Journal of Advanced Sciences and Nanotechnology

Journal homepage: https://joasnt.com

Intelligent Information Retrieval (IIR) Based Clustering Algorithms: State-of-The-Art

Azal Qussay Saeed [1,*], Essa Ibrahim Essa [1]

[1] College of Computer Science and Information Technology, University of Kirkuk, Iraq.

Article information

Article history:
Received: March, 24, 2023
Accepted: May, 10, 2023
Available online: Sept., 16, 2023

Keywords:
Intelligent Information Retrieval System,
Clustering algorithms,
Big data,
Document,
Query.

Corresponding Author:
Azal Qussay Saeed
stch21m002@uokirkuk.edu.iq

Citation : Saeed, A. Q. ., & Essa, E. I. Intelligent Information Retrieval (IIR) Based Clustering Algorithms: State-of-The-Art. Journal of Advanced Sciences and Nanotechnology, 2(2), 215–225.

Abstract

The great spread of the Internet and the steadily increasing volume of data constituted a great motivation for researchers to explore new and efficient ways to manage, access, and benefit from them in a way that ensures the reduction of time, effort, and cost. As Intelligent information retrieval techniques (IIR) contributed to this end as one of the most important fields of information science that are concerned with searching, indexing, and retrieving the required information. To achieve easy and fast access to the enormous data should be divided into groups containing the same objects. Clustering is a useful data mining tool for dealing with IIR systems that can be clustered utilizing any of the clustering algorithms. In order to have a deeper understanding of these topics and to know the latest findings of the researchers in this regard, this research paper reviews the literature on IIR systems and their related clustering algorithms for the past eight years. Intending to form an optimal retrieval system that uses the best clustering algorithm to optimally retrieve the data required by the query according to the user needs.

DOI : https://doi.org/10.55945/joasnt.2023.2.2.215-225,
ISSN: 2791-0903/

1. Introduction

Information intelligence retrieval is an important subject and introduces a framework for huge data queries in the following subsection we put the main parameters in this area.

A. Intelligent information retrieval

In our time, the volume and variety of information that is available have grown as a result of the World Wide Web's (WWW) development. Currently, the Internet is a massive information archive, stored in a structured manner like databases and unstructured manners like HTML and text files, so the question now is how to properly enhance the accessibility of this information. Users need the help of systems designed to find documents that fulfill their unique requirements such a system is renowned as an IIR system, it is the process of looking in documents for information, seeking documents directly, seeking metadata that defines documents, or looking for

text, images, sound, or data in databases [1].

B. Intelligent information retrieval architecture

As shown in Figure 1 the main IIR architecture and working method of such systems can be summarized in this regular scenario when a user types a search query into a search engine, and the search engine returns results as a list of documents ranked by relevance, this process consists of two essential functions indexing and querying, during the indexing phase, the system creates structures for the documents so that their contents become searchable, the querying phase uses those indexing structures with the retrieval algorithms on the user's query to find and produce the related documents in a ranked list., the indexing phase usually includes two processes: (text transformation and index creation). The text transformation method is to convert documents into terms by multi-preprocessing steps and create indexed terms. While the (query transformation and ranking) are two essential components of the querying process, because a raw query may not fully represent the linguistic variety of the information demands, query modification is critical for ultimate retrieval performance [2].

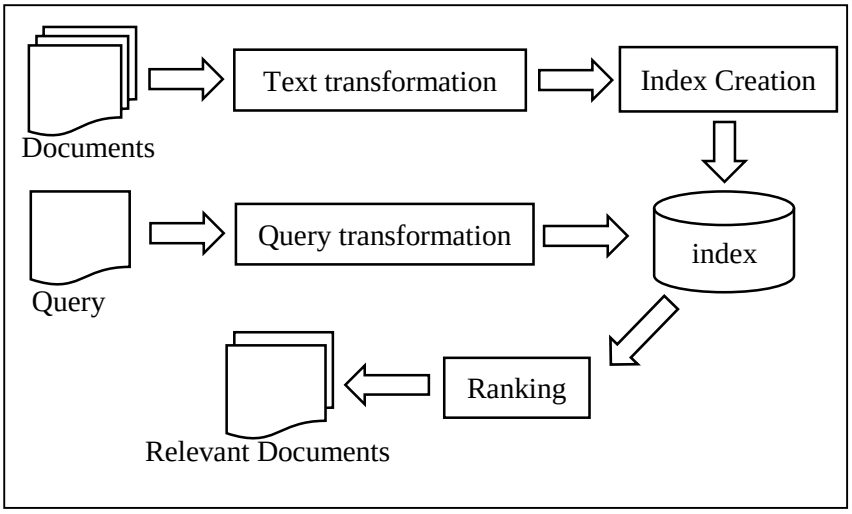

Figure 1. IIR architecture [2].

