Open access Journal International Journal of Emerging Trends in Science and Technology
IC Value: 76.89 (Index Copernicus) Impact Factor: 4.219 DOI: https://dx.doi.org/10.18535/ijetst/v4i9.29

Product Data Clustering using Weighted Similarity Measure

Authors
Dr. S. Aquter Babu
Assistant Professor, Department of Computer Science
Dravidian University
Kuppam, Chittoor District, Andhra Pradesh, India

Abstract
In market basket analysis user ranking of products is very important in addition to the values of attributes of objects. Similarity based comparison between objects play a very important role in many business operations such as ranking of objects with respect to the preferences of customers, finding a set of top-k objects in ranking order, and finding a set of k-nearest neighbor objects in ranking order and so on present study proposes a new similarity finding measure between objects. This measure computes weighted sums of values of attributes and priority values of respective values of attributes. These weighted sums are computed using a linear function formula. Finally a new clustering technique is proposed for clustering market basket analysis products using newly proposed similarity search measure between two objects new clustering technique based on new similarity finding measure is very useful in many real time applications and in many very large database operations including query execution.

Index Terms - Items clustering, attribute values, customer opinions, threshold, weighted sums, linear function, similarity measures

1. Introduction

Database operations, particularly large database operations and are crucial in many real life database critical operations in market data analysis, web searching, scientific data analysis and research and so on finding similarity between objects play a crucial role in clustering as well as in data analysis. Existing methods for finding similarity between objects are completely based only on the values of attributes of objects. Various objects' similarity finding measures are having been proposed in the literature of similarity estimations. For instance it is used to find pages or documents with similar words over the web [1] or in order to detect customers with abnormal behavior based on the products they buy [2]. Estimation of the similarity between objects is a fundamental operation in data management [3]. In the literature many different similarity metrics have been proposed for evaluating the similarity between two data items, such as the Euclidean distance and the cosine similarity [3].

Some of the important techniques that are used for finding similarity between objects are:

1. Euclidean distance
2. Longest common sub sequence (LCSS)
3. Cosine similarity
4. Edit distance
5. k-nearest neighbor measure
6. Kernel distance measure

Similarity computations can be performed for the detection of similar conversations and comments between users of the social networks (i.e., comments on Facebook, tweets on Twitter) [4]. In order to perform such kind of similarity computations, we exploit a query type, termed a reverse top-k query [4]. In contrast to a top-k query that returns the k products with the best score for a specific customer, the result of a reverse top-k query is the set of customers for whom a given product belongs to their top-k set [5].

All these methods estimate objects' similarity measures between objects based on only values of attributes. In reality, similarity values are computed more accurately and more generally when priority values of attributes are taken into consideration in addition to the values of attributes. These types of requirements are very useful in business applications and many critical database operations. For example, in production management attribute values and opinions of customers are both very useful for ranking the products based on the customers' preferences. The proposed technique is also useful for executing many of the database query operations efficiently, effectively and optimally.

Instead of considering only values of attributes new linear functions based technique is proposed for product clustering with respect to opinions of customer's. Linear function computes similarity values using both values of attributes of the product and the respective priorities or opinions of values of those attributes. This linear functions is denoted and represented as

f-weighted(product p) =

p[1]*priority[1]+p[2]*priority[2]+....+p[n]*priority[n]

where, p is the product, p[1] is the value of the first attribute of the product p, p[2] is the value of the second attribute of the product p, priority[1] is the preference value of the first attribute, priority[2] is the preference value of the second attribute, priority[n] is the preference value of the nth attribute.

2. Problem Definition

All the business items, their values of attributes and the corresponding opinions of attributes specified in the respective tables. In the previous methods only distance measures are used for finding similarity between objects. Present study proposes a new similarity finding measure between any two comparable data items or objects. This new measure is a weighted combination of values of objects and their respective preferences or opinions of those attribute values. Authors addresses the problem of measuring the quality of top-k result sets returned by an information retrieval system, as is the case of comparing search engine results [6].

