Handling Transactional Data Features via Associative Rule Mining for Mobile Online Shopping Platforms

Maureen Ifeanyi Akazue¹, Sebastina Nkechi Okofu², Arnold Adimabua Ojugo³, Patrick Ogholuwarami Ejeh⁴, Christopher Chukwuunaya Odiakose⁵, Frances Uche Emordi⁶, Rita Erhovwo Ako⁷, Victor Ochuko Geteloma⁸

Department of Computer Science, Delta State University, Abraka, Nigeria¹
Department of Marketing and Entrepreneurship, Delta State University, Abraka, Nigeria²
Department of Computer Science, Federal University of Petroleum Resources, Effurun, Nigeria³, ⁷, ⁸
Department of Computer Science, Dennis Osadegba University Anwai-Asaba, Nigeria⁴, ⁵
Department of Cybersecurity, Dennis Osadegba University Anwai-Asaba, Nigeria⁶

Abstract—Transactional data processing is often a reflection of a consumer’s buying behavior. The relational records if properly mined, helps business managers and owners to improve their sales volume. Transaction datasets are often ripped with the inherent challenges in their manipulation, storage and handling due to their infinite length, evolution of product features, evolution in product concept, and oftentimes, a complete drift away from product feat. The previous studies’ inability to resolve many of these challenges as abovementioned, alongside the assumptions that transactional datasets are presumed to be stationary when using the association rules—which have been found to also often hinder their performance. As it deprives the decision support system of the needed flexibility and robust adaptiveness to manage the dynamics of concept drift that characterizes transaction data. Our study proposes an associative rule mining model using four consumer theories with RapidMiner and Hadoop Tableau analytic tools to handle and manage such large data. The dataset was retrieved from Roban Store Asaba and consists of 556,000 transactional records. The model is a 6-layered framework and yields its best result with a 0.1 value for both the confidence and support level(s) at 94% accuracy, 87% sensitivity, 32% specificity, and a 20-second convergence and processing time.

Keywords—Association rule mining; online shopping platforms; feature evolution; concept drift; concept evolution; shelf placement

I. INTRODUCTION

Data connotes everything we can manipulate [1]. It can exist in structured and unstructured forms. During processing, data can be tracked as it mutates from one form to another [2]. It can also be quantified or mined by removing unwanted feats therein (i.e. noise) [3], and analyzed to reveal its hidden relations and patterns [4]. Informatics processing needs has today transformed our society [5] with tools that advance effective resource sharing with its inherent benefits [6]. These also yield a range of threats and complications to the normal operations of systems deployed to ease living at every frontier [7]. With the advances in Internet penetration, businesses constantly decentralize [8], as means to reshape/refocus her processes via data warehousing, to ease transaction accessibility and availability [9]. Business owners have become aware of their responsibilities to consumers [10], and the management of business transactions that now heavily rely [11] on their capacities to adequately manage transactions of all forms with its allied processes [12].

Transactions are processed via two modes namely: (a) batch processing that allows a large volume of transactions processed simultaneously [13], making it more economical. E.g. include bill/report generation [14], credit-card transactions [15], image processing, etc. [16], and (b) real-time processing allows many consumers to process and simultaneously perform a variety of transactions [17]. Also termed stream processing [18], examples include point-of-sales unit [19], online purchase and ticketing, reservation [20], and traffic controls [21], etc.

With the daily volume of data generated [22], it is critical to find better ways to effectively retrieve patterns from processed data and to unveil hidden relations in stored repositories [23]. This quest of mining meaningful data requires in-depth analysis with decision-making skills [24] that can only be efficiently achieved via mining [25]. Classifying transactional stream data [26] – is ripped with a range of issues including (a) the infinite length-size of data for continuous, real-time transactions with no bounds [27], (b) concept drift is an issue for which a consumer shifts decision to purchase an item [28], (c) concept evolution for which a new product acts as a close-substitute/replacement to a class of old products [29] to evolves a data stream, and lastly, (d) feature evolution in which various data-streams for newer product feasts occur regularly, and such instances occurs with the corresponding increase in the data-streams due to the increase in the product replacement feature [30].

