International Journal on
"Technical and Physical Problems of Engineering"
(IJTPE)
Published by International Organization of IOTPE

ISSN 2077-3528
IJTPE Journal
www.iotpe.com
ijtpe@iotpe.com

December 2023 | Issue 57 | Volume 15 | Number 4 | Pages 123-133

WEB MINING ALGORITHMS APPLIED TO UNIVERSITY RESEARCHERS PROFILES: A CASE STUDY

D.M.A. Al Kerboly [1] M.M. Hamad [2] O.A. Dawood [2]

1. Computer Science Department, College of Computer Science and Information Technology, College of Education for Pure Science, University of Anbar, Anbar, Iraq, doreyedm@uoanbar.edu.iq
2. Computer Science Department, College of Computer Science and Information Technology, University of Anbar, Anbar, Iraq, dr.mortadha61@uoanbar.edu.iq, omar-abdulrahman@uoanbar.edu.iq

Abstract- The study of big data is fundamental to modern academic and practical research. The data sets available in today's digital world are diverse and multifaceted. They encompass various types of information, such as online transactions, emails, movies, audios, images, sensor data, logs, posts, search queries, health records, social networking interactions, science data, sensors, and mobile phones with their associated apps. The increasing volume of multimedia data has a significant impact on internet backbone traffic, which is projected to rise by 70% by 2013. In this era of big data, only Google boasts a worldwide distribution of over a million servers. Meanwhile, new websites are being launched at a rapid pace, with 571 going live every minute of the day. Google Scholar profiles contain a wealth of information about academic research, including authors, publication statistics, and keywords related with each piece. Collecting and analyzing massive volumes of data can yield major insights into the academic world. Google Scholar Profiles provide writers with an easy approach to promote their scholarly achievements. The web scraper collects required online links, extracts required data from the source links, and saves the data in a csv file. By merging data from Google Scholar profiles with k-means clustering, we may construct a powerful tool for analyzing the academic research environment and uncovering relevant ideas. implying that the items in the rule are not commonly encountered in the dataset. This means that the items in the rule appear in only 2.25% of the transactions in the dataset. The rules' confidence level of 1.0 indicates that all transactions that have the antecedent also have the consequent. The high level of confidence indicates a strong association between the components in the rule.

Keywords: Big Data, Google Scholar Profiles, Web Scraper, Education Data Mining, K-Means Clustering.

1. INTRODUCTION

Modern academic and practical initiatives are centered around big data research. However, this type of research comes with a host of challenges that must be addressed in order to enhance service quality. Google processes 20 petabytes of data per day and watches 7.2 billion pages per day, and Yahoo built Hadoop, an open-source data storage and processing system, in 2006. Big data is a term used to describe a collection of huge and sophisticated data sets that are tough to process. Big Data analytics aids in the detection of anomalous activity in linked devices as well as the identification of faulty transformers [1-3].

Web mining is a sort of data mining that collects and extracts information from web data using at least one or more structural or usage data. Recommender systems gather information about the user using a variety of methods and sources in order to forecast which items the user requires and provide recommendations or advice based on the discoveries of the study the analytical process [4]. Big data has four characteristics: big volume, wide variety, high velocity, and high veracity. Deep learning has made great progress in the learning of large data characteristics during the previous few years. The study of how educators and students interact with educational systems is known as EDM, or Education Data Mining. Web mining is a method for discovering hidden connections among massive amounts of data. Many data mining companies utilize recommender systems to create recommendations for items like music, articles, movies, and even research papers. Web usage mining refers to the process of extracting valuable insights from web data access information. Web data mining and other technologies have resulted in improved web agent technology [5-7].

Google Scholar Profiles allow writers to easily highlight their scholarly output. You can also make your profile public so that it appears in Google Scholar results when others search for your name. When Google Scholar discovers new citations to your work on the web, the citation metrics are generated and updated automatically, and you may then add groups of relevant articles rather than just one at a time. Big data is rapidly becoming prevalent in education, creating new challenges for web mining and information processing. Web mining is a technique for uncovering hidden relationships in vast amounts of data by looking at the most popular items for example on a Google Scholar webpage or reviewing

previous academics' work [5]. Using many Python libraries, such as Pandas. Pandas is an open-source library designed to work with relational or labeled data. It includes several data structures and methods for working with time series and numerical data. Pandas is quick and offers users exceptional performance and efficiency [8].

In the next section, "Literature Survey: This section provides a comprehensive review of the existing literature on the topic, highlighting the relevant theories, concepts, and research findings. This section will reference previous research that has been done on the topic and provide background information for the study.

3. Motivation: This section explains the motivation behind the research, including why the research question is important and relevant. This section will explain why the research is needed and what gaps in previous research the study aims to address.

4. The section on research methodology provides an overview of the techniques employed in the study, such as the research design, data collection methods, and data analysis techniques. This section will detail the methods used to collect and analyze data.

5. The First Phase: This section describes the research's first phase, which includes data collection and preparation.

6. The Second Phase: This section describes the second phase of the research, which involves data analysis. This section contains subsections on dataset preprocessing and manual classification, as well as a description of the methods used to manually classify and prepare the dataset for further analysis.

7. The Third Phase: This section describes the third phase of the research, which involves developing a K-means clustering model and an Apriori algorithm to find associations among the data.

8. Result Analysis: the section offerings the research results and offers a comprehensive examination of the results. This section will detail the findings of the study and how they contribute to the research question.

2. LITERATURE SURVEY

The objective of this research was to determine the presence or absence from a collection of studies k-means and association rule.

1) Dharshinn, et al., 2019 [9], The studies were using WEKA, an open-source program, and utilized the Apriori algorithm as well as k-means because it allows for a variety of data mining methods comprise a range of techniques, including data preprocessing, clustering, regression, classification, visualization, and feature selection. This study integrates the Apriori algorithm as well as the k-means clustering technique to find out how each impacts the other. According to the results of the tests, by utilizing the grouping the k-means with Apriori algorithms, it is possible to obtain more detailed information in less time than with the Apriori method alone, through the total computation period of 21.93 minutes besides 17.41 minutes for the Apriori besides k-means combination.

2) Agrawal, et al., 2020 [10], The studies used K-means and Apriori algorithm. Creation of a machine learning dataset from a database management system for association and clustering. The result is a dataset has 13 characteristics and 150 records. They have used unsupervised machine learning methods like clustering and association analysis on a created dataset. The Elbow Method is used to establish the ideal number of clusters in clustering.

