Vol 7. No. 3 - September, 2014
African Journal of Computing & ICT

www.ajocict.net

Frequent Pattern and Association Rule Mining from Inventory Database Using Apriori Algorithm

K.S. Adewole, A.G. Akintola & A.R. Ajiboye
Department of Computer Science
University of Ilorin
Ilorin, Kwara State, Nigeria
adewole.ks@unilorin.edu.ng, abimbolar@yahoo.com, ajibraheem@unilorin.edu.ng

S.O. Abdulsalam
Department of Computer, Library and Info. Science
Kwara State University
Malete, Kwara State, Nigeria
sulepy@yahoo.com,

ABSTRACT

Recently, data mining has attracted a great deal of attention in the information industry and in a Society where data continue to grow on a daily basis. The availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge is the major focus of data mining. The information and knowledge obtained from large data can be used for applications ranging from market analysis, fraud detection, production control, customer retention, and science exploration. A record in such data typically consists of the transaction date and the items bought in the transaction. Successful organizations view such databases as important pieces of the marketing infrastructure. This paper considers the problem of mining association rules between items in a large database of sales transactions in order to understand customer-buying habits for the purpose of improving sales. Apriori algorithm was used for generating strong rules from inventory database. It was found that for a transactional database where many transaction items are repeated many times as a superset in that type of database, Apriori is suited for mining frequent itemsets. The algorithm was implemented using PHP, and MySQL database management system was used for storing the inventory data. The algorithm produces frequent itemsets completely and generates the accurate strong rules.

Keywords: Apriori Algorithm, data mining, database, strong rules & inventory.

African Journal of Computing & ICT Reference Format:
K.S. Adewole, A.G. Akintola & A.R. Ajiboye & S.O. Abdulsalam (2014). Frequent Pattern and Association Rule Mining from Inventory Database Using Apriori Algorithm. Afr J. of Comp & ICTs. Vol 7, No. 3. Pp35-42.

1. INTRODUCTION

Data mining is an interdisciplinary sub-field of computer science. It is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a dataset and transform it into an understandable structure for further use [1]. Advances in data collection and storage technology have led organizations to store vast amounts of data pertaining to their business activities. Extracting useful information from such huge data collections is of importance in many business decision making processes. Such an activity is referred to as Data Mining or Knowledge Discovery in Databases (KDD).

Data mining includes tasks such as Classification, Similarity Analysis, Summarization, Sequential Pattern Discovery, and Association Rule Mining among others [2]. Data mining or Knowledge discovery in databases (KDD) is the process of discovering previously unknown patterns form the huge amount of data stored in flat files, databases, data warehouses or any other type of information repository. Database mining deals with the data stored in database management systems such as MySQL [3].

Vol 7. No. 3 - September, 2014
African Journal of Computing & ICT

www.ajocict.net

The task of association rule mining is to find correlation relationships among different data attributes in a large set of data items, and this has gained lot of attention since its introduction. Such relationships observed between data attributes are called association rules [4]. A typical example of association rule mining is the market basket analysis. Consider a retail store that has a large collection of items to sell. Business decision regarding discount, cross-selling, grouping of items in different aisles, and so on needs to be made often in order to increase the sales and hence the profit. This inevitably requires knowledge about past transaction data that gives the buying habits of customers. The association rules in this case will be of the form "customers who bought item A also bought item B," and association rule mining is to extract such rules from the given transaction data history.

Market basket analysis, also known as association rule mining, is a method for discovering consumer purchasing patterns by extracting associations or co-occurrences from the stores' transaction database. The most well-known methodology called Apriori, was introduced by Agrawal et al. [4] which, as all algorithms for finding large itemsets, can be stated as follows. Given two non-overlapping subsets of product items, X and Y, an association rule in form of $X \Rightarrow Y$ indicates a purchase pattern that if a customer purchases X then he or she also purchases Y. Two measures, support and confidence, are commonly used to select the association rules [4].

