Intelligent Computation to build a Novel Recommender of Products through (PageRank-Clustering and DgSpan- FBR)

Samaher Al-Janabi [0000-0003-2811-1493[
Department of Computer Science, Faculty of Science for Women (SCIW), University of Babylon, Babylon, Iraq,
Samaher@itnet.uobabylon.edu.iq

Noor K.
Department of Computer Science, Faculty of Science for Women (SCIW), University of Babylon, Babylon, Iraq,

2020 International Conference on Decision Aid Sciences and Application (DASA) | 978-1-7281-9677-0/20/$31.00 ©2020 IEEE | DOI: 10.1109/DASA51403.2020.9317269

Abstract— Data is considered as one of the worthiest existent resources in the world. The fast development of technology provides us with huge/big data that needs deep analysis to understand and extraction useful information\knowledge from it. Thus, analysing such kind of data resembles a great challenge for researchers in the field of intelligent computation. The proposed model consists of three stages: the first stage is the collecting and preparing data (i.e., includes handle missing values and data convert into graphs). The second stage generates communities from graphs by developing one of the graph mining algorithms called (PageRank Clustering & Verification). Third stage includes building association patterns to the optimal clusters (subgraphs) and validating from it through developing one of the association pattern algorithms called Develop gSpan-Forward\ Backward Rules (DgSpan-FBR). The design model called Novel Recommender of products based on Graph Mining (NRGM) is distinguished by compressing the lost time of searching the best products that users need as well as increasing the accuracy of selecting of products. NRGM model is applied using on one of huge/big database have six parts different in number of samples, NRGM processed database to generated a set of communities through PageRank-Clustering. Then verification from these communities by apply two measures called (i.e., Modularity and Silhouette) to find optimal community. We test different number of communities (i.e., 2, 4, 6, 8,10 and 12) Then found two communities are the optimal where the Modularity of it is 0.679 while, silhouette is 0.699. Finally, a set of association patterns is found from the optimal community through both traditional association graph mining method called (gSpan) and developed method (DgSpan-FBR), where, the traditional method generates 3000 pattens for each community. While DgSpan-FBR generates 262 patterns. As a result, NRGM is consider as pragmatic model to deal with hug/big dataset, at the same time, it reduces the searching time and increase the accuracy of results.

Keywords— Big Data, DgSpan, FBR, , Recommendation System, Graph Mining Techniques, PageRank-clustering, gSpan.)

I. INTRODUCTION (*HEADING 1*)

The development of technology and the widespread use of social media has led to an increase the volume of data, in addition, the means of social media have become the most used techniques to promote the various goods established and, therefore, it is difficult for the user to choose the product that meets his needs, one of the main challenges in data is how can analysis it and discovery important knowledge from it.

This paper follows an intelligent computation method to simplification of user (customer) to choose the product satisfy their requirements, where we will be building Novel Recommender of products based on Graph Mining (NRGM) model, where products are present as recommendations to save the time for them. Before, we forward to this field the reader must know the main materials will used.

Delivers groceries database is a huge data needed to intelligent and deep computation to extract useful pattern from it. The input of that database (collection of datasets and each dataset have different features), while the output optimal patterns shows the recommendations of customers results of NRGM. Main advantages of delivers groceries database can cover and give a complete picture for huge\ big database. It is very difficult to understand their details of that data but relationships between datasets and their features can establish our multi rules simplified understanding that data. Graph is considered as a statistical model from which values of multi characteristic measures which described the behaviors of that graph can be extracted.

PageRank clustering PageRank using to find the rank web pages in their search engine results, this technique developed for use to discover optimal number of communities called PageRank clustering.

gSpan developed of gSpan through use the optimal community for the inputs of gSpan, this community discovered from validation by modularity and silhouette .

Forward /backward rules these rules use to find optimal list of recommendation through multi rules compare with other measures (i.e., confidence) that use one rule.

In this study's database there are some tables (i.e., products, department and aisles) that don't contain a primary key obviously but only orders dataset that contain a primary key represented by (customer-id). While the products dataset table is contained (department and aisles). Thus, we are obliged to joiner between orders table and products table to get complete dataset to building the recommender.

II. MAIN CONCEPT

In this section, we will show the main concept used for definition and solve problem.

A. Knowledge Discovery in Database Stages

This concept consists of multi steps, which are describes in the next sections.

B. Data Integration

After collecting data from multiple sources, for study the natural of that data if it suffers from any problem to process it. To avoid duplications and inconsistencies in the dataset, Data integration includes some operations including attribute correlation analysis, variables and domains identification and unification, tuples duplication, and finally detecting conflicting values from the different sources [7].

C. Data Preprocessing

Preprocessing the data is an important phase includes handle missing data, outliers, and imbalance and find the features most effect in take the decision. We inspected the data for missing values and random errors such as duplicate identifiers for two or more products and same order identifiers associated with different customers. [7].

D. Graph Mining Techniques

Graph Mining is the field of knowledge discovery in database contains many techniques used to analyze the properties of real-world graphs, predict how the structure and properties of a given graph might affect some application, and develop models that can generate realistic graphs that match the patterns found in real-world graphs of interest [19].