3) Kusak, et al., 2020 [11], The studies used Apriori algorithm, K-means algorithm. The Aim is to build an index map of the landslide that happened in Karahacl Region in January 2019 and to assess the region considerations conventional statistics and geographic information mining.

4) Naser, et al., 2021 [12], The methods employed in the study include Hierarchical Clustering, K-means Algorithms, Planned Clustering Algorithm, the CURE Clustering Algorithm, and Partition Clustering. The purpose of this research is to look into an artificial clustering method that divides a large number of objects into smaller sets (clusters). Things within one set are related to one another in this instance, rather than objects within other sets. Automatic clustering does not involve predefining the clusters number; instead, it is determined created on number of groupings in the actual dataset structure. As a result, this research will utilize the proposed algorithm as a reference to develop an approach that determines clusters number created on the distance between apices. The research aims to demonstrate that this approach can effectively identify the partitions of clustering for entire data set.

5) Ying Zhou, 2022 [13], the studies used short text clustering, K Means, FP-Growth. These papers aim developed a new FP-Growth-based enhanced K-means-based short text data mining technique. This research approaches the association strength between words rather than the similarity between short texts, like LSH, by clustering short text sets with frequent item sets. The study showed that we are able to beat out the conventional short text clustering techniques in terms of inter-class distance.

Table 1. A Results are compared of studies that K-means and association rule, and feature selection strategies

Citations	Techniques used	Aim	Result
Dharshinn and et.al, 2019 [9]	In the WEKA, Apriori algorithm and k-means	This study combines Apriori and k-means clustering to investigate how each strategy affects the other.	Apriori approach outperforms Apriori and K-means algorithms in info detail with the time commuting.
Agrawal and et.al, 2020 [10]	K-means and Apriori	Creating a machine learning dataset from a database rephrase management system.	Creating dataset has 13 characteristics and 150 records.
Kusak and et al., 2020 [11]	Apriori, K-means	The goal is to build an index map of the landslide in Karahacl Region	Traditional statistics and geographic information are used to evaluate regions.
Naser and et. al. 2021[12]	K-means Hierarchical Clustering	Clusters are automatically clustered to mirror the dataset structure	The proposed method can efficiently find clustering partitions for the full dataset
Ying Zhou, 2022 [13]	K Means, FP-Growth	K-means-based short text data mining prioritizes association strength over similarity.	Inter-class distance algorithms outperformed traditional short text clustering algorithms.

These studies led us to the conclusion that allows for a wide range of the data mining process includes techniques for example data pre-processing, data clustering, data regression, data classification, data visualization, besides feature collection, several studies used the open-source application WEKA with the Apriori method and K-means. Some researchers combine the Apriori algorithm and the k-means clustering method to examine how they interact. In clustering, the Elbow Method is used to determine the optimal number of groups. These studies made use of the Apriori and K-means algorithms. As a result, this work will serve as motivation for the construction of an algorithm in this project that determines the clusters number is determined based on distance between vertices.

3. MOTIVATION

Anbar University is a leading research and teaching institution with a wide community of researchers working in a variety of subjects. Regardless of the presence of many excellent academics, the university's research output and ranking might be improved further. One method is to encourage collaboration among scholars in the same field of study. Collaboration research teams can improve the quality and quantity of research outputs while also providing a platform for academics to exchange their knowledge and expertise. This work suggests the use of combination k-means and Apriori algorithms to support the development of collaborative research teams at Anbar University. These algorithms are frequently used in web mining and clustering, and they can be used to discover possible research collaborators based on their research interests, outputs, and other relevant factors in research collaboration. With these algorithms, Anbar University researchers can build research teams that are associated with their research interests and aims, improving the quality and impact of their research results. This paper proposes the construction of features are found, including research topic working groups, assessment working groups, and other relevant committees, in addition to enabling the formation of research teams. These groups can provide as a forum for researchers to discuss their findings, get input, and share their knowledge and skills. The groups can also review the quality of research outputs and provide suggestions on how to enhance them, which can help Anbar University's research output and ranking overall.

4. RESEARCH METHODOLOGY

This study presents the suggested system options supplied by the educational application in order to build and implement an integrated, combined technique for big data educational applications that utilizes web mining technologies. The project presented here involves a specific technique that may be implemented by following these steps: Web scraping enables us to gather a variety of study features. Through using University of Anbar Researcher as an example, Google Scholar was utilized to produce a dataset for academic usage, which was collected and stored as a CSV file. On the subgroup dataset, K-means clustering should be utilized. The prior algorithm's output had taken into account the Apriori technique's input, as shown in Figure 1, we divided this model into three phases, and the following figures illustrate the contents of each phase.

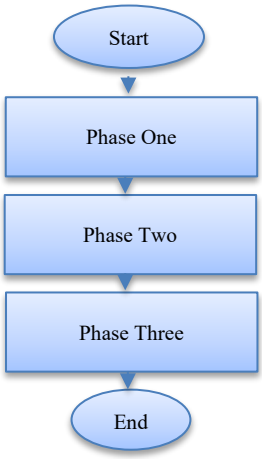

Figure 1. Proposed model

5. THE FIRST PHASE

As illustrated in Figure 2, the first Phase content makes use of Google Scholar for Anbar University researchers' profile web links using the "uoanbar.edu.iq" domain as input to web scraping stages. Crawlers primarily consist of downloaders, information extractors, schedulers, and crawl queues. The planner will seed the Allowing to download, and the downloader will collect page information from the web. Crawlers are classified into generic web crawler, focusing web crawlers, distributed crawlers and parallel crawlers [14].

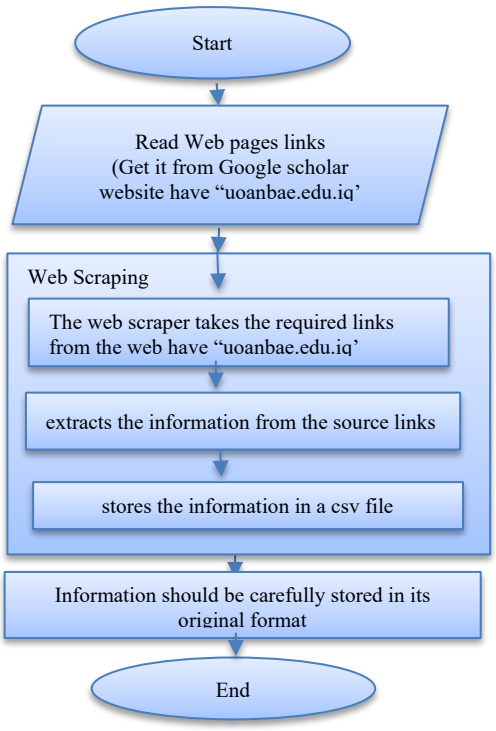

Figure 2. The first phase of proposed model

Web Crawler functions as a software application that assists in web scraping. Many search engines use a web crawler, sometimes known as a spider or an internet bot. Web crawlers collect all the pages from a given set of URLs and their links using this all-purpose crawling technology. This method aids the crawler in collecting several pages from various sites. A focused crawler serves to gather information just on the subject at hand. The primary goal of this focused crawler is to search specific web sites that are relevant to a given set of issues [6], [15], [16]. Web scraping is a computer approach for gathering huge amounts of information from websites. Web scraping is not illegal, but it is possibly unethical. This paper highlights data extraction from webpages using different research-backed web scraping strategies [17].

