IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-8, Aug 2019 E-ISSN: 2347-2693

Clustering as a Tool for Categorization of Unstructured Data

Ngor Gogo[1*], E. O. Bennett[2]

[1,2]Department of Computer Science, Rivers State University, Port Harcourt, Nigeria

[*]*Corresponding Author: stegongor70@gmail.com, Tel.: +2348037546747*

DOI: https://doi.org/10.26438/ijcse/v7i8.116121 | Available online at: www.ijcseonline.org

Accepted: 17/Aug/2019, Published: 31/Aug/2019

***Abstract*—** The volume of information untapped are locked up in huge volume of text documents (unstructured data) that could aid the economy, government, individuals and corporate organisation to improve on the state of life and develop better working system cannot be overemphasized, therefore the need to extract this information and give a structure that will facilitate its proper storage and access when required becomes so important. The target of this research is to explore Clustering as a Tool for Categorizing Unstructured Data (Text document). The K-Prototype Algorithm was applied for the purpose of clustering these unstructured data to give structure to it. There are two major phases involved in this: first is the pre-processing phase (Tokenization, Stemming, and Stop Word Removal) and secondly the clustering phase. The system built performed better as shown from the result, that it can be use to categorise text documents for proper and easy storage and accessibility.

***Keywords*—** Unstructured data, Clustering, Categorisation, K-Prototype Algorithm, pre-processing

I. INTRODUCTION

The volume of data produced and shared by various organizations, businesses, industries, etc. and the problems that are associated with managing, accessing and storage of information echoes the need for business organizations to pay a closer attention to the relevance of unstructured data in today's intense competitive world. Goutam in this publication (Analysis of unstructured data) revealed that significant volume of the digital world is dominated with unstructured data. Unstructured data occupies a very huge portion in digital space; an estimate of 80% of the quantity in comparison to only 20% of structured data [1]. On daily basis, there has been a continuous rise in the volume of unstructured data (text documents) and the volume of information locked inaccessible. Therefore, the need to explore efficient Categorization tool that will enhance fast and appropriate group of these text documents that will consequently facilitate easy access and storage of vital information.

Clustering is the process or technique applied in grouping data objects on the basis of some aspects of relationship existing between the objects in the group called clusters. Clustering has diverse areas that it can be applied such as: pattern recognition, data mining and firmness, machine learning, etc. [2]. Clustering is a formal acquisition of the knowledge of techniques and algorithm for grouping or clustering data items (elements) in accordance with measured or identifiable intrinsic attributes or similarity. Cluster analysis does not apply category label that identifies data elements with earlier identifiers, such as group labels. The unavailability of class information differentiates data clustering (unsupervised learning) from classification analysis (supervised learning). The purpose of clustering is to search into nature as to discover structure in data. The rich history of clustering and its application in various scientific fields dates as far back as before 1955, when k-means clustering algorithm was first published. Going forward from then, scientists have developed quite a number of other clustering algorithms. This informed the challenge faced in designing a clustering algorithm that can serve for a general purpose [3].

Clustering is the method of creating groups in scattered cases, fragmenting a single, diverse set of cases into several subsets of similar cases depending on similarity of their properties; it emphasizes on finding a form (structure) in a collection of untagged data. Clustering implies the collection of objects which are similar between them and are different from objects belonging to other clusters. The similarity measure is distance: two or more objects belong to the same cluster, if they are intimate based on a given distance (geometrical distance). This is referred to as distance-based clustering. Clustering is viewed as one of the most unsupervised learning problems. It falls into the category of data driven data mining; this is used in unfolding the connections between attributes in unknown data. "I don't know what I don't know", seems the approach [4].