In computer science and data mining, Apriori is a classic algorithm for learning association rules. Apriori is designed to operate on databases containing transactions such as collections of items bought by customers, or details of a website frequentation. Other algorithms are designed for finding association rules in data having no transactions or having no timestamps [2]. In this information age, due to the use of sophisticated technologies such as computers, satellites, etc., we have been collecting tremendous amounts of information like business transactions, scientific data, medical data, satellite data, surveillance video & pictures, world wide web repositories to name a few. With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, if not necessary, to develop powerful means for analysis and perhaps an interpretation of such data and for the extraction of interesting knowledge that could help in decision-making.

The kinds of patterns that can be discovered depend upon the data mining tasks employed. By and large, there are two types of data mining tasks: descriptive data mining tasks that describe the general properties of the existing data, and predictive data mining tasks that attempt to make predictions based on inference on available data. One of the popular descriptive data mining techniques is Association Rule Mining (ARM), owing to its extensive use in marketing and retail communities in addition to many other diverse fields. Mining association rules is particularly useful for discovering relationships among items from large databases [2].

Association rule mining deals with market basket database analysis for finding frequent itemsets and generate valid and important rules. Various association rules mining algorithms have been proposed by Agrawal et al. [4]. These algorithms include Apriori, Apriori-TID and Apriori Hybrid. Other algorithms for finding frequent itemsets include pincer search [3], FP (frequent pattern) tree [5]. A frequent pattern tree generates frequent itemsets without candidate generation.

1.1 Market Basket Analysis

Market basket analyses are an important component of analytical system in retail organizations. There are several definitions of market basket analysis. In a broader meaning, market basket analysis targets customer baskets in order to monitor buying patterns and improve customer satisfaction [6]. The following analytics can be used: attachment rates, demographic baskets, brand switching, customer loyalty, core items, items per basket, in-basket price, revenue contribution, shopper penetration and others.

In a narrower meaning, market basket analysis gives us the answer to the following question: which goods are sold together within the same transaction or to the same customer? By analyzing this information, we try to find out persistent patterns in order to offer related goods together and therefore increase the sales. We can track related sales on different levels of goods classifications or on different customer segments. Market basket analysis can also help retailers plan which items to put on sale at reduced prices [2]. Figure 1 shows market basket analysis.

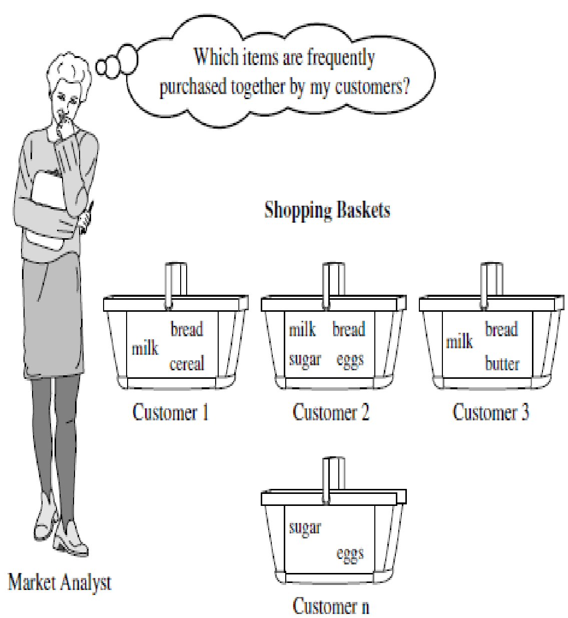

Figure 1: Market basket analysis adopted from [2]

Vol 7. No. 3 - September, 2014
African Journal of Computing & ICT

www.ajocict.net

The remaining part of this paper is organized as follows. Section 2 discusses the relevant related research carried out by the previous researchers in association rule mining. Section 3 focuses on methodology with main emphases on the Apriori algorithm used in this paper. Section 4 discusses the results obtained while Section 5 concludes the research work.