Table 2. Comparison between web scraping vs web crawling [18]

Web crawling	Web Scrapping
Web crawlers scan links on websites and index the content they find, creating a search index that search engines can use	Web scrapers extract information from a website and save it in a database or CSV file in an organized fashion
Data is saved by crawling.	Data is extracted by scraping.
Only requires a crawler or crawl agent	A crawler or crawl agent with a parser is required
Often performed on a vast scale	It is feasible at any scale.
Deduplication is required, and the crawler listens to the robots.txt file	Deduplication is optional. Scraper ignores the robots.txt file and views itself as a search engine.

6. THE SECOND PHASE

The second phase's content as illustrated in Figure 3, read the acquired data set, then preprocess it, then manually classify these data sets according to the field of study, after which save the outputs dataset as a CSV file.

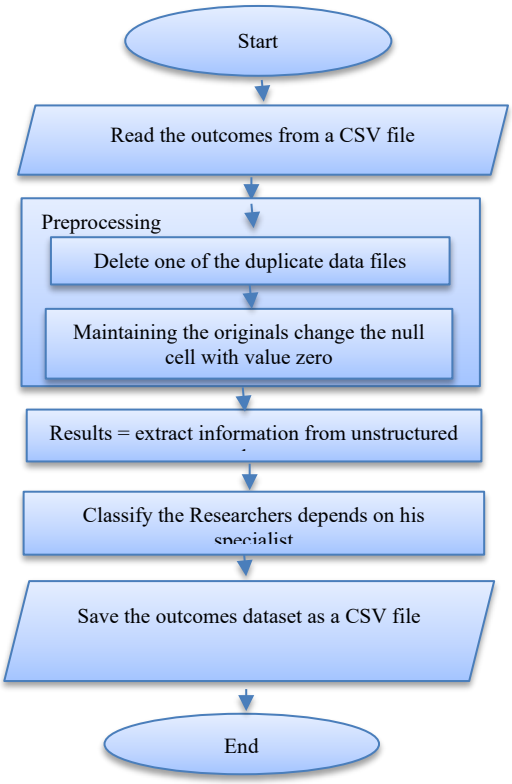

Figure 3. The second phase of proposed model

6.1. Dataset Preprocessing

Following the initial insertion of our dataset in a CSV file, the data was received and processed using Python's Panda module, including an examination to check if any data was duplicated or missing. The categorization model cannot be constructed unless each of these procedures is completed appropriately. Figure 4 shows the information about original dataset without preprocessing.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1301 entries, 0 to 1300
Data columns (total 5 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Researcher Name       1301 non-null   object
 1   Citation              893 non-null    float64
 2   Affiliation           1265 non-null   object
 3   Researcher Interested 1190 non-null   object
 4   LinK Of Researcher    1301 non-null   object
dtypes: float64(1), object(4)
memory usage: 50.9+ KB
```

Figure 4. Information about original dataset without preprocessing

a) Data Duplication: We delete all duplicate researcher accounts and select only the one with the highest number of citations across all of its accounts. Data duplication within the original data has grown so clear that it may be easily detected, which is why we destroyed them. They are usually the least interesting in terms of the information they can offer

b) Missing Value: Most datasets have some missing values. There are numerous methods to this problem that can be taken. Any null data in our example is usually changed to 0 (zero) as shown in Figure 5.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1197 entries, 0 to 1196
Data columns (total 5 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Researcher Name       1197 non-null   object
 1   Citation              1197 non-null   object
 2   Affiliation           1197 non-null   object
 3   Researcher Interested 1197 non-null   object
 4   LinK Of Researcher    1197 non-null   object
dtypes: object(5)
memory usage: 46.9+ KB
```

Figure 5. Information about preprocessing DataSet

6.2. Classify our Dataset

When we turned our unstructured data to structured data, the dataset grouped in various ways. The first one is named (1), which refers to the employee of the University of Anbar, while the others are named (2) depending on the affiliation of the researcher. Another group depends on the researcher's interests ['Computer Science', 'Mathematics', 'Chemistry', 'Physics', 'Biology', 'Engineering', 'Islamic Science', 'History', 'medical', 'Dental', 'Pharmacy', 'Teaching Methods', 'Geography', 'Arabic', 'English', 'Sociology', 'Law', 'Agricultural', 'Sport', 'Geology'] have these numbers consequently [3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22]. Take 0 for any other interested researchers or the researcher who has not registered his interest on his Google Scholar account.

7. THE THIRD PHASE

The third phase content reads the CSV file saved as output from previous implementations, then implements k-means, and finally uses a csv file output regarding the K-means algorithm, input to an Apriori Algorithm, as illustrated in Figure 6.

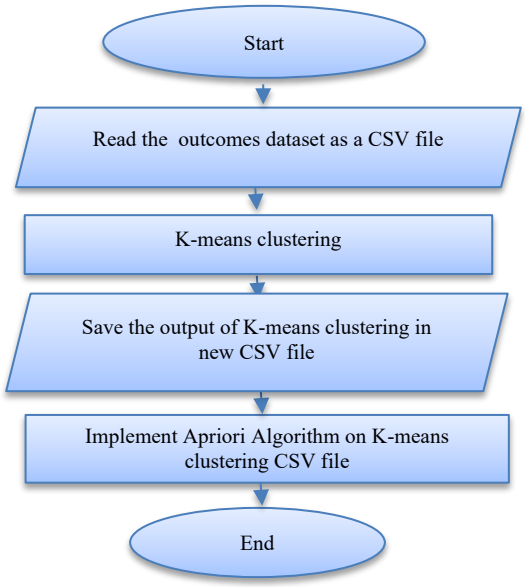

Figure 6. The third phase of proposed model

7.1. K-Means Clustering

K-means clustering basic and simple unsupervised data mining strategies for dealing with the well-known clustering challenge. J.B. MacQueen developed the approach, which is used for data mining and pattern identification. The K-means method has difficulty determining the optimal number of clusters, but it becomes more effective as the distance between clusters grows. The K-means model attempts to arrange a set of samples into this K group in such a way that the predictions of Equations (1) and (2) are validated. The minimal Euclidean distance of each sample from the center of each class or category is used in this method to allocate each sample to a class. The findings of elbow determine the optimal number of clusters for K-means [6], [19-21]. During the solvation process, Equation (1) is often used at the beginning of the technique to identify the center of each class [21].