The rest of the paper is organized as followed: section I contains introduction of Clustering as a Tool for

Categorization of Unstructured Data, Section II contains the Related Work of different document clustering system, section III explains the methodology and use case diagram of the system, section IV contains the system architectural and the essential steps taken in the system, section V describes the result and discussion of the system are presented, section VI concludes the research work

II. RELATED WORK

An Efficient Algorithm for Document Clustering and Classification is a necessity that intends addressing problems and challenges associated with proper organization, storage and retrieval of unstructured data. Unstructured data according to [5] is information, whose forms differ in many ways, and doesn't fit into the conventional data models; and obviously isn't a suitable for a mainstream relational database. Appreciably is the emergence of other platforms for storing and managing such data, it is progressively prevalent in Information Technology systems and is used by organizations in diverse operational intelligence and analytics applications. In dealing with unstructured data, there is a need to look at the subject of Big Data, which is a broad umbrella under which unstructured data is a type. Big data is not a recent concept. History is furnished with volumes of data/information overflow characterizing, as a result of the advent of social transformation and the introduction of new technologies. Within this period, individuals, various administrative powers, organizations, firms, and industries have come-up with essential data sets, organized at the time in sequence that are logical and coherent. To be translated into relevant information about the past, making it as useful as necessary in present times [6].In Pattern Based Clustering, the issue of pattern-based clustering dates back to the early years of data mining. A pattern-based flat clustering model that applies association rules to generate a hyper-graph of pattern (that is, using atomic patterns as vertices and rules applied to form hyper-edges). A more reliable hyper-graph partitioning algorithms is used to get pattern clusters. In addition, instances are clustered by designating each instance to its best pattern cluster. This model was further applied in later applications such as topic identification [7] and web image clustering [8]. In another vein [9] used efficient search space pruning techniques to achieve a universal summary set that has one of the longest frequent patterns for each transaction. Clusters are later formed from this set. There are many disadvantages associated with this approach, such as dependence on the minimum support threshold.

In Hierarchical Clustering [10] replying on global item sets suggested an earlier pattern-based hierarchical clustering model. This model was later modified upon by [11] and [12] that modified various aspects of the clustering process. [13] In tackling this, a different approach was applied; first mined globally significant maximum hyperlink (that is, high h-confidence) patterns, and all applicable pattern clusters are associated with instance. Hierarchical agglomerative clustering was as well applied to unify these clusters. This was also used by [11] to fuse together top-level nodes. Result from [13] reveals that this model produced clustering quality that is similar to UPGMA, in addition to being able to automatically identifying cluster label, which is an added advantage. A modification was carried out on the existing model by applying closed interesting item sets as globally significant patterns used for clustering, and choosing hierarchical relationship efficiently by applying interestingness measure. It was revealed that this approach outmatched both existing pattern-based hierarchical clustering algorithm and the best known agglomerative (UPGMA: Unweighted Pair Group Method with Arithmetic Mean) and partition-based algorithm on commonly used data sets. [8].

Subspace clustering is an offshoot of traditional clustering that looks forward to find clusters in different subspace within a data set. It places the search for germane dimensions allowing them to locate clusters that reside in multiple, possibly overlapping subspaces. These algorithms use a top-down or bottom-up search technique. The top down algorithm locates a first clustering in the full set of dimensions and evaluate the subspace of each cluster, using a series of iterations to make better the results. Transposable, bottom-up locates dense regions in low dimensional space and merge them to form clusters [14]. In trying to further enhance existing clustering algorithm, a Contrast Pattern-based Clustering (CPC) algorithm was proposed by Fore, N. K. to build up clusters without a distance function, by concentrating on the quality and density or fullness of contrast pattern, that contrast the clusters in a clustering. CPC seeks to optimize the Contrast Pattern-Based Clustering Quality (CPCQ) index which recognizes the expert-determined classes are most appropriate cluster for many data sets in the UCI repository. Experiments reveal that CPCQ scores are higher using UCI data sets for clustering gotten by CPC than other well-known clustering algorithms. In addition, CPC is able to regain expert clustering with higher accuracy, the expert clustering from these data sets than those algorithms [15]. Document clustering is the bringing of documents together on the grounds of their similarity. Documents are separated into different groups; documents with similarities are grouped together in the same group. Clustering expedites the process of knowledge discovery [16].

