WishDish: "Recipe Prediction using K-Medoids Clustering Technique for Big Data Analytics"

Twisha Phirke
Department of Computer Engineering
VESIT, University of Mumbai, India

Pushkara Dighe
Department of Computer Engineering
VESIT, University of Mumbai, India

Darshana Rathi
Department of Computer Engineering
VESIT, University of Mumbai, India

Jayalakshmi Iyer
Department of Computer Engineering
VESIT, University of Mumbai, India

Nupur Giri
Department of Computer Engineering
VESIT, University of Mumbai, India

ABSTRACT
The project involves developing a web application called "WishDish" which proves as a one-stop destination for shopping food items, getting recipe suggestions for those food items and planning daily meals for the week on a calendar, all at one place. Recipe suggestions will be made using k-medoids clustering technique on weighted recipes. Aim is to make relevant and distinct recipe suggestions for food items bought every day. A part of this project also involves evaluating performance of the application based on the algorithm used, for different database sizes and accuracy, using R.

General Terms
Recommendation system, k-medoids, data analytics, k-means, clustering techniques.

Keywords
Big data, data analytics, k-medoids, recommendation system, recipe suggestions, food shopping, food planning, meal planning.

1. INTRODUCTION
This project primarily involves developing an application which analyzes customer's shopping behavior for predicting different recipe options, based on user's item selection. It is a content based recommendation system that evaluates similarities in a recipe's properties. Data is analyzed in such a way that the recipe is suggested in the order of relevance with respect to the items in the user's shopping cart. Project consists of four main

1.1 Shopping cart
The application allows a user to shop for items from various categories such as fruits and vegetables, meat products, dairy and bakery products etc.

1.2 Recipe suggestions
Recipe suggestions are made based on the items added to the shopping cart. Suggestions can be filtered on the basis of cuisine, ease of cooking, preferences (veg/non-veg/vegan/gluten-free), food allergens etc.

1.3 Planning calendar
An option to plan meals for the week is provided via a calendar. Meals can be categorized into breakfast, lunch and dinner. This calendar can then be shared with family members as the week's meal plan.

1.4 Item suggestions
The website includes a section where a user can browse through recipes from various cuisines. It is possible to generate a detailed list of ingredients needed for that recipe and buy the available food items from that list.

The flow of the system can be seen in Fig 1.

modules:

2. IMPLEMENTATION STRATEGIES
A large amount of data was mined from the Internet using a Google extension tool called "Web Scraper". Details such as name, brand, price, weight and images of 2207 food items were collected from a single site [5]. Around 967 recipes were scraped from three sites [6], [7], [8] with details such as recipe name, cuisine, nutritional value, ingredients, and method to cook etc. This data was processed and refined using Google's OpenRefine tool. The user interface for this application was developed as a website using HTML, CSS and JavaScript. Databases are stored in XAMPP and linked using PHP code.

Recipes for suggestions based on the food items selected by the user are obtained by implementing pattern matching in MySQL. Suggested recipes are ordered in terms of their relevance by assigning weights. This dataset of weighted recipes is further clustered to reduce its size. The final result obtained is a fixed number of recipes utilizing the maximum possible number of food items bought by the user, rather than the user getting lost among multiple suggestions.

2.1 Assigning weights to the recipes
The items added to the shopping cart by the user are used to query for recipes containing those items as ingredients. Recipes are weighted according to the number of items indexed with it. Recipes with maximum matching items are given the maximum weight.

2.2 Assigning weights to the items
Weighted recipes are selected. Items in these recipes, which have not been added to the shopping cart by the user yet, are given weights. These weights are decided on the number of

suggested recipes they occur in. Item which occurs the most in the list of weighted recipes is given the maximum weight. These items are then displayed to the user as "suggested items" according to their importance, in order to complete a recipe.