C. Intelligent information retrieval models

The following are the most important IIR models [1]:
Extended Boolean Model.
Boolean Matching/Model.
Vector Space Model (VSM).
Fuzzy Set Model.
Probabilistic Modeling.
Latent Semantic Indexing.
Neural Networks.

D. Evaluation of Intelligent information retrieval systems

There are several methods for evaluating the efficiency, effectiveness and quality of IIR systems. The relevance of the document to the user's needs is traditionally the most important factor in determining how effective a retrieval is, this property can be measured using a variety of criteria, the most important and common being ***Precision***, ***recall***, and ***F-measure***.

Precision is the percentage of the relevant documents the IIR system retrieves in response to a query and the overall number of documents retrieved.

$$\textbf{Precision} = \frac{\textbf{Relevant} \cap \textbf{Retrieved}}{\textbf{Retrieved}} \quad \text{..... (1)}$$

Recall is the ratio of the number of documents that were found to be relevant to the query to the total number of documents that were found to be relevant to the query in the collection of documents.

$$\textbf{Recall} = \frac{\textbf{Relevant} \cap \textbf{Retrieved}}{\textbf{Relevant}} \quad \text{..... (2)}$$

F-measure is also known as F-score, the F-measure is a harmonic range of Precision and recall.

$$\textbf{F-measure} = \frac{2 * \textbf{Precision} * \textbf{Recall}}{\textbf{Precision} + \textbf{Recall}} \quad \text{..... (3)}$$

Precision depends on the retrieved documents, whereas recall is dependent on the relevant documents in the collection. And high recall is the ability to identify all of the retrieved records as relevant, whereas high precision is just retrieved records [3, 4].

E. Clustering algorithms

The process of grouping data objects into a number of clusters based on their similarities and differences is known as clustering. The ability to identify a group of items based on their similarities and produce a pattern from an unsupervised (unlabeled) dataset is regarded as one of the most important techniques. The primary goal is to allocate data points, such as that there should be a large inter cluster distance and a small intra cluster distance [5]. The fundamental benefit of clustering is that it does not require prior data knowledge. Clustering algorithms are widely used in data science and data mining, where the goal is to group information that shares common characteristics and determine the ideal number of clusters [6].

F. Common Clustering Algorithms

Hierarchical and partitioned clustering are the two main subtypes of clustering. While partitioning clustering creates data partitions by using any optimal criterion, hierarchical clustering discovers the clusters by splitting the data in either a top-down or bottom-up method in a recursive manner [7]. Table 1 shows the most prominent clustering algorithms.

Table 1. The benefits and drawbacks of clustering algorithms.

Clustering Algorithm	Work papers	Advantages	Disadvantages
Partitional clustering	[8, 9]	➢ It needs to specify the number of clusters from the beginning of the work, so it is suitable for big datasets. ➢ Very helpful when the clusters are of convex shape having a similar size.	➢ Only appropriate for spheroid separation. ➢ The problem of optimization (Not guaranteed lowest value)
Hierarchical clustering	[10, 11]	➢ Because of the disability predicting the number of clusters from the start it is suitable for small data set . ➢ It partitioned a dataset into multiple levels of partitioning.	➢ the problem of how to find the optimal number of clusters. ➢ Not valid for big datasets.
Density-Based clustering	[8, 12]	➢ They are very helpful for mining big data sets because they can facilely distinguish noise. ➢ It can address clusters of arbitrary shape.	➢ It need a lot of calculation ➢ The main drawback is the a priori definition density function.
Grid-Based clustering	[13, 14]	➢ Quick processing time. ➢ Efficient process spatial data.	➢ less activeness when used with noisy and complex topologies datasets.
Spectral clustering	[15, 16]	➢ Deal with complex data and creates	➢ massive space and time

	arbitrary shapes clusters. ➢ Suitable for clustering graph matrix.	complexity. ➢ Low efficient in large-scale issues.

G. Clustering-based intelligent information retrieval

The first process in the procedure for document IIR is to scan every document in a collection and then score each document according to how relevant it is to the user's query. The much more relevant documents are then chosen (those that scored the highest) using a ranking function, and they are displayed to the user. It may take a while to respond to queries on huge document collections when using this method. To bypass this obstacle, clustering-based algorithms have been used to improve the efficiency of document IIR. These methods all revolve around the preprocessing step, which creates clusters out of related documents from a collection of documents. Then, to answer a query, the system chooses the cluster that is most significant to the query and limits the search to only documents in those clusters. Clustering methodologies can be much faster than traditional document IIR methods because they do not need to scan an entire set of documents to find an answer to a question [17].