Data mining techniques are very much useful in many real time applications. Most important data mining applications are:

1. Classification
2. Clustering
3. Association
4. Outlier detection
5. Web mining
1. Text mining
2. Graph mining and so on

Present paper proposes a new clustering technique for clustering business items using both values of attributes and their believable factors of preferences. This clustering technique uses newly proposed similarity measure between two objects. This clustering technique of clustering business items is very useful in many real time business critical applications.

ITEM	Cost($)	QUALITY(points)
I-1	25	44
I-2	16	66
I-3	86	94
I-4	45	26
I-5	60	80
I-6	80	90
I-7	90	60
I-8	36	58
I-9	26	40
I-10	76	20
I-11	90	80
I-12	40	60
I-13	30	40
I-14	80	90

Table-1: ITEMS DETAILS

Customer	Priority (cost)	Priority (quality)
C-1	0.30	0.70
C-2	0.90	0.10
C-3	0.40	0.60
C-4	0.10	0.90
C-5	0.60	0.40
C-6	0.80	0.20
C-7	0.50	0.50
C-8	0.20	0.80
C-9	0.70	0.30
C-10	0.40	0.60

Table-2 Priority for ITEM-1 by all customers

Customer	Priority (cost)	Priority (quality)
C-1	0.40	0.60
C-2	0.30	0.70
C-3	0.90	0.10
C-4	0.60	0.40
C-5	0.10	0.90
C-6	0.50	0.50
C-7	0.45	0.55
C-8	0.25	0.75
C-9	0.45	0.55
C-10	0.55	0.45

Table-3 Priority for ITEM-2 by all customers

Customer	Priority (cost)	Priority (quality)
C-1	0.35	0.65
C-2	0.25	0.75
C-3	0.30	0.70
C-4	0.40	0.30
C-5	0.70	0.30
C-6	0.90	0.10
C-7	0.85	0.15
C-8	0.65	0.35
C-9	0.45	0.55
C-10	0.50	0.50

Table-4 Priority for ITEM-3 by all customers

Customer	Priority (cost)	Priority (quality)
C1	0.40	0.60
C2	0.60	0.40
C3	0.80	0.20
C4	0.20	0.80
C5	0.35	0.65
C6	0.45	0.55
C7	0.65	0.35
C8	0.75	0.25
C9	0.85	0.15
C10	0.10	0.90

Table-5 Priority for ITEM-4 by all customers

Customer	Priority (cost)	Priority (quality)
C1	0.90	0.60
C2	0.60	0.40
C3	0.80	0.20
C4	0.75	0.80
C5	0.45	0.65
C6	0.15	0.55
C7	0.40	0.35
C8	0.35	0.25
C9	0.65	0.15
C10	0.70	0.90

Table-6 Priority for ITEM-5 by all customers

Customer	Priority (cost)	Priority (quality)
C1	0.40	0.60
C2	0.70	0.30
C3	0.80	0.20
C4	0.20	0.80
C5	0.50	0.50
C6	0.30	0.70
C7	0.10	0.90
C8	0.55	0.45
C9	0.60	0.40
C10	0.45	0.55

Table-7 Priority for ITEM-6 by all customers

Customer	Priority (cost)	Priority (quality)
C1	0.40	0.60
C2	0.10	0.90
C3	0.90	0.10
C4	0.70	0.30
C5	0.50	0.50
C6	0.20	0.80
C7	0.30	0.70
C8	0.80	0.25
C9	0.60	0.40
C10	0.25	0.75

Table-8 Priority for ITEM-7 by all customers

Customer	Priority (cost)	Priority (quality)
C1	0.60	0.40
C2	0.20	0.80
C3	0.90	0.10
C4	0.10	0.90
C5	0.80	0.20
C6	0.30	0.70
C7	0.40	0.60
C8	0.65	0.35
C9	0.45	0.55
C10	0.50	0.50