$$z_j = \frac{\sum_{x \in c_j} x}{\#c_j} \quad \text{for } j = 1...k \tag{1}$$

where, create K classes or groups with the names "C_1, C_2, ..., C_k" to cluster m samples from set M. Using Equation 1. Determine the vector Z_j, which also represents the center or representative of each C_j category. $\#C_j$ is the number of samples in the C_j class, and x is the vector of a C_j member instance. To represent the center of each group, K samples are selected at random. Calculate the optimal solution resulting from the classification of "C_1, C_2, ..., C_k." Using Equation (2), get the total of the distances of the tests from the categories' centers.

Minimize the optimal solution of Equation (2) to find the suitable categorization on the dataset M with the stated number of K groups [21].

$$f(C_1, C_2, C_k) = \sum_{j-1}^{k} \sum_{x \in c_j} |X - z_j|^2 \tag{2}$$

Example: Group the thirty points below (with (x, y) representing locations) into three clusters $C_1(2, 10)$, $C_2(2, 5)$, $C_3(8,4)$, $C_4(5, 8)$, $C_5(7, 5)$, $C_6(6, 4)$, $C_7(1, 2)$, $C_8(4, 9)$, $C_9(6,9)$, $C_{10}(9,2)$, $C_{11}(5,2)$, $C_{12}(6,5)$, $C_{13}(4,8)$, $C_{14}(5,5)$, $C_{15}(6,4)$, $C_{16}(7,2)$, $C_{17}(9,9)$, $C_{18}(2,7)$, $C_{19}(1,5)$, $C_{20}(8,2)$, $C_{21}(6,4)$, $C_{22}(9,6)$, $C_{23}(2,8)$, $C_{24}(1,9)$, $C_{25}(8,7)$, $C_{26}(5,4)$, $C_{27}(4,6)$, $C_{28}(8,7)$, $C_{29}(7,4)$, and $C_{30}(6,4)$

1. The initial cluster centers are $C_1(2, 10)$, $C_4(5, 8)$, and $C_7(1, 2)$.
2. The distance function between two points $P_1=(x_1,y_1)$ and $P_2=(x_2,y_2)$ is defined as:

$$D_p(P_1, P_2) = |x_2 - x_1| + |y_2 - y_1|$$

3. Apply the K-means algorithm to determine the three cluster centers after the second iteration.
4. The K-means clustering algorithm discussed earlier will be utilized.

For the first iteration:

- Calculate the distance between each point and the three cluster centers using the provided distance function, the distance between point $C_1(2,10)$ and each of the three cluster centers
- Calculation of DistanceMean1 between first point $C_1(2,10)$ and first cluster center (2,10)

$$\begin{aligned} D_p(C_1, \text{Cluster Center 1}) &= |x_2 - x_1| + |y_2 - y_1| \\ &= |2-2| + |10-10| \\ &= 0 \end{aligned}$$

- Calculation of DistanceMean2 between first point $C_1(2,10)$ and second cluster center (5,8)

$$\begin{aligned} D_p(C_1, \text{Cluster Center 2}) &= |x_2 - x_1| + |y_2 - y_1| \\ &= |5-2| + |8-10| \\ &= 3+2=5 \end{aligned}$$

- Calculation of DistanceMean3 between first point $C_1(2,10)$ and third cluster center (1,2)

$$\begin{aligned} D_p(C_1, \text{Cluster Center 3}) &= |x_2 - x_1| + |y_2 - y_1| \\ &= |1-2| + |2-10| \\ &= 1+8=9 \end{aligned}$$

- Calculation of DistanceMean1 between second point $C_1(2,5)$ and first cluster center (2,10)

$$\begin{aligned} D_p(C_1, \text{Cluster Center 1}) &= |x_2 - x_1| + |y_2 - y_1| \\ &= |2-2| + |5-5| \\ &= 0+5=5 \end{aligned}$$

- Calculation of DistanceMean2 between second point $C_1(2,5)$ and second cluster center (5,8)

$$\begin{aligned} D_p(C_1, \text{Cluster Center 2}) &= |x_2 - x_1| + |y_2 - y_1| \\ &= |5-2| + |8-5| \\ &= 3+3=6 \end{aligned}$$

– Calculation of DistanceMean3 between second point $C_1(2,5)$ and third cluster center (1,2)

$$D_p(C_1, \text{Cluster Center 3}) = |x_2 - x_1| + |y_2 - y_1|$$
$$= |1-2| + |2-5|$$
$$= 1+3 = 4$$

Similarly, we calculate the DistanceMean1, DistanceMean2, and DistanceMean3 of other points from Determine the centers of the three clusters. Then, create a table containing all the computed distances. Utilize the table to give respectively idea to its neighboring cluster based on center with the shortest distance. The $C_1(2,10)$ became cluster 1 because the Distance Mean1 for these points is the smallest among all Distance Means. The $C_2(2,5)$ became cluster 3 because the Distance Mean3 for these points is the smallest among all Distance Means. Similarly, we check the smallest Distance Mean for other points to get the clusters shown in the Table 3.

