Improving Shopping Experience of Customers using Trajectory Data Mining

Tanuja. V and Govindarajulu. P

Department of Computer Science, S. V. University, Tirupati, Andhra Pradesh, INDIA

Abstract

The up-to-date tracking tools, such as RFID (Radio Frequency Identification), Wi-Fi sensing brought a tremendous breakthrough in business analytics. These current trends can be used to record the paths taken by shoppers in a supermarket which form a trajectory data. By mining these trajectories it is possible to uncover shopper's behavior including how they move among shelves of different product types, the time they spend on specific shelves, and other information that can be used to make a product arrangement that improves customers' shopping experience and hence attract customers. Trajectories of shoppers in a grocery store as recorded from RFID tags or with the records of attendant located on their shopping ride forms the basis for the analysis. By using a clustering algorithm, we can profile shopping paths by neighborhoods visited and then by time spent to uncover the common paths of different categories of shoppers such as those who are looking to grab a few important items and leave. The clustered shoppers' trajectories of items uncover the customer behavior that can be used to sketch groups of shoppers based on their space use and assess the impact of the spatial configuration of a store layout on shoppers' behavior. In this paper, we propose a novel method of shopping super market items of trajectory clustering algorithm which groups similar transactions of super market item sets (trajectory) for shop which elucidates customer shopping behavior.

A new similarity measure between transaction trajectories of super market items to group market basket item sequences of trajectories is developed. Our proposed method can find the dominant super market shopping path sequences of item sets that are capable of identifying the hotspots where most of the customers' visits are made and the least visited paths as well. This can understand shopper behavior in a store. We used the real dataset of a supermarket in Nellore to apply the proposed methodology, as case study, to demonstrate the advantages and usefulness of the method.

Key words:

Data mining, Market Basket Analysis, Clustering, LCSS clustering, Sequence of items, purchasing patterns

1. Introduction

The goal of floating customer loyalty in any business requires an emphasis on one-to-one marketing and personalized services and it is the heart of CRM (Customer Relationship management). It is essential to understand individual customer preferences for products, to recommend the most appropriate product. The identification and recommendation of such products are all based on the use of customer's real-time buying activities such as viewing, basket placement, and purchasing of products. Bartholomaus Endeand Rudiger Brause et al. proposed new mutual information based clustering approach and outlined its implications for the example of user profiling [1]. Ching-Huang Yun, Kun-Ta Chuang and Ming-Syan Chen et al. developed an efficient clustering algorithm for market basket data analysis of purchased data items to minimize the small large ratio in each group [3]. Ching-Huang Yun, Kun-Ta Chuang and Ming-Syan Chen et al. developed a new measurement in view of the features of market-basket data; this measure is called the category-based adherence, and utilizes this measurement to perform the clustering [4].

Market basket analysis is one of the data mining methods focusing on discovering purchasing patterns by extracting associations or co-occurrences from a store's transactional data [17].Agglomerative hierarchical clustering creates a hierarchy of clusters which are represented in a tree structure called a Dendorogram [21]. Association rule mining is a method commonly used for Market Basket Analysis. In retail, Market Basket Analysis helps the retailer to find commonly purchased products so as to identify cross-selling opportunities, optimize store layout, and manage inventory [26].

The motivation for mining datasets is the possibility of realizing inherent information, helping to gain understanding of the fundamental phenomenon of movement [20]. Data mining techniques have attracted a growing amount of attention by many industrial companies due to its wide ability to improving marketing strategies with an opportunity of major revenues [17]. A common architecture for very large-scale applications is a cluster of compute nodes [14].

Association rule mining (ARM) is an important core data mining technique to discover patterns/rules among items in a large database of variable-length transactions. The goal of ARM is to identify groups of items that most often occurs together [12]. The data that represent the movement of an object have been referred to using several different terms including trace data or traces, movement data, and mobility data [10]. Savitha S. Kadiyala and Alok Srivastav

Manuscript received February 5, 2017
Manuscript revised February 20, 2017

et al. introduced a framework for identifying appropriate data mining techniques for various CRM activities [12].
In today's buyers 'market situation the goal of retailers is to increase the gross profit margin through sales improvement and cost reduction. This needs improving the efficiency of operation and providing attractive services for customers. Especially, the market focus of discount stores has been only low-price strategy with the expansion of many branch stores. Market basket analysis or regional analysis based on customer purchase history and demographic information forms the basis for new strategy. A methodological approach to analyzing in-store customer behavior with a view to optimizing space and store performance, understanding these customer processes and movement patterns thus helped the retail collaborator maximize the performance of the store. If we can find the areas where most sales activities occur and where customers tend to stay for a long time in the store, store manager can understand where to display products and how to build an effective store environment. A more effective store environment can provide convenient services for customers and hence increases sales. Up to now, however, store managers have relied on experiences of the high-sales locations and those where customers tend to stay for a long time. Based on their experience, they decided where to display products and how to change in-store layout. In- Chul Jung et al. have proposed a shopping path clustering algorithm with sequence pattern matching method, LCSS (longest common subsequence) method, which groups have similar moving path for shop in order to understand characteristics of shopping [8]. By adopting the longest common subsequence (LCSS) method as the basic idea and expanding on it, authors have developed the main shopping path clustering algorithm that is capable of identifying the hotspots where most of the customers' visits are made and the dead spots with few visits [8]. The proposed similarity measure based on LCS is a better measure for moving paths that have different moving length compared to Euclid distance.