III. METHODOLOGY

A. Constructive Research Methodology

The constructive research methodology was adopted in this research because it targets resolving practical issues as well as generating theoretical contributions that are academically acceptable. This term construct points to an important plus

being created, which includes: a framework, algorithm, model, theory and software. Acquiring expanse knowledge of the problem domain and associated theories could form a strong foundation that may enhance construction of a desired solution [17].

B. Use Case Diagram for the Proposed System

The Use Case Diagram for the proposed system is a representation of the relationship and interactions between the actor (user) and the system when it is fully developed.

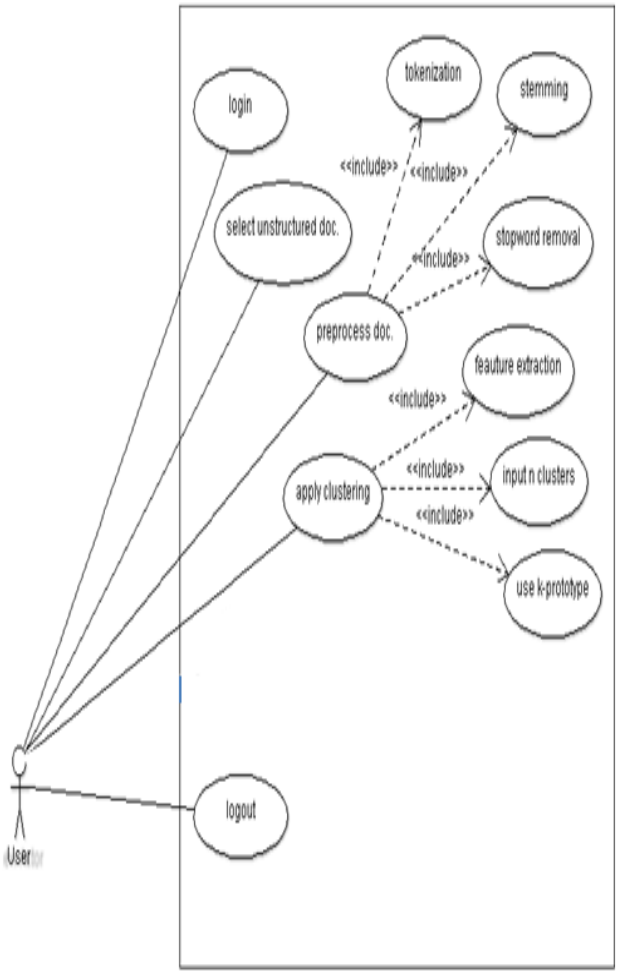

Figure 1: Use Case Diagram for the proposed system

IV. SYSTEM ARCHITECTURE

Architectural design the proposed system is the very initial step in the modeling of the software development process. It functions as the major connection between the design and requirements engineering, as it always indicates the main structural parts in the system and the relationships between them. It is a high level representation of the proposed system.

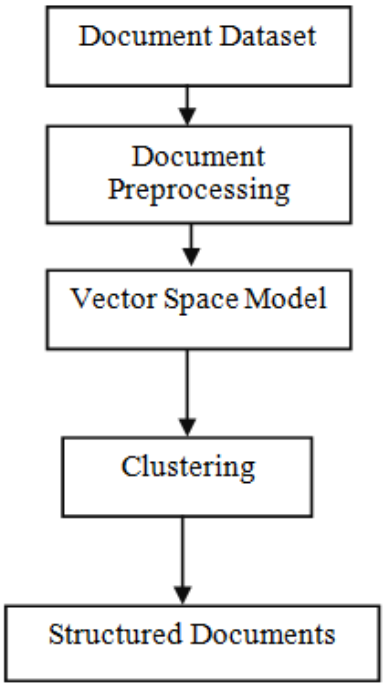

Figure 2: Proposed System Architecture

The proposed system accepts text documents (unstructured data) as it input dataset. The dataset under-goes a pre-processing Phase:

- Tokenization involves the process of breaking down word / characters into smaller parts or crunches while the security and integrity of the word is maintained.
- Stemming involves returning words to their root or morpheme state (by either removing their prefix or suffix).
- Stop word removal involves removing words which would appear to be less important in selecting document that would be compatible to meeting a user's prerequisite; these words are completely expelled from the vocabulary. This technique increments the effectiveness and efficiency of the result of clustered documents. Example of the stop words are: a, is, that, those, then, the, when, etc.