Table 3. Clustering algorithm such as K-means clustering

			(2, 10)	(5, 8)	(1, 2)	
	Point		DistanceMean1	DistanceMean2	DistanceMean3	clusters
C_1	2	10	0	5	9	1
C_2	2	5	5	6	4	3
C_3	8	4	12	7	9	2
C_4	5	8	5	0	10	2
C_5	7	5	10	5	9	2
C_6	6	4	10	5	7	2
C_7	1	2	9	10	0	3
C_8	4	9	3	2	10	2
C_9	6	9	5	2	12	2
C_{10}	9	2	15	10	8	3
C_{11}	5	2	11	6	4	3
C_{12}	6	5	9	4	8	2
C_{13}	4	8	4	1	9	2
C_{14}	5	5	8	3	7	2
C_{15}	6	4	10	5	7	2
C_{16}	7	2	13	8	6	3
C_{17}	9	9	8	5	15	2
C_{18}	2	7	3	4	6	1
C_{19}	1	5	6	7	3	3
C_{20}	8	2	14	9	7	3
C_{21}	6	4	10	5	7	2
C_{22}	9	6	11	6	12	2
C_{23}	2	8	2	3	7	1
C_{24}	1	9	2	5	7	1
C_{25}	8	7	9	4	12	2
C_{26}	5	4	9	4	6	2
C_{27}	4	6	6	3	7	2
C_{28}	8	7	9	4	12	2
C_{29}	7	4	11	6	8	2
C_{30}	6	4	10	5	7	2

7.1.1. Iteration Two

Following that, we must recalculate the updated means for the new clusters. We achieve this by averaging all of the points contained within each cluster are as follows:

- Cluster 1 comprises of ((2+2+2+1)/4, (10+7+8+9)/4) = (1.75, 8.5)
- For Cluster 2, we have (8+5+7+6+4+6+6+4+5+6+9+6+9+8+5+4+8+7+6)/19, (4+8+5+4+9+9+5+8+5+4+9+4+6+7+3+6+7+4+4)/19) = (6.26, 5.89)
- For Cluster 3, we have ((2+1+9+5+7+1+8)/7, (5+2+2+2+2+5+2)/7) = (4.71, 2.85)
- As previous iteration we get new clusters.
- These steps are repeated until we get the same clusters in each iteration.

The actual WCSS and silhouette score values for the dataset and clustering arrangement specified are returned.

$$WCSS = \sum_{i \in n} (X_i - Y_i)^2 \quad (3)$$

The WCSS was subsequently compared against the clusters K number to calculate the number of clusters required. WCSS uses the formula below to calculate the sum of data' distances from cluster the centroids. WCSS uses the formula above to calculate the sum of data' distances from cluster the centroids. where Y_i represents the finding the center of X_i [22]. Within-Cluster-Sum-of-Squares (WCSS): The clustering's WCSS value is 57.4218. This number indicates how densely the clusters are clustered. The clusters become much more tightly bound as the WCSS ratio decreases. The Silhouette Score, the mean silhouette coefficient of all samples is calculated, and the silhouette coefficient of each sample is defined as:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (4)$$

where, $a(i)$ [23] is the average length between the ith samples and all other points within the exact cluster as i and $b(i)$ is shortest average distance between ith example and any other cluster. The silhouette score is a number between -1 and 1, with higher values suggesting a better fit to the cluster. The silhouette score is 0.1089. It is calculated by comparing the similarity in clustering, the similarity of a data point is determined by comparing it with other data points in the same cluster. The similarity is then compared to data points in other clusters to determine the data point's assigned cluster. A score of 1 indicates that the data item is well-suited to its cluster, whereas a score of -1 means it is not. The silhouette score ranges from -1 to 1, with 1 representing that the data point is not well suited to its cluster.

7.2. Apriori Algorithm

Agrawal and Srikant's theoretical model is frequently employed (1994). This approach generates a set of association rules (p, q) that characterize the links between the components of the criticism pattern collection. The apriori method is an excellent tool for developing Boolean association rules by mining frequently occurring item sets. A "bottom-up" strategy is used by apriori by raising the frequency of subgroups [24]. In the field of machine learning, an association rule is an unsupervised learning strategy. Its purpose is to find interesting association rules that indicate linkages between variables in a dataset. It has two critical standards for evaluating the quality of a rule by measuring its strength: assurance and assistance [5]. The technique of searching for hidden patterns in data that can be detected but are difficult to categorize is known as data mining. Machine learning techniques are used in association rules mining to discover novel correlations or the co-occurrence of variables in large datasets. As evidenced by publications, numerous researchers have investigated association rule mining.

A variety of academic and commercial applications have employed association rules to analyze the co-occurrence of web sites. Association rule mining can be used to gather frequently visited websites into a server session. Association rules can serve as a stimulus for pre-fetching documents, which can decrease the perceived latency for users when loading a page from a remote website [9]. In web usage mining, association criteria are used to detect which pages are accessed together within a single network visit. The process of discovering links between variables is known as association rule learning. Although there are no direct hyperlinks connecting these pages, they are linked through relationships [6], [25].

The efficiency and outcome of association analysis is closely tied to the Apriori algorithm performance, it is a widely recognized process for association rules [18]. The Apriori method is an effective way for discovering association rules in frequently occurring item sets. It uses a "bottom-up" technique in which existing subgroups are expanded one item at a time. This has applications in fields such as research and scientific field analysis [19]. When a researcher adds an item to their field study list, the algorithm analyses its features and decides its support before providing it to the researcher. If the predicted support exceeds or equals the minimal support, the item-set will be listed as a frequent item-set. Apriori searches (*k+1*) datasets containing k items using an iterative layer-by-layer search process. A dataset is uncommon if it does not exceed the minimal support level, and a subset of it cannot be frequent [26-28]. The volume of data pouring in from all directions has increased, making it more difficult to organize data, uncover correlations and relationships, and assess large data sets.

As a result, multiple new technologies are being developed to extract useful data from vast amounts of text content using various text mining approaches. Different data mining and analysis techniques, including tools, algorithms, and strategies, perform differently on different data sets. Choosing the best algorithm for a variety of analytics tasks may be tough. The same business task can be accomplished using many algorithms, but each approach yields a distinct result, and some algorithms can generate a number of outputs [24]. An association rule is a machine learning unsupervised learning strategy. Machine learning techniques are used in association rules mining to discover novel correlations or the co-occurrence of variables in large datasets. It contains two key benchmarks for assessing a rule's quality by measuring its strength: confidence and support. The Apriori method is an effective way for discovering association rules in frequently occurring item sets. It uses a "bottom-up" technique in which existing subgroups are expanded one item at a time. When a researcher adds a field study item to their list, the algorithm checks the item's details. In association rule learning, the Apriori technique is used to extract common item sets from huge datasets. It operates by iteratively lowering the minimal support criterion until it finds the required number of item sets.

Algorithm 1. The Apriori algorithm might be applied to a text data column

Input: Text data, minimum support threshold Output: Frequent item sets Goal: Goal: An Apriori Algorithm is Applied to a Text Tata Column.
Start Read CSV File (output the k-means Algorithm) and chose the column to manipulate Convert the text data to a numerical representation (e.g., list of unique words or n-grams) Set the minimum support threshold for the item sets Find all item sets that have at least one support item by counting the occurrences of each item set in the numerical representation of the text data Remove any subsets of other item sets Repeat steps 3 and 4, lowering the minimum support threshold each time, until the required number of item sets is found or a minimum support threshold of 0 is reached The final item sets can then be analyzed or further processed. End

In these algorithms, we use the k-means.csv file (the previous algorithm's output) by way of involvement to the Apriori process, besides the results are shown in Figure 8. At the end saved these data frame as csv file (Apriori .csv) that showed in Figure 7b.