2. Related Work

Market Basket Analysis (MBA) allows researchers to uncover non-obvious and usually hidden and counterintuitive associations between products, items, or categories. This methodological approach allows researchers to identify those items that co-occur (i.e., appear together) on a frequent basis and assess the extent to which they co-occur. MBA has been used to understand consumer behavior regarding types of books that are purchased together (as purchased on Amazon.com) as well as different types of wines that the same individual is likely to purchase (as purchased on VirginWines.com) (Berry & Linoff, 2004). Because MBA originated in the field of marketing and was initially used to understand which supermarket items are purchased together (i.e., placed together in the same "basket"), the technique adopted the name market basket analysis [7].
Some researchers have studied customer behaviors using direct observation or questionnaires. In-store advertisements and promotions have proven records to amplify the magnitude of unplanned purchasing among consumers (McClure and West, 1969). Cox (1964) measured relationship between shelf space and product sales. Dickson and Sawyer (1986) investigated consumers' knowledge and use of price information at the supermarket point of purchase (POP) and Hoyer (1984) provided a view of decision making based on the notion that consumers are not motivated to engage in a great deal of in-store decision making at the time of purchase when the product is purchased repeatedly and is relatively unimportant. Radio frequency identification (RFID) and clustering techniques to analyze customers shopping path. Many research mainly have used to clustering algorithm among data mining methods to detect the main shopping path patterns. Larson et al. (2005) and Hui et al. (2009) tried to identify the major shopping path using the k-medoids clustering algorithm. However, a clustering algorithms using Euclidean distance similarity measure on shopping path suffers from two problems. First, during the process of clustering, the clusters which are divided at the location of obstacles such as sales shelves can converge into the same cluster group. Because people cannot walk cross obstacles such as shelves and merchandise stands, the store's physical environment and obstructions (shelves, merchandise stands, etc.) should be considered as a constraint for the shopping path clustering. Second, the length of a shopping path must be constant in order to apply a clustering algorithm to the shopping path data; however, this length is actually variable in a store [8]. An algorithm is proposed for clustering using fitness values. A fitness value is assigned to each tuple using the fitness function. Based upon this fitness value the tuples will be assigned to the clusters. If the fitness value of the tuple is equal to or nearly equal to the threshold value of the generated set of random clusters then only the tuple will be assigned to the cluster otherwise tuple is assigned to the outlier cluster. If there are many clusters in the outlier cluster then a similarity is calculated among these clusters and outlier is detected. In this approach, tie can also occur i.e. if atuple belongs to two clusters then we can arbitrarily assign this tuple to any one cluster [24].
Techniques of Data mining for market baskets: There are many data mining techniques and algorithms that are available to discover meaningful pattern and rules.

There are many different data mining techniques that are available for data management and some of the techniques are given below:

***Classification*:** In classification, first examine the features of newly presented object and assign it to a predefined class for example classifies the credit applicants as low, medium or high risk.

***Association*:** The main goal of association is to establish the relationship between items which exist in the market. Typical examples of association modeling are Market basket Analysis and cross selling programs. The tools used for association rule mining are apriori algorithm and weka tool kit.

***Prediction*:** In this functionality, prediction of some unknown or missing attributes values based on other Information. For example: Forecast the sale value for next week based on available data.

***Clustering*:** In this, Data Mining organizes data into meaningful sub-groups (clusters) such that points within thegroup are similar to each other, and as different as possible from the points in the other groups. It is an unsupervised classification. An effective dynamic unsupervised clustering algorithmic approach for market basket analysis has been proposed by Verma et al.

***Outlier Analysis*:** In this, Data Mining is done to identify and explain exceptions. For example, in case of Market Basket Data Analysis, outlier can be some transaction which happens unusually [22].

3. Trajectory Analysis

3.1 Trajectory analysis in general

The main goal of the trajectory pattern data mining is to find hot regions from the trajectories, and then find sequential relationships among the hot regions, and then use these results in many real time applications such as clustering, classification, association and many other applications. There exist many prediction techniques that use vector based, pattern based, and association based models in order to predict location of users at the specific time given. Some authors have used association rules for storing movement behaviors of users and at the same time these association rules are represented in the trajectory pattern tree. Signature tree is also used by some authors as an indexing structure for deriving and managing sequential pattern relationship. In the literature of trajectory data management, there exist many methods and plans to discover the movement behavior details of users. Present study explains movement behaviors of users and then how these movement behavior details of users are used for clustering or classifying the users into groups.

With the help of many modern technologies such as global positioning systems (GPS), smart phone sensors and mobiles it is very easy to collect and store very large scale trajectory data of tracking traces of moving objects - vehicles, persons, animals, and vessels etc. Examples for moving objects are - people vehicles, animals, and hurricanes etc. Many trajectory data mining based applications are beneficial to the universities, telecommunication industries, government, common people, business people and many commercial organizations. Storing and managing very large trajectory training data sets is very difficult. Comparing trajectories of two different objects is the fundamental function in trajectory data mining and as such it is very difficult to find a similarity metric for comparing trajectory sets of two different users because trajectories of different objects are constructed with different sampling strategies and sampling rates. Constructing a similarity metric for comparing trajectories of two different objects is very difficult because trajectories belonging to two different objects are not homogeneous. Important layers in the trajectory data mining framework are - *data collection*, *trajectory data mining techniques*, *applications* [27].