Big data often refers to a large collection of data consisting of (un)structured data [31], stored in a repository/warehouse [32]; It requires a more critical, authentic-time investigation in a bid to reveal the relations between the data items. This helps us to better understand the varied levels of abstraction [33] and in-depth knowledge patterns that can be revealed behind various hidden values in data as they are stored in repositories [34]. The nature of many basket tasks is that the transactions are handled in real-time making it apparently tedious and quite difficult to manage [35]. With such transactions, items are either purchased alone, or as a combination of itemset(s) to form a basket [36]. Thus, storing and managing such data, yields a plethora of issues ranging from concept evolution, feature evolution, infinite data length, and concept drift [37].
Many online shops yield mobile smart device users – a basket experience for which items are purchased directly from (in real-time) [38] via an online shop or platform. Thus, such physical acquisition of items is said to be a purchase from an e-store, web shop, virtual shop, or online shop via a market basket [39]. Thus, it becomes imperative to employ a data mining in extracting useful data from such a voluminous amount of data [40]. A consumer can make a series of purchases – and these can also yield an infinite number of changes in the buyer’s preferences over time – called concept drift in the consumer’s purchasing pattern or behavior [41].

These benefits are not without challenges, and it include (but not limited to): (a) there is a great need to find better means to handle the daily, continuous volume of data generated [42] – as many of these data can either exist in either their structured [43] or unstructured formats [44], (b) previous studies on data stream classification modes – have sought to address the issues of conceptual drift and infinite length challenges with little success [45]. It is found that such models often employ apriori mode and frequency growth patterns in the transactional data stream [46]. But, in cases where the model has used association rule mining – they have often assumed transactional data [47] are stationary, which is not the case, and (c) the assumed stationary nature of transactional data does not yield the required flexibility [48], robustness, and adaptiveness needed for association rule method [49] to be used in resolving the inherent issues of both features evolution and concept evolution as rippled across transaction data streams.

Our study explores germane theories of consumer purchase patterns fused with association rule mining (on the one hand), and fused with frequency growth pattern (as hybrid framework and method) to address the inherent issues with concept evolution, concept drift, and feature evolution amongst itemset basket placement; These, and other complications as present in the basket transaction data streams – are challenges that the study wishes to address.

Section I introduces the study with a view to unveiling the meaning of data, big data, transactions and others. Section II details the problem formulation in handling transactional data streams and expressing the issues of feature evolution, concept evolution and concept drift with itemset (basket data analysis) as well as leveraging a variety of consumer purchasing pattern theories. Section III details result found as evidence to support the decision during discussions of the findings, and conclusion.

II. METHODS AND MATERIALS

A. Problem Formulation

A market basket problem can be defined as a search for joint values of variables in X = (x1, x2, …, xq) with the highest frequencies in binary-valued data The variables Xk represent consumer purchases transactions [50] – and are usually a total of all itemsets sold by a store. The observation with each variable xk is assigned one-of-two values (0 or 1) [51], and represented as in Equation 1 [52] below:

\[ X_{ok} = \begin{cases} 0, & \text{if no purchase or transaction is made} \\ 1, & \text{if } k - \text{item is purchased in a transaction} \end{cases} \] (1)

Variables that are frequently purchased together have a joint value of 1. And, ensures the inventory system is automatically updated for re-stock [53], cross-selling [54], shelf and product placement cum location [55], catalog design, cross-marketing sales promotions, and consumer segmentation on purchasing on [56]. If we represent each purchase by the consumer using x1, x2, etc as binary variables respectively [57]; Then, mining the data will seek to find a subset of integers K = {1-to-N} [58] such that the dataset becomes large, Eq. (2) holds true as thus [59]:

\[ P(\prod_{k \in K} X_k = 1) \] (2)

K represents an itemset (i.e. the number of items in a basket or cart), and N is the size of an itemset. The probability that agrees with Eq. (2) is called a support S of the itemset K, which is computed as in Eq. (3) [60]:

\[ z(\prod_{k \in K} X_k = 1) = \frac{1}{N} \sum_{o=1}^{N} \prod_{k \in K} X_{ok} \] (3)

The observations o for which \( \prod_{k \in K} X_{ok} = 1 \) contain k-items [61]. With a lower bound value of l, the basket algorithm seeks all itemset kl with support greater than this lower bound l (i.e. \( \{kl \mid S(kl) > 1 \} \) [62]. This yields the model in Eq. (3), and also represents our formalization of the market basket problem [63], which consists of the following, and agrees with [64]:

1) First, frequency of purchased itemset is determined and analyzed using a given threshold value [65] – calculated as the Cartesian product of all similar items Xn. If its support is greater than the established lower bound as in Eq. (3), the algorithm halts and recommendations are suggested to the customer [66].

