visual context: patterns, trends, and correlations that may not be detected in text-based data in a graph or a chart [9] [110].

When it comes to today data are generating much faster than earlier. Therefor those analyzing methods also can be changed a bit. We can see that visualization techniques can be used to visualize the data based on their behaviors. For a few examples, we can use different colors, different shapes or different size to differentiate the data. Therefor viewer can understand the differences between those data easily [111].

There are attributes which are important to consider while designing a visualization.

- Color
- Length
- Width
- Orientation
- Shape
- Size
- Position

**Color**

In data, visualization coloring holds a significant part. Colors can differentiate those different categories easily. Based on different colors human eye understand them easily. For example, if we think the pie chart, we give different colors to those different areas Therefore we can identify those areas differently. Another example is a line chart. We give different colors to lines to differentiate those lines among others.

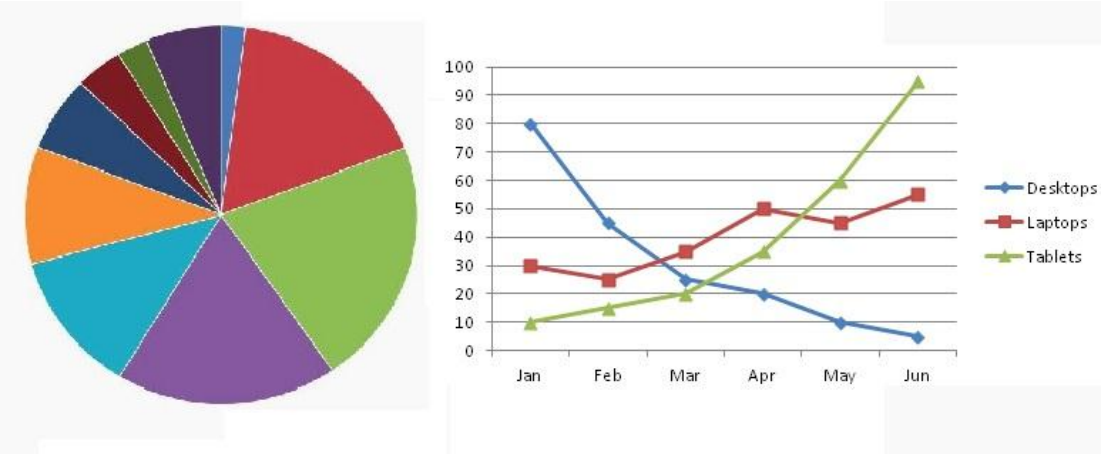


Figure 3.21: Color related visualizations

**Length and width**

Length and width also using widely when it comes to representing data. As an example get a bar chart. Length of the bar chart basically represents the amount of data. Based on different widths we can easily identify those higher data and the lesser data. Also in the Gantt charts, we can see the time taken by the action based on the width of that element.

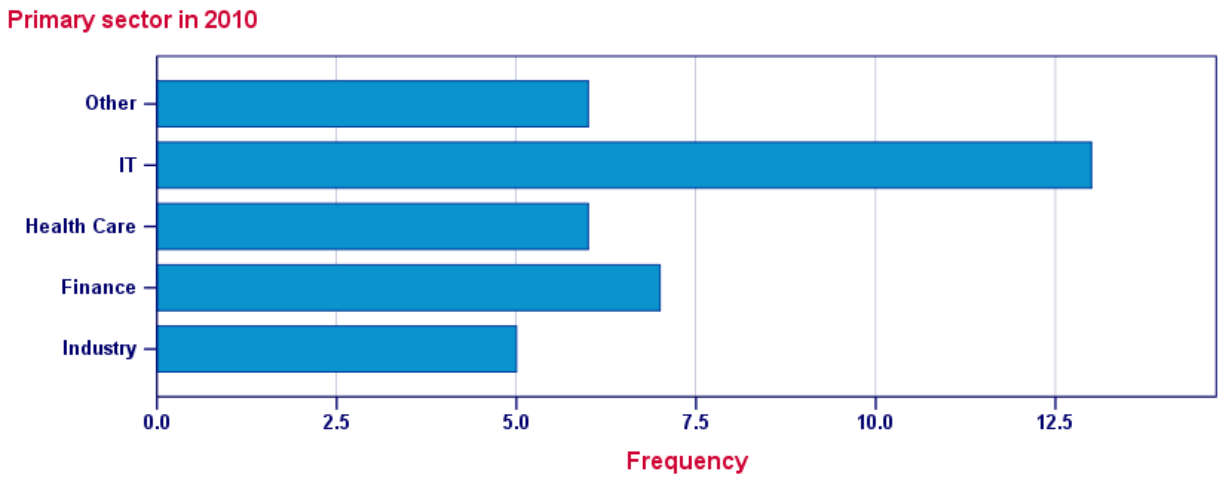


Figure 3.22: Length and width related visualizations

**Orientation**

In visualization orientation of that visualization can make it harder or easier to understand the data. For example a pie chart with a key. Therefor you can understand those categories by looking at the key. If those categories were written on the graph itself, it can be overlapped and harder to understand the data. Also in Area or line charts need to give proper descriptions about the data to understand them.

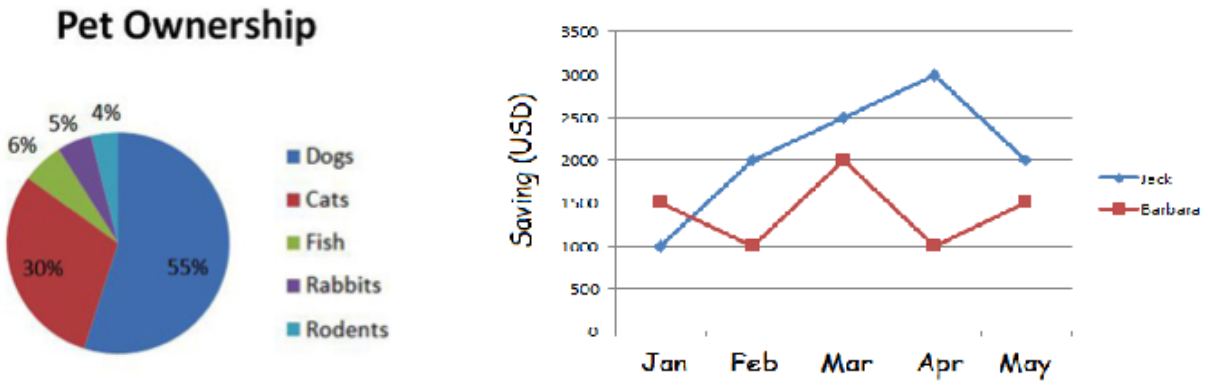


Figure 3.23: Oriented visualizations

**Shapes**

In the visualization to show the deference between categories of the data, we can use a different type of shapes also. For example, in a scatter plot we can use a different type of shapes like squares, triangles or circles. Therefor user can differentiate those data very easily.

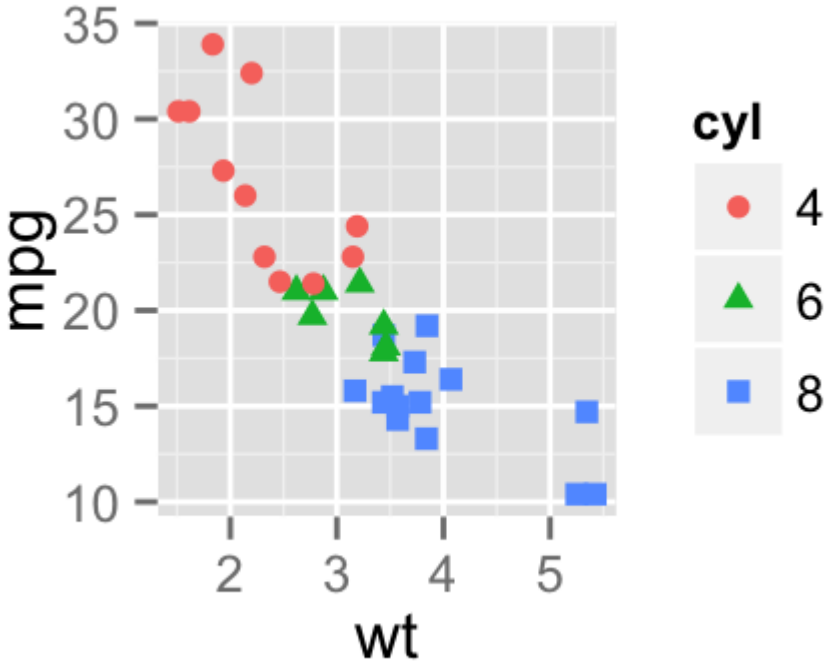


Figure 3.24: Shapes related visualizations

**Size**

Different sizes of the representation may address the value of those data. For example bubble charts bigger the bubble it will represent higher value and smaller the radius of the bubble it will represent smaller values.