Potential applications of data mining are retailing, banking, credit card management, insurance, telecommunications, telemarketing and human resource management. There exist many applications of similarity search on spatio - temporal trajectories. A suitable distance function is used for finding similarity between uncertain trajectories. Trajectory clustering algorithms for moving objects are very useful in finding traffic jams, important location identification, and facilities available at a particular location at particular time etc. Some of the applications of trajectory data mining are Route Recommendation, Animal Migration, Transportation Management, Real time Traffic Information Details of Transport Organization, and Tourism. Some of the useful trajectory data mining applications are path discovery, shortest path discovery, individual behavior prediction, group behavior prediction, location prediction, service prediction and so on. Applications of trajectory data mining can be classified based on domain of application as follows:

***Path Discovery*:** Path discovery is also known as route discovery. There exist many ways for finding a path between any two given nodes. Sometimes there is a need to find most frequent path between two locations in a certain time period. In many real time applications there is a need to find an efficient and correct path. For some applications, shortest path is needed, some applications require shortest path, another set of applications need cheapest or most popular path. Path discovery must find at least one path. Most frequent paths are better than the fastest paths or shortest paths in many real time applications .In terms of public transportation, people's real demand for public

transportation are employed to identify and optimize existing flawed bus routes, thus improving utilization efficiency of public transportation**.** Dai *et al.* proposed a recommendation system that chooses different routes for drivers with different driving preferences. This kind of *personalized route recommendation* avoids flaws of previous unique recommendation and improves quality of user satisfaction. Previous experiences showed that human mobility as extraordinary regular and thus predictable [27].

3.2 Trajectory Similarity

A common problem in measuring the similarity of trajectories and trajectory clustering is the handling of sub-trajectories [9]. Measuring similarity between any two trajectories is definitely one of the most important tasks in trajectory data maintenance since it serves as the basic foundation of many advanced trajectory data analyses such as similarity search, clustering, outlier detection and classification.

In the trajectory literature approximately ten of trajectory similarity measures have been proposed. Some of the important similarity measures are: Euclidean distance (ED), distance based on Longest Common Subsequence (LCSS), Edit Distance with Real Penalty (ERP), Dynamic Time Warping (DTW), and Edit Distance on Real sequence (EDR). Many of these works and some of their ex-tensions have been widely cited in the literature and applied to facilitate query processing and data mining in trajectory data [5].

3.3 Trajectory data analysis of a market basket data in specific

Wireless networks and GPS are the two main sources of trajectory data for moving objects. The technologies like GPS provide a considerably more precise positioning. The behavior of the mobility of people changes over time. For instance, a new place of work, opening and closure of shops or changed means of transportation normally influence the mobile behavior. It is therefore important that algorithms can easily incorporate structural changes and adapt to new patterns in mobile behavior. Due to physical and faster an object moves, the more frequently an object's technical limitations during data collection and storage, uncertainty will arise which is an inherent characteristic of spatiotemporal data. While it can be broadly assumed that time is delivered with high accuracy, uncertainty of location varies with the applied technology between a few meters (GPS) and kilometers (GSM). In addition, the accuracy is greatly influenced by the sampling rate. The location is to be reported to sustain a given level of spatial uncertainty. Background knowledge as well as certain assumptions about movement behavior helps to reduce the uncertainty in data [25].

Trajectory data based clustering methodology is very useful for dividing trajectories into groups with similar movement patterns. Discovery of trajectory patterns is very useful in learning interactions between moving objects. Mobility based data clustering is essentially forming the similarity groups of moving objects such as vehicles, animals, people, and cell phones and so on. Many general frameworks exist for mining communities from multiple sources of trajectories. Moving objects are clustered based on trajectory related information such as semantic meaning of trajectories, weights of locations, movement velocity, feature movements, feature characteristics, temporal duration, and spatial dispersion. The mobility-based clustering is less sensitive than the density-based clustering to the size of trajectory dataset. Huey-Ru and Wu et al. proposed an algorithm called Divclust for finding regional typical moving styles by dividing and then clustering the trajectories. Panagiotakis C et al. proposed a method for trajectory segmentation and sampling based on the representativeness of the trajectories in the moving object databases [28].

CRM is recommended for establishing exceptional relationships with customers and for adding more value to goods and services than what is possible through traditional transaction practices. Traditional marketing was focused in gaining customers. Today it is time to retain the customers. The new CRM paradigm reflects a change in the traditional marketing. Customer retention is essential through great service, trust and, relationship. Then relationship marketing is not only about the 4Ps ((product, price, place, and promotion) but also long-term relationships with people with pace. An early definition of relationship marketing is provided by Gronroos, ''The role of relationship marketing is to identify, establish, maintain and enhance relationships with customers and other stakeholders, at a profit, so that the objectives of all other parties involved are met; and that this is done by a mutual exchange and fulfillment of promises'' [29].

Clustering is a division of data into groups of similar items. Representing the data by fewer clusters necessarily loses certain fine details, but achieves simplification. It models data by its clusters. Data modeling puts clustering in a historical perspective rooted in mathematics, statistics, and numerical analysis. From a machine learning perspective clusters correspond to hidden patterns, the search for clusters is unsupervised learning, and the resulting system represents a data concept. From a practical perspective clustering plays an outstanding role in data mining applications such as scientific data exploration, information retrieval and text mining, spatial database applications, Web analysis, CRM, marketing, medical diagnostics, computational biology, and many others. From

a machine learning perspective clusters correspond to hidden patterns, the search for clusters is unsupervised learning, and the resulting system represents a data concept. Therefore, clustering is unsupervised learning of a hidden data concept.