2. Literature Review

Numerous academics have made significant contributions to the field of IIR and clustering algorithms used on it in earlier scientific studies. In the sections that follow, we will describe some of their work, highlight its advantages and drawbacks, and explain how it might be applied. Two sections of the review procedure were used, one for the IIR and the other for clustering algorithms:

A. Intelligent Information Retrieval (IIR)

The IIR system should retrieve documents for users in a ranked list according to the degree of relevance to the user's information needs. In order to rank the retrieved context information affording to information relevance and to improve the relevance of retrieved documents. Agbele et al. [18] proposed an adaptive Document Ranking Optimization (DROPT) algorithm for IIR in an Internet-based or selected databases environment. The DROPT method adapts itself to single user information requests based on environment and search context. The novel DROP approaches include appropriate methods for identifying user characteristics and exploring user interests to improve the performance of the IIR system. Evaluation of DROPT technology by expert users shows performance results improvement using ‘precision at position n’ over the chosen baseline algorithms methods.

Ahmed [19] design a platform that organizes, retrieves and discovers information in a helpful way from the information space in a more obvious and intelligently way this achieve by semantic based IIR that combines current information retrieval system and semantic web technologies such as the xml and ontology description of the data that improve performance of IIR in spite of the past ways of the IIR such as keyword based method.

Siregar et al. [20] compare the most important term weighting techniques used in IIR systems and also suggest a new method to improve them by introducing a global weight value that converts log 10 to log 2 and increases the global term weighting IDFP by 1. The results showed an improvement in the values of precision and NAIP.

Gahlawat et al. [21] concluded in their research that the IIR system's ability to effectively retrieve information is affected by a variety of factors including (User behavior, User Efficacy, User cognitive skills, and System Reliability), and system's productivity can be boosted by using these factors. These factors were combined to generate User Characteristic Data (UCD), which would be distinct for each user. The system will be using customer feedback to refresh the UCD as the user's action may alter over time, the model will use this data to derive the ideal user outcome.

Khin & Yee [22] recommend a tried-and-true IIR system that uses a Web Query Classification Algorithm (WQCA) to obtain documents that are more pertinent to the user's needs. This system can categorize the taxonomies of web searches (query attributes) and analyze confusing domain phrases. Traditional IIR systems' inadequacy of semantic connection can be fixed by the proposed WQCA. A group of pertinent documents is presented by the suggested search engine using conceptual retrieval. There are two drawbacks to this system. This system is unable to appropriately identify a user query if it is overlong and has spelling problems. This system needs to train several categories and test for spelling issues in the preprocessing stage in order to get around these constraints.

Alahmadi [23] analyzes and discusses the architecture and ranking algorithms of search engines which are

utilized for study examples of how IIR systems from disparate databases can be used to deliver the right results. On the most widely used search engines, Yahoo, Google, and Bing, query experiments were conducted. To evaluate the performance of the selected engines recall and precision measures have been used, Consequently, it was determined that Yahoo outperformed Google and Bing in terms of precision value and delivered the maximum relevant documents relative to users' assumptions.

Aygun & Benesova [24] are emphasized the important aspects of multimedia retrieval methods. The study presents a retrieval task called "retrieve every media (REM)" that looks at whether a query to recover every item of multimedia data in the system exists. and suggests page-oriented precision as a substitute performance indicator for multimedia information systems.

Azzopardi et al. [25] submit an implement (“cwl_eval”) that combines a number of metrics used to evaluate the IIR systems. This paper explains the C/W/L framework measurements are defined by a unique function from which a number of associated measurements can be derived, such as the expected usefulness of each item, expected whole utility, the anticipated cost of each unit, anticipated overall cost, and anticipated depth.

A novel perspective of analyzing the fairness problems in IIR was suggested by Gao & Shah [26] proffer a framework that first describes the solution space on any given dataset by estimating optimal solution values and theoretical boundaries, and then the proposed model used the solution space to ease different of analysis and decision making which would else be extremely time and resource consuming.

Wahyudi et al. [27] designed a system for retrieving documents or files in the (JavaScript Notation format) or (JSON). According to the findings of this study, The vast number of query words has no impact on this retrieval system's ability to quickly seek information from the contents of JSON files, the simplicity of word searches because of the keyword-stemming procedure, and The output of the Vector Space Model approach, known as Cosine Similarity, is utilized as a benchmark when ranking texts.

In order to obtain faster and more exact searches and comments on the data files of most concern to the user. Li et al. [28] applied the machine learning of the primary model employing unsupervised learning for dispersed leaf nodes to the search engine structure with the distributed cluster designing. Then the search engine's data association source is created utilizing a new inverted index technique based on the class label. And after that, there is an efficient connection between the unstructured data's basic information, labelled information, and the final tally algorithm's outcome. As a result, it builds an index list of data file grouping that allows for efficient unstructured data categorization and querying in massive data streams.