Table-9 Priority for ITEM-8 by all customers

Customer	Priority (cost)	Priority (quality)
C1	0.20	0.80
C2	0.45	0.65
C3	0.85	0.15
C4	0.60	0.40
C5	0.20	0.80
C6	0.80	0.20
C7	0.40	0.60
C8	0.70	0.30
C9	0.80	0.20
C10	0.40	0.60

Table-10 Priority for ITEM-9 by all customers

Customer	Priority (cost)	Priority (quality)
C1	0.60	0.40
C2	0.40	0.60
C3	0.70	0.30
C4	0.50	0.50
C5	0.60	0.40
C6	0.50	0.50
C7	0.90	0.10
C8	0.30	0.70
C9	0.40	0.60
C10	0.30	0.70

Table-11 Priority for ITEM-10 by all customers

Customer	Priority (cost)	Priority (quality)
C1	0.40	0.60
C2	0.60	0.40
C3	0.80	0.20
C4	0.80	0.20
C5	0.30	0.70
C6	0.60	0.40
C7	0.40	0.60
C8	0.50	0.50
C9	0.70	0.30
C10	0.40	0.60

Table-12 Priority for ITEM-11 by all customers

Customer	Priority (cost)	Priority (quality)
C1	0.90	0.10
C2	0.80	0.20
C3	0.60	0.40
C4	0.40	0.60
C5	0.80	0.20
C6	0.90	0.10
C7	0.30	0.70
C8	0.90	0.10
C9	0.30	0.70
C10	0.60	0.40

Table-13 Priority for ITEM-12 by all customers

Linear weighted function is

FW = I_1[cost] * priority + I_1[quality] * priority[quality]

Product	I_1[cost] * priority + I_1[quality] * priority[quality]
I_1C_1	25*0.30 + 44 * 0.70 = 7.50 + 30.80 = 38.30
I_1C_2	25*0.90 + 44 * 0.10 = 22.50 + 4.40 = 26.90
I_1C_3	25*0.40 + 44 * 0.60 = 10.0 + 26.40 = 36.40
I_1C_4	25*0.10 + 44 * 0.90 = 2.50 + 39.60 = 42.10
I_1C_5	25*0.60 + 44 * 0.40 = 15.0 + 17.60 = 32.60
I_1C_6	25*0.80 + 44 * 0.20 = 20.0 + 8.80 = 28.80
I_1C_7	25*0.50 + 44 * 0.50 = 12.5 + 22.0 = 34.50
I_1C_8	25*0.20 + 44 * 0.80 = 5.0 +35.20 = 40.20
I_1C_9	25*0.70 + 44 * 0.30 = 17.5 + 13.20 = 30.60
I_1C_{10}	25*0.40 + 44 * 0.50 = 10.0 + 26.40 = 36.40

Table-14 Weighted sums for item I_1

Sum value = 346.90, average of 346.90 = 34.6

Preference list of customers who have item I_1 in their preference list is = {C_2, C_6, C_9, C_5, C_7}

Product	I_1[cost] * priority + I_1[quality] * priority[quality]
I_2C_1	16*0.40 + 66 * 0.60 = 6.40 + 39.60 = 46.0
I_2C_2	16*0.30 + 66 * 0.70 = 4.80 + 46.20 = 51.0
I_2C_3	16*0.90 + 66 * 0.10 = 14.0 + 6.60 = 21.0
I_2C_4	16*0.60 + 66 * 0.40 = 9.60 + 26.40 = 36.0
I_2C_5	16*0.10 + 66 * 0.90 = 1.60 + 59.40 = 61.0
I_2C_6	16*0.50 + 66 * 0.50 = 8.0 + 33.0 = 41.0
I_2C_7	16*0.45 + 66 * 0.65 = 7.2 + 42.90 = 50.10
I_2C_8	16*0.25 + 66 * 0.75 = 4.0 +49.50 = 53.50
I_2C_9	16*0.45 + 66 * 0.55 = 7.2 + 36.30 = 43.50
I_2C_{10}	16*0.55 + 66 * 0.45 = 8.80 + 29.70 = 38.50