Data mining deals with large databases that impose on clustering analysis additional severe computational requirements. These challenges led to the emergence of powerful broadly applicable data mining clustering methods surveyed below. Market basket data has this form. Every transaction can be presented in a point-by-attribute format, by enumerating all items j, and by associating with a transaction the binary attributes that indicate whether j-items belong to a transaction or not. Such representation is sparse and two random transactions have very few items in common. This is why similarity (sub-section Proximity Measures) between them is usually measured by Jaccard coefficient. Important source of high dimensional categorical data comes from transactional (market basket) analysis [19].

One of the key techniques used by the large retailers is called Market Basket Analysis (MBA), which uncovers associations between products by looking for combinations of products that frequently co-occur in transactions(frequent item sets). Market Basket Analysis allows retailers to identify relationships between the products that people buy.

Market baskets are real-time sources of data to study customer shopping behavior. Understanding the behavior of the customer guides the business strategy.

Retailers can use the insights gained from MBA in a number of ways, including:

1. Grouping products that co-occur in the design of a store's layout to increase the chance of cross-selling;
2. Driving online recommendation engines ("customers who purchased this product also viewed this product"); and
3. Targeting marketing campaigns by sending out promotional coupons to customers for products related to items they recently purchased.

To carry out an MBA we need a data set of transactions. Each transaction represents a group of items or products that have been bought together and often referred to as an "item set". For example, one item set might be: {pencil, paper, staples, rubber} in which case all of these items have been bought in a single transaction.

In an MBA, the transactions are analyzed to identify rules of association. For example, one rule could be: {pencil, paper} => {rubber}. This means that if a customer has a transaction that contains a pencil and paper, then they are likely to be interested in also buying a rubber. Before acting on a rule, a retailer needs to know whether there is sufficient evidence to suggest that it will result in a beneficial outcome. We therefore measure the strength of a rule by calculating the following three metrics (note other metrics are available, but these are the three most commonly used):

Support: the percentage of transactions that contain all of the items in an item set (e.g., pencil, paper and rubber). The higher the support the more frequently the item set occurs. Rules with a high support are preferred since they are likely to be applicable to a large number of future transactions.

Confidence: the probability that a transaction that contains the items on the left hand side of the rule (in our example, pencil and paper) also contains the item on the right hand side (a rubber). The higher the confidence, the greater the likelihood that the item on the right hand side will be purchased or, in other words, the greater the return rate you can expect for a given rule.

Market Basket Analysis is a vital tool for retailers who want to better understand the relationships between the products that people buy. There are many tools that can be applied when carrying out MBA and the delicate aspects to the analysis are setting the confidence and support thresholds in the Apriori algorithm and identifying which rules are worth pursuing. Typically the latter is done by measuring the rules in terms of metrics that summarize how interesting they are, using visualization techniques and also more formal multivariate statistics. Ultimately the key to MBA is to extract value from your transaction data by building up an understanding of the needs of your consumers. This type of information is invaluable when we are interested in marketing activities such as cross-selling or targeted campaigns.

Sequences of item sets in the market basket can be assumed as trajectories. A customer 'basket represent a trajectory, which provides the information about the sequence of products picked by the customer for his /her basket. Multiple visits of a customer to the grocery shop/supermarket accumulates the trajectory data of that customer .This accumulation forms the basis to study the individual behavior of a customer. The collection of data from all customers visiting the shop constitutes the way to study the group behavior of the customers. The groups can be formed by comparing the sequences/subsequences of individual behavior and ultimately forms the way to find the group behavior.

4. Case Study

4.1 Collection of shopping data

The data is collected from a supermarket at Nellore, Andhra Pradesh. The data consists of Transaction ID, list of items picked into the basket in sequence. The market basket data sample is collected for 100 customers including three to five visits of each customer. The data of each visit is considered as a trajectory sequence of purchased items. In this way for each customer a group of trajectories are picked.

Customer Id	Tid	Sequence of items picked
C0001	101	AB
C0001	102	AC
C0001	103	BC
C0001	104	BD
C0001	105	ABC
C0001	106	ACD
C0001	107	ABCD
C0002	201	AB
C0002	202	AC
C0002	203	BD
C0002	204	ABC
C0002	205	ACD
C0002	206	ABCD
C0003	301	BC
C0003	302	CD
C0003	303	CE
C0003	304	DE
C0003	305	BCD
C0003	306	CDE
C0003	307	CEF
C0003	308	BCDE
C0004	401	CD
C0004	402	CE
C0004	403	CF
C0004	404	DE
C0004	405	CDE
C0004	406	CEF
C0004	407	DEF
C0005	501	EF
C0005	502	EG
C0005	503	FH
C0005	504	EFG
C0005	505	EFH
C0005	506	EGH
C0005	507	EFGH
C0005	508	EGH
C0005	509	EFGH
C0006	601	EF
C0006	602	EG
C0006	603	FH
C0007	701	IJ
C0007	702	IL
C0007	703	JK
C0007	704	KL
C0007	705	IJK
C0007	706	JKL
C0008	801	IJ
C0008	802	JK
C0008	803	KL
C0008	804	JKL
C0009	901	JK
C0009	902	KL
C0009	903	JKL
C0010	1001	KM
C0010	1002	KL
C0010	1003	LM
C0010	1004	LN
C0010	1005	MN
C0010	1006	KMN
C0011	1101	KM
C0011	1102	KL
C0011	1103	MN
C0012	1201	LM
C0012	1202	LN
C0012	1203	MN
C0012	1204	LMN