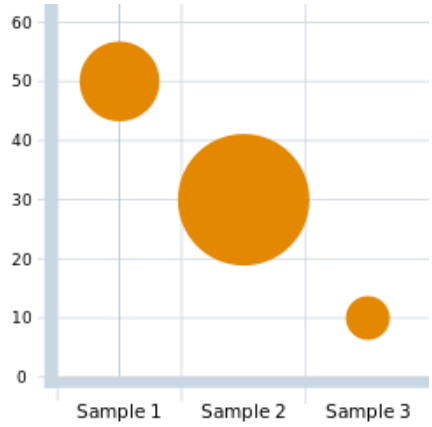


Figure 3.25: Size related visualizations

**Position**

The position of the data visualization also important when understanding the data representation. Here, for example, let's take scatter plot with x and y co-ordinates. In scatter plot x and y values will

give exact point to those data, and those values itself will represent the data. Also in the 3D plot here also 3 values altogether give the position to the data to represent them.

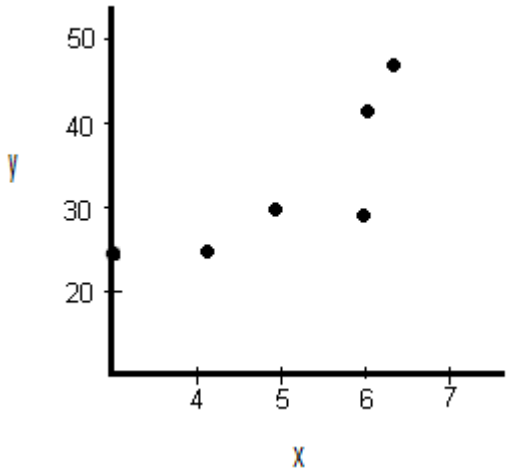


Figure 3.26: Position related visualizations

Moreover, we can categorize the visualization based on the following categories also.

- Points
- Lines
- Areas
- Angle

In a scatter plot, we can use points to show the data values and its representation. If we are drawing a chart which changes with the time that variation can be shown using lines. Ares of the chart can be used to represent data. Based on area data value can be changed and the human eye can easily identify area deference. Then in the pie or doughnut charts, we mainly use the angle of the data representation to get the amount of data. Based on the angle we can decide the data value amount or the percentage.

## 3.10 FEEDBACK VALIDATION

There are different types of feedbacks. Such as user gives their feedbacks intentionally or unintentionally. Therefor the outcome can depend on those feedbacks. After every test run, we use those feedbacks to improve the system. If those given feedbacks are not valid accuracy of the system also gets lower.

Therefor we need to concern above situation and have to come up with a solution. Here as the solution training, data will not affect right after giving feedbacks. First, they will store the data in

the database before the improvement of the training data. Then validate those data based on a number of variables and context type and then only add the data in to training data set.

# Chapter 4

# IMPLEMENTATION

## 4.1 INTRODUCTION

This chapter briefly explains the key implementation details. First, explain the installation of the machine learning module in to the system. Then will discuss the other main module rule-base component. Then other sub components such as grouping methods, visualization component, and context identification component also will be explained.

## 4.2 MACHINE LEARNING MODULE

### 4.2.1 Setup the module

Machine learning component plays a major role in this system and before start with the module we need to configure the PHP-ML library in to the system.

In this system, we have used PHP as the programming language because it comes easier when deploying the system in any environment. No need of installing new modules or setting up the environment. Therefor we selected PHP-ML as our machine learning library to do our ML operations. PHP-ML is a fresh approach to Machine Learning in PHP. Algorithms, Cross Validation, Neural Network, Preprocessing, Feature Extraction and much more in one library.

Before starts the machine learning process we need to configure a few things.

- Import libraries using composer
- Update PHP version to 7.1

First of all, to run the machine leering in the system we have to create library files and paths exactly PHP-ML needed. For that, we need few commands using composer in the command line. Following is the composer command first need to run in the directory where PHP-ML code included [116].

```
composer require php-ai/php-ml
```

After running this command library path and folders will copy in to the system, and now we can run the PHP-ML code.

Running the PHP-ML code also need to be done using the command line. It needs to run PHP code using the command line. Also, need to pass the values as arguments to the file.

```
PHP ml.php argument1 argument2
```

This command will run the machine learning code. Then the exe command will be used to run the machine learning file because it needed to be run using the command line.

```
exec(PHP ml.php argument1 argument2)
```

Updating the PHP version is easier if cpannel of the host provided that functionality. If not we can use the command line to update the PHP version in the server.

```
sudo apt-get install php7.0
```

### 4.2.2 Module Pseudocode code

Following are the terminology used in Pseudocode.

- *training_features*: features data set of the trained data
- *training_lables*: labled data of trained data
- *input_features*: data features needed to classify

**Pseudocode 2** Machine Learning

```
Require: Input Dataset, the Training data set
Ensure: Select a suitable chart type
1.  Start
2. set classifier DecisionTree
3. train(training_features, training_lables)
4. predict(input_features)
5.
6.  remove(lables_with_predicted)
7.  get(new_training_data)
8.  train(new_training_data_features, new_training_data_lables)
9.  predict(input_features)
10.
11.   End
```

Now we have the next best recommendation for the given text data. Likewise, we can go for more iterations to get more recommendations. If we are going for the third recommendation, we need to remove the rows related to both recommendation1 and recommendation2.

## 4.3 RULE-BASED APPROACH

Rule-base systems usually define outcomes or suggestions based on previously collected knowledge. Here in this chart recommendation process rules are defined based on the intention of the visualization, a number of variables, data types and constrains we previously defined.

### 4.3.1 The theory behind the Rule-base representation

Rules are mainly considered the intention of the visualization and number of variables in the dataset. Later based on possible data types and predefined constrains all the outcomes will be ordered.

- IF intention == “relationship” AND variables == 2 ORDER BY (datatypes == {Numeric} AND constraint = {C1,C2 })

THEN chart = { Scatter plot}

- IF intention == “relationship” AND variables == 3 ORDER BY (datatypes == {Numeric} AND constraint = { C3, C2})
  THEN chart = { Bubble chart}
- IF intention == “relationship” AND variables == 2 AND datatype=INCLUDE (Date-time) ORDER BY (datatypes == { Date-time, Nominal } AND constraint = { C4, C6})
  THEN chart = { Gantt chart }
- IF intention == “relationship” AND variables == 2 ORDER BY (datatypes == {Norminal} AND constraint = { C4, C7})
  THEN chart = { Tree diagram }
- IF intention == “relationship” AND (variables == 2 OR variables == 3) ORDER BY (datatypes == {Norminal} AND constraint = { C5, C8})
  THEN chart = { Network diagram }
- IF intention == “relationship” AND variables ==2 ORDER BY (datatypes == {Norminal} AND constraint = { C1})
  THEN chart = { Hierarchical edge binding }
- IF intention == “comparison” AND variables =>2 ORDER BY (datatypes == { Numeric, Nominal, Location, Percentage } AND constraint = { C4, C10})
  THEN chart = { Parallel coordinates }
- IF intention == “comparison” AND variables ==2 ORDER BY (datatypes == { Nominal, Numeric } AND constraint = { C4, C11})
  THEN chart = { Bar chart }
- IF intention == “comparison” AND variables >2 ORDER BY (datatypes == { Numeric, Nominal, Location, Percentage } AND constraint = { C4, C10})
  THEN chart = { Radar chart }
- IF intention == “comparison” AND variables >2 ORDER BY (datatypes == { Nominal, Numeric, Percentage } AND constraint = { C4, C12})
  THEN chart = { Column chart }
- IF intention == “comparison” AND variables >2 ORDER BY (datatypes == { Numeric, Nominal, Location, Percentage } AND constraint = { C5, C13})
  THEN chart = { Table }
- IF intention == “comparison” AND variables ==2 AND datatype=INCLUDE (Date-time) ORDER BY (datatypes == { Date-time, Numeric, Nominal } AND constraint = { C1, C6, C21})
  THEN chart = { Single line chart }
- IF intention == “comparison” AND variables ==2 ORDER BY (datatypes == { Date-time, Nominal, Numeric, Percentage } AND constraint = { C4, C12})
  THEN chart = { Column chart }
- IF intention == “comparison” AND variables ==2 ORDER BY (datatypes == { Date-time, Numeric } AND constraint = { C15})
  THEN chart = { Radial bar chart }
- IF intention == “distribution” AND variables ==3 ORDER BY (datatypes == { Numeric } AND constraint = { C1, C16})
  THEN chart = { 3D chart }