Table-15 Weighted sums for item I_2

Sum value = 441.60, average of 441.60 = 44.1

Preference list of customers for threshold 45 who have item I_2 in their preference list is = {C_3, C_4, C_6, C_9, C_{10}

Product	I_1[cost] * priority + I_1[quality] * priority[quality]
I_3C_1	86*0.30 + 94 * 0.65 = 30.1 + 61.10 = 91.20
I_3C_2	86*0.90 + 94 * 0.75 = 21.50 + 70.4 = 92.00
I_3C_3	86*0.40 + 94 * 0.70 = 25.8 + 65.8 = 91.60
I_3C_4	86*0.10 + 94 * 0.60 = 34.4 + 56.4 = 90.80
I_3C_5	86*0.60 + 94 * 0.30 = 60.2 + 29.4 = 89.60
I_3C_6	86*0.80 + 94 * 0.10 = 77.4 + 9.4 = 86.80
I_3C_7	86*0.50 + 94 * 0.15 = 73.1 + 14.1 = 87.20
I_3C_8	86*0.20 + 94 * 0.35 = 55.9 +32.9 = 88.80
I_3C_9	86*0.70 + 94 * 0.65 = 38.7 + 61.1 = 99.80
I_3C_{10}	86*0.40 + 94 * 0.50 = 43.0 + 47.0 = 90.00

Table-16 Weighted sums for item I_3

Sum value = 907.80, average of 907.80 = 90.78

Preference list of customers for threshold 91.0 who have item I_3 in their preference list is = {C_4, C_5, C_6, C_7, C_8, C_{10}}

Product	I_1[cost] * priority + I_1[quality] * priority[quality]
I_4C_1	45*0.40 + 26 * 0.60 = 18.0 + 15.60 = 33.60
I_4C_2	45*0.60 + 26 * 0.40 = 27.0 + 10.40 = 37.40
I_4C_3	45*0.80 + 26 * 0.20 = 36.0 + 5.20 = 37.20
I_4C_4	45*0.20 + 26 * 0.80 = 9.0 + 20.80 = 29.80
I_4C_5	45*0.35 + 26 * 0.65 = 15.75 + 16.9 = 32.65
I_4C_6	45*0.45 + 26 * 0.55 = 20.25 + 14.3 = 34.55
I_4C_7	45*0.65 + 26 * 0.35 = 29.25 + 9.1 = 38.35
I_4C_8	45*0.75 + 26 * 0.25 = 33.75 +6.5 = 40.25
I_4C_9	45*0.85 + 26 * 0.15 = 38.25 + 3.9 = 42.15
I_4C_{10}	45*0.10 + 26 * 0.90 = 4.50 + 23.40 = 27.9

Table-17 Weighted sums for item I_4

Sum value = 353.85, average of 353.85 = 35.385

Preference list of customers for threshold 36.0 who have item

I_4 in their preference list is = {C_1, C_4, C_5, C_6, C_{10}}

Similarly remaining computation details are shown below:

Weighted computations for product I_5 are

Sum value = 685.0, average of 685.0 = 68.5 ≈ 69

Preference list of customers for threshold 69.0 who have item I_5 in their preference list is = {C_1, C_2, C_3, C_4, C_9, C_{10}}

Weighted computations for product I_6 are

Sum value = 854.0, average of 854.0 = 85.4 ≈ 86.0

Preference list of customers for threshold 86.0 who have item I_6 in their preference list is = {C_1, C_2, C_3, C_5, C_8, C_9, C_{10}}

Weighted computations for product I_6 are

Sum value = 854.0, average of 854.0 = 85.4 ≈ 86.0

Preference list of customers for threshold 86.0 who have item I_6 in their preference list is = {C_1, C_2, C_3, C_5, C_8, C_9, C_{10}}

Weighted computations for product I_7 are

Sum value = 742.50, average of 742.50 = 74.25 ≈ 75.0

Preference list of customers for threshold 86.0 who have item I_7 in their preference list is = {C_1, C_2, C_5, C_6, C_7, C_{10}}