2. RELATED WORK

Data mining in education using association rule mining technique for identifying students' failure patterns was proposed by [7]. This approach identifies hidden relationship between the failed courses and suggests relevant causes of the failure to improve the low capacity students' performances. The researchers observed that the execution time of the proposed approach is inversely proportional to minimum support, since it increases as minimum support decreases, which confirmed increase in system complexity and response time as the minimum support decreases. It was also observed that to have a less complex system and a constructive, interesting and relevant patterns, the minimum confidence and support should be large enough to trash out coincidence patterns. In their research, 19 frequent itemsets and 114 rules were generated.

The researchers concluded that all the rules with confidence 1, are very strong rules, which implies that if a student failed the determinant (antecedent) course(s), such student will surely fail the dependent (consequent) course(s). It was recommended that this rule should be placed in curriculum structure. Also, if the rule support is higher, it means that all the courses involved are failed together by most of the considered students. Their proposed approach assists in the curriculum structure and modification in order to improve students' academic performance and trim down failure rate.

Bin, Wynne and Yiming [8] proposed hybridized mining algorithms for rule generation and for classification. The proposed algorithm is called CBA (Classification Based on Associations) which consists of two parts; a rule generator (CBA-RG) and a classifier builder (CBA-CB). The rule generator is based on Apriori algorithm for finding association rules. The CBA-RG generates set of class association rules (CARs) which consists of all the possible rules (PRs) that are both frequent and accurate. This is done by making multiple passes over the data. The proposed CBA-CB algorithm uses CARs to produce the best classifier out of the whole set of rules. This involves evaluating all the possible subsets of the training data and selecting the subset with the right rule sequence that gives the least number of errors. The researchers concluded that the classifier built this way is more accurate than that produced by the state-of-the-art classification system like C4.5. Furthermore, this integration helps to solve a number of problems that exist in the current classification systems.

Pragya, Madan, and Nupur [9] carried out a study on Apriori algorithm. An association driven mining application developed in Java was proposed. This application can be used to manage retail businesses that provide retailers with reports regarding prediction of product sales trends and customer behavior. The proposed approach improved Apriori algorithm in terms of running time, number of database scan, memory consumption and the interestingness of the rules over the classical Apriori Algorithm. The researchers proposed a mining algorithm for incremental large itemsets generation. Frequent itemsets discovered depends on value of parameters like support and number of transactions read at a time. Thus, execution time of the algorithm depends on transactional datagroups and minimum support value.

Margaret, Yongqiao, Le and Zahid [10] conducted an extensive survey of association rule mining. In their research, previous studies on association rule mining were explored and a brief classification strategy and performance comparison was carried out. A Genetic Algorithm for mining quantitative association rules was proposed by [11]. This algorithm was named QuantMiner. QuantMiner is a genetic-based algorithm for mining quantitative association rules. In QuantMiner, an item is either an expression $A = v$, where A is a categorical (also called qualitative) attribute and v is a value from its domain, or an expression $A \in [l, u]$ where A is a quantitative attribute.

According to the researchers, QuantMiner works directly on a set of rule templates. A rule template is a preset format of a quantitative association rule, either chosen by the user or computed by the system. For each rule template, the algorithm looks for the best intervals for the numeric attributes occurring in that template, using a Genetic Algorithm. Figure 2 shows the proposed QuantMiner algorithm together with its fitness function.