Details of customer transaction sequences from 13to 24 are shown below

O,P,Q,R,OP,OQ,PQ,PR,OPR,OPQ,OQR
O, P, Q, T, OP,OQ,PQ,OPQ
O, P,Q, OP,OQ,PQ,OPQ
O, P,S,OP,OT,PT,OPT,OPS,OTS
T,U,V,W,TU,TV,VW,TUW,TVW
T,U,V,W,X,TU,VW,WX,TUW
T,V,U,TU,TV,VW,TUW
V,W,X,VW,WX
W,X,Y,Z,WX,WY,XY,YZ,WXY,WYZ,XYZ,WXYZ
X,Y,Z,A,XY,YZ,XYZ
O,P,T,U,OP,OT,PT
T,U,V,W,X,TU,TV,TW,UW,VW,TVX,UWX,VWX

After applying the clustering algorithm, final clusters are shown below

1) A,AB,ABC,ABCD,AC,ACD,B,BC,BCD,BCDE,BD,C,CD, CDE,CE,CEF,CF,D,DE,DEF,E,F
2) E,EF,EFG,EFGH,EFH,EG,EGH,F,FH,G,H,I,J
3) I,IJ,IJK,IL,J,JK,JKL,K,KL,KM,KMN,L,LM,LMN,LN,M,M N,N
4) O,OP,OPQ,OPR,OPS,OPT,OQ,OQR,OT,OTS,P,PQ,PR,PT, Q,R,S,T,U
5) T,TU,TUW,TV,TVW,TVX,TW,U,UW,UWX,V,VW,VWX, W,WX,X
6) A,W,WX,WXY,WXYZ,WY,WYZ,X,XY,XYZ,Y,YZ,Z

4.2 Algorithms

4.2.1 Existing Longest Common Sub Sequence (LCSS) algorithm [8]

1. randomly select starting sequences of k-individual customers
2. for each individual sequence of a customer
3. for each k in K-groups
4. cluster_index = argmax(LCSS_similarity(sequence,k))
5. insert the sequence into the cluster(cluster_index)
6. endfor
7. endfor

4.2.2a) Existing Algorithm (Breadth First Algorithm) for representing trajectories of items of one customer in an efficient tree data structure

Input: 1) A set of trajectories of items of one customer, T, Represented as transformed trajectories.
2) A minimum conditional Probability threshold (minimum probability).
3) A minimum support threshold (minimum support).

Output: A sequential tree structure that represents transactions of items details of a single customer.

1. root = null //A root node with null entry
2. S_0 ={root} //Initially tree contains only root node
3. k = 0
4. While ($S_k \neq 0$) do //While the set S_k contains items in the transactions then perform
5. S_{k+1} = 0 // S_{k+1} is to store item names in the next level
6. For each node s in the set S_k do
7. Find frequent items and then create conditional table of node s
8. For each sequence in frequent hot regions do
9. If the sequence is in conditional table of node is same as s then
10. Create a new trajectory set of s preceded with each sequence s and
11. sequence is a child of s , so add node s concatenated with sequence into S_{k+1}
12. End if
13. End for
14. End for
15. k = k + 1
16. End while

b) Proposed Algorithm for Clustering Trajectories of Market Basket transactions of Customers

Input: A set of 'n' number of customers' data in the form of trees, each tree represents a root consisting of a set of trajectories of purchased items by a single customer in the market basket

Output: A set of clustered trajectories of purchased items of customers buying behavior

1. Create initially 'n' number of individual clusters such that each cluster consisting of a set of trajectories of items of one customer in the market basket.

2. Store full details of all the initial 'n' clusters of 'n' customers in the appropriate data structures.
3. For each individual cluster i = 1 to n in terms of 1 do
4. For each individual cluster number j = i + 1 to n in terms of 1 do
4.1 Find similarity measure between cluster i and j and store it in the appropriate efficient data Structure for further processing.
End-of-loop (j for loop)
5. Convert all computed similarity measures into Normalized measures for ease and uniform Processing.
6. Sort all the normalized measures and then select the One with highest similarity measure value for clustering The two corresponding previous clusters.
7. End-of-loop (i for loop)
8. Repeat steps 3 through 7 until specified number of final clusters of trajectories of customers buying behavior are formed.

4.3. Results

Similarity of two customers increases as the length of Longest Common Sub Sequence (LCSS) of two customer increases

Consider the following set of trajectory paths of four customers.