Weighted computations for product I_8 are

Sum value = 469.30, average of 469.30 = 46.93 ≈ 47.0

Preference list of customers for threshold 86.0 who have item I_8 in their preference list is = {C_1, C_3, C_5, C_8, C_9, C_{10}}

Preference list of customers for item I_9 in their preference list is = {C_2, C_3, C_4, C_6, C_7, C_8}

Preference list of customers for item I_{10} in their preference list is = {C_2, C_4, C_6, C_8, C_9, C_{10}}

Preference list of customers for item I_{11} in their preference list is = {C_1, C_2, C_5, C_6, C_7, C_8, C_{10}}

Product	Customer list of preferences
P1	C_2, C_6, C_9, C_5, C_7
P2	C_3, C_4, C_6, C_9, C_{10}
P3	C_4, C_5, C_6, C_7, C_8, C_{10}
P4	C_1, C_2, C_3, C_4, C_9, C_{10}
P5	C_1, C_2, C_3, C_4, C_9, C_{10}
P6	C_1, C_2, C_3, C_5, C_8, C_9, C_{10}
P7	C_1, C_2, C_5, C_6, C_7, C_{10}
P8	C_1, C_3, C_5, C_8, C_9, C_{10}
P9	C_2, C_3, C_4, C_6, C_7, C_8
P10	C_2, C_4, C_6, C_8, C_9, C_{10}
P11	C_1, C_2, C_5, C_6, C_7, C_8, C_{10}
P12	C_1, C_2, C_5, C_6, C_8
P13	NIL
P14	NIL

Table-18 Clusters of customers with respect to products

Preference list of customers for item I_{12} in their preference list is = {C_1, C_2, C_5, C_6, C_8}

New clustering algorithm

1. Read details of ‘n’ number of items
2. Read priorities of values of attributes by customers
3. Compute weighted sums of all items with respect to their attribute values and priorities
4. Store lists of priorities of all customers separately based on item type
5. For each pair of items I and j estimate similarity measure between items I and j
6. Group items based on the specified highest threshold values of similarity measures
7. Repeat the steps 5 and 6 for the ungrouped items until a specified condition is true
8. Print all final clusters of items

Intersection divided by union measure computations are shown below:

$$P_1\,\partial\,P_2 = \frac{|P_1 \cap P_2|}{|P_1 \cup P_2|}$$

$$= \frac{|\{C_{6,}C_9\}|}{|\{C_2, C_3, C_4, C_5, C_6, C_7, C_9, C_{10}\}|}$$

$$= \frac{2}{8} \qquad = 0.25$$

$$P_1\,\partial\,P_3 = \frac{|P_1 \cap P_3|}{|P_1 \cup P_3|}$$

$$= \frac{|\{C_{6,}C_5, C_9\}|}{|\{C_2, C_4, C_5, C_6, C_7, C_8, C_9, C_{10}\}|}$$

$$= \frac{3}{8} \qquad = 0.375$$

$$P_1\,\partial\,P_4 = \frac{|P_1 \cap P_4|}{|P_1 \cup P_4|} = \frac{|\{C_{6,}\ C_5\}|}{|\{C_1, C_2, C_4, C_5, C_6, C_7, C_9, C_{10}\}|}$$

$$= \frac{2}{8} \qquad = 0.25$$

$$P_1\,\partial\,P_5 = \frac{|P_1 \cap P_5|}{|P_1 \cup P_5|}$$

$$= \frac{|\{C_2, C_9\}|}{|\{C_1, C_2, C_3, C_4, C_5, C_6, C_7, C_9, C_{10}\}|}$$

$$= \frac{2}{9} \qquad = 0.22$$

$$P_1\,\partial\,P_6 = \frac{|P_1 \cap P_6|}{|P_1 \cup P_6|}$$

$$= \frac{|\{C_{2,}C_5, C_9\}|}{|\{C_1, C_2, C_3, C_5, C_6, C_7, C_8, C_9, C_{10}\}|}$$