www.ajocict.net

```
Function Fitness (A ⇒ B)
  TempFitness = Gain(A ⇒ B)
  if TempFitness ≥ 0 then
     foreach interval I in A ⇒ B do
        //favor small intervals
        TempFitness * = (1-Prop(I))^2
        if Support(A ⇒ B) < MinSupp then
           //penalize low support rules
           TempFitness - = Nbtuples
  return TempFitness
```

```
Algorithm 2: QUANTMINER
  Input: A dataset composed of NbTuples, PopSize,
         GenNb, CR, MR, MinSupp, MinConf
  Output: Quantitative association rules R
  Select a set of attributes
  Let R_t a set of rule templates defined on these attributes
  Compute the set of frequent itemsets on categorical
  attributes in R_t
  R = ∅
  foreach r ∈ R_t do
     Generate a random population POP of PopSize
     instantiated rules following the template r
     i=1
     while i ≤ GenNb do
        Form the next generation of population by
        mutation and crossover w.r.t. MR and CR.
        Keep PopSize rules in POP with the best Fitness
        values
        i++
     R = R ∪ Argmax_{R∈POP} Fitness(R)
  return R
```

Figure 2: QuantMiner algorithm proposed by [11]

The researchers concluded that QuantMiner is very useful as an interactive data mining system that provides a better optimization criterion based on both support and confidence for keeping only high quality and interested rules.

3. METHODOLOGY

3.1 Association Rule Mining

Association Rule: An association rule is an implication of the form $X \Rightarrow Y$, where *X* and *Y* are sets of items called itemsets and $X \cap Y = \emptyset$. Here, X is called the antecedent and Y is consequent. There are two important measures for association rules, these are support (*s*) and confidence ($\propto$) [10]. An itemset that contains ***k*** items is a ***k***-itemset. The set {bread, butter, nylon} is a **3**-itemset. The occurrence frequency of an itemset is the number of transactions that contain the itemset. This is also known as the frequency, support count, or count of the itemset [2].

Support: The support (*s*) of an association rule is the ratio of the records that contain $X \cup Y$ to the total number of records in the database. Therefore, support ($X \Rightarrow Y$) = $P(X \cup Y)$. If a rule has 6% support value, this means that 6% of the total records contain $X \cup Y$.

Confidence: The confidence ($\propto$) of a given number of records is the ration of the number of records that contain $X \cup Y$ to the number of records that contain X. Therefore, confidence ($X \Rightarrow Y$) = $P(Y|X)$. Mathematically we have:

Confidence ($X \Rightarrow Y$) =

$$P(Y|X) = \frac{support\ (X \cup Y)}{support(X)} = \frac{support\_count(X \cup Y)}{support\_count(X)} \dots\dots\dots\dots\dots \quad (1)$$

For example, if a rule has a confidence of 80%, this means that 80% of the records containing *X* also contain *Y*.

3.2 Apriori Algorithm

Apriori algorithm was proposed by Agrawal and Srikant 1994 as an algorithm for mining frequent itemsets for Boolean association rules. This algorithm works on the prior knowledge of frequent itemset properties such that k-itemsets are used to explore *(k + 1)*-itemsets. The set of frequent k-itemsets where $k \geq 1$ is found by scanning the database to accumulate the count for each item and later collects those items that satisfy the minimum support which is denoted by L_1. The set L_1 is then used to find L_2 which is the set of frequent 2-itemsets. This process continues until no more frequent k-itemsets can be found. According to [2], Apriori property is used to improve the efficiency of the level-wise generation of frequent itemsets. This property states that "*all nonempty subsets of a frequent itemset must also be frequent*". For instance, if an itemset *I* is not frequent, that is, it does not satisfy the minimum support threshold, then any other resulting itemset composed by adding an item say *O* to itemset I (i.e $I \cup O$) cannot occur more frequently than *I*.

This means that $P(I \cup O) < min\_sup$ where *min_sup* is the minimum support threshold. Apriori algorithm follows a two-step process called join and prune. Below is the detail of the Apriori algorithm adapted from [2]. This algorithm was implemented in this paper using PHP and MySQL to discover frequent itemsets for mining Boolean association rules.