- IF intention == “distribution” AND variables ==1 ORDER BY (datatypes == { Numeric } AND constraint = { C4})
  THEN chart = { Histogram }
- IF intention == “distribution” AND variables ==2 ORDER BY (datatypes == { Numeric } AND constraint = { C1, C2})
  THEN chart = { Scatter plot }
- IF intention == “composition” AND variables =>2 AND datatype=INCLUDE (Date-time) ORDER BY (datatypes == { Date-time, Numeric } AND constraint = { C1, C2, C6})
  THEN chart = { Area chart }
- IF intention == “composition” AND variables =>2 ORDER BY (datatypes == { Date-time, Numeric } AND constraint = { C4})
  THEN chart = { Stacked column chart }
- IF intention == “composition” AND variables =>2 AND datatype=INCLUDE (Date-time) ORDER BY (datatypes == { Date-time, Numeric } AND constraint = { C1, C2, C6})
  THEN chart = { Stacked area chart }
- IF intention == “composition” AND variables ==2 ORDER BY (datatypes == { Numeric, percentage } AND constraint = { C1, C2})
  THEN chart = { Area chart }
- IF intention == “composition” AND variables ==2 ORDER BY (datatypes == { Numeric, percentage, Nominal } AND constraint = { C4, C17, C18})
  THEN chart = { Pie chart }
- IF intention == “composition” AND variables ==3 ORDER BY (datatypes == { Numeric, Nominal } AND constraint = { C4, C20})
  THEN chart = { Sun burst }
- IF intention == “composition” AND variables ==2 ORDER BY (datatypes == { Numeric, Nominal } AND constraint = { C4, C19})
  THEN chart = { Doughnut chart }

In this system we expect more matching rules because we need more suggestions as outputs. For some systems this might be a problem. But in some situations there can be exceptions or no matching rules according to the given data. In those situations we consider the rules without including the intention of visualization.

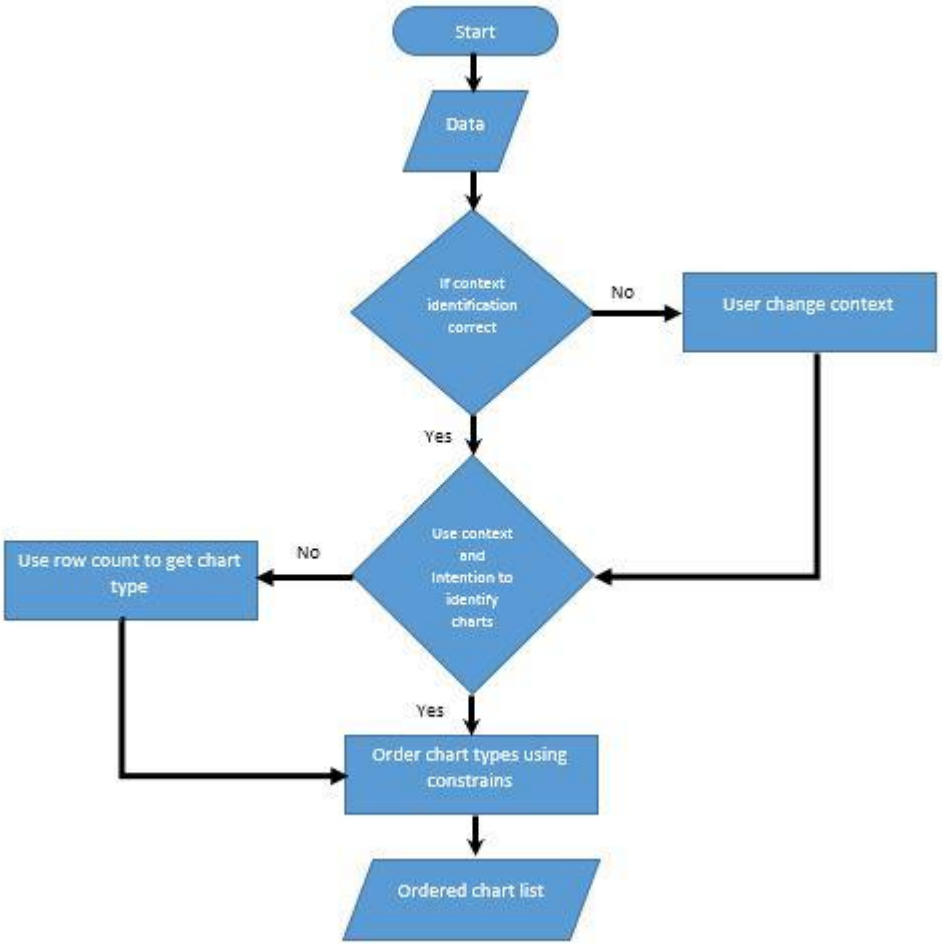


Figure 4.1: Rule-Base Flow

Pseudocode code of the chart selection based on pre-collected knowledge:

- var_num: number of variables after columns are joined
- context_arry: array which contains all the context data related to each column
- time: context type which can be included inside context_arry
- user_intention: intention of the user

**Pseudocode** RB Mapping chart types

**Require:** Input Dataset, Mapped chart types, db table

**Ensure:** Select suitable chart type

**Return:** chart types

1. Start
2. if (*user_intention*=comparision) then

```
3.   if (in_array(time, context_arry)) then
4.     return { Single Line, Multi Line, column, Radial Bar }
5.   else if
6.     return {Table, Bar, column, parallel coordinates, Radar chart }
7. else if (user_intention=composition) then
8.   if (in_array(time, context_arry)) then
9.     return { Area, Stacked column}
10.   else if
11.     return {Pie, Area, Doughnut, Sunburst }
12. else if (user_intention=relationship) then
13.   if (var_num=2) then
14.         return {Scatter, Hierarchical edge binding, Gantt }
15.   else if (var_num=3) then
16.     return {Bubble, Network, Tree}
17. else if (user_intention=distribution) then
18.   if (var_num=1) then
19.         return {Histtogram}
20.   else if (var_num=2) then
21.     return {Scatter}
22. else if (var_num=3) then
23.     return {3D chart}
24.
25.  End
```

Code of the system will work based on extracted rules from the database. First, it will look for the user intention. Based on user intention there are different choices. For that system looks for temporal variable existence and the number of variables. Based on those data next step of the rules will apply to get the final recommendations.

Table 4.1: Intention of data representation

| | Comparison | Relationship | Composition | Distribution |
|---|---|---|---|---|
| definition | Set one set of variables apart from another, and display how those two variables interact | Show a connection or correlation between two or more variables | Collect different types of information that make up a whole and display them together | Collection of related or unrelated information simple to see how it correlates |
| Features/ properties | It places objects to be compared in the same space Therefore that | Constrained to either explicit or implicit relations | Displaying data in several different ways may benefit user cognition. | Easy to extract data from distribution fields. |

| | | | | |
|---|---|---|---|---|
| | differences can be detected as a low-level visual feature | | Each representation allowing the user to focus on different aspects of the data | |
| Type of data | Dynamic data | Complex data | Primitive data, recursive data | Desecrate and continuous data, Spatial data |
| Possible chart types | Bar chart, column chart, Radar chart, Line chart, Table | Scatter, Bubble, Tree, Network | Pie chart, Sunburst chart, Doughnut chart, Area chart, Stacked column chart | 3D chart, Scatter plot, Histogram |
| References | [117] | [118] | [119] | [120], [121] |