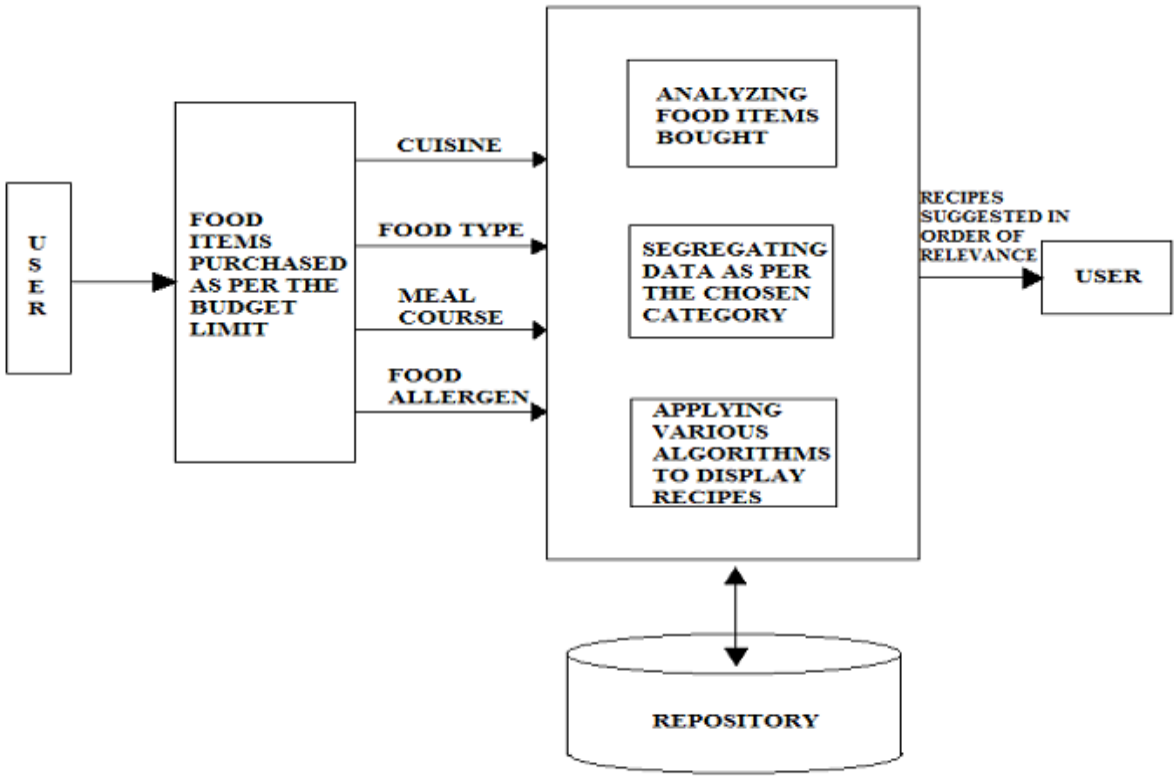

Fig 1: System block diagram

All weights are assigned dynamically. Weighted recipes are sent for clustering. This technique helps in creating a better database for the clustering techniques to work on.

2.3 Selecting a clustering technique

Earlier, the project involved clustering using k-means clustering technique for data mining. This returns a floating point value which cannot be tracked back into a fixed database to relate to a recipe. This method was not effective since the centers of the clusters formed had to be searched within the dataset and not predicted newly.

Hence, the new approach involves clustering using k-medoids clustering technique. K-medoids clustering technique provides a function, which returns the indices of the medoids formed. These indices relate to the recipe IDs in the dataset. As database used for clustering is fixed and structured, it is helpful when data points are returned as cluster centers.

2.4 K-medoid clustering technique

CLARA (Clustering Large Applications) is a data mining algorithm which partitions around medoids. CLARA and k-medoids return the same result for small size datasets. CLARA has been used over the basic k-medoids algorithm as the dataset sent for clustering is of variable size and in some cases it could be large i.e. more than 100 entries.

Algorithm:

1) Initialize: Randomly select (without replacement) *k* of the *n* data points as the medoids
2) Associate each data point to the closest medoid
3) While the cost of the configuration decreases:
4) For each medoid *m*, for each non-medoid data point *o*:
5) Swap *m* and *o*, re-compute the cost (sum of distances of points to their medoid)
6) If the total cost of the configuration increased in the previous step, undo the swap

The recipes with maximum and second maximum weights are sent to RStudio for clustering. This is done using user session IDs and RMySQL package. Once required data is acquired, algorithm for CLARA can be performed in R using existing data mining packages and functions available for cluster analysis. As the dataset has been refined and ordered and only then linked to RStudio, CLARA works effectively and provides relevant results.

The centers of the clusters formed will then be the most relevant recipe to be used for suggestion, distinct in their properties. The number of clusters to be made has been limited to 7. Over multiple iterations, it has been observed that it is optimal to form 7 clusters. This number is larger than the minimum size of the dataset that can be sent for clustering. It is also not too small for larger datasets. Recipes in the cluster having maximum data points can be suggested as "related recipes", giving the user a variety of recipes to choose from.

3. EXAMPLE SET

Consider the following sample set:

The item names can be seen in Table 1.Weights are assigned to each recipe as per the items indexed. Weights change dynamically with every new addition to the shopping cart. Recipes having maximum and second maximum weights are sent for clustering. These recipes are shown in Table 2.