Procedure Apriori_Algorithm()
Input: *D*, a database of transactions; *min_sup*, the minimum support count threshold
Output: *L*, frequent itemsets in *D*

African Journal of Computing & ICT

www.ajocict.net

Steps:

(1) L_1 = find_frequent_1-itemsets(D);
(2) **for** (k = 2; $L_{k-1} \neq \phi$; k++) f
(3) C_k = apriori_gen(L_{k-1});
(4) **for each** transaction $t \in D$ { // scan D for counts
(5) C_t = subset(C_k, t); // get the subsets of t that are candidates
(6) for each candidate $c \in C_t$
(7) c.count++;
(8) **endfor**
(9) $L_k = \{ c \in C_k \mid c.count \geq min\_sup \}$
(10) **endfor**
(11) **return** $L = \cup_k L_k$;

Procedure apriori_gen(L_{k-1}: frequent (k-1)-itemsets)

(1) **for each** itemset $l_1 \in L_{k-1}$
(2) **for each** itemset $l_2 \in L_{k-1}$
(3) **if** $(l_1[1] = l_2[1]) \wedge (l_1[2] = l_2[2]) \wedge ::: \wedge (l_1[k-2] = l_2[k-2]) \wedge (l_1[k-1] < l_2[k-1])$ then {
(4) $c = l1 \times l2$; // join step: generate candidates
(5) **if has_infrequent_subset**(c, Lk-1) then
(6) **delete** c; // prune step: remove unfruitful candidate
(7) else add c to C_k;
(8) **endif**
(9) **return** C_k;

Procedure has_infrequent_subset(c: candidate k-itemset; Lk-1: frequent (k-1)-itemsets); //use prior knowledge

(1) **for each** (k-1)-subset s of c
(2) **if** $s \notin L_{k-1}$ then
(3) **return** TRUE;
(4) **return** FALSE;

4. RESULTS AND DISCUSSION

Figure 3 shows the interface used for setting minimum support and confidence threshold for the association rules generation. Item to be bought is typed inside the text box and the system searches the database to suggest items. The item is added to the transactional data by clicking on "Make Transaction" button.

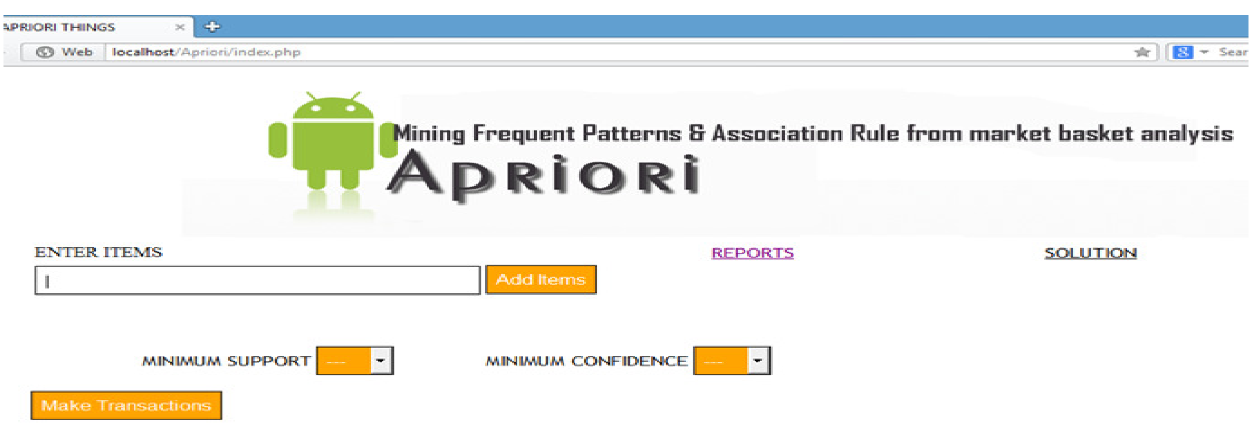

Figure 3: Interface for setting min_support and min_confidence

The solution menu is used to perform the join and prune steps of the Apriori algorithm as shown in Figure 4. The initial stage shows the transactional data used for the association mining followed by the frequent itemsets generation stages. Three itemsets were generated from the initial transactional data. These represent those itemsets, which satisfied the minimum support count threshold.