$$= \frac{3}{9} \qquad = 0.333$$

$$P_1\,\partial\,P_7 = \frac{|P_1 \cap P_7|}{|P_1 \cup P_7|} = \frac{|\{C_{2,}C_6, C_5, C_7\}|}{|\{C_1, C_2, C_5, C_6, C_7, C_9, C_{10}\}|}$$

$$= \frac{4}{7} \qquad = 0.57$$

$$P_1\,\partial\,P_8 = \frac{|P_1 \cap P_6|}{|P_1 \cup P_6|}$$

$$= \frac{|\{C_{5,}, C_9\}|}{|\{C_1, C_2, C_3, C_5, C_6, C_7, C_8, C_9, C_{10}\}|}$$

$$= \frac{2}{9} \qquad = 0.22$$

$$P_1\,\partial\,P_9 = \frac{|P_1 \cap P_9|}{|P_1 \cup P_9|} = \frac{|\{C_{2,}C_6, C_7\}|}{|\{C_2, C_3, C_4, C_5, C_6, C_7, C_8, C_9\}|}$$

$$= \frac{3}{8} \qquad = 0.375$$

$$P_1\,\partial\,P_{10} = \frac{|P_1 \cap P_{10}|}{|P_1 \cup P_{10}|}$$

$$= \frac{|\{C_{2,}C_6, C_9\}|}{|\{C_2, C_4, C_5, C_6, C_7, C_8, C_9, C_{10}\}|}$$

$$= \frac{3}{8} \qquad = 0.375$$

$$P_1\,\partial\,P_{11} = \frac{|P_1 \cap P_{11}|}{|P_1 \cup P_{11}|}$$

$$= \frac{|\{C_{2,}C_5, C_6, C_7\}|}{|\{C_1, C_2, C_5, C_6, C_7, C_8, C_9, C_{10}\}|}$$

$$= \frac{4}{8} \qquad = 0.5$$

$$P_1\,\partial\,P_{12} = \frac{|P_1 \cap P_{12}|}{|P_1 \cup P_{12}|} = \frac{|\{C_{2,}C_7, C_6\}|}{|\{C_1, C_2, C_5, C_6, C_7, C_8, C_9\}|}$$

$$= \frac{3}{7} \qquad = 0.43$$

Items P_1, P_7, P_{11} and P_{12} are grouped into one cluster based on the highest threshold values. After first iteration remaining products are:

In the second iteration same process is repeated and ITEM-2 is compared with all the un-clustered items.

Product	Customer list of preferences
P2	C_3, C_4, C_6, C_9, C_{10}
P3	C_4, C_5, C_6, C_7, C_8, C_{10}
P4	C_1, C_2, C_3, C_4, C_9, C_{10}
P5	C_1, C_2, C_3, C_4, C_9, C_{10}
P6	C_1, C_2, C_3, C_5, C_8, C_9, C_{10}
P8	C_1, C_3, C_5, C_8, C_9, C_{10}
P9	C_2, C_3, C_4, C_6, C_7, C_8
P10	C_2, C_4, C_6, C_8, C_9, C_{10}

Table -19

$$P_2\,\partial\,P_3 = \frac{|P_2 \cap P_3|}{|P_2 \cup P_3|}$$

$$= \frac{\left|\{C_{4,}C_6, C_{10}\}\right|}{|\{C_3, C_4, C_5, C_6, C_7, C_8, C_9, C_{10}\}|}$$

$$= \frac{3}{8} \qquad = 0.375$$

$$P_2\,\partial\,P_4 = \frac{|P_2 \cap P_4|}{|P_2 \cup P_4|} = \frac{\left|\{C_{4,}C_6, C_{10}\}\right|}{|\{C_1, C_3, C_4, C_5, C_6, C_9, C_{10}\}|}$$