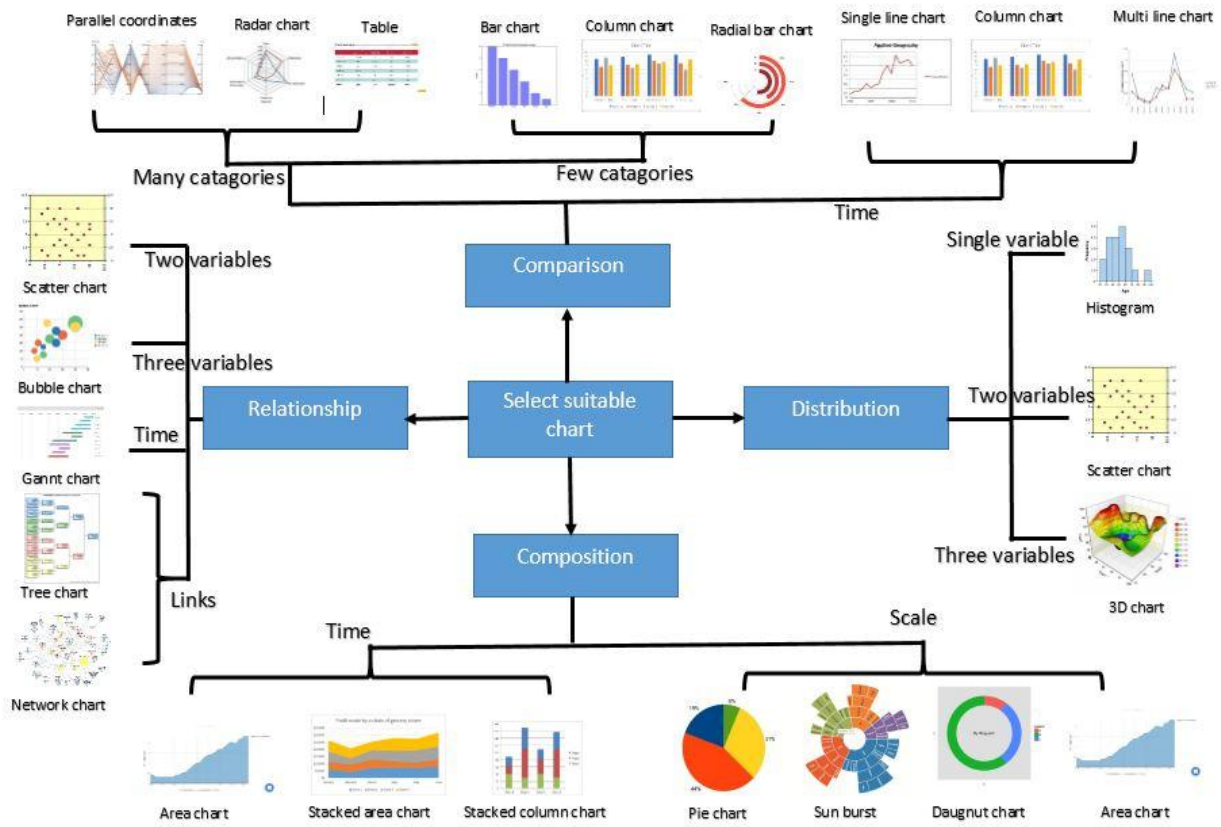


Figure 4.2 Chart Selection diagram

Above diagram shows all chart types consider in the rule-base in the system. They are mapped based on user intention available data types and also based on a number of variables.

## 4.4 INTEGRATING RECOMMENDATION

Recommendation module consists of two parts. They are rule-based component and machine learning component. When integrating recommendations, common recommendations will consider mainly in the process.

Following are the terminology used in Pseudocode.

- ml_recommendations : Recommendations given by machine learning component
- rb_recommendation: Recommendations given by rule-based component

```
Pseudocode Recommendation
Require: Input Dataset, context data, database
Ensure: Find suitable data types
1. Start
2.
3. get(ml_recommendations)
4. get(rb_recommendation)
5. remove_duplicates(ml_recommendations, rb_recommendation)
6.
7. get(previous_feed)
8. order using previous_feed
9. return recommendations
10.
```

```
11. End
```

## 4.5 COLUMN GROUPING

Some datasets come with multiple columns which need to be grouped together, such as address fields, date fields, etc. Those grouping can be categorized based on the type of data included. Following are the considered categories.

- Group as string using dash (-)
- The group as a string using a slash (/)
- The group as a string using a comma (,)
- Arithmetic addition
- Average

In this grouping, step user can provide the column numbers of grouping, type of the grouping and the new header name of the grouped data.

## 4.6 VISUALIZATION COMPONENT

Here in the visualization component works mainly based on two steps. As the first step right after the user selects the chart type component loads options based on the selected chart. In the next step, the component will load the provided dataset based on the selected chart and options provided.

### 4.6.1 Chart option provider

The first user needs to select a chart type from given recommendations. Then chart option provider's task is to give suitable options for the selected chart type.

For example, let say user selected bar chart as a chart type and in the next step the module will give the suitable options to select.

In a Bar chart

- Select Y axis
- Select X axis (bars)

Here dropdowns will appear to select suitable columns in the dataset. After the user selects, the options process will move to the next step.

### 4.6.2 Chart visualizer

After the user selects the correct options next module will load the dataset within the selected chart. There are html files for each chart type in the system. The module will pass the selected options and the dataset into those files and load that html into a specific  in the code to display the visualization.

### 4.6.3 Pseudocode for visualization

Following are the terminology used in Pseudocode.

- selected_chart : User selected chart
- select_opt: User selected chart options

```
Pseudocode Visualization
Require: recommended charts, the data set
Ensure: visualizing the charts
  27. Start
  28. get(selected_chart)
  29. call chart_options
  30. load select_opt
  31.
  32. Visualize(data, selected_chart, select_opt)
  33. return visualization
  34. End
```

The first user selects the chart from the given list or from the other chart list. Then system calls chart options related to the selected chart type. Then forward the user selected options to the visualization module along with the selected chart and the data to be visualized. Finally, visualization will provide based on the provided data to the module.

## 4.7 BACKEND ADMIN CONTROLLERS

In this backend, there are a few modules which help the main process of the system. Here we can add a new chart to the system, add a new rule to the rule-base and can view details of previously added datasets.

### 4.7.1 Chart adder

The main option of this module is adding new chart types for the system. Here we can upload the HTML file related to a specific chart. Moreover, here we provide the editing options to the visualization mapping js file and the chart option provider js file.

Uploading chart html files are mainly followed D3 javascript library to generate graphs. Bacically those HTML will be loaded in to specific  to display the data of the selected dataset.

### 4.7.2 Rule adder

In this sub module mainly responsible for adding new rules to the system. When adding a new chart to the system, there should be a specific rule or rule set for that chart. Therefore this is where those rules are adding to the main system. Rules will consider the data types of a data set and a number of variables user intention and previously defined constrains. With those conditions, we can add a new chart in to the system.

### 4.7.3 Dataset viewer

Here we can view all the test runs related to previously uploaded datasets. After selecting a dataset, the results related to selected dataset will be displayed. It will include a final recommended chart, if that chart is another chart or not, processing time, cpu usage, memory usage and time of the test run.


Figure 4.3: Elapsed time for test runs

Finally, the variation of the Elapsed time is displayed with each test runs. Therefor we can determine the average time taken to process the dataset.

Here we can change the dimensions to check how other variables vary with the time. Here we provide Elapsed time, cpu usage and memory usage as other variables. Furthermore, we can change the color of the line as well.

**Elapsed time**

Here the time is calculated for each process in the system and sum those times altogether to come up with a final elapsed time. In the above graph, we recorded the Elapsed time respect to the recorded time of each test run related to the selected dataset.

Data_uploading_time= time taken to upload the data

Preprocessing_time= time taken to preprocess and arrange data

Context_identify_time= time taken to identify the context of the data

Recommendation_time= time taken by the recommendation module

Visualization_time=time taken to visualize the selected graph

Elapsed_time = (Data_uploading_time + Preprocessing_time + Context_identify_time + Recommendation_time+ Visualization_time)

**Memory usage**

In PHP we can calculate the maximum memory usage in each process. After each process, we calculate the current memory usage. Later we compare all those usages and get the maximum memory usage as the memory usage during the whole process.

**CPU usage**

Like in memory usage we calculate CPU usage in every process and look for the maximum percentage of the CPU usage while in each process. Maximum CPU usage will be considered as the CPU usage of the system.

# Chapter 5

# EXPERIMENTS AND RESULTS

This chapter details the experiments, results, and analysis with regards to the research carried out.

## 5.1 USABILITY STUDY

We can evaluate the tool using the System Usability Scale (SUS) method. Questions will be asked based on user experience to study the usability of the tool [113].

The system usability (SUS) is a reliable way of measuring the usability. This method will call as a "quick and dirty." There are 10 questions based on the usability of the system. There are five response options for those questions from strongly agree to disagree strongly.

There are assign values for each question based on positivity or negativity of the question. Those 10 values will be added together to come up with one value. Finally, that value will be multiplied by 2.5 to convert the scale from 0-40 to 0-100.

There are many types of research to find out about this SUS score. They have found that score above 68 considers as above average and below that will be below average. This scores can normalize to produce a percentile ranking; therefore, we can interpret all scores as one.

Following are the questions.