The output of the preprocessing phase is a Vector Space Model (Weighted Matrix). The Vector Space Model of a text data can be regarded as a word-by-document matrix, whose rows are the words and columns are document vectors; where each entry W_j depicts the weight of word ***i*** in the document j. The weight W_j can be determined in many ways. The frequency computes the number of occurrences of a term t_i in the document j.

$t_{ij} = f_{ij}$ = *frequency of word in document j*

This phase is concluded by creating weight matrix, this becomes the input to the Clustering Component (Phase)

$$\begin{pmatrix} T1 & & T1 \cdots & & Ti \\ D1 & w11 & w12 \ldots w1i & & c1 \\ D2 & w21 & w22 \cdots w2i & & c2 \\ : & : & : & : & : \\ Dj & wj1 & wj2 \ldots wji & & ck \end{pmatrix}$$

Figure 3: Vector Space Model (Weighted Matrix)

A. Clustering Component

The Clustering Component is responsible for the following activities: Euclidean Distance between two documents, Feature Extraction and Feature Clustering using K-Prototype algorithm.

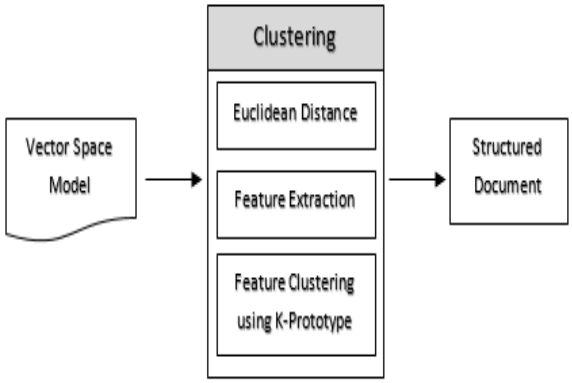

Figure 4: Clustering Component

- *Feature Extraction:*

This is used for extraction of features (important words and phrases in this case) from the documents. It can be achieved by using the Named-Entity tagger and frequency of unigrams and bigrams to extract the important words from the documents. The given document is read by the Named-Entity tagger and words that appear to form the main base of the document are extracted and are further used for the calculation of the similarity of a given set of documents.

- *Euclidean Distance:*

This is the most important phase in which the extracted features are clustered based on their co-occurrence. Here, similarity of documents is measured using the Euclidean distance measure, documents that are similar tend to have a very low value in as their distance measure while exact documents have a Euclidean distance measure of zero (0) meaning that they are the same. After the calculation of similarity, the k-prototype clustering algorithm is then used to cluster the documents based on this similarity, the k-prototype is used due to its advantages over the k-mean and k-mode clustering algorithms and also its ability to handle large data sets.

Euclidean distance is a standard metric for geometric probability. It is the distance between two points and can easily be measure with a ruler in two- or three-dimensional space.

Measuring distance between two documents d_x and d_y represented by their term vectors $\overrightarrow{t_x}$ and $\overrightarrow{t_y}$ respectively, the Euclidean distance of the two documents is defined as

$$D_E(\overrightarrow{t_x}, \overrightarrow{t_y}) = \left(\sum_{t=1}^{m} |w_{t,x} - w_{t,y}|^2\right)^{1/2} \quad (1.0)$$

Where the term set T = {t_1, …., t_m), and $w_{t,x} = tfidf(d_x t)$

B. K-prototype Function

This is the main function in the clustering phase of the algorithm, its input is a set of documents whose similarity has already been computed using the Euclidean distance measure, and a user defined input of k-clusters being the total number of clusters to be generated by the algorithm. The algorithm is built upon three processes, initial prototypes selection, initial allocation, and re-allocation. The first process simply randomly selects k objects as the initial prototypes for clusters. The second Starting from a set of initial cluster prototypes, assigns and current clusters of the object are updated. Variable moves and records the number of objects which have changed clusters in the process. Thus it clusters the given set of documents into the k clusters and plots a scattered plot as its output to show the documents in various clusters.