2) Secondly, if a consumer purchases an item, the system provides similar itemsets, and also recommends the same for other customers with similar purchasing patterns [67].

B. Basket Transaction Theoretical Frameworks

To resolve the issue(s) – association rule mining is used on transaction dataset(s) to generate numerous itemsets that yields the purchasing behavior for various customers [68]. We thus, adapt the theories below and their corresponding relevance thus:

1) The theory of Reasoned Action emphasizes behavior that is dependent on a consumer’s attitude, behavioral choices, and public opinion [69]. It thus implies that a consumer’s decisions to purchase is constantly influenced by his/her intents, choice, and personal beliefs. All of which, aligns with Fig. 1 [70]. The theory’s relevance is such that a consumer can purchase item(s) if presented with specific expected results. S(he) can also change his/her decision, which in turn will yield attitudinal changes in relation to his/her trust and confidence about the item [71]. These are shocks gained from either experience, or can result as the influence on a consumer by friends with precious data about the product; Which, in turn – yields a concept drift [72]. It thus, ensures that a consumer’s action is based on purpose – making each
consumer more rational as his/her choice is poised to serve their best interest and intentions [73], which agrees with [74].

2) Planned Behaviour Theory states that attitude towards a behavior, subjective norms, and perceived control often shapes a consumer's behavioral intents and in turn, his/her actions. This theory improves the analytical capability of reasoned actions via the perceived control of behaviors. Since not all behavior is subject to a consumer's control – it is expedient we add perceived behavioral control which implies that irrespective of the action taken – a consumer's behavior is determined both by attitude, subjective norm, and their perception/firm belief they are in control [76].

3) Engel, Kollet, and Blackwell extends the reasoned action by focusing on the consumer’s mental state before his/her decision to purchase [77]. It bolsters the reasoned action through a planned set of behaviors [78] as thus: (a) that a consumer absorbs advertised information and knowledge as presented by the vendor [79], (b) that a consumer may process the retrieved knowledge about a product, and also can leverage on previous experience to compare between the observed versus expected outcome [80], and (c) that a consumer decides either to accept or reject the purchase of an item [81] – yielding a choice or decision reached from balanced insight through mental synthesis. Thus, with data input as its greatest prize [82] – the product manufacturers must equip managers with adequate knowledge in place of the product line that will eventually drive consumers to keep buying the item; And in turn – this will shore up and push up sales volume of the product [83]. This theory unveils the underlying feats that may cause purchase shift in the consumer behavior [84] – such that where and if a consumer is not adequately informed, (s)he can reject the purchase of an item as means to normalize with the online data cum knowledge available [85]. Thus, external shocks (i.e. friends, item review ratings etc) can or may influence a consumer choice and decision to either accept or reject the purchase of a product [86].

4) Impulse Theory – Here, purchase decisions are influenced by an impulse to suddenly buy a product; thus, such buys do not serve any purpose. They are grouped into (a) pure impulse, (b) reminded impulse, (c) suggested impulse, and (d) planned impulse (if the consumer knows the item they wish to buy – even if they are unsure of it). Its relevance is that it yields an irrational behavior pattern in purchase drift; But, it embellishes the marketability of the product – from packaging and displays over the shelf with greater emphasis laid on the various attributes of the product such as its cost, etc. These influence a buyer’s impulse – and note that an electronic description of the product should be sensitive enough to trigger such purchase drift on the consumer to like and accept the product – irrespective of their premonitions [87].

Our framework hinges on the relations between various components in transaction analysis – emphasizing consumer purchasing-pattern. The issues of feature and concept evolution arise from the manufacturer's quest to meet consumer needs and buying patterns; and in turn, yield concept drift [88]. To resolve these, association rule mining is carried out on a transactional basket (appropriate) dataset to generate a variety of itemsets (basket) that adequately represents a consumer buy pattern. It justifies our adoption of the adapted consumer behavior theories as in Fig. 1, with adopted TRA/TPB that directly explains our research problem. To derive meaningful data via these theories, we visualized the consumers’ behavior to help us resolve the issue of concept drift.