E. Clustering

Clustering is a type of unsupervised learning. Clustering searches the complete space and splits the data into groups based on some similarity measures. The result of clustering should make each two similar samples fall in the same group where the dissimilar samples fall in different groups. Different clustering algorithms might give different results and also the same algorithm used with different parameters might return different results [13].

TABLE I. COMPARE AMONG THE CLUSTERING TECHNIQUES

Techniques of Clustering	Advantages	Disadvantages
Spectral Clustering	▪ Does not make strong assumptions on the statistics of the clusters ▪ Easy to implement. ▪ Good clustering results. ▪ Reasonably fast for sparse data sets of several thousand elements	▪ May be sensitive to choose of parameters ▪ Computationally expensive for large datasets
Marko Clustering (MCL)	▪ Simple and elegant. ▪ Widely used in Bioinformatics because of its noise tolerance and effectiveness	▪ Very slow because takes 1.2 hours to cluster a 76K node social network. ▪ Prone to output too many clusters.
PageRank Clustering	• It is top-down and it can have a holistic view of the graph. • It is fast enough to be employed for clustering very large real-world networks. • Is very flexible because of its top-down approach • Has the potential to be used distributed which can make it an ideal solution for clustering at a scale that is infeasible with • Any single-machine implementation	• This algorithm partitions each sub graph into two sub graphs.

TABLE 2. SUMMARIZES THE CHARACTERISTICS CLUSTERING METHODS [26].

Method	General Characteristics
Partitioning methods	▪ Best fit for clusters of spherical shapes. ▪ Use the distance as the measure that assigns each sample to its cluster. ▪ Each cluster is represented by either mean or median of its data samples. ▪ Not fits large datasets.
Hierarchical methods	▪ Clustering the data into multiple levels and the best level is used. ▪ Not error tolerant to wrong merges or wrong splits. ▪ Micro clustering can be used with such methods.
Density-based methods	▪ Better than partitioning methods as it can find clusters of any shapes. ▪ Likely to assign each sample to the highest density cluster. ▪ Useful for filtering outliers.
Grid-based method	▪ Splits the data into a grid of samples. ▪ Scalable and faster than previous methods due to being independent to the number of samples.

There are many techniques for each type of cluster, but we will take four techniques and compare them in terms of advantage and disadvantages of each technique. In this work, we will deal with clustering from types hierarchical represented by PageRank clustering.

F. Graph Clustering Techniques

There are many types of Techniques of Clustering that will summarize the Advantages and Disadvantages by Table 2.

Algorthim#1: NRGM

Input: Database called delivers groceries have five parts (F1, F2, F3, F4 and F5)

Output: List of recommendations.

// Pre-Processing Stage

1: F6= call Joiner (F1, F2)

2: Ds'= F6

3: HDs' = Call handle missing values (Ds')

4: Gi=Call convert HDs' to graph

// Build NRGM Model

5: Split datasets into training &testing datasets

// PageRank Clustering

6: **For** each training dataset

7: Call PPRV // Personal PageRank Vector

8: Call DPR // Distance PageRank

9: Call PRC //PageRank Clustering

10: End for

// Validation of Clustering

11: Call *Evaluation* base on Modularity

12: Call *Evaluation* base on Silhouette Validity

// Generating Association Pattern by gSpan

13: **For** all nodes &edges in optimal cluster (subgraph)

14: Call gSpan

15: End for

16: For each pattern generating by gSpan

17: Call *Evaluation* base on forward /backward edges

18: End for

End NRGM

III. NOVEL RECOMMENDATION SYSTEM (NRGM)

In this paper, building novel recommender of products based on graph mining (NRGM) includes four stage; the first stage is including data collection from an open-source dataset published by Instacart. The second stage, it is preprocessing the dataset from missing values and covert datasets into graph. The third stages are building model, this stage split into two phases, the first phase clustering the graph into sub graphs based on PageRank clustering, the second phase consider with extraction useful pattern for each subgraph using gSpan algorithm that based on depth first search.

The final stages include evaluation the clustering using two measures (i.e., Modularity and Silhouette Validity) and verification from patterns based on forward & backward rules. Figure 1 explain block diagram of NRGM model while the main steps of that model show in algorithm 1.

IV. EXPERIMENT

This section shows the results for each stage in NRGM. Also, justification that results.

A. Description of Database

The database was published in 2017 for research purposes. It is containing over 3 million grocery orders from more than 200,000 customers. Database is anonymized and does not contain any customer information. It only contains a unique identifier (id) for each customer. The data is available in multiple comma-separated-value (csv) files. The first file contains product information that includes product id, product name, aisle id and the department id. The aisle id represents the identifier that signifies, where the product is placed in the store and department id signifies the category to which the product belongs to. Table 3 description of database.

B. Preprocessing Stage

This stage consists of three basic steps until the dataset is ready for use.

Data Joint

In this paper used joined to connected between two dataset (i.e., the dataset of order and the dataset of products) the dataset of order consists of seven columns while the dataset of products consists of four columns, so the final dataset consists of eleven columns, the algorithm 2 explained joiner steps.