C. System Requirement / Set Up

- The software (EADCC) is a web-based application and was built using HTML 5, CSS 3 and JavaScript (ECMA 6) as its front end while PHP (7.2.3) was used to handle the backend.
- Google Chrome browser was used during the testing phase of the software development.
- The system was tested with a dataset of 737 BBC sports text documents having 5 natural classes from http://mlg.ucd.ie/datasets/bbc.html.
- BBC Sports Dataset:
- Documents: 737, Terms: 4613, Natural Classes: 5 (athletics, cricket, football, rugby, tennis).

The details of the dataset are given in Table 1 below.

Table 1: Dataset of BBC Sport

Natural Classes	Number of Documents	Total Words	Total Unique Words
Athletics	101	1018	458
Cricket	124	1032	287
Football	265	997	389
Rugby	147	977	531
Tennis	100	589	162
Total	737	4613	1827

D. Clustering Output

Figure 5. Sample Clustering output page (Scattered Plot)

E. Documents Clustered into Folders

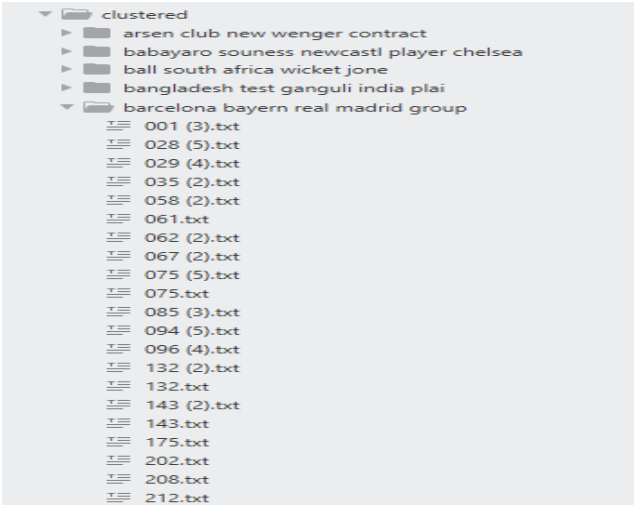

Figure 6: Documents Clustered into Folders

V. RESULTS AND DISCUSSION

The proposed system was developed with the adoption of K-prototype algorithm which was tested with a dataset from BBC sports repository. The Dataset were absolutely text document that were upload into the system. It was pre-processed and finally subjected to clustering. The number of clusters expected is predefined by the user (actor) of the system. The name of each cluster is derived from the system, using the five (5) most frequent occurring words that appeared in the cluster but not in other clusters and stored in folders that are created based on the number of predefined number of clusters.

Figure 5 presents a scattered plot which is the output of the developed clustering system showing the various clusters of documents and each dot represents a document with some overlapping others. The overlap is as a result of their degree of similarity. The system also displays the various cluster folders, their labels and text documents in that particular cluster labelled folder as shown in Figure 6.

Table 2, displays the various clusters and the number of documents in them, this is demonstrated with the aid of a column chat on figure 7. Reading from the chart the cluster with the highest number of documents was “ball, south” and “babayaro, souness” had the least number of text document in its cluster.

Finally, Table 3 shows the time performance and efficiency of our developed system, when compared to three (3) other existing systems. Our developed system clustered a thousand text documents at the least time of 197 seconds; this shows a significant performance and efficiency in the developed system with reference to clustering time. This was also graphically displayed on a line graph in Figure 8.