There are 7 Relation derived. frozenset({'BiotechnologyPlant', 'PlantBiotechnology'}) frozenset({'BiotechnologyPlant', 'PlantMolecular'}) frozenset({'GeneticsMolecular', 'MolecularBiology'}) frozenset({'MedicalMicrobiology', 'MicrobiologyMolecular'}) frozenset({'Molecularbiology', 'MicrobiologyMolecular'}) frozenset({'PlantBiotechnology', 'PlantMolecular'}) frozenset({'BiotechnologyPlant', 'PlantBiotechnology', 'PlantMolecular'})	(a) Relation Derived
pair= frozenset({'BiotechnologyPlant', 'PlantBiotechnology', 'PlantMolecular'}) Title1 Title2 Support Confidence Lift 0 BiotechnologyPlant PlantBiotechnology 0.0172 1.0 58.0 1 BiotechnologyPlant PlantMolecular 0.0172 1.0 58.0 2 GeneticsMolecular MolecularBiology 0.0258 1.0 16.571 3 MedicalMicrobiology MicrobiologyMolecular 0.0172 0.6666 19.333 4 Molecularbiology MicrobiologyMolecular 0.0172 0.5 19.333 5 PlantBiotechnology PlantMolecular 0.0172 1.0 58.0 6 BiotechnologyPlant PlantBiotechnology 0.0172 1.0 58.0	(b) Data Frame as Csv File (Apriori. Csv)

Figure 7. Apriori algorithms

Apriori is a transactional data mining algorithm that finds frequent item sets and association rules. It is commonly used for market basket analysis, but can also be applied to a dataset with text data. To apply the Apriori algorithm to a dataset containing text data, first tokenize the text and then represent the text data in a binary format. The algorithm's primary parameters are minimum support and leverage.

Minimum support is occurrences number of an item set that must occur for it to be considered a frequent item set. Leverage is the difference between the observed and expected levels of support for the rule if the items were independent. Conviction is a metric used to assess how strongly two items are associated. The parameters user choose will depend on the details of the dataset and the issue they're trying to solve. Users can find the ideal setup for their data by experimenting with various parameter settings and evaluating the outcomes. The Apriori algorithm is a well-liked technique for identifying frequent item sets in a dataset and deriving association rules from them. It is applied to a dataset saved in a pandas Data Frame by importing the apyori library and using it to do so. A specific column is chosen from the Data Frame, converted to a list, and then the Apriori method is applied to the list. The resulting association rules and four evaluation metrics are then printed. The rule is the actual association rule in the form of "item A -> item B", the percentage of interactions in the dataset that contain the item set is measured by this metric (The antecedent and the consequent segments of the rule).

More frequent item sets are indicated by a higher support. The lift is a statistic that evaluates the level of association between the elements in comparison to their independent. These evaluation metrics are commonly used to assess the power of association rules and find the most interesting or relevant rules for the dataset. It is an association rule that says "item A -> item B," which means that item A and item B are connected. The antecedent and consequent of the rule serve as its representation. Support is defined mathematically as the percentage of dataset transactions that contain the item set (The antecedent and the consequent segments of the rule). It is mathematically represented as follows [29]:

$$Support(A => B) = = \frac{\text{Number of transactions that contain both } A \text{ and } B}{\text{Total number of transactions}} \quad (6)$$

Confidence is described mathematically as the percentage of events that include both the antecedent (item A) and the consequent (item B). It has the following mathematical representation [29]:

$$Confidence(A => B) = = \frac{\text{Number of transactions that contain both } A \text{ and } B}{\text{Number of transactions that contain } A} \quad (7)$$

$$Confidence(A, B => C) = = \frac{\text{Number of transactions that contain both } A, B \text{ and } C}{\text{Number of transactions that contain } A, B} \quad (8)$$

Lift: According to mathematics, it is the degree of association between the elements in comparison to their independence. It is determined by dividing the confidence by the confidence that would be expected if the components were independent. It has the following mathematical representation [30]:

$$Lift(A => B) = \frac{Confidence(A => B)}{Support(B)} \quad (9)$$

where, A and B, respectively, are the rule's antecedent and consequent.

It is important to remember that the Support, Confidence, and Lift are interconnected. As an example, a rule with high support will also have high confidence and lift since a large proportion of the transactions will have both A and B. Together, these measures are used to assess an association rule's strength and choose which rules are most interesting or relevant to the dataset.

8. RESULT ANALYSIS

The primary purpose of a system is to cluster a researcher's interests using Association Rules for researchers who used Anbar University and the "uoanbar.iq" domain and had similar research interests. K-means clustering is a mathematical method for identifying patterns in numerical data. It uses unsupervised learning and does not have precise labels for the input points. The silhouette score is a well-known K-means evaluation tool for determining how close an object is to its own cluster in comparison to other clusters. The method also provides inertia ratings for different cluster counts in order to discover the ideal number of clusters for the dataset. The assumption is that if the clusters are compact and well defined, the distances between the cluster centers will be shorter. The output of the K-means clustering algorithm (CSV file) is input to Apriori algorithm The elements in the rule are rare in the dataset, appearing in only 2.25% of transactions. The confidence level of 1.0 for the rules indicates that all transactions that have the antecedent also have the consequent. The lift of a rule refers to the ratio of the expected level of support for the rule to the observed level of support with a silhouette score of 0.1089, the data is not well clustered, and the value is often close to zero, indicating that the clusters are not well separated. The evaluation step of the K-means clustering algorithm is shown in Figure 8.

```
+============================+
        WCSS       : 57.09178850573161
 Silhouette Score : 0.1221383785625252
+============================+
```

Figure 8. K-me|ans clusters evaluation

Items A and B have some association rules that the algorithm has discovered. Low support values (0.0224) and high confidence values describe these rules (8.9, 5.9333, 8.6666, 29.6666, and 44.5). The percentage of events in the dataset that contain the items in the rule is represented by the support values. With a support value of 0.0224, just 2.24% of the connections in this instance hold both item A besides item B. The degree of confidence that item B will be acquired if item A is ordered is indicated by the confidence values. The high confidence levels in this situation (8.9, 5.9333, 8.6666, 29.6666, and 44.5) imply that there is a substantial correlation between item A and item B and that choosing item A is a good predictor of purchasing item B. The text also makes reference to the lift values, which evaluate the degree of correlation between item A and item B. A lift value larger than 1 denotes the existence of a significant and unexpected association between items A and B, indicating that they are not independently related to one another.