Fig. 1. The reasoned action theory with its various components (Source: [75]).

C. Data Gathering / Sample Population

The dataset was retrieved from Roban Stores, and contains about 982,980 records – representing transactions for the period of 18 months (i.e. 2017-2018). Training records for framework have the selected features as: (a) basket itemsets, (b) unit price, (c) item quantities, (d) total itemset price, (e) invoice number, and (f) date of transaction. These were adopted to address the issues of concept drift for each consumer, and for each of the requisite transaction.

Dataset consists of consumer profiles with demographics (i.e. age, sex, and status) as seen in Table I – all of which aid in studying the customer buying pattern and behaviors.

D. ItemSet Data Description

The Roban Stores (RS) transaction data contain itemsets of single and combined itemset purchases as a basket. An itemset as used here, describes data-streams measures and dimensions. Example description of the dimensions for bread as snacks:

<table>
<thead>
<tr>
<th>ItemSet Description for RS.Snacks</th>
</tr>
</thead>
<tbody>
<tr>
<td>RS Bread.Snacks</td>
</tr>
<tr>
<td>Bread.Snacks = RSB_E ∩ RSB_W ∩ RSB_F ∩ RSB_M</td>
</tr>
</tbody>
</table>

For Each Selected Bread.Snack do

| RSB_E.Bread.Snacks = ES ∩ EM ∩ EL |
| RSB_W.Bread.Snacks = WWB ∩ SFWB |
| RSB_F.Bread.Snacks = FM ∩ FLS ∩ FLUS |
| RSB_M.Bread.Snacks = MS ∩ MLS ∩ MLUS |

End For Each

The semantic library has the following keys with the bread category grouped into three (3) as thus:
RSB_E = Roban_Stores Bread Enriched-set (ES, EM, EL)  
= Enriched small, Enriched Medium, Enriched large

RSB_W = Roban_Stores Bread Wheat-set (WWB, SFWB)  
= Whole wheat bread, sugar-free wheat bread

RSB_F = Roban_Stores Bread Fruit-set (FM, FSL, FLUS)  
= Fruit Medium, Fruit large (sliced/unsliced)

RSB_M = Roban_Stores Bread Malt-set (MS, MLS, MLUS)  
= Malt Small, Malt large (sliced/unsliced)

TABLE I. DATASET DESCRIPTION, DATA TYPES, AND FORMAT

<table>
<thead>
<tr>
<th>Features</th>
<th>Data Type</th>
<th>Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>Invoice_Number</td>
<td>Long Int.</td>
<td>1234</td>
</tr>
<tr>
<td>Quantity</td>
<td>Short Int.</td>
<td>1234</td>
</tr>
<tr>
<td>Unit Price</td>
<td>Float</td>
<td>123.45</td>
</tr>
<tr>
<td>Transaction Time</td>
<td>Time</td>
<td>D:M:Y</td>
</tr>
<tr>
<td>Weekly Transaction</td>
<td>Int</td>
<td>1234</td>
</tr>
<tr>
<td>Monthly Transaction</td>
<td>Int</td>
<td>1234</td>
</tr>
<tr>
<td>Freq. Trans. Types</td>
<td>Int</td>
<td>1234</td>
</tr>
</tbody>
</table>

E. Association Rule Mining Calibration

To ensure that only accurate data is processed, we needed to calibrate the association rule mining for each basket using Hadoop tableau visualizer for Calib and Hovitz-Thompson estimator. It ensures that only appropriate rules for transactions are generated via the frequent-pattern growth algorithm [89]. The generated rules are analyzed using RapidMiner v8.1 and were used to effectively calibrate the customer profile dataset via the simple random sampling without replacement (srswor) distribution as in the algorithm listing 1.

The algorithm listing 2 extends customer profile calibration via random sampling without replacement (srswor) distribution for the bread itemset combination.