TABLE 3. DESCRIPTION OF DATABASE

Name of Dataset	# Features	# Samples
Order Dataset	7	1048575
Orders_ products _prior	4	1048575
Orders_ products _train	4	1048575
Aisles	2	134
Department	2	21
Product	4	49688

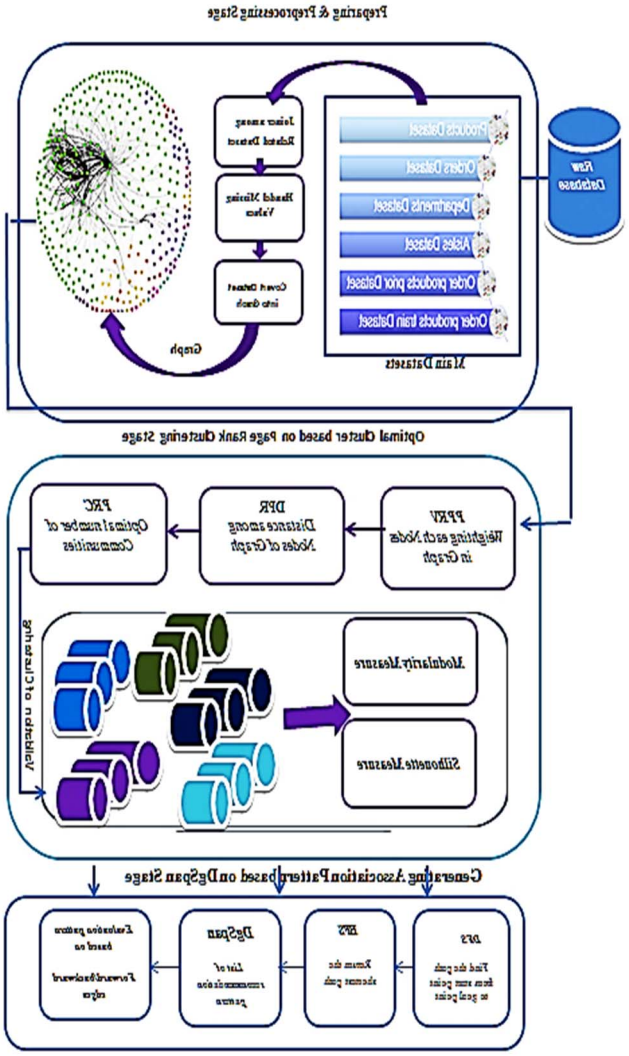

Fig. 1. Block Diagram of Recommender

Algorithm #2: Joiner

Input: Datasets (i.e., two files from different dataset)

Output: Joiner the Two File

1: Read F1 and F2
2: Compare feature of two files
3: Fetch content feature of two files
4: Compare features length of F1 with feature length of F2
5: **IF** feature F1. Length>=feature F2. Length
6: **For** i=0 to feature F1. Length
7: **For** j=0 to feature F2. Length
8: **If** feature F1 [i]. Equals feature F2 [j]
9: count=1
10: **Else**
11: **If** count==0
12: Diff count. Add=i
13: Repeat Compare feature when feature of F2 is larger
14: Generated array of merge content the two file in single file
15: **If** content F1.size () <content F2.size()
16: **For** i=0 to content F1.size ()
17: Line1=fetch of F1 feature
18: Line2=fetch of F2 feature
19: from line2, some data is required
20: **For** j=0 to diff count. Size ()
21: connect the two array line 1 and line2 in line1
22: count=i
23: Now complete from F2
24: **For** i=count to content F2.size
25: **For** j=0 to feature F1.length
26: line_1 [j] =" "
27: from line2, some data is required
28: **for** k=0 to diff count. Size ()
29: Joint the two array line 1 and line2 in line1
30: Repeat steps when, If content F2.size<content F1.size
End Joiner

It combines two datasets tables. The join is based on the joining columns of both tables. The result of joint between the orders and products dataset that content 11 features by combine 7 features from order with 4 features from products. The purpose of this stage is preparing the dataset to extract useful information from it.

Missing values

As a result of the process joiner, we note that these data are suffering from missing value such as (Days since prior order is suffer missing value =**0.062** result from missing 3086 samples of total sample is 49688, aisle id and department id are suffered missing value =**0.067** result from missing 3336 samples of total sample is 49688, so we use mean equation. the result is handled missing values by mean. The purpose of this stage is preparing the dataset to using in convert dataset into graph.

Convert dataset into graph

In this step, converted dataset obtained from the previous two steps into three types of graph (random graph, indirect graph, direct graph). To explain this step, we will take example contain of 6 orders.

Algorithm 3: Page Rank Clustering (PRC)