1) A,B,C,D,E,F
2) E,F,A,B,D,C
3) A,B,E,F
4) D,F,C,A,B

Longest Common Sub Sequence (LCSS) computations for clustering are shown below:

Similarity between (1 and 2) = length of LCSS/(length of string 1 + length of string 2)
= LCSS length of (A,B,D)/(length of (A,B,C,D,E,F) + length of (E,F,A,B,D,C)
= 3/(6 +6) =3/12 = 0.25

Similarity between (1 and 3) = 4/(6+4) = 0.4

Similarity between (1 and 4) = 2/(6+5) = 0.18
Similarity between (2 and 3) = 2/(6+4) = 0.2
Similarity between (2 and 4) = 3/(6+5) = 0.27
Similarity between (3 and 4) =2/(4+5) = 0.22
Sequences 1 and 3 are most similar and they are clustered.
Proposed clustering method is explained as follows:
Similarity between (1 and 2) = (sequence1 ∩ sequence2)/ (sequence1Usequence2)=(ABCDEF∩EFABCD)/ (ABCDEF U EFABCD) = ABCDEF/ABCDEF = 1/1 =1.0
Similarity between (1 and 3) = 4/6 =0.66
Similarity between (1 and 4) = 5/6 = 0.83
Similarity between (2 and 3) = 4/6 = 0.66
Similarity between (2 and 4) = 5/6 = 0.83
Similarity between (3 and 4) = 3/6 = 0.5
In the proposed method 1 and 2 are most desirable sequences for clustering. When LCSS method is used for clustering the trajectories of purchased items of customers, the clusters formed are the hot regions of the shopping area. These regions guide the management in identifying the shopping behavior of the customers by which the relationship with the customers can be enriched. The shopping behavior shows the means to reorganize the item placements in order to improve the shopping experience of the customers and improvement in the sales as well.
In the proposed algorithm multiple records of a single customer can be taken and processed which will provide better behavior of a customer. By considering and evaluating the clustering results the shop management is in a position to identify the group of items which are frequently bought by the customer. This identification guide the shop management to place such group of products in a same place or nearby places. This type of arrangement is for the customer to locate the items of their interest. This type of arrangements improve the customer relationship.

5. Comparisons

5.1 Comparison of proposed algorithm with existing algorithms:

All the algorithms in the literature focused on minimum distance or maximum matching of objects.
Some of the popular algorithms include:
1) Cosine similarity.
The cosine similarity between two vector is a measure of the cosine of the angle between them, and the value is between [-1, 1]. Here we use it to measure the similarity between the directions of two sub-trajectories.
2) LCSS always returns the sub-sequence in common between two trajectories.

To define and find the similarity between trajectories of market basket items is a challenging task, since buying patterns of different people vary randomly. One important class of trajectory analysis is computing trajectory similarity. This paper introduces and compares four of the most common measures of trajectory similarity: longest common subsequence (LCSS), the similarity of trajectories is defined based on both geographic and semantic features of movements and this approach is used to detect trajectory clusters and infer future locations of moving objects [13]. A recent approach which detects similarity in semantic trajectories is proposed. The approach uses the longest common subsequence (LCSS) algorithm to find the similarity mainly on the semantics[13].The semantic ratio between two trajectories measures a degree of semantic similarity between them, and is defined as similarity Ratio (tra1, tra2) = LCSS(tra1, tra2) / min(|tra1|, |tra2|)
3) Fréchet distance,
4) Dynamic time warping (DTW), and
5) Edit distance.
Notice that many clustering algorithms using Euclidean similarity measure are inadequate for in-store actual shopping path real pattern grouping details.There are many variations among the items purchasing patterns of customers because of many reasons such as product cost, demand, distance, quality, and space and so on. Hence, all the shopping trajectory pattern details must be normalized for convenient use and correct decision making.Notice that a whole trajectory may not always work in identifying clusters. Definitely an efficient and effective supermarket environment always provides convenient services for customers.
In terms of the measures used to find the similarity the proposed algorithm is showing ease and phase in computations. The measure used in the existing algorithm is LCSS (longest common subsequence).The proposed algorithm is using different ratios and normalizations in a better way.
In the case of market basket analysis, the concern is on associations among item sets. For finding associations it is depended on frequent item set mining. Very few used clustering on market basket data. Here we used a new approach of clustering of trajectory data of market baskets.

For clustering trajectory data it is needed to find the similarity between two clusters based on similarity grouping as follows: Euclidean distance and Dynamic Time Wrapping are two methods for finding similarity measure between two trajectories and these methods are failed in finding similarity of trajectories because they require multiple parameters to be set. In general, a trajectory is defined as a sequence of multi-dimensional points. A stay point is an important point in the trajectory with respect to processing or servicing where the user or object stayed for a while.

In our proposed methodology trajectories are represented in the form of trees. As trees are efficient in storage and computation comparing to other forms used in the literature, it is implicit that the proposed algorithm is better in computation and performance. Comparisons can be made with tree data structure with in O (log n) time where other forms need more time.

5.2 Discussion of Results

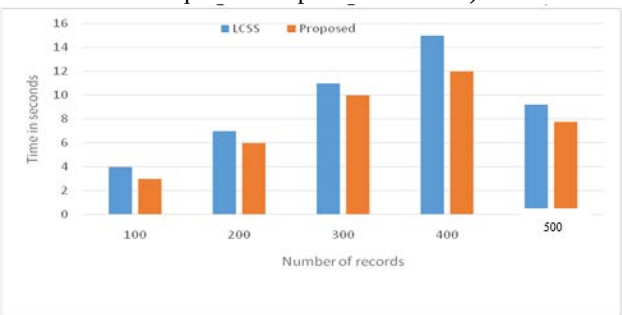
The clusters formed under traditional approaches: The existing approaches used one record for each customer to get the clustering done.