Algorithm 1: Calibrate data variables in each stratum

Cat ("stratum 1/n"): Stratum 1
data1 = data[data$element=='a',]
x1 = x[data$element=='a']
total1 = calib((resp(1, n_row(data1))) %*% X1)
sr1 = sr$sstratum==1, }
xs1 = X[sr1$ID_Bread.Snacks]
d1 = 1/(sr$prob*sr$prob_resp)
g1 = calib(xs1, d1, total1, method = "linear"
check calibration (xs1, d1, total1, g1)
Sreport
[1] "the calibration is done"
Sresult = [1]true
Svalue = [1]1e-06
Cat("stratum 2/n"): Stratum 2
data2 = data[data$element=='ab',]
x2 = x[data$element=='b']
total2 = calib((resp(1, n_row(data2))) %*% x2)
sr2 = sr$sstratum==2, }
xs2 = X[sr2$ID_WWB]
d2 = 1/(sr$prob*sr$prob_resp)
g2 = calib(xs2, d2, total2, method = "linear"
check calibration (xs2, d2, total2, g2)
calibration cannot be done  max estimate is given by ‘value’

Algorithm 2: Calibrate 1 with strata

Xs = X [sr$IDBread, ]
d = 1/(sr$prob*sr$prob_resp)
Compute: g = calib (Xs, d, total, method = "linear")
For Each Selected Parameter to Calibrate do
1. summary (g)
2. output  w = d * g
3. check calibration (Xs, d, total, g)
4. Sreport
5. [1] “the calibration is done”
7. Svalue = [1]1e-06
End For Each

Fig. 2 shows the architecture employed towards resolving the issues of concept drift, feat evolution, and concept evolution for basket analysis. It comprises of six-data layers as adapted from the elixir architecture – incorporating these ingestion, collection, processing, visualization, sources, and storage. The collection and ingestion layers have been combined to form the pipelining layer [90].

III. RESULTS AND FINDINGS DISCUSSION

A. Performance Evaluation of the Framework

We used three (3) types of tests as below [91]:

1) Alpha testing helps a programmer identify errors in the product before its release for public use. It focuses on finding weaknesses before beta tests and seeks to ensure users employ black-box/white-box testing modes.

2) Beta test is before the release of software for commercial use. It is usually the final test and often includes program system distribution to experts – seeking means to improve on the product. We sent the product to the store for the beta test [92].
3) Unit testing often requires individual units or components of software to be tested. This phase/stage of software development often seeks to corroborate and ensure that each part of the software performs according to its design specification. The smallest testable part of any software is known as the unit test. It has few inputs with a single output [93].

Tables II and III respectively show the summary result of both the alpha tests and unit testing for the various execution time taken to yield the requests.

Table II shows that the performance of Frequency-growth pattern (FP) using the minimum support value of a 0.1, and a confidence level of 0.1. This yields and shows that 0.79 (i.e. 79%) of consumers preferred to buy bread and drinks throughout their transactions as analyzed. And, average convergence time it took for the algorithm to compile was within 20 seconds.

Table III shows performance of the Apriori ARM algorithm using minimum support of 0.1 and a confidence level of 0.1, which shows that 79% of consumers preferred buying bread and drink of the entire transactions analyzed. And the time it took for the algorithm to compile was within 26 seconds.

### TABLE II. RESULT OF THE FREQUENT-PATTERN GROWTH ALGORITHM

<table>
<thead>
<tr>
<th>Association Rules</th>
<th>Support Level</th>
<th>Confidence Level</th>
<th>Execution in Secs</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM Enriched Large Bread, DM Whole Wheat Bread → 7UP Pet Drink 50CL</td>
<td>0.026</td>
<td>0.194</td>
<td></td>
</tr>
<tr>
<td>DM Fruit Malt Bread, DM Enriched Cake Bread Large → C-Way Peach 500ML</td>
<td>0.006</td>
<td>0.214</td>
<td>18secs</td>
</tr>
<tr>
<td>DM Enriched Large Bread; DM Enriched Cake Bread Large → Nutella Ferrero Hazelnut Spread 350g</td>
<td>0.006</td>
<td>0.214</td>
<td></td>
</tr>
</tbody>
</table>