Input: G: Graph
jc: Jumping constant E : constant
Output: Optimal Number of Clustering
1: For all v ∈ G
2: Compute PPRV (jc,s)
3: End for
4: Find the roots of ϕ (jc) (There can be more than one root if G has a layered clustering structure.)
5: For all roots jc do
6: Compute ϕ (jc) according to equation
ϕ (jc) =dv||PPRV (jc, s) D^-1/2-PPRV(jc, PPRV(jc,s))D^-1/2||^2
7: **IF** ϕ (jc) < = € then
8: Compute ψ (jc) according to equation
ψ (jc) = dv|| PPRV (jc, PPRV(jc,v))D^-1/2-πD^-1/2||^2
9: **Else**
10: Go to the next jc
11: **End** if
12: **IF** k < ψ(jc)- 2 - € then
13: Go to the next jc
14: **Else**
15: Select c log n sets of k potential centers, randomly cho sen according to π
16: **End** if
17: **For** all sets S = {v1; : : : ; vk} do
18: Let C be the set of centers of mass where ci = pr(_jc; vi).
19: Compute μ(C) and ψα(C) according to equation

$$\mu(C)=\sum_{v\in V}\left|\left|\frac{PPRV(jc,v)D^{-1}}{2}-\frac{PPRV(jc,PPRV(jc,v))D^{1}}{2}2\right|\right|^{2}$$

$$\psi\alpha(C)=\sum_{c\in C} vol(Rc)\,||\frac{PPRV(jc,v)D^{-1}}{2}-PPRV(jc,PPRV(jc,v))D\hat{}1/2||\hat{}$$

20: **IF** | μ(C) - ϕ(jc) |<= € and |ψjc(C)- ψ(jc) |<= € then
21: Determine the k according to the DPR algorithm using C and
turn them.
22: **End** if
23: **End** for
24: **End** for
End PRC

A. The first type random graph for 6 orders as shown

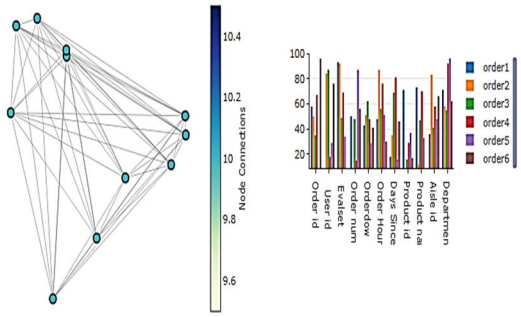

Fig. 2. Example of Random Graph to Six Orders for Joiner Dataset

B. the second type indirect graph for 6 orders, the results of this method is used to generate communities as shown in the Figure 3.

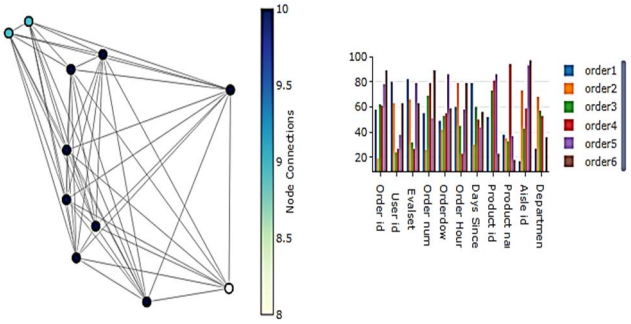

Fig. 3. Example of Undirect Graph to Six Orders after Joiner Datase

C. The third type direct graph for 6 orders explained in Figure 4.

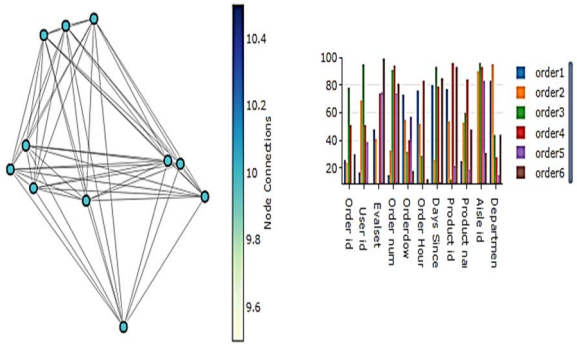

Fig. 4. Example of Direct Graph to Six Orders after Joiner Dataset

C. Results of Finding Optimal Community

In this stage, the output of prepressing stage is used as inputs to the (PageRank clustering –verification), where grouping our products according to its orders represented by set of communities / clusters pass as input to other techniques represent by DgSpan-FBR. This stage consists of three steps. The main details and parameters of PRC represent in Algorithm 3

- ***First step:***

Find the weight for each node, after determined the values of two parameter (i.e., amounts of sampling and jumping constant), as explained in chapter three in the algorithm3.5. Where sv= number of sampling in G and jc=constant. Where choose (sv=49688 and jc=0.85). We note the results of PPRV to generate the weight for each node in graph, Table 4 explain results of PPRV.

TABLE 4. WEIGHTS GENERATED BY PPRV

Name of nodes	Weights
Node1	0.09392
Node2	0.09392
Node3	0.07749
Node4	0.08558
Node5	0.09392
Node6	0.09392
Node7	0.08558
Node8	0.09392
Node9	0.09392
Node10	0.09392
Node11	0.09392

TABLE 5: RESULTS OF EACH SUB COMMUNITY

#Sub Community	# Samples	#Samples	#Samples	#Samples	#Samples	#Samples
Sub_C #1	11015	18001	3265	8622	10956	2184
Sub_C #2	38673	8622	8053	5098	6065	2711
Sub_C #3		14720	764	12341	7509	7161
Sub_C #4		8345	9429	1859	5029	8986
Sub_C #5			24813	9618	1740	2840
Sub_C #6			3364	4979	2391	1981
Sub_C #7				4430	7894	2990
Sub_C #8				2740	1368	1994
Sub_C #9					4814	6097
Sub_C #10					1922	1468
Sub_C #11						2184
Sub_C #12						9899
Sum	49688	49688	49688	49688	49688	49688

To verify the validity of the results of this algorithm, we notice that the sum of the weights is equal to the number 1 and this result confirms the accuracy of this algorithm.