1. I think that I would like to use this system frequently.
2. I found the system unnecessarily complex.
3. I thought the system was easy to use.
4. I think that I would need the support of a technical person to be able to use this system.
5. I found the various functions in this system were well integrated.
6. I thought there was too much inconsistency in this system.
7. I would imagine that most people would learn to use this system very quickly.
8. I found the system very cumbersome to use.
9. I felt very confident using the system.
10. I needed to learn a lot of things before I could get going with this system.

We have evaluated the tool for System Usability Scale (SUS), with 20 users in the age group of 20-30 years, who have basic data visualization knowledge. An average SUS score of 75.125 out of 100 was observed for the usability of the presented tool in this paper. In the open-ended questionnaire, the majority of the users have agreed with the advantages of the tool. Among the positive reviews, the users have indicated that the tool is easy to use as all the steps are automated. Figure 5 shows the n-gram opinion analysis for the statements relevant to the positive reviews of the tool. Here, 86% of the users have mentioned the tool is easy to use; 81% of the users have satisfied with the tool function integration and felt confident while using the tool; 77% of users have stated that the tool can be learnt to use quickly and 58% of users have liked to use the tool frequently.

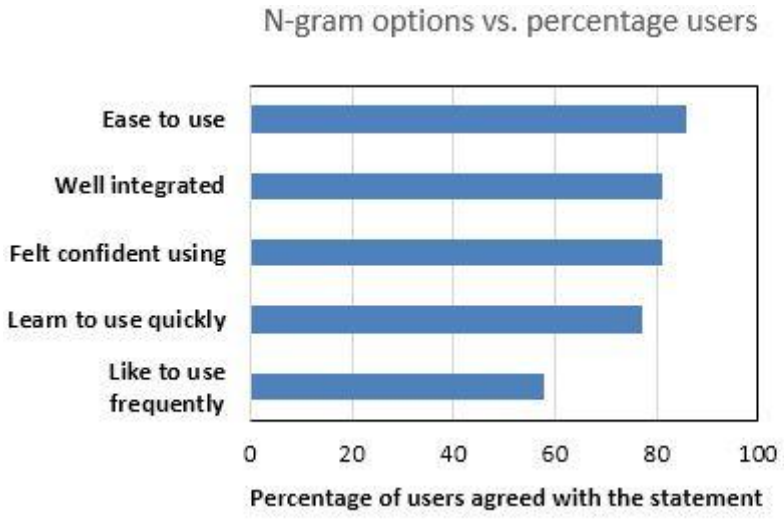


Figure 5.1: Analysis of N-gram opinion

When considering the user feedback for the negative statements of the tool, only one user has mentioned that the tool is complex to use and need some prior knowledge. None of the users have agreed with the remaining open-ended questionnaire, that includes the users need technical support, the system is inconsistent and cumbersome. Thus, the SUS gives positive results for the usability of the tool.

## 5.2 THE ACCURACY OF THE SYSTEM

Accuracy of the system can be measured as one whole system. Rather looking as a whole system, we can look component wise accuracy. We can look for the accuracy of the recommendation component with and without the machine learning sub component.

Likewise the feedback module with and without using feedback data to improve the training data. From that, we can ensure how important it is the feedback module and the machine learning module in to this system [113].

For calculations, we can use the following equations.

Precision = |{relevant information} ∩ {retrieved information}| / |{ retrieved information} |

Recall = |{relevant information} ∩ {retrieved information}| / |{relevant information} |

F1 = 2 * (Precision * Recall)/(Precision + Recall)

**Datasets**

Following are the datasets that we selected to evaluate the visualization outcomes together with user feedback.

Table 5.1: Datasets and features

| | Data set name | Number of rows | Column names included in the dataset | URL or reference/year |
|---|---|---|---|---|
| 1 | YouTube rating dataset with details on category, view, like and dislike data for videos | 7998 | Category/ nominal<br>Views/ numeric<br>Likes/ numeric | https://www.kaggle.com/data |
| 2 | Pit stop data of each driver. This includes the position in the stops and duration and time. | 6252 | DriverID/ nominal<br>Stop/ numeric<br>Lap/ numeric<br>Duration/ time | https://www.kaggle.com/data |
| 3 | Formula race data that shows each player's position and time for each lap. Data of one racer is considered here. | 58 | driverId/ nominal<br>lap/ numeric<br>position/ numeric<br>time/ date time | https://www.kaggle.com/data |
| 4 | Data related to players. Winnings and losses of each player are considered for the study. | 100 | Name/ nominal<br>Win/ numeric<br>Loss/ numeric | https://www.kaggle.com/data |
| 5 | Data related to FIFA players. This study analyses how the stamina level varies with age. | 17995 | Stamina/ numeric<br>Age/ numeric | https://www.kaggle.com/data |

| | | | | |
|---|---|---|---|---|
| 6 | Cocoa quality related data. | 1796 | Specific bean/ nominal<br>Rating/ percentage | https://www.kaggle.com/data |
| 7 | Crime records for 3 major districts in Sri Lanka (Colombo, Gampaha, Kaluthara). Also, include different types of crimes. | 23 | Crime type/ nominal<br>Crimes Colombo/ numeric<br>Crimes Gampaha / numeric<br>Crimes Kaluthara / numeric | http://www.data.gov.lk/search/type/dataset |
| 8 | Literary percentages for each Province in Sri Lanka. It shows literary percentages of male, female and total. | 9 | Area/ location<br>Male/ percentage<br>Female/ percentage<br>Total/ percentage | http://www.data.gov.lk/search/type/dataset |
| 9 | Accident records for each district in Sri Lanka. | 25 | Districts/ location<br>Total accidents/ numeric | http://www.data.gov.lk/search/type/dataset |
| 10 | Tea export quantities and earned revenue for selected years (2006-2012). | 7 | Year/ date-time<br>Tea-Kg/ numeric<br>Tea-$/ numeric | http://www.data.gov.lk/search/type/dataset |
| 11 | Departure for foreign employment | 18 | Year/date-time<br>Self Basis Male/Numeric<br>Self Basis female/Numeric<br>Through Agencies Male/Numeric<br>Through Agencies Male/Numeric | http://www.data.gov.lk/search/type/dataset |

| | | | | |
|---|---|---|---|---|
| 12 | Exported goods values by type of good | 96 | Year/Date-time<br>2011/Numeric<br>2012/Numeric<br>2013/Numeric | http://www.data.gov.lk/search/type/dataset |
| 13 | Bandwidth Averages for April and may 2018 | 742 | SiteName/Norminal<br>MayAvarage/Numeric<br>AprilAvarage/Numeric | http://www.data.gov.lk/search/type/dataset |
| 14 | Number of registered motor vehicles and motor cars per 1000 people 0 | 3 | Item/Nominal<br>2007/Numeric<br>2008/Numeric<br>2009/Numeric<br>2010/Numeric<br>2011/Numeric<br>2012/Numeric | http://www.data.gov.lk/search/type/dataset |

Accuracy is the closeness of the measured value to standard or known value. Precision, recall, and F1 are widely used accuracy measures. Precision is a number of correct outcomes from all outcomes received. The recall is a number of correct ones among the total number of relevant instances. F1 is the measure that joins both precision and recall. It will be harmonic mean of precision and recall. Table 5.2 summaries the precision, recall, and F1-measures of the obtained recommender data.

Datasets:

1. YouTube rating dataset with details on category, view, like and dislike data for videos
2. Pit stop data of each driver. This includes the position in the stops and duration and time.
3. Formula race data that shows each player's position and time for each lap. Data of one racer is
4. Data related to players. Winnings and losses of each player are considered for the study.
5. Data related to FIFA players. This study analyses how the stamina level varies with age.
6. Cocoa quality related data.

7. Crime records for 3 major districts in Sri Lanka (Colombo, Gampaha, Kaluthara). Also, include different types of crimes.
8. Literary percentages for each Province in Sri Lanka. It shows literary percentages of male, female and total.
9. Accident records for each district in Sri Lanka.
10. Tea export quantities and earned revenue for selected years (2006-2012).
11. Departure for foreign employment
12. Exported goods values by type of good
13. Bandwidth Averages for April and may 2018
14. Number of registered motor vehicles and motor cars per 1000 people

Table 5.2 Accuracy Analysis

| Data set | Accuracy | | |
|---|---|---|---|
| | Precision | Recall | F1- measure |
| 1 | 0.62 | 0.8333 | 0.7109 |
| 2 | 0.7142 | 0.8333 | 0.7691 |
| 3 | 0.8 | 0.8888 | 0.842 |
| 4 | 0.9285 | 0.9285 | 0.9285 |
| 5 | 0.8181 | 0.9 | 0.857 |
| 6 | 0.8333 | 0.909 | 0.8695 |
| 7 | 0.9285 | 0.9285 | 0.9285 |
| 8 | 0.923 | 1 | 0.9599 |
| 9 | 0.9166 | 1 | 0.9564 |
| 10 | 0.909 | 0.909 | 0.909 |
| 11 | 0.9166 | 0.9166 | 0.9166 |

It can be seen that the accuracy level is above 0.8 in most of the cases. The chart recommendations are provided using a combined method of machine learning and rule based components. The methodology

supports complex data processing such as aggregating data columns. User feedback is also used to increase the accuracy of the recommendations. The evaluation has shown an accuracy over 0.8.