Vol 7. No. 3 - September, 2014
African Journal of Computing & ICT

www.ajocict.net

Mining Frequent Patterns & Association Rule from market basket analysis
Apriori

APRIORI during iterations (<<BACK) REPORTS STRONG RULE

ID	ITEMS
1	Camera WashingMachine GassCooker
2	DVDPlayer ElectricKettle Refrigerator
3	GassCooker Freezer Iron
4	TV SamsungS4 iPhone
5	TV WashingMachine Refrigerator
6	TV Generator TVStand
7	TV Generator TVStand
8	TV Generator TVStand
9	TV WashingMachine Refrigerator
10	GassCooker Freezer Iron
11	CDPack ExtensionBox
12	Camera WashingMachine GassCooker
13	Camera WashingMachine GassCooker

ID	ITEMS	SUPPORT COUNT
1	Camera	3
2	WashingMachine	5
3	GassCooker	5
4	DVDPlayer	1
5	ElectricKettle	1
6	Refrigerator	3
7	Freezer	2
8	Iron	2
9	TV	6
10	SamsungS4	1
11	iPhone	1
12	Generator	3
13	TVStand	3
14	CDPack	1
15	ExtensionBox	1

ID	ITEMS	SUPPORT COUNT
1	GassCooker	5
2	WashingMachine	5
3	Refrigerator	3
4	TV	6
5	Generator	3
6	TVStand	3

ID	ITEMS	SUPPORT COUNT
1	TV Generator	3
2	TV TVStand	3
3	Generator TVStand	3

Figure 4: Apriori algorithm during iterations

Figure 5 shows the strong rules generated after the computation of confidence on the three most frequent itemsets as satisfied by the threshold. This shows that if a customer purchases TV, there is 70% confidence that he/she will purchase TV Stand and Generator. Conversely, there is 85% confidence that a customer who purchases Generator will purchase both TV and TV Stand as well.

Mining Frequent Patterns & Association Rule from market basket analysis
Apriori

BACK REPORTS

STRONG RULES	
RULES	**CONFIDENCE**
TV --> Generator TVStand	70%
Generator --> TV TVStand	85%

Figure 5: Generated strong rules

African Journal of Computing & ICT

www.ajocict.net

5. CONCLUSION

This paper focuses on Apriori implementation for association rule mining. Frequent itemsets generated by Apriori algorithm completely depend on a minimum support threshold. It was observed that Apriori wastes running time due to candidate itemsets generation, hence, the need for a more robust hybrid algorithm for association rule mining is inevitable. However, for a transactional database where many transaction items are repeated many times as a superset in that type of database, Apriori is suited for mining frequent itemsets. Thus, this algorithm produces frequent itemsets completely. The proposed system can be used to determine the buying patterns of customers with greater rate of accuracy and subsequently improve daily sales.

REFERENCES

[1] Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases, American Association for Artificial Intelligence. Retrieved 20th July, 2013 from http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf

[2] Han, J. and Kamber, M. (2006). Data Mining: Concepts and Techniques, Morgan Kaufmann. USA: Elsevier Inc.

[3] Lin, D., Kedem, Z. M. (2002). Pincer-Search: An Efficient Algorithm for Discovering the Maximum Frequent Set, IEEE Transaction on Knowledge and Data Engineering, 14(3), pp. 553-566.

[4] Agrawal, R. Imielinski, T. and Swami, A. (1993). Mining Association Rules between Sets of Items in Large Databases, Proceedings of the ACM SIGMOD International Conference on Management of Data.

[5] Han, J., Pei, J., and Yin, Y. (2000). Mining Frequent Patterns without Candidate Generation, Proc. ACM SIGMOD.

[6] Microstrategy (2003). Business Intelligence in the Retail Industry, Microstrategy World 2003 Conference, Las Vegas.