- ***Second***

In this section, get the distance between nodes depend on PPRV. Where the number of samples is 49688, this step is important to generate communities.

- ***Third step:***

Generating the main communities/clusters according to its orders, before need to determine many parameters effect in their results. Each one of that parameter effects on result obtained. In general, the main parameters used here; G= undirected graph, jc, number of clusters 2, 4, 6,8,10 and 12 for the same values of other parameters. Table 5 represents results obtained from different numbers of cluste

We will use different number of clusters (i.e., sub community) to test the validation and accuracy of algorithm. To find the optimal sub community (i.e., sub graph) we will use two basic measures (i.e., modularity and silhouette).

- **Modularity Measure**

The result of modularity measure explains in Table 6.

TABLE 6: RESULTS OF MODULARITY MEASURE

Number of Communities	Value of Modularity
Two community	**0.679**
Four community	0.417
Six community	0.298
Eight community	0.236
Ten community	0.202
Twelve community	0.174

We note through the table 6 the optimal result by two sub communities shown in bold. The best value of modularity convergent from 1 while the worst value convergent of zero.

- **Silhouette Measure**

The result of silhouette measure explains in Table 7

TABLE 7: RESULTS OF SILHOUETTE MEASURE

Number of Communities	Value of Silhouette
Two Community	**0.699**
Four Community	0.635
Six Community	0.466
Eight Community	0.310
Ten Community	0.318
Twelve Community	0.326

We note through the table 7 the optimal result by two sub communities shown in bold. The best value of silhouette convergent from 1 while the worst value convergent of zero.

D. Results of Finding Association Patterns

This section shows the results of both traditional and develop algorithm for find the association patterns (i.e., gSpan and DgSpan-FBR) also, we will explain all the parameters of both algorithms.

1) Implementation of Traditional gSpan

In this section, this algorithm is apples to all communities result from applied (PageRank Clustering) on database contains over (3 million grocery orders) from more than (200,000) customers. So it will produce 3,000 patterns as explain in appendix. In general, because the number of patterns is very huge, we will explain 50 patterns of it choose randomly. These results explain in Table 8.

Table 8 explains all community (i.e., two community, four community, six community, eight community, ten community and twelve community) results from applied PRC, and through apply traditional gSpan on these communities' results set of patterns.

Implementation DgSpan-FBR

In this section, DgSpan is applied to optimal community (i.e., two community) result from applied (PageRank Clustering and validation) by two measures (i.e., modularity and silhouettes measures), the optimal cluster divided in two sub communities (i.e., sub community number 1 and sub community number 2). Algorithm 4 Cleary the main points to implementation the DgSpan-FBR.

The first step, we apply new algorithm called (DgSpan) onto the sub community number 1 contain of **11015** samples result from Table 9. This algorithm (DgSpan) resulted (**294**) Patterns in the form:

IF Conditions Then Action

In general, some pattern results are shown from apply DgSpan onto the sub community number 1 explain below.

Pattern #1: IF (order_id is between (2819074.95202 - 3178291.98642)) AND (user_id is between (84 - 90)) AND (order_number is between (4 - 8)) AND (order_dow is between (3 - 3.37273)) AND (order_hour_of_day is between (14 - 15)) AND (product_id is between (1239.00001 - 1335.00001)) AND (aisle_id is between (85 - 112)) AND (department_id is between (109 - 128)) THEN ProductName is Organic Edamame & Mung Bean Fettuccine.

Pattern #2: IF (user_id is between (0 - 132)) AND (eval_set is between (- prior)) AND (product_id is between (0 - 1898)) AND (department_id is between (0 - 81)) THEN ProductName is Klondike Petite Potatoes

Pattern #46: IF (user_id is between (15 - 33)) AND (eval_set is between (test - train)) AND (order_number is between (1 - 47)) AND (order_hour_of_day is between (4 - 23)) AND (department_id is between (1.00002 - 129.99999)) THEN product_name is Infant's Blend Probiotic.

Pattern #47: IF (user_id IS 26.68966) AND (eval_set IS train) AND (order_hour_of_day IS 12.86207) AND (days_since_prior_order IS 5.16834) AND (product_id IS 371.34483) AND (department_id IS 9.82759) THEN product_name is Non-Fat Blueberry on the Bottom Greek Yogurt.

Pattern #201: IF (user_id is between (103 - 131)) AND (eval_set is between (test - prior)) AND (product_id is between (1558.00003 - 1888.99997)) AND (department_id is between (1 - 91)) THEN product_name is Pomegranate Blueberry Pistachio Plus Antioxidants 1.4 oz. Fruit & Nut Bars.

Pattern #202: IF (user_id IS 24.0375) AND (eval_set IS train) AND (order_number IS 10.1) AND (department_id IS 11.2375) THEN product_name is Original Sprouted Grains Protein & Fiber Hot Oatmeal.