Table 1. Items in the shopping cart

ITEM_ID	ITEM_NAME
1	Onion
2	Bitter Gourd
3	Beetroot
4	Papaya
8	Ginger
9	Potato
15	Cabbage
39	Black Brinjal
40	Lemon
43	Mango

Table 2. Assignment of weights to the recipes

RECIPE_ID	RECIPE_NAME	WEIGHTS
8	Saag Paneer	4
12	Baked Veggie Samosas	4
15	Beef kofta curry with fluffy rice, beans & peas	4
24	Chicken Curry	4
28	Fish Tikka Curry	4
32	Carrot Salad	4
668	Mango Pizza	4
710	Veg. Spaghetti	5
736	Baked and Layered Casserole	4

3.1 Comparison between clustering techniques

The means for this dataset are:

710.0 736.0 28.0 12.0 21.5 668.0 8.0

The medoids for this dataset are:

8 12 15 32 668 710 736

Thus, it can be seen that the means returned do not correlate to the database whereas the medoids returned are the recipe IDs, which can be used as relevant suggestions to the user.

3.2 Calculating precision, recall and accuracy

For the above example, precision, recall and accuracy values can be calculated as follows:

Precision = good items recommended / all recommendations

= TP / (TP+FP)

= 7/7 = 100%

Recall = good items recommended / all good items

= TP / (TP+FN)

= 7/7 = 100%

Accuracy = (good items recommended + bad items discarded)/ all items

= (TP+TN) / ALL

= (7+2) / 9 = 100%

Similarly, 10 example sets were considered. These example sets produced different sizes of datasets, to undergo clustering. Variation of accuracy across these 10 datasets of different sizes is shown in Fig 2. The average accuracy of the system, calculated over 10 example sets was found out to be 89.34%.

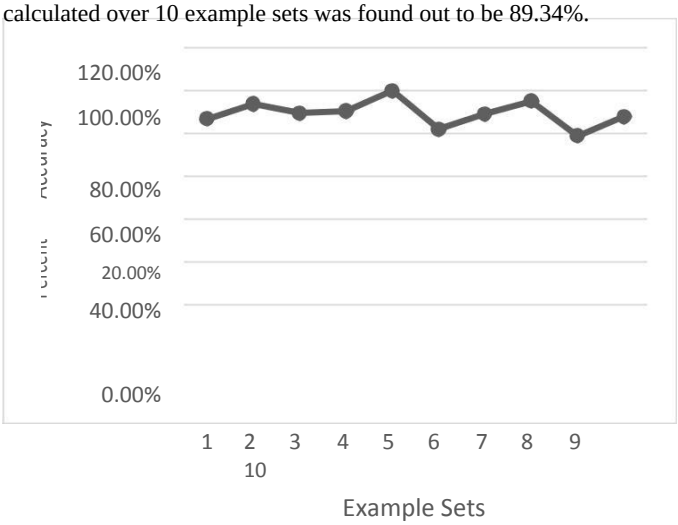

Fig 2: Accuracy graph

4. CONCLUSION AND FUTURE SCOPE

- The data collected from this application can be monitored and analyzed to study the user behavior based on frequent food item/recipe selection.
- Graphs can be formed using the data collected from the website, which will help in managing and updating the inventory with respect to the most popular item, most popular recipe, knowing the peak period for sales in a week etc.
- Collected data can also be used to study trends in health industry and food consumption habits of users i.e. veg/ vegan/ non-veg etc.
- Suggestions for extra food items can be made by analyzing the quantity of an item required in the recipes planned for the week.
- Quantity/weight of the food items can be considered for suggesting recipes. In case a food item is not completely utilized by the recipes planned in the calendar, other recipes which use the remaining quantity of the food item can be suggested.
- History of the items bought recently by a user can be maintained based on the shelf life of the food items. Items having a valid shelf life will also be considered while suggesting recipes for items bought in the future. This feature will help us make recipe suggestions similar to real life situations.
- Comparisons of various clustering techniques and analysis of their results will further help in optimizing the application to provide much more relevant results.

5. REFERENCES

[1] Mugdha Jain, Chakradhar Verma, "Adapting k-mean for Clustering in Big Data" International Journal of Computer Applications (0975 – 8887) Volume 101– No.1, September 2014.

[2] Sakuntala Gangaraju, "Recipe Suggestion Tool".

[3] Jeremy Cohen, Robert Sami, Aaron Schild, Spencer Tank, "Recipe Recommendation"- May 2013.

[4] Ji-Rong Wen, Jian-Yun Nie, Hong-Jiang Zhang, "Clustering User Queries of a Search Engine".

[5] https://grocermax.com

[6] http://www.jamieoliver.com

[7] http://www.tarladalal.com

[8] http://www.sanjeevkapoor.com

[9] https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm

[10] https://stat.ethz.ch/R-manual/R-devel/library/cluster/html/clara.object.html

[11] http://www.fi.muni.cz/~xpelanek/PV254/slides/evaluation.pdf

[12] http://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/