Rule: BiotechnologyPlant -> PlantBiotechnology
Support: 0.017241379310344827
Confidence: 1.0
Lift: 58.0
==========
Rule: BiotechnologyPlant -> PlantMolecular
Support: 0.017241379310344827
Confidence: 1.0
Lift: 58.0
==========
Rule: GeneticsMolecular -> MolecularBiology
Support: 0.02586206896551724
Confidence: 1.0
Lift: 16.57142857142857
==========
Rule: MedicalMicrobiology -> MicrobiologyMolecular
Support: 0.017241379310344827
Confidence: 0.6666666666666666
Lift: 19.333333333333332
==========
Rule: Molecularbiology -> MicrobiologyMolecular
Support: 0.017241379310344827
Confidence: 0.5
Lift: 19.333333333333332
==========
Rule: PlantBiotechnology -> PlantMolecular
Support: 0.017241379310344827
Confidence: 1.0
Lift: 58.0
==========
Rule: BiotechnologyPlant -> PlantBiotechnology
Support: 0.017241379310344827
Confidence: 1.0
Lift: 58.0
==========

Figure 9. Apriori algorithms evaluation result

The Apriori algorithm has discovered association rules that denote that item A and item B are related. Each rule has a low value for support, 0.0224, and a high amount of confidence, 8.9, 5.9333, 8.6666, 29.6666, and 44.5. The lift values of the rules suggest a strong and unexpected relationship between the items in the rule, suggesting that they are not independent from one another. The values of support, confidence, and lift should be understood in the context of the particular use case because they depend on the particular dataset and the minimal support and confidence criteria employed in the algorithm. The paragraph concludes by stating that the basic support and confidence requirements established in the algorithm, as well as the specific use case, determine how support, confidence, and lift values should be interpreted. This means that the individual problem being solved and the data being studied should be carefully considered while evaluating the Apriori algorithm's outcomes.

8. CONCLUSIONS

Big data, which includes online transactions, emails, movies, audios, photos, sensor data, data from diverse sources such as logs, posts, search queries, health records, social networking activities, scientific data, sensors, mobile phones, and associated applications, plays a crucial role in contemporary academic and practical research. Google Scholar profiles are a goldmine of academic research material. Education Data Mining is an effective method for assessing the academic research environment and identifying useful ideas.

The high level of confidence suggests a strong relationship between the rule's components. The lift of the rule is 8.9, 5.933, 8.9, 29.6666, 29.6666, and 44.5, which is the ratio of the rule's expected and observed levels of support. Finally, this work demonstrates the capabilities of grouping k-means besides the Apriori algorithm cutting-edge facilitating a creation of collaborative research teams at Anbar University. Implementing the recommended recommendations will allow the university to improve overall research output and ranking while also contributing to the growth of knowledge in a wide range of fields.

9. FUTURE WORKS

We want to make our system more effective. To do this, we'll also use data analytics and machine learning techniques.

- In the future, for generate recommendations that are more personalized and accessible for users, we will integrate natural language processing (NLP) methods with recommendation algorithms.
- By adding stream processing tools like Apache Kafka and Apache Spark streaming, we will also focus on developing real-time recommendation systems.

REFERENCES

[1] N. Deepa, et al., "A Survey on Blockchain for Big Data: Approaches, Opportunities, and Future Directions", Journal of Future Generation Computer Systems, Vol. 131, pp. 209-226, February 2022.

[2] N.H.N. Abdullah, S. Sanusi, E. Savitri, "The Role and Implications of Big Data on Strategic Management Accounting Practices: A Case Study in a The Role and Implications of Big Data on Strategic Management Accounting Practices: A Case Study in a Malaysian Manufacturing Company", Management and Accounting Review, Vol. 21, No. 1, Malaysia, April 2022.

[3] A. Elomari, L. Hassouni, A. Maizate, "Deep Learning for Optimization of Chunks Placement on HADOOP/HDFS", International Journal on Technical and Physical Problems of Engineering (IJTPE), Issue 49, Vol. 13, No. 4, pp. 194-200, December 2021.

[4] M.H. Mohamed, M. Khafagy, M.H. Ibrahim, "Recommender Systems Challenges and Solutions Survey", International Conference on Innovative Trends in Computer Engineering (ITCE), Egypt, February 2019.

[5] L. Oughdir, A. Ibriz, K. Dahdouh, A. Dakkak, "Improving Online Education Using Big Data Technologies", Intech, The Role of Technology in Education, p. 13, March 2020.

[6] P.S. Sharma, D. Yadav, "Web Page Ranking Using Web Mining Techniques: A Comprehensive Survey", Graph-based Intelligence for Industrial Internet-of-Things, Vol. 1, No. c, 2022.

[7] B. Jabir, N. Falih, "Big Data Analytics Opportunities and Challenges for the Smart Enterprise", International Journal on Technical and Physical Problems of Engineering (IJTPE), Issue 47, Vol. 13, No. 2, pp. 20-26, June 2021.

[8] V. Singh, "The 4 Differences you Should Know", https://2u.pw/93N9A.