TABLE 8: RESULTS OF SILHOUETTE MEASURE

# Case Study	#Communities	#Patterns
Case Study #1	Sub_C#1	1246
	Sub_C#2	1754
Case Study #2	Sub_C#1	1536
	Sub_C#2	341
	Sub_C#3	832
	Sub_C#4	291
Case Study #3	Sub_C#1	337
	Sub_C#2	513
	Sub_C#3	233
	Sub_C#4	619
	Sub_C#5	1052
	Sub_C#6	346
Case Study #4	Sub_C#1	426
	Sub_C#2	324
	Sub_C#3	937
	Sub_C#4	139
	Sub_C#5	573
	Sub_C#6	232
	Sub_C#7	215
	Sub_C#8	154
Case Study #5	Sub_C#1	692
	Sub_C#2	317
	Sub_C#3	424
	Sub_C#4	340
	Sub_C#5	127
	Sub_C#6	136
	Sub_C#7	402
	Sub_C#8	112
	Sub_C#9	322
	Sub_C#10	128
Case Study #6	Sub_C#1	174
	Sub_C#2	207
	Sub_C#3	361
	Sub_C#4	464
	Sub_C#5	216
	Sub_C#6	134
	Sub_C#7	223
	Sub_C#8	132
	Sub_C#9	305
	Sub_C#10	113
	Sub_C#11	184
	Sub_C#12	487

Pattern #293: IF (user_id is between (35 - 63)) AND (eval_set is between (test - prior)) AND (product_id is between (487 - 955)) AND (aisle_id is between (3 - 92)) AND (department_id is between (1 - 128)) THEN product_name is Thick Gel Toilet Bowl Cleaner.

Pattern #294: IF (order_id is 3164014.00007) AND (user_id is 150) AND (eval_set is train) AND (order_number is 14) AND (order_dow is 4) AND (order_hour_of_day is 13) AND (days_since_prior_order is 30) AND (product_id is 2251) AND (aisle_id is 112) AND (department_id is 128) THEN product_name is Organic Garam Masala.

The second step, we apply new algorithm called (DgSpan) onto the sub community number 2 contain of **38673** samples result from Table 9. This algorithm (DgSpan) resulted (**306**) Patterns.

Pattern #1: IF (order_id IS 681304.32011) AND (user_id IS 859.2676) AND (eval_set IS train) AND (order_hour_of_day IS 13.26761) AND (product_id IS 13696.73238) AND (aisle_id IS 100) THEN product_name is Grandmas Brownie Bc 2.875 Oz.

Pattern #2: IF (user_id IS 116.97674) AND (eval_set IS train) AND (order_hour_of_day IS 15.33762) AND (product_id IS 1720.20931) THEN product_name is Pimiento Stuffed Greek Olives.

Pattern #49: IF (user_id IS 859.38333) AND (eval_set IS train) AND (order_hour_of_day IS 13.18333) AND (days_since_prior_order IS 4.66568) AND (product_id IS 13713.53333) THEN product_name is Original Diced Tomatoes and Green Chilies.

Pattern #50: IF (order_id IS 2998683.46922) AND (user_id IS 87) AND (eval_set IS train) AND (order_number IS 6) AND (order_dow IS 3.18636) AND (order_hour_of_day IS 14.5) AND (product_id IS 1287.00001) AND (aisle_id IS 98.5) AND (department_id IS 118.5) THEN product_name is Organic Edamame & Mung Bean Fettuccine.

Pattern #100: IF (user_id IS 857.31395) AND (eval_set IS train) AND (order_number IS 14.94186) AND (product_id IS 13648.65116) THEN product_name is Silver Grade Pacific Arame. **Table 8:** Result of Traditional gSpan of all Communities

Pattern #101: IF (user_id IS 137.33987) AND (eval_set IS train) AND (order_number IS 12.5098) AND (product_id IS 1985.77124) AND (aisle_id IS 24.77778) AND (department_id IS 14.16993) THEN product_name is Natural Uncured Beef Hot Dog.

Pattern #200: IF (user_id IS 860.05405) AND (eval_set IS train) AND (order_number IS 65.78378) AND (order_hour_of_day IS 13.13514) AND (days_since_prior_order IS 4.08108) AND (product_id IS 13738.21621) THEN product_name is Heirloom Navel Orange.

Pattern #201: IF (user_id IS 142.10345) AND (eval_set IS train) AND (order_hour_of_day IS 14.43678) AND (days_since_prior_order IS 4.30426) AND (product_id IS 2103.4138) AND (department_id IS 14.16092) THEN product_name is Fudge Sticks Jumbo Peanut Butter Cookies.

Pattern #300: IF (user_id IS 857.80612) AND (eval_set IS train) AND (order_number IS 15.92857) AND (product_id IS 13661.94898) AND (department_id IS 16.90816) THEN product_name is Dark Chocolate Fudge Stripe.

Pattern #301: IF (user_id IS 149.50193) AND (eval_set IS train) AND (product_id IS 2259.37838) AND (aisle_id IS 106.89189) AND (department_id IS 13.15444) THEN product_name is Mint Chip.

Pattern #305: IF (user_id IS 73.35455) AND (order_number IS 7.31818) AND (days_since_prior_order IS 27.28182) AND (product_id IS 1110.79091) AND (department_id IS 12.52727) THEN product_name is Linguine No 7 Pasta.