## 5.3 TIME CONSUMPTION

The processing time of the system is another major fact that we have addressed. Here we got the different size of data and checked the time that takes to process the data. Time is measured for the time taken to read and process data and to provide suggestions. Full time duration can take as a collection of all following processes.

Data_uploading_time= time taken to upload the data

Preprocessing_time= time taken to preprocess and arrange data

Context_identify_time= time taken to identify the context of the data

Recommendation_time= time taken by the recommendation module

Visualization_time=time taken to visualize the selected graph

Elapsed_time = (Data_uploading_time + Preprocessing_time + Context_identify_time + Recommendation_time+ Visualization_time)

Following table will display the number of rows and execution time taken to execute those rows.

Table 5.3 Time Analysis

| Number of rows (Data record size) | Time (ms) |
|---|---|
| 58 | 1815 |
| 742 | 2023 |
| 1796 | 2163 |
| 7998 | 2415 |
| 17995 | 2989 |

The following chart will show the variation of the processing time with the number of records.

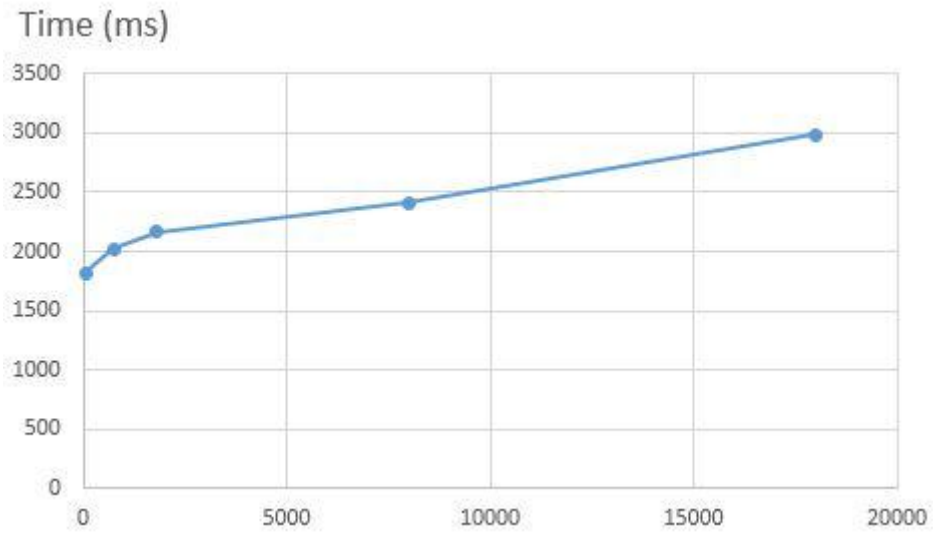


Figure 5.2 Processing times

According to Figure 5.2, the processing time is increased with the growth of the data set. When the feedback count is increased, the system requires more database access time and processing time. Thus the recommendation algorithm can be improved to reduce the processing time with the increase of the dataset records.

## 5.4 MEMORY AND CPU USAGE

Memory and the cpu usage of each process can be calculated. Following codes can be used to get those values in PHP.

**Memory usage**

```
memory_get_peak_usage()
```

**CPU useage**

```
function current_cpu_usage() {
  $cmd = 'typeperf  -sc 1  "\Processor(_Total)\% Processor Time"';
  exec($cmd, $lines, $retval);
  if($retval == 0) {
    $values = str_getcsv($lines[2]);
    return floatval($values[1]);
  } else {
    return false;
```

```
        }
    }
```

Following table shows calculated memory usages to our previous datasets mentioned when testing the accuracy of the system.

| Dataset | Size (Rows) | Memory usage | CPU useage |
|---|---|---|---|
| 1 | 7998 | 2173208 B | 35.717556% |
| 2 | 6252 | 1922920 B | 47.562425% |
| 3 | 58 | 469280 B | 6.887847% |
| 4 | 100 | 478392 B | 20.52352% |
| 5 | 17995 | 3840880 B | 48.184449% |
| 6 | 1796 | 678152 B | 15.323028% |
| 7 | 23 | 469256 B | 9.134617% |
| 8 | 9 | 469264 B | 11.438758% |
| 9 | 25 | 469224 B | 11.474494% |
| 10 | 7 | 469280 B | 35.653271% |
| 11 | 18 | 469392 B | 14.204358% |
| 12 | 96 | 469256 B | 18.042024% |
| 13 | 742 | 667800 B | 16.238022% |
| 14 | 3 | 469344 B | 7.667024% |

Here we can see the CPU usage and memory usage are depend on the size of the dataset. When data row count increases, the CPU and memory usages are going up.

# Chapter 6

# CONCLUSIONS

Today world when taking our daily decisions unintentionally recommender systems are supporting us. Therefore it is better that we understand those recommendations properly. Still, there are drawbacks in those provided suggestions and the way of presenting data to the user. It is better to present in a chart based method Therefore it will be easy to understand by any one. Still, there should be a good mapping technique to map the outcome with the input data. In currently many tools use many methods to give a recommendation. This study has presented a toolkit to recommend data visualization methods based on the context of data. Information retrieval and knowledge engineering based methods are applied to identify the context and patterns of the considered datasets. The methodology supports complex data processing such as aggregating data columns. The chart recommendations are provided using a combined method of machine learning and rule based components that incorporate computational intelligence techniques. Additionally, user preferences for the chart types are also recorded to improve the accuracy of the leaning process. The evaluation of the tool has shown accuracy over 0.8.

Rule-base will provide accurate suggestions in the early stages. Here in the system, we can add new chart types and new rules. Only the data behaviors can't predict recommendations. At this situation, the user involves interactively with the system to give feedbacks. Those feedbacks will improve the quality of the final output. Still, those feedbacks can be wrong. For a further improvement feedback validation technique which will be helpful while improving machine learning data. Here in the system, we have used the Decision tree method to classify the data in machine learning component. For more improvement, we can go for a neural network process for more accuracy I the system. This can be addressed as further development.

A possible extension is applying data aggregation in row level. Moreover, multiple recommender algorithms based on the data type can be used to increase the accuracy of this soft computing solution for the recommendation of data visualization. Furthermore, other quality attributes such as load balancing, resource utilization can be addressed with the increase of the data recommendation process in the long-run.

# REFERENCES


[1] L. Sharma and A. Gera, "A Survey of Recommendation System : Research Challenges," *Int. J. Eng. Trends Technol.*, vol. 4, no. 5, pp. 1989–1992, 2013.

[2] X. Song, C. Lin, B. Tseng, and M. Sun, "Modeling Evolutionary Behaviors for Community-based Dynamic Recommendation," *Proc. 2006 SIAM Int. Conf. Data Min.*, pp. 558–562, 2006.

[3] J. Kehrer and H. Hauser, "Visualization and Visual Analysis of Multifaceted Scientific Data : A Survey," vol. 19, no. 3, pp. 495–513, 2013.

[4] P. Kaur, M. Owonibi, and B. Koenig-ries, "Towards Visualization Recommendation – A Semi-Automated Domain-Specific Learning Approach," *27 th GI-Workshop Found. Databases (Grundlagen von Datenbanken)*, pp. 30–35, 2015.

[5] S. Amershi, J. Fogarty, and D. Weld, "ReGroup: interactive machine learning for on-demand group creation in social networks," *Proc. 2012 ACM Annu. Conf. Hum. Factors Comput. Syst. - CHI '12*, p. 21, 2012.

[6] C. Science and J. Ye, "MASTER ' S THESIS The author would like to express her special gratitude to :" *MASTER ' S THESIS*, 2015.

[7] M. Montaner, "A Taxonomy of Personalized Agents on the Internet," *Group*, pp. 1–65, 2001.

[8] C. Plaisant, "The Challenge of Information Visualization Evaluation," 2004.

[9] C. Ware, *Information Visualization: Perception for Design: Second Edition*. 2004.

[10] D. M. De Lima, J. F. Rodrigues, and A. J. M. Traina, "Graph-Based Relational Data Visualization," *17th Int. Conf. Inf. Vis.*, pp. 210–219, 2013.