Table 2: Number of Documents Clustered in each Cluster Topic (Label)

Cluster Topic	Number of documents
Arsen, Club	18
Babayaro, Souness	12
Ball, south	150
Bangladesh, Test	120
Barcelona, Bayern	42

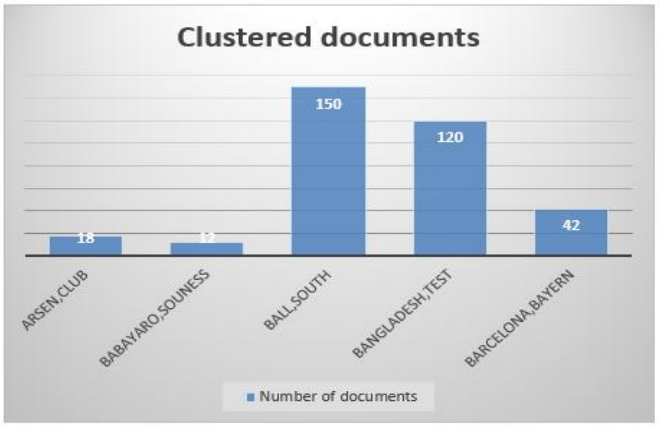

Figure 7: Column Chart Displaying the number of Documents Clustered in each Cluster Topic in the Developed Systems in Table 2.

Table 3: Clustering Time Comparison of Systems

Systems	Time Taken To Cluster 1000 Documents (sec.)
System 1 (Jyotismita Goswami, 2015)	360
System 2 (Anna Huang,2017)	384
System 3 (Santra & Josephine, 2012)	281
New System	194

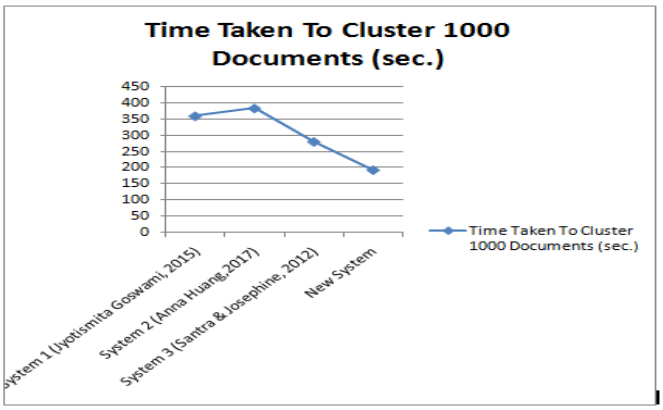

Figure 8. Line Graph Displaying **Time** Comparison Document Clustering Systems in Table 3

VI. CONCLUSION

The serious challenges associated with storage and accessibility of unstructured data (text documents) and the very pressing need to efficiently and correctly cluster and classify these text documents that make up and hold 70% of the information that are relevant to the world so that it can be structured in order to make it easier to extract, store and access such information. In Nigeria, for example like many nations in the world, most information that can aid research and the development of the nation are trapped in volumes of text documents that are unstructured, thereby, making it hard to properly store and retrieve this information when needed. The employment of K-Prototype Algorithms to this research work has aided in the production of a system with an efficient text document clustering capability.

This makes it easier for unstructured text documents to be sorted, clustered based on their similarities and stored for easy access to relevant information contained in them.

REFERENCE

[1]. Chakraborty, Goutam, Murali Pagolu, and Satish Garla. Text mining and analysis: practical methods, examples, and case studies using SAS. SAS Institute, **2014**.

[2]. Praveen, P., and B. Rama. "A k-means Clustering Algorithm on Numeric Data." International Journal of Pure and Applied Mathematics Vol.**117**, Issue.**7**, pp.**157-164**, **2017**.

[3]. Jain, Anil K. "Data clustering: 50 years beyond k-means." In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp**.3-4**. Springer, Berlin, Heidelberg, **2008**.

[4]. Bhambri, . M. A. & Gupta, D. An Analysis of Document Clustering Algorithm,in ICCCCT-10, IEEE 2010, pp.**402-406**, **2013**.