$$= \frac{3}{7} \qquad = 0.43$$

$$P_2\,\partial\,P_5 = \frac{|P_2 \cap P_5|}{|P_2 \cup P_5|} = \frac{\left|\{C_3, C_{4,}C_9, C_{10}\}\right|}{|\{C_1, C_2, C_3, C_4, C_6, C_9, C_{10}\}|}$$

$$= \frac{4}{7} \qquad = 0.57$$

$$P_2\,\partial\,P_8 = \frac{|P_2 \cap P_8|}{|P_2 \cup P_8|}$$

$$= \frac{\left|\{C_{3,}C_9, C_{10}\}\right|}{|\{C_1, C_3, C_4, C_5, C_6, C_8, C_9, C_{10}\}|}$$

$$= \frac{3}{8} \qquad = 0.375$$

$$P_2\,\partial\,P_9 = \frac{|P_2 \cap P_8|}{|P_2 \cup P_8|}$$

$$= \frac{\left|\{C_{3,}C_4, C_6\}\right|}{|\{C_2, C_3, C_4, C_6, C_7, C_8, C_9, C_{10}\}|}$$

$$= \frac{3}{8} \qquad = 0.375$$

$$P_2\,\partial\,P_{10} = \frac{|P_2 \cap P_{10}|}{|P_2 \cup P_{10}|} = \frac{\left|\{C_{4,}C_6, C_9, C_{10}\}\right|}{|\{C_2, C_3, C_4, C_6, C_8, C_9, C_{10}\}|}$$

$$= \frac{4}{8} \qquad = 0.375$$

Items P_2, P_4, P_5 and P_{10} items grouped into one cluster based on the highest threshold values. After second iteration remaining ungrouped preference lists of customers are

Product	Preference lists of customers
P3	C_4, C_5, C_6, C_7, C_8, C_{10}
P6	C_1, C_2, C_3, C_5, C_8, C_9, C_{10}
P8	C_1, C_3, C_5, C_8, C_9, C_{10}
P9	C_2, C_3, C_4, C_6, C_7, C_8

Table-20

Final clusters are

Cluster No	Products
Cluster1	P_1, P_7, P_{11}, P_{12}
Cluster2	P_2, P_4, P_5, P_{10}
Cluster3	P_3, P_6, P_8, P_9

Table 21

Conclusion

Clustering method based on the similarity measure estimation between objects is a very important technique in data mining as well as machine learning techniques. This technique is very useful in many real time applications such as spatial database maintenance, product management, sales forecasting of items, product priority estimation purpose, and ranking of items in many database queries, finding k-nearest neighbor items, cross selling of products in the estimated potential sales area, improving of market basket sales of products. Also, clustering has many applications in research, science, and other applications. In the feature there is a scope for clustering data values containing uncertain values in the product specifications, distance measures and so on.

References

1. Rajaraman and J. D. Ullman, Mining of Massive Datasets. Cambridge,U.K.: Cambridge Univ. Press, 2012.
2. K. Georgoulas and Y. Kotidis, “Towards enabling outlier detection in large, high dimensional data warehouses,” in Proc. Scientific Statistical Database Manage., 2012, pp. 591–594.
3. Konstantinos Georgoulas, Akrivi Vlachou, Christos Doulkeridis, and Yannis Kotidis User-Centric Similarity Search, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 29, NO. 1, JANUARY 2017
4. H. Becker, M. Naaman, and L. Gravano, “Learning similarity metrics for event identification in social media,” in Proc. 3rd ACM Int.Conf. Web Search Data Mining, 2010, pp. 291–300.
5. Vlachou, C. Doulkeridis, Y. Kotidis, and K. Nørva g, “Reverse top-k queries,” in Proc. IEEE Int. Conf. Data Eng., 2010, pp. 365–376.
6. R. Fagin, R. Kumar, and D. Sivakumar, “Comparing top-k lists,”SIAM J. Discrete Math., vol. 17, no. 1, pp. 134–160, 200

OPENION FOR ITEM-2

Customer	Priority (cost)	Priority (quality)
C-1	0.40	0.60
C-2	0.60	0.40
C-3	0.80	0.20
C-4	0.20	0.80