[11] S. Carpendale, "Evaluating information visualizations," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 4950 LNCS, pp. 19–45, 2008.

[12] O. Pietquin and M. Lopes, "Machine Learning for Interactive Systems: Challenges and Future Trends," *Proc. of WACAI*, 2014.

[13] "data set," 2016. [Online]. Available: http://whatis.techtarget.com/definition/data-set.

[14] "Datasets for Data Science and Machine Learning," 2017. [Online]. Available: https://elitedatascience.com/datasets.

[15] T. E. Smith, "1. Areal data analysis," pp. 1–6, 2015.

[16] "Difference Between Discrete and Continuous Data." [Online]. Available: https://keydifferences.com/difference-between-discrete-and-continuous-data.html.

[17] N. Pradhan and K. kumar Pandey, "An Analytical and Comparative Study of Various Data Preprocessing Method in Data Mining," vol. 4, no. 10, pp. 174–180, 2014.

[18] D. Tomar and S. Agarwal, "A Survey on Pre-processing and Post-processing Techniques in Data Mining Divya," vol. 7, no. 4, pp. 99–128, 2014.

[19] J. M. Hellerstein, "Quantitative Data Cleaning for Large Databases," *United Nations Econ. Comm. Eur.*, p. 42, 2008.

[20] T. Bogers and A. Van Den Bosch, *Collaborative and content-based filtering for item recommendation on social bookmarking websites*, vol. 532. 2009.

[21] J. Bennett and S. Lanning, "The Netflix Prize," *KDD Cup Work.*, pp. 3–6, 2007.

[22] "content awareness." [Online]. Available: https://www.gartner.com/it-glossary/content-awareness.

[23] G. Linden, B. Smith, and J. York, "Amazon.com recommendations: Item-to-item collaborative

filtering," *IEEE Internet Comput.*, vol. 7, no. 1, pp. 76–80, 2003.

[24] P. Melville and V. Sindhwani, "Recommender systems," *Encycl. Mach. Learn.*, pp. 829–837, 2010.

[25] S. Bostandjiev, J. O. Donovan, and T. Höllerer, "TasteWeights: A visual interactive hybrid recommender system," *Proc. 6th ACM Conf. Recomm. Syst.*, pp. 35–42, 2012.

[26] L. Xiang, "Hulu ' s Recommendation System On-line Architecture," pp. 1–10, 2011.

[27] B. Gretarsson, J. O'Donovan, S. Bostandjiev, C. Hall, and T. Höllerer, "SmallWorlds: Visualizing social recommendations," *Comput. Graph. Forum*, vol. 29, no. 3, pp. 833–842, 2010.

[28] X. Ning, C. Desrosiers, and G. Karypis, "A comprehensive survey of neighborhood-based recommendation methods," *Recomm. Syst. Handbook, Second Ed.*, pp. 37–76, 2015.

[29] "How to Create a Simple Knowledge Base that Empowers Customers." [Online]. Available: https://www.salesforce.com/hub/service/create-knowledge-base/.

[30] S. Bouraga, I. Jureta, S. Faulkner, and C. Herssens, "Knowledge-Based Recommendation Systems," *Int. J. Intell. Inf. Technol.*, vol. 10, no. 2, pp. 1–19, 2014.

[31] J. Chen, Y. Tang, J. Li, C. Mao, and J. Xiao, "Community-Based Scholar Recommendation Modeling in Academic Social Network Sites," pp. 325–334, 2014.

[32] B. Mutlu, E. Veas, and C. Trattner, "VizRec: Recommending Personalized Visualizations," *ACM Trans. Interact. Intell. Syst.*, vol. 6, no. 4, pp. 1–39, 2016.

[33] J. Kwon and S. Kim, "Friend recommendation method using physical and social context," *Int. J. Comput. Sci. ...*, vol. 10, no. 11, pp. 116–120, 2010.

[34] S. Gutta and K. Kurapati, "Four-way recommendation method," vol. 1, no. 19, 2003.

[35] M. Iguchi, "User-Profile Based web page Recommendation System and User-Profile Based web page Recommendation method," vol. 1, no. 19, 2007.

[36] T. H. Roh, K. J. Oh, and I. Han, "The collaborative filtering recommendation based on SOM cluster-indexing CBR," *Expert Syst. Appl.*, vol. 25, no. 3, pp. 413–423, 2003.

[37] K. Choi, D. Yoo, G. Kim, and Y. Suh, "A hybrid online-product recommendation system: Combining implicit rating-based collaborative filtering and sequential pattern analysis," *Electron. Commer. Res. Appl.*, vol. 11, no. 4, pp. 309–317, 2012.

[38] M. J. Pazzani, "A Framework for Collaborative , Content-Based and Demographic Filtering," no. Lang 1995, pp. 393–394, 2000.

[39] A. Holzinger, "Interactive Machine Learning for Health Informatics," *Springer Brain Informatics*, pp. 1–12, 2016.

[40] M. Gales and S. Young, "The Application of Hidden Markov Models in Speech Recognition," *Found. Trends® Signal Process.*, vol. 1, no. 3, pp. 195–304, 2007.

[41] J. Brownlee, "Supervised and Unsupervised Machine Learning Algorithms," 2016. [Online]. Available: https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/.

[42] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques," *Informatica*, vol. 31, pp. 249–268, 2007.

[43] "Classification."[Online].Available: https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/classify.htm#DMCON004.

[44] “Regression.”[Online].Available: https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/regress.htm#DMCON005.

[45] A. Coates, A. Arbor, and A. Y. Ng, “An Analysis of Single-Layer Networks in Unsupervised Feature Learning,” *Aistats 2011*, pp. 215–223, 2011.

[46] “Clustering.”[Online].Available: https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/clustering.htm#DMCON008.

[47] “Association.”[Online].Available: https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/market_basket.htm#DMCON009 .

[48] M. Kabra, A. A. Robie, M. Rivera-Alba, S. Branson, and K. Branson, “JAABA: interactive machine learning for automatic annotation of animal behavior,” *Nat. Methods*, vol. 10, no. 1, pp. 64–67, 2012.

[49] B. C. Wallace, K. Small, C. E. Brodley, J. Lau, and T. a. Trikalinos, “Deploying an interactive machine learning system in an evidence-based practice center,” *Proc. 2nd ACM SIGHIT Symp. Int. Heal. informatics - IHI ’12*, p. 819, 2012.

[50] A. Mulloni, “Interactive Machine Learning System for Automated Annotation of Information in Text,” vol. 1, no. 19, 2014.

[51] A. Kapoor, B. Lee, D. Tan, and E. Horvitz, “Learning to Learn : Algorithmic Inspirations from Human Problem Solving,” *Proc. Twenty-Sixth AAAI Conf. Artif. Intell.*, pp. 1571–1577, 2008.

[52] K. Lee, J. Moore Myers, E. Treasure, R. Herring, S. McNulty, and D. Stotts, “Integrating GIS Visualization Tools for Ecosystem Management,” *GEOProcessing 2014, Sixth Int. Conf. Adv. Geogr. Inf. Syst. Appl. Serv.*, no. c, pp. 122–128, 2014.

[53] N. Gehlenborg *et al.*, “Visualization of omics data for systems biology,” *Nat. Publ. Gr.*, vol. 7, no. 3s, pp. S56–S68, 2010.

[54] T. O. S. U. Craige Roberts, “Information Structure in Discourse: Towards an Integrated Formal Theory of Pragmatics1,” no. 1967, pp. 1–53, 1993.

[55] M. Aoki, “Hrizontal vs. Vertical Information Structure of the Firm,” *Am. Econ. Rev.*, vol. 76, no. 5, pp. 971–983, 1986.

[56] J. Chen and D. Lopresti, “Model-based tabular structure detection and recognition in noisy handwritten documents,” *Proc. - Int. Work. Front. Handwrit. Recognition, IWFHR*, pp. 75–80, 2012.

[57] H. Fujisawa, Y. Nakano, and K. Kurino, “Segmentation Methods for Character Recognition: From Segmentation to Document Structure Analysis,” *Proc. IEEE*, vol. 80, no. 7, pp. 1079–1092, 1992.

[58] L. Toma, “Spatial Data Structures,” 2008. [Online]. Available: http://www.bowdoin.edu/~ltoma/teaching/cs340/spring08/.

[59] “Spatial (space) Structures.” [Online]. Available: http://www.setareh.arch.vt.edu/safas/009_introduction_01_ss.html.