[5]. Goswami, J. A comparative Study on clustering and classification Algorithms, International Journal of Scientific and Applied Science (IJSEAS) Vol.**1**, issue.**1**, June **2015** ISSN: **2395-3470** pp.**170-177**, **2015**.

[6]. Fredrick, J. & Leonardo S. Data Clustering, its application and benefits, Semantic Scholar, **2017**.

[7]. Clifton, Chris, Robert Cooley, and Jason Rennie. "Topcat: Data mining for topic identification in a text corpus." IEEE transactions on knowledge and data engineering Vol.**16**, Issue.**8** pp.**949-964**, **2004**.

[8]. Malik, Hassan H., and John R. Kender. "Clustering web images using association rules, interestingness measures, and hypergraph partitions." In Proceedings of the 6th international conference on Web engineering, pp.**48-55**, **2006**.

[9]. Wang, J. & Karypis, G., Efficient Summarizing Transactions for Clustering, In Proceedings of the Fourth IEEE International Conference on Data Mining, **2014**.

[10]. Beil, Florian, Martin Ester, and Xiaowei Xu. "Frequent term-based text clustering." In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pp.436-442. **2002**.

[11]. Fung, Benjamin CM, Ke Wang, and Martin Ester. "Hierarchical document clustering using frequent itemsets." In Proceedings of the 2003 SIAM international conference on data mining, Society for Industrial and Applied Mathematics, pp.**59-70**. **2003**.

[12]. Yu, Hwanjo, Duane Searsmith, Xiaolei Li, and Jiawei Han. "Scalable construction of topic directory with nonparametric closed termset mining." In Fourth IEEE International Conference on Data Mining (ICDM'04), pp. 563-566. IEEE, 2004.

[13]. Xiong, Hui, Michael Steinbach, Pang-Ning Tan, and Vipin Kumar. "HICAP: Hierarchical clustering with pattern preservation." In Proceedings of the 2004 SIAM International Conference on Data Mining, pp.**279-290**. Society for Industrial and Applied Mathematics, **2004**.

[14]. Parsons, Lance, Ehtesham Haque, and Huan Liu. "Subspace clustering for high dimensional data: a review." Acm Sigkdd Explorations Newsletter Vol.**6**, Issue.**1,** pp.**90-105**, **2004**.

[15]. Fore, Neil Koberlein. "A Contrast Pattern Based Clustering Algorithm for Categorical Data." **2010**.

[16]. Osinski, Stanislaw, and Dawid Weiss. "A concept-driven algorithm for clustering search results." IEEE Intelligent Systems Vol.**20**, Issue.**3,** pp.**48-54**, **2005**.

[17]. Oyegoke, Adekunle. "The constructive research approach in project management research." International Journal of Managing Projects in Business, Issue.**4,** pp.**4573-595**, **2011**.

Authors Profile

Mr. G. Ngor pursed Bachelor of Science from Rivers State University of Science and Technologies, Nigeria in 1999. He is currently pursuing Master of Science from Rivers State University, in the Department of Computer Science, Nigeria. He is currently working as Assistant Chief Systems Analyst/ Programmer, Rivers State University, Nigeria since 2003. He is a member of computer professionals of Nigeria (CPN) since 2010. This is his first publication and its on Clustering Algorithms. He has 1 year of Research Experience.

Dr. E O Bennett pursed Bachelor of Science from Rivers State University of Science and Technologies, Port Harcourt in year 1998 and Master of Science in year 2008 and Ph.D in year 2014 from University of Port Harcourt, Nigeria. He is currently working as a Lecturer in Department of Computer Science, Rivers State University, Nigeria since 2012. He is a member of computer professionals of Nigeria (CPN). He has published more than 20 research papers in reputed international journals and conferences and is also available online. His main research work focuses on Algorithms, Big Data Analytics, Data Mining. He has 7 years of teaching experience and 6 years of Research Experience.