[7] Oladipupo, O. O. and Oyelade, O. J. (2010). Knowledge Discovery from Students' Result Repository: Association Rule Mining Approach, International Journal of Computer Science & Security (IJCSS), 4(2), pp.199-207.

[8] Bing, L., Wynne, H. and Yiming, M. (1998). Integrating Classification and Association Rule Mining, American Association for Artificial Intelligence. Retrieved 20th June, 2013 from http://www.cs.rit.edu/~dmrg/dm_winter/reading/integ_class_arm.pdf.

[9] Pragya, A., Madan, L. Y. and Nupur, A. (2013). Study on Apriori Algorithm and Its application in Grocery Store, International Journal of Computer Applications, 74(14), pp. 1-8. Retrieved 14th July, 2013 from http://www.ijcaonline.org.

[10] Margaret, H. D., Yongqiao, X., Le, G. and Zahid, H. (2009). A Survey of Association Rules. Retrieved 14th July, 2013 from http://www.coursehero.com/file/2789589/grue-assoc

[11] Ansaf, S., Christel, V., and Cyril, N. (2007). QuantMiner: A Genetic Algorithm for Mining Quantitative Association Rules. Retrieved 20th June, 2013 from http://www.eecs.wsu.edu

African Journal of Computing & ICT

www.ajocict.net

Authors' Biography

Adewole Kayode Sakariyah is a lecturer of at the Department of Computer Science, University of Ilorin, Ilorin, Kwara State, Nigeria. He obtained a National Diploma in Computer Science with distinction at the Federal Polytechnic, Ede, Osun State in 2005, a B.Sc Computer Science with first class honours at the University of Ilorin, Ilorin, Nigeria in 2009, a Master of Science Degree in Computer Science at the University of Ilorin, Ilorin, Nigeria. His research interests include Information Security, Artificial Intelligence, Data Mining and Mobile Computing. He is highly proficient in application software development using a wide range of programming languages. He can be reached by phone on +2348038544348 and through E-mail adewole.ks@unilorin.edu.ng

Akintola Abimbola Ganiyat is a Lecturer in the Department of Computer Science, Faculty of Communication and Information Sciences. University of Ilorin. Ilorin. She obtained a Bachelor degree (B.Sc.) in Computer Science from the Bowen University Iwo, Osun State, Nigeria in 2007, where she graduated with a Second Class Honours (Upper Division). A Master degree (M.Sc.) in Computer Science from the University of Ilorin in 2011. She is currently pursuing a doctoral degree (Ph.D) in Computer Science at University of Ilorin. Her research interests include Biometrics, Data Mining and Software Engineering. She can be reached through abimbolar@yahoo.com

Abdulsalam Sulaiman Olaniyi holds a Bachelor of Science (B.Sc.) Degree and Master of Science (Degree) in Computer Science both from University of Ilorin, Ilorin, Nigeria. He is currently a lecturer in the Department of Computer, Library and Information Science, Kwara State University, Malete, Nigeria. He is a member of Nigeria Computer Society (MNCS). His research areas include Data Mining, Software Engineering and Artificial Intelligence. He can be reached by phone on +2348030785287 and through E-mail sulepy@yahoo.com

Ajiboye Adeleke Raheem obtained B.Sc.(Computer Science) from the Department of Computer Science, University of Ilorin, Ilorin, Kwara State, Nigeria. He obtained a Master of Science in Computer Science at the same department in 2010. He joined the department as an Assistant Lecturer in 2011. In 2013, he commenced a PhD degree in Computer Science at the Universiti Malaysia Pahang. He is a member of Nigeria Computer Society (NCS) and Computer Professionals Registration Council of Nigeria (CPN). His research interests include: Data Mining, Computational Intelligence and Image Processing. Ajiboye has five publications to his credit in reputable International Journals. He can be contacted through e-mails: ajibraheem@unilorin.edu.ng ; ajibraheem@live.com