[9] N.P. Dharshinni, F. Azmi, I. Fawwaz, A.M. Husein, S.D. Siregar, "Analysis of Accuracy K-Means and Apriori Algorithms for Patient Data Clusters", J. Phys. Conf. Ser., Vol. 1230, No. 1, 2019.
[10] P. Agrawal, H. Tolani, A. Memon, S. Oza, "Generation of Machine Learning Dataset from RDBMS for Clustering and Association Analysis", Gradiva, Vol. 8, No. 5, pp. 868-874, May 2020.
[11] L. Kusak, F.B. Unel, A. Alptekin, M.O. Celik, M. Yakar, "Apriori Association Rule and K-Means Clustering Algorithms for Interpretation of Pre-Event Landslide Areas and Landslide Inventory Mapping", Open Geosci., Vol. 13, No. 1, pp. 1226-1244, 2021.
[12] S.O. Ajmi Al Shuwaili, S. Obied Redywi, M.A. Naser, "A Hybrid Approach for Text Clustering", Mater. Today, Proc., June 2021.
[13] Y. Zhou, "Application of K -Means Clustering Algorithm in Energy Data Analysis", Wirel. Commun. Mob. Comput., Vol. 2023, September 2022.
[14] L. Yu, Y. Li, Q. Zeng, Y. Sun, Y. Bian, W. He, "Summary of Web Crawler Technology Research", J. Phys. Conf. Ser., Vol. 1449, No. 1, 2020.
[15] S.S. Bhamare, "Web Crawler: A Survey", International Journal of Innovative Science and Research Technology, Vol. 7, No. 8, pp. 613-615, 2022.
[16] D.H. Hameed, S.H. Hashem, "Web Pages Retrieval by Using Proposed Focused Crawler", J. Al Nahrain Univ., Vol. 19, No. 2, pp. 154-164, 2016.
[17] A.S. Bale, "Web Scraping Approaches and their Performance on Modern Websites", The 3rd Int. Conf. Electron. Sustain. Commun. Syst. (ICESC 2022), pp. 956-959, India, September 2022.
[18] C. Kenny, "Web Scraping vs Web Crawling", Zyte (formerly Scrapinghub) #1 Web Scraping Service, 2022. www.zyte.com/learn/difference-between-web-scraping-and-web-crawling/
[19] M.K. Chyad, M.M. Hamad, "A Proposed Movie Recommender System to Solve Sparsity, Cold Start and Diversity Problems using Clustering Algorithms", Solid State Technol., Vol. Solid Stat, No. 6, 2020.
[20] A.E. Ezugwu, A.K. Shukla, M.B. Agbaje, O.N. Oyelade, A. Jose Garcia, J.O. Agushaka, "Automatic Clustering Algorithms: A Systematic Review and Bibliometric Analysis of Relevant Literature", Neural Computing and Applications, Vol. 33, No. 11, 2021.
[21] A. shirazy, A. Hezarkhani, A. Shirazi, S. Khakmardan, R. Rooki, "K-Means Clustering and General Regression Neural Network Methods for Copper Mineralization probability in Char-Farsakh, Iran", Geol. Bull. Turkey, Vol. 64, No. 2021, pp. 79-92, 2021.
[22] D.N. Vasundara, S. Naini, N. Venkata Sailaja, S. Yeruva, "Classification of Skin Diseases Using Ensemble Method BT", The Second International Conference on Advances in Computer Engineering and Communication Systems, 2022.
[23] M. Shutaywi, N.N. Kachouie, "Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering", Machine/Statistical Learning and Modeling with Potential Applications in Entropy, Information Theory, and Artificial Intelligence, Vol. 23, No. 6, p. 1-17, 2021.
[24] S.R. Joseph, H. Hlomani, K. Letsholo, "Data Mining Algorithms: An Overview", Int. J. Comput. Technol., Vol. 15, No. 6, pp. 6806-6813, 2016.
[25] S.S.E. Alqallaf, W.M. Medhat, T.A. El Shishtawy, "A Hybrid Recommender Framework for Selecting a Course Reference Books", J. Theor. Appl. Inf. Technol., Vol. 100, No. 4, pp. 1004-1014, 2022.
[26] Y. Xiang, C. Shuai, Y. Li, Y. Zhang, "Information Reconstruction of Student Management Work Based on Association Rules Mining", Computational Intelligence, Internet of Things and Artificial Intelligence-Based Smart and Sustainable Healthcare Systems, Vol. 2022, Article ID 2318515, 2022.
[27] N. Hasimah, I. Teo, "Market Basket Analysis Using Apriori Algorithm: Grocery Items Market Basket Analysis Using Apriori Algorithm", no. November, 2022.
[28] Y. Li, "The Teaching Evaluation System of College Students under the Background of Big Data", International Journal of Frontiers in Engineering Technology, Vol. 4, No. 6, pp. 56-61, 2022.
[29] N. Domadiya, U.P. Rao, "ElGamal Homomorphic Encryption-Based Privacy Preserving Association Rule Mining on Horizontally Partitioned Healthcare Data", J. Inst. Eng. Ser. B, Vol. 103, No. 3, pp. 817-830, 2022.
[30] A.A. Majid, C.S. Pramudyo, "Association Rules for Layout Design and Promotion Strategy", The Second Asia Pacific International Conference on Industrial Engineering and Operations Management Surakarta, pp. 112-121, Indonesia, September 2021.

BIOGRAPHIES

Name: **Doreyed**
Middle Name: **Muhammed Ahmed**
Surname: **Al Kerboly**
Birthday: 04.09.1981
Birthplace: Anbar, Iraq
Bachelor: Computer Science, College of Computer Science and Information Technology, University of Anbar, Anbar, Iraq, 2003
Master: Computer Science, University of Al Bayt, Mafraq, Jordan, 2015
Scientific Publication: 3 Papers
The Last Scientific Position: Assist. Teacher, University of Anbar, Anbar, Iraq, Since 2016
Research Interests: Big Data, Web Minin, Recommender Systems

Name: **Murtadha**
Middle Name: **Mohammed**
Surname: **Hamad**
Birthday: 05.04.1961
Birthplace: Anbar, Iraq
Bachelor: Computer Science, College of College of Sciences, University of Mosul, Mosul, Iraq, 1983
Master: Computer Science, College of Sciences, University of Baghdad, Baghdad Iraq, 1991
Doctorate: Computer Science, Department of Computer

Science, University of Technology, Baghdad, Iraq, 2004
The Last Scientific Position: Prof., Computer Science, College of Computer Science and Information Technology, University of Anbar, Anbar, Iraq, Since 2004
Research Interests: Data Warehouse, Data Mining, and Big Data
Scientific Publication: 60 Papers, 10 Project, 34 Thesis and Dissertation
Scientific memberships: Member of Ministerial Committee for Development of Information Systems Curricula, Iraqi Ministry of Education - Member of Science and Technology Journal Committee, University of Wasit, Al Kut, Iraq - Chairman of Promotion Committee, College of Computer Science and Information Technology, University of Anbar, Anbar, Iraq

Name: **Omar**
Middle Name: **Abdulrahman**
Surname: **Dawood**
Birthday: 30.07.1986
Birthplace: Anbar, Iraq
Bachelor: Computer Science, College of Computer Science and Information Technology, University of Anbar, Anbar, Iraq, 2008
Master: Computer Science, College of Computer Science and Information Technology, University of Anbar, Iraq, 2011
Doctorate: Computer Science, Department of Computer Science, University of Technology, Baghdad, Iraq, 2015
Scientific Publication: 35 Papers, 7 Thesis
The Last Scientific Position: Assist. Prof, Computer Science, College of Computer Science and Information Technology, University of Anbar, Anbar Iraq, Since 2019
Research Interests: Network Security, Cryptography, Number Theory, Cipher Design