[60] B. A. Loiselle, V. L. Sork, J. Nason, and C. Graham, “Spatial genetic structure of a tropical understory shrub, *Psychotria officinalis* (Rubiaceae),” *Am. J. Bot.*, vol. 82, no. 11, pp. 1420–1425, 1995.

[61] “Temporaldata.”[Online]. Available: http://pro.arcgis.com/en/pro-app/help/mapping/time/temporal-data.htm.

[62] “TemporalData.”[Online].Available: http://blogs.oregonstate.edu/geo599spatialstatistics/2014/04/28/temporal-data-spatial-autocorrelation/.

[63] “Tree Structure.” [Online]. Available: http://searchdatamanagement.techtarget.com/definition/tree-structure.

[64] U. Cengiz Turker and S. Balcisoy, “A visualisation technique for large temporal social network datasets in Hyperbolic space,” *J. Vis. Lang. Comput.*, vol. 25, no. 3, pp. 227–242, 2014.

[65] B. Luo and J. Xia, “A novel intrusion detection system based on feature generation with visualization strategy,” *Expert Syst. Appl.*, vol. 41, no. 9, pp. 4139–4147, 2014.

[66] “Scatter Plots.” [Online]. Available: http://mste.illinois.edu/courses/ci330ms/youtsey/scatterinfo.html.

[67] “Line Graph.” [Online]. Available: https://www.smartdraw.com/line-graph/.

[68] S. Chart, “3D Surface Plots,” vol. 2, pp. 1–10.

[69] D. Holten, “Hierrchical Edge Bundles: Visualizaiton of Adjacency Relations in Hierarchical Data,” *IEEE Trans. Vis. Comput. Graph.*, vol. 12, no. 5, pp. 741–748, 2006.

[70] Y. Jia and M. Garland, “Hierarchial Edge Bundles for General Graphs,” *Work*, no. June, 2009.

[71] “Histograms.” [Online]. Available: https://statistics.laerd.com/statistical-guides/understanding-histograms.php.

[72] “Bubble chart.” [Online]. Available: http://www.bubblechartpro.com/content/what_are_bubble_charts.php.

[73] “Area chart.” [Online]. Available: https://study.com/academy/lesson/what-is-an-area-chart-definition-examples.html.

[74] “Column chart.” [Online]. Available: https://www.merriam-webster.com/dictionary/column chart.

[75] “What to consider when creating stacked column charts.” [Online]. Available: https://blog.datawrapper.de/stacked-column-charts/.

[76] “Stacked Column Charts.” [Online]. Available: https://help.salesforce.com/articleView?id=chart_column_stacked.htm&type=5.

[77] “Geo map.” [Online]. Available: https://docs.thoughtspot.com/4.4/end-user/search/about-geo-charts.html.

[78] M. Welling, “A First Encounter with Machine Learning,” 2011.

[79] R. Y. Zhong, S. T. Newman, G. Q. Huang, and S. Lan, “Big Data for supply chain management in the service and manufacturing sectors: Challenges, opportunities, and future perspectives,” *Comput. Ind. Eng.*, vol. 101, pp. 572–591, 2016.

[80] “in-tag.” [Online]. Available: https://www.codefuel.com/in-tag/.

[81] “LensKit.” [Online]. Available: http://lenskit.org/.

[82] “Duine tool.” [Online]. Available: https://sourceforge.net/projects/duine/.

[83] “Plotly.” [Online]. Available: https://plot.ly/.

[84] “Polymaps.” [Online]. Available: http://polymaps.org/.

[85] M. Barbacci and Others, “{Q}uality {A}ttributes,” *IEEE Software*, vol. 18, no. CMU/SEI-95-TR-021, 1995.

[86] “Accuracy and Precision.” [Online]. Available: https://labwrite.ncsu.edu/Experimental Design/accuracyprecision.htm.

[87] “Usability Evaluation Basics.” [Online]. Available: https://www.usability.gov/what-and-why/usability-evaluation.html.

[88] “What are outliers in the data?” [Online]. Available: https://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm.

[89] M. Unger, A. Bar, B. Shapira, and L. Rokach, “Towards latent context-aware recommendation systems,” *Knowledge-Based Syst.*, vol. 104, pp. 165–178, 2016.

[90] I. R. P. Rodrigues, “UCAT: Ubiquitous Context Awareness Tools for The Blind,” 2013.

[91] E. Gilman, *Exploring the use of rule-based reasoning in ubiquitous computing applications*. 2015.

[92] “Rule-Based Expert Systems.” [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-21004-4_7.

[93] “How to Choose the Best Chart for Your Data.” [Online]. Available: https://lifehacker.com/5909501/how-to-choose-the-best-chart-for-your-data.

[94] S. Few, “Effectively Communicating Numbers Selecting the Best Means and Manner of Display,” *Book*, no. November, 2005.

[95] R. Length and R. Width, “Designing science graphs for data analysis and presentation The bad , the good and the better part4,” pp. 38–68, 1999.

[96] “Effects of visualizing statistical information – an empirical study on tree diagrams and 2 × 2 tables.” [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4549558/.

[97] “Understanding Network Diagrams – Everything YOU Need to Know.” [Online]. Available: https://www.askwillonline.com/2012/01/understanding-network-diagrams.html.

[98] “Better Know a Visualization: Parallel Coordinates.” [Online]. Available: http://www.juiceanalytics.com/writing/writing/parallel-coordinates.

[99] “How to Choose the Right Chart - A Complete Chart Comparison.” [Online]. Available: https://www.edrawsoft.com/chart/choose-right-chart.php.

[100] J. Gulbis, “Data Visualization – How to Pick the Right Chart Type?,” 2016. [Online]. Available: https://eazybi.com/blog/data_visualization_and_chart_types/.

[101] “Choosing the right chart type: Column charts vs Stacked Column Charts.” [Online]. Available: https://www.fusioncharts.com/blog/choosing-the-right-chart-type-column-charts-vs-stacked-column-charts/.

[102] “Breaking down hierarchical data with Treemap and Sunburst charts.” [Online]. Available: https://www.microsoft.com/en-us/microsoft-365/blog/2015/08/11/breaking-down-hierarchical-data-with-treemap-and-sunburst-charts/.

[103] “Pie Chart VS Doughnut Chart: When to use each.” [Online]. Available: https://ux.stackexchange.com/questions/105837/pie-chart-vs-doughnut-chart-when-to-use-each.

[104] “Machine Learning What it is and why it matters.” [Online]. Available: https://www.sas.com/en_us/insights/analytics/machine-learning.html.

[105] “Machine Learning: What it is and Why it Matters.” [Online]. Available: https://www.simplilearn.com/what-is-machine-learning-and-why-it-matters-article.

[106] “Decision Tree - Classification.” [Online]. Available: http://www.saedsayad.com/decision_tree.htm.

[107] “What is a Decision Tree? How does it work?” [Online]. Available: https://clearpredictions.com/Home/DecisionTree.

[108] R. Njeri, “What Is A Decision Tree Algorithm?” [Online]. Available: https://medium.com/@SeattleDataGuy/what-is-a-decision-tree-algorithm-4531749d2a17.

[109] “D3.” [Online]. Available: https://github.com/d3/d3/wiki.

[110] W. Paper, *Principles of Data Visualization - What We See in a Visual*. .

[111] C. Chen, *Handbook of Data Visualization*. .

[112] L. G. Williams, D. Ph, C. U. Smith, D. Ph, and C. U. Smith, “Performance Evaluation of Software Architectures Performance Evaluation of Software Architectures,” no. 303, 1998.

[113] J. Brooke, “SUS - A quick and dirty usability scale.”

[114] J. Akosa, “Predictive Accuracy : A Misleading Performance Measure for Highly Imbalanced Data Classified negative,” pp. 1–12, 2017.

[115] P. Jalote, B. Murphy, M. Garzia, and B. Errez, “Measuring Reliability of Software Products,” *Int. Syposium Softw. Reliab. Eng.*, pp. 1–8, 2004.

[116] “PHP-ML - Machine Learning library for PHP.” [Online]. Available: https://github.com/php-ai/php-ml.

[117] M. Gleicher, D. Albers, R. Walker, and J. C. Roberts, “Visual Comparison for Information Visualization,” pp. 1–29, 2011.

[118] M. Dörk, S. Carpendale, and C. Williamson, “Visualizing Explicit and Implicit Relations of Complex Information Spaces.”

[119] W. Javed, “Exploring the Design Space of Composite Visualization.”

[120] J. Leigh, A. Johnson, L. Renambot, V. Vishwanath, T. Peterka, and N. Schwarz, “Visualization of Large-Scale Distributed Data.”

[121] L. Dungan, “Visualizing Distribution Data For this paper , we extract,” 2018.