### TABLE III. RESULT OF THE ASSOCIATION RULE MINING ALGORITHM

<table>
<thead>
<tr>
<th>Association Rules</th>
<th>Support Level</th>
<th>Confidence Level</th>
<th>Execution in Secs</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM Enriched Large Bread, DM Whole Wheat Bread → 7UP Pet Drink 50CL</td>
<td>0.062</td>
<td>0.294</td>
<td></td>
</tr>
<tr>
<td>DM Fruit Malt Bread, DM Enriched Cake Bread Large → C-Way Peach 500ML</td>
<td>0.009</td>
<td>0.412</td>
<td>26secs</td>
</tr>
<tr>
<td>DM Enriched Large Bread; DM Enriched Cake Bread Large → Nutella Ferrero Hazelnut Spread 350g</td>
<td>0.009</td>
<td>0.412</td>
<td></td>
</tr>
</tbody>
</table>

B. Result Findings: Analysis of Consumption Pattern

Fig. 3 shows the itemsets summary frequently purchased by a consumer, his/her consumption pattern, and how much in revenue percentage such consumer has contributed to the store using our consumption history. This analysis can be tracked to display (the daily, weekly, and monthly averages) consumption and spending pattern of each consumer.

Fig. 4 – iView shows a snippet of all selected objects and their status at the time of viewing the analysis report. It yields a consumer frequency of purchased itemset(s) with a live-stream analysis of various transactions. The iView aids managers to track, view and trace each transactional object property in real-time (i.e. s/he can do this as new transactions trickle in and as transactions data streams change).

To view the data-stream reports, the manager logs in to view the object summary visitation frequency and spending. It shows the summary of how frequently consumers visit the store and is tracked daily, weekly, monthly, and/or annually. And this agrees with [94].

![Summary analysis of an individual’s consumption.](image)

![Analysis of consumer consumption history.](image)

Fig. 5 shows the concept drift – and by extension, the consumer's consumption pattern summary for itemsets either in single or combination that is purchased together. As with our example in the report, we see the itemset combination of bread and drink was more than any other.

Fig. 6 shows restock option (that is, percentage) of all the item(s) currently left in the inventory. These are automatically updated with each consumer transaction, and in turn – reduce the error encountered with the traditional mode of inventory restocking and stock-taking, currently available in the store.

C. Discussion of Findings

Results show that the association rule mining trained with the frequent-pattern growth algorithm performed better than the Apriori algorithm (with transactions generated on the frequency of itemsets purchased). The frequent items represent
consumer purchasing patterns and behavior for the system being modeled. With association rules (mined/generated) – the framework seeks to induce the basket analysis to study consumer purchasing patterns and their frequency over time by resolving the issues of concept drift, concept evolution, and features evolution inherent in real-time transaction data streams [95].

![Consumption pattern summary](image)

**Fig. 5.** Concept drift consumption summary for itemset.

![Bread status](image)

**Fig. 6.** The Restock module with the percentage of the item(s).

This study agrees with [96] in provisioning consumer buying theories that sought to recognize reasons that contribute to a consumer's decision to purchase an itemset or product. These theories formed the basis to resolve the challenges presented in data streams by concept drift and its association with basket analysis – which previous studies did not try to resolve [97]. The study notes that to resolve the issues of concept drift with market baskets analysis – it is critical to use an enormous volume of transactional stream datasets [98] collected over time. This will help the proposed system train the association rules to accurately predict the consumer purchasing/buying pattern cum behavior [99] – even with the occurrences of a drift. The study agrees with [100] in our use of big-data analytics tools such as Spark to study customer behaviors in market basket analysis.

IV. CONCLUSION

In resolving the issues inherent in transaction data streams for real-time processing and concerning its use with market baskets – it is imperative to use multiple sources of the dataset to effectively visualize a consumer’s drift to purchase item(s) and products within a store over a period. The model yields the best result with a 0.1 value for both the confidence and support level(s) at 94% accuracy, 87% sensitivity, and a specificity of 32% with a 20-second convergence and processing time. Our framework’s data visualizer displayed both the consumer’s consumption pattern vis-à-vis the inventory stock with the consumer’s profile. Such data have been found to provide and yield new means to a transaction that could be stored on other databases for retrieval and further studies such as the Amazon RedShift.

REFERENCES