Pattern #306: IF (order_id IS 2657460.00625) AND (user_id IS 857.79487) AND (eval_set IS train) AND (order_number IS 14.74359) AND (product_id IS 13663.41025) THEN product_name is Organic Red Lentil Mini Fettucini.

After that, forward/ backward rules are applied on the patterns obtain by (DgSpan) on the optimal sub community before that, we need to know which of that sub community is optimal, therefore we return to applied modularity measure on both it. Where we found the value of modularity in sub community number 1=0.51and the value of modularity in sub community number 2=0.83.

So, sub community number 2 enters to (DgSpan-FBR) is used, it is found that the number of patterns obtain is (262), so the pruning in number of patterns is very low because we applied that algorithm on the optimal graph. In other word the patterns are more Coherent and investigative concept for optimal patterns. While, if the same algorithm is applied on all graph the pruning is increase, this prove the accuracy of suggest model, but since it is not logic to display that entire pattern to the user after they become number (262) pattern. So the five patterns are pressed as recommendation for any customer from the optimal pattern based on most previous orders.

Algorithm #4: gSpan-FBR

***Input**:* Optimal Cluster

Output: S set of patterns (recommenders)

```
1: S =
2: For each optimal cluster
3:   s = Call DFS
4:   Insert s into S
5:   Call BFS
6:   Find all the edges e such that s can be right-most extended
     insert in c
7:   If the trees have max number of edges
8:     Remove duplication tree
9:   Else
10:    kept the tree has minimal number of edges
11:  End if
12: End for
//forward & backward Rules
13: For each value of set patterns (S)
14:   If v1 of e1 < v2 of e1
15:     If v1 of e2 < v2 of e2
16:       If v1 of e2 <= v2 of e1 & v2 of e2 == v2 of e1+1
17:         this is an acceptable e2
18:       Else
19:         e2 being considered isn't valid
20:     Else
21:       e2 is a backward edge
22:     If v1 of e2 == v2 of e1 & v2 of e2 <v1 of e1
23:       this is an acceptable e2
24:     Else
25:       e2 being considered isn't valid
26:   Else
27:     e1 is a backward edge
28:     If v1 of e2 < v2 of e2 (forward edge)
29:       If v1 of e2 <= v1 of e1& v2 of e2 == v1 of e1+1
30:         this is an acceptable e2
31:       Else
32:         e2 being considered isn't valid
33:     Else
34:       e2 is a backward edge
35:       If v1 of e2 == v1 of e1& v2 of e1<v2 of e2
36:         this is an acceptable e2
37:       Else
38:         e2 being considered isn't valid
39: End forward/backward edge
End gSpan-FBR
```

V. Discussions and Conclusions

One of main problem related to data through integration between two datasets (i.e., the orders data and products data) is solved, this performs by applied joiner. We get complete data set to building the recommender.

Succeeds to convert dataset into graph, this step considers the core to preparing the dataset for the (PRC-V).

PRC-V approves the ability to extraction optimal cluster from the content graph. It not only by clustering keywords of the products, but also by determining the product for each order in the delivery database. That information is very useful for using in gSpan-FBR.

Validation measures of the community give good indication of cluster connection inside each community through Modularity and Silhouette.

GSpan-FBR technique is a powerful tool for generated association patterns compare with the traditional methods, as explain in experiment more powerful and faster, GSpan-FBR is consider method to be used with recommendation system because it led to true recommender.

The integration between the two-fields, (PRC-V) & (gSpan-FBR) presented a new image to view and analyze the hug dataset. Integration also gives a good justification related to cluster network.

The knowledge discovery by the NRSGM determines the most important products that needed the customer in the faster time and high accuracy.

In the future, we can use the proposed model to generated recommendation in different datasets such as (i.e., movies, papers, books and etc.) to test performance of recommendation systems. Also, order's time-stamp, customer's demographics and geographic details to further improve the recommendation system could be used.

References

[1] Al-Janabi, S. & Alkaim, A.F. A nifty collaborative analysis to predicting a novel tool (DRFLLS) for missing values estimation, Springer, Soft Comput (2020), Volume 24, Issue 1, pp 555–569. DOI 10.1007/s00500-019-03972-x

[2] Kaoutar Makdad, Rafik Lasri and Abdellatif El Abderrahmani , "Important Method of Exchange and Sharing of Massive Data Between Connected Objects," SBD 53, pp. 280–287, 2019. doi.org/10.1007/978-3-030-12048-1_28

[3] Robert de Graaf ,"MANAGING YOUR DATA SCIENCE PROJECTS," Springer ,2019 . doi.org/10.1007/978-1-4842-4907-9

[4] Ning Wang , Hui Zhao, Xue Zhu and Nan Li, "The Review of Recommendation System", Springer, CCIS 980, pp. 332–342, 2019. doi.org/10.1007/978-981-13-7025-0_34.

[5] Mohit Thakkar, "Beginning Machine Learning in iOS.," Springer,2019. doi.org/10.1007/978-1-4842-4297-1.