Khalifi et al. [29] they studied how to increase the efficiency of IIR technology and discussed the topic through five elements: (query expansions, advanced text processing techniques, machine learning classification, user preference optimization, and dataset organization). The researchers studied the effect of each element on a number of stages of the IIR system like (merging stemming, terms classification, vector representation, documents clustering, terms weighting, query expansions, query classification, and finally feedback consideration), this study has incorporated IIR system enhancements.

Vector space-based model is one of the generally popular algorithms for IIR systems, document classification, and other text mining applications, at the heart of these models, are measures of distance, or measures of similarity. Eminagaoglu [30] proposed a novel similarity measure that can be effectively used for vector space models and related algorithms such as Rocchio and k-nearest neighbours (k-NN) as well as some clustering algorithms such as K-means. The new measure might be used within all appropriate algorithms, models, and methods for document classification text mining, and relevant knowledge management systems.

For rising retrieval performances and improving Arabic Stemmer, Alnaied et al. [31] suggested a new approach named Arabic Morphology Information Retrieval (AMIR) In order to produce or retrieve stems, many criteria concerning the connection between Arabic letters must be used in order to locate the root or stem of the corresponding words utilized as indexing term. The proposed approach AMIR was compared against FARASA, LUCENE, and no-stem approaches. The results of the precision mean average show that AMIR stem algorithm outperforms others.

Hammache & Boughanem [32] have spoken about a new language model (LM) for IIR that includes the word position. The major concept is to give the phrases near the beginning of a text more weight. There are two suggested methods for term position: (PosFirst) for a term's first appearance in a document and (PosAll) for all instances of a term within. The phrase "position-based document model" refers to the two elements in a formal way (Position LM), According to the experiments, the Position LM by itself is more reliable and efficient than the unigram LM.

Joby [33] retrieved information from the large set of data available on the internet by using the latent-SA. The method employs semantic-based analysis instead of retrieving documents based on the keywords. By using the WordNet, the proposed mechanism found the synonym and the thesaurus. The results of the experiments showed the effectiveness of the system in retrieving relevant information from the Internet accurately for a wide range of queries.

To improve IIR systems' efficiency, Neji et al. [34] studied the linguistic meaning of the terms and offer a hybrid ranking documents scheme. The suggested model uses WordNet ontology to extract concepts to effectively represent documents and provides a formula to assess the level of similarity between the query and the terms in the document in order to rank the documents.

Jain et al. [35] prepared IIR systems using a fuzzy ontology for the expansion of query, In this study, a concepts dictionary is designed for a particular domain, external ontology, and Fuzzy membership is allocated utilizing ConceptNet (Global Ontology). The query is expanded by using the suggested fuzzy membership among the various concepts, and the most related concepts in the particular domain could identify. a semantic web can be built to deal with the context of the query by merging the query expansion method with the existing search engines. after retrieving a number of documents on a search engine the proposed model was evaluated by different parameters such as (MRR, precision@10, MAP), and the results show that the number of documents is reduced with the support of query expansion about 1/1000 documents numbers for every query on various web search engines.

In order to create an IIR system for retrieved sets of documents published in the English language, recommended Joe & Jack [36] use a cluster-based IIR system. The system operation went through two key stages. First, the clustering creation and analysis process is performed by grouping similar documents into clusters. Second, the clusters are ranked according to the user's query by the IIR system in order to obtain the most documents related to the query. The results are then evaluated using assessment criteria. Precision and recall (P@5, P@10) of the obtained results P@5 equaled 0.660, while P@10 equaled 0.655.

Ye [37] conducted a comprehensive study of current IIR systems as a framework that includes organizing and indexing information, query, retrieval models, and personal techniques. And show the prominent weaknesses in these systems, in order to know how to deal with networks as an information resource. Also proposed a secondary filtering model. By binding the (attribute, content, and structure) three-query filtering approach, carrying out a Bayesian for improved classification, then outputting the final results the ID3 learning is used. The experimental findings reveal that this study approach has a greater retrieval performance of 80% and a retrieval time of 1.9 minutes shorter.

Lechtenberg et al. [38] proposed a new approach to develop the IIR function of citation databases and scientific abstracts by utilizing a query-by-documents method, applied and examined on the Scopus® database employing two example studies. This study contributed to the inclusion of the Monte Carlo sampling process at the stage of creating a set of queries, which came to two results: (i) the interference of human skill (costly resource) decreased, and (ii) keep away from human bias. The outcomes of the retrieval process are fast and fine, such as the elevating values of recall and actually relevant scientific papers clarify by the reference studies.

To bridge the gap between queries and documents in different languages Zhang et al. [39] introduce a Hierarchical Knowledge Enhancement (HIKE) model for the Cross-Lingual Information Retrieval (CLIR) task. HIKE presented an external