The clusters formed under proposed methodology: The proposed algorithm made use of multiple records for each customer leading to polished results. Therefore it is more generalized than the existing algorithm.

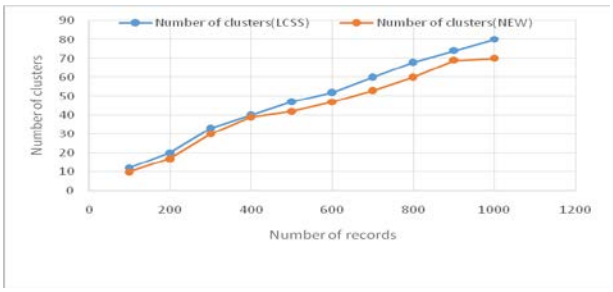
Computational Time Comparison

The k-means clustering algorithm depends on the chosen value of number of selected initial clusters where the proposed algorithm has no such restriction and has the freedom to move to higher accuracy. In other words the proposed algorithm is more generalized version of the existing algorithm and it can handle many item sequences of each customer.

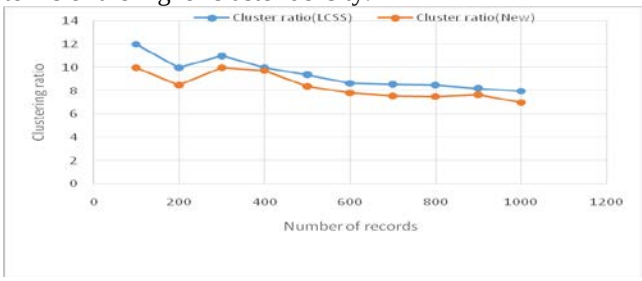
The traditional method is consuming 15% more time than the proposed methodology (based on 5 sample runs for each customer sequences of purchased items).



It is observed from the time comparison chart that the new method is saving up to 15% of the execution time.



It is observed from the above chart that both the algorithms go hand in hand in terms of the number of clusters generated but the new algorithm is a little bit sharper in terms of the higher cluster density.



From the above chart it is observed that the clustering ratio of both algorithms went hand in hand, but the new algorithm is showing upper hand.

6. Conclusion

A novel and new clustering process on super market items of trajectory data sets representing market baskets is discussed and used on a real time super market data. The same is compared with the traditional methods of market basket analysis. Existing Longest Common Sub Sequence (LCSS) clustering algorithm is not scalable and computationally very high in terms of floating point multiplications. Proposed new clustering algorithm is very high scalable and efficient in terms integer computations only when compared with the existing LCSS clustering algorithm. Significant improvements are observed in terms of methodology, computational time complexity in the experimental results. The new clustering strategy of analyzing market baskets will show a way towards new generation market basket analysis in future.

References

[1] BartholomausEnde & RudigerBrause ,E-Finance Lab, Dept. of Computer Science and Mathematics,Johann Wolfgang Goethe-University,60054 Frankfurt, Germany, “Mutual Information based Clustering of Market Basket Data for Profiling Users”

[2] Bob Price, Russ Greiner, Gerald Häubl& Alden Flatt , Computing Science , University of Alberta, “Automatic Construction of Personalized Customer Interfaces”,

[3] Ching-Huang Yun, Kun-Ta Chuang and Ming-Syan Chen, Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan, ROC, "An Efficient Clustering Algorithm for Market Basket Data Based on Small Large Ratios".

[4] Ching-Huang Yun, Kun-Ta Chuang+ and Ming-Syan Chen, Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan, "Using Category-Based Adherence to Cluster Market-Basket Data", ROC, Graduate Institute of Communication Engineering, National Taiwan University, Taipei, Taiwan,

[5] Haozhou Wang, Han Su, Kai Zheng, Shazia Sadiq and Xiaofang Zhou, School of Information Technology and Electrical Engineering, The University of Queensland, Australia "An Effectiveness Study on Trajectory Similarity Measures" Proceedings of the Twenty-Fourth Australasian Database Conference (ADC 2013), Adelaide, Australia.

[6] Herman uinis,Lura E. Forcum,HarryJoo "Using Market Basket Analysis in Management Research", Journal of Management,Vol. 39 No. 7, November 2013 1799-1824

[7] Hechen Liu and Markus Schneider, "Similarity Measurement of Moving Object Trajectories", National Science Foundation (NSF) under the grant number NSF-IIS-0812194

[8] In-Chul Jung and Young S. Kwon &Yung-Seop Lee, Department of Industrial and Systems Engineering, South Korea, "A Sequence Pattern Matching Approach to Shopping Path Clustering", Proceedings of the 2012 International Conference on Industrial Engineering and Operations Management ,Istanbul, Turkey, July 3 – 6, 2012

[9] Jae-Gil Lee, Jiawei Han, and Kyu-Young Whang, "Trajectory clustering: a partition-and group framework". In ACM SIGMOD, PAGES 593-604, 2007.

[10] Jean Damascene Mazimpaka, Sabine Timpf , and Alter Postweg, "Trajectory data mining - A review of methods and applications " Paper under review, Journal of Spatial Information Science

[11] Jeffrey S. Larson , Eric T. Bradlow, Peter S. Fader, The Wharton School, The University of Pennsylvania, USA, "An exploratory look at supermarketshopping paths", International Journal of Research in Marketing 22 (2005) 395–414,0167-8116, Elsevier B.V.