[6] Al_Janabi S., Razaq F. (2020) A Novel Tool DSMOTE to Handel Imbalance Customer Churn Problem in Telecommunication Industry. In: Farhaoui Y. (eds) Big Data and Networks Technologies. BDNT 2019. Lecture Notes in Networks and Systems, vol 81. Springer, Cham. doi.org/10.1007/978-3-030-23672-4_4

[7] Al-Janabi, S., Alkaim, A.F. & Adel, Z. An Innovative synthesis of deep learning techniques (DCapsNet & DCOM) for generation electrical renewable energy from wind energy. Soft Comput (2020). https://doi.org/10.1007/s00500-020-04905-9

[8] Sebasti´an A. R´ıosa and Ivan F. Videla–Cavieresb. "Generating groups of products using graph mining techniques", Elsevier, pp. 730 – 738,2014. http://creativecommons.org/licenses/by-nc-nd/3.0/.

[9] Jevin D. West, Ian Wesley-Smith and Carl T. Bergstrom. " A recommendation system based on hierarchical clustering of an article-level citation network,". IEEE TRANSACTIONS ON BIG DATA, Volume:2 , Issue: 2 ,pp. 113 –123, June 1 2016 .**DOI:** 10.1109/TBDATA.2016.2541167

[10]

[11] Ukrit Marung , Nipon Theera-Umpon and Sansanee Auephanwiriyakul . "Top-N Recommender Systems Using Genetic Algorithm-Based Visual-Clustering Methods," *Symmetry,* Received: 6 April 2016; Accepted: 17 June 2016; Published: 24 June 2016.doi.org/10.3390/sym8070054.

[12] G. Krishna Kishore and D. Suresh Babu. "Recommender System based on Customer Behavior for Retail Stores,".Journal of Computer Engineering, Volume 19, Issue 3, Ver. PP 06-17 I May.-June. 2017. DOI:10.9790/0661-1903010617

[13] Pipsa Harno . "Techniques for Mining Transactional Data for Personalized Marketing Actions", 23.05.2017 Business Technology.

[14] Hong Cheng and Jeffrey Xu Yu, "Graph pattern mining; Subgraph mining,",Springer,2018.

[15] Bhumika Bhatt, Premal J Patel and Hetal Gaudani, "A Review Paper on Machine Learning Based Recommendation System",IJEDR | Volume 2, Issue 4 | ISSN: 2321-9939.2014. https://www.ijedr.org/papers/IJEDR1404092.pdf

[16] Rus M. Mesas and Alejandro Bellog´ın, "Exploiting recommend dation confidence in decision-aware recommender systems".Springer ,2018.doi.org/10.1007/s10844-018-0526-3.

[17] Martin Hilbert and Priscila López , "The World's Technological Capacity to Store, Communicate, and Compute Information", Page 110.1126/science.1200970 ,2011. doi:10.1126/science.1200970

[18] Tom Breur, "Statistical Power Analysis and the contemporary ''crisis'' in social sciences," Journal of Marketing Analytics ,pp. 61–65 ,2016. doi:10.1057/s41270-016-0001-3.

[19] Sabeur Aridhia and Engelbert Mephu Nguifob, "Big Graph Mining: Frameworks and Techniques," Elsevier, 2016. doi.org/10.1016/j.bdr.2016.07.002

[20] S. Meenakshi and R. Renukadevi, "A Review on Spectral Clustering and its Applications," International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 8, August 2016. DOI: 10.15680/IJIRCCE.2016. 0408050

[21] Mohammad Soruri, Javad Sadri and S. Hamid Zahiri, "Gene clustering with hidden Markov model optimized by PSO algorithm, " Springer-Verlag London Ltd., part of Springer Nature 2018. doi.org/10.1007/s10044-018-0680-9.

[22] Fan Chung and Alexander Tsiatas, "Finding and Visualizing Graph Clusters Using PageRank Optimization", Springer- pp. 86-97, 2010.https://link.springer.com/chapter/10.1007/978-3-642-18009-5_9

[23] Charu C. Aggarwal, "Association Pattern Mining," Online, springer 14 April 2015. DOI 10.1007/978-3-319-14142-84.

[24] Garcia,S.;Luengo,J. and Herrera,F. " Data preprocessing in Data Mining, " , Switzerland :Springer, International Publishing,2015

[25] Wang and Yang Xu, "Mining Graph Pattern Association Rules," Online springer LNCS 11030, pp. 223–235, 2018.

[26] M.Kavitha and S.T.Tamil Selvi , "Comparative Study on Apriori Algorithm and Fp Growth Algorithm with Pros and Cons," International Journal of Computer Science Trends and Technology (I JCS T) – Volume 4 Issue 4, 2016.

[27] Samaher Al-Janabi, Mahdi Abed Salman and Ahmed Fanfakh, 2018. "Recommendation System to Improve Time Management for People in Education Environments". Journal of Engineering and Applied Sciences, 13: 10182-1019. **DOI:** 10.3923/jeasci.2018.10182.10193

[28] Liangfu Lu, Xiaoxu Ren, Lianyong Q, Chenming Cu and Yichen Jiao,"Tar get Gene Mining Algorithm Based on gSpan,"Springer , LNICST 268, pp.518–528, 2019.doi.org/10.1007/978-3-030-12981-1_36.

[29] Al-Janabi, S., Mohammad, M. & Al-Sultan, A. A new method for prediction of air pollution based on intelligent computation. Soft Comput 24, 661–680 (2020) doi:10.1007/s00500-019-04495-1