[12] Jianwei Li , Ying Liu, Wei-keng Liao &AlokChoudhary, Northwestern University, "Parallel Data Mining Algorithms for Association Rules and Clustering"

[13] Josh Jia-Ching Ying, Eric Hsueh-Chan Lu,Wang-Chien Lee, Tz-ChiaoWeng, and Vincent S.Tseng. "Mining user similarity from semantic trajectories". In LBSN, pages 19–26, 2010.

[14] Jure Leskovec, AnandRajaraman& Jeffrey D. Ullman, MilliwayLabs,Stanford Univ. "Mining of Massive Data sets ",copyright c 2010, 2011, 2012, 2013, 2014 AnandRajaraman, Jure Leskovec

[15] Kevin Toohey, Matt Duckham, "Trajectory Similarity Measures" Published in Newsletter SIGSPATIAL Special SIGSPATIAL Homepage archieveVolume 7 Issue 1, March 2015, pages 43-50, ACM New York, NY,

[16] KhairilAnnuar B. Abdul Kadir, University Putra Malaysia, "Clustering Algorithm for Market Basket Analysis: The underlying concept of Data Mining Technology", FSKTM 2003

[17] Loraine Charlet Annie M.C. and Ashok Kumar , Department of Computer Science, Government Arts College, Trichy, India, "Market Basket Analysis for a Supermarket based on Frequent Item set Mining", International Journal of Computer Science Issues, Vol. 9, Issue 5, No 3, September 2012, ISSN (Online): 1694-0814

[18] Manpreet Kaura, Shivani Kanga, "Market Basket Analysis: Identify the changing trends of market data using association rule mining", International Conference on Computational Modeling and Security (CMS 2016) , Published by Elsevier B.V.

[19] PavelBerkhin , "Survey of Clustering Data Mining Techniques", Accrue Software, Inc.

[20] Petterkihlstrom, "Literature Study and Assessment of Trajectory Data Mining Tools ", SoMEX KAND 2015-05 ,

[21] RujataSaraf, Prof. SonalPatil, "Market-Basket Analysis Using Agglomerative Hierarchical Approach for Clustering a Retail Items", International Journal of Science and Research (IJSR) Volume 4 Issue 3, March 2015

[22] SaurkarAnand V, Bhujade V, Bhagat P, Khaparde A. "A Review Paper on various Data Mining Techniques" .International Journal of AdvancedResearch in Computer Science and Software Engineering 2014:4(4):98-101.

[23] Savitha S. Kadiyala&Alok Srivastava , Georgia State University , " Data Mining For Customer Relationship Management" , International Business & Economics Research Journal Volume 1, Number 6

[24] SheenuVerma, Research Scholar, &SakshiBhatnagar, Assistant Professor,Ambala College of Engineering and Applied Research, Mithapur, Haryana, India, "An Effective Dynamic Unsupervised Clustering Algorithmic approach for Market Basket Analysis", International Journal of Enterprise Computing and Business Systems, ISSN (Online) : 2230-8849,Volume 4 Issue 2 July 2014,

[25] SusantaSatpathy, Lokesh Sharma, Ajaya K. Akasapu, Netreshwari Sharma, "Towards Mining Approaches for Trajectory Data", International Journal of Advances in Science and Technology Vol. 2, No.3, 2011

[26] SweeChuan Tan, Jess Pei San Lau, SIM University, School of Business, 535A Clementi Road, Singapore {jamestansc, "Time Series Clustering: A Superior Alternative for Market Basket Analysis "

[27] Tanuja.V MCA, M.Tech and Prof. P.Govindarajulu, M.Tech, Ph.D, S.V. University, Tirupati" A Survey on Trajectory Data Mining", International Journal of Computer Science and Security, Vol.10, Issue 5, 2016

[28] Tanuja .V MCA, M.Tech and Prof. P.Govindarajulu, M.Tech, Ph.D, S.V. University, Tirupati "Application of Trajectory Data Mining Techniques in CRM using Movement Based Community Clustering ", International Journal of Computer Science and Network Security, Vol. 16 , N0.11, November'2016

[29] Tanuja .V MCA, M.Tech and Prof. P.Govindarajulu, M.Tech, Ph.D, S.V. University, Tirupati "Application of Trajectory Data Clustering in CRM : A Case Study " International Journal of Computer Science and Network Security, Vol. 17 , N0.1, January' 2017

[30] Yu Zheng. “Trajectory data mining: An overview”, ACM Trans. Intelligent System Technol. 6, 3, Article 29 (May2015), Microsoft Research, 41 pages

V. TANUJA received Master of Computer Applications degree from Sri Venkateswara University, Tirupati, AP and Master of Technology degree in Computer Science & Engineering from Acharya Nagarjuna University, Guntur, AP. She is a research scholar in the department of Computer Science, Sri Venkateswara University, Tirupati, AP, India. Her research focus is on Data Mining in Customer Relationship Management. .

P. GOVINDARAJULU, Retd. Professor, Department of Computer Science, Sri Venkateswara University, Tirupathi, AP, India. He received his M. Tech., from IIT Madras (Chennai), Ph. D from IIT Bombay (Mumbai), His area of research are Databases, Data Mining, Image processing, Intelligent Systems and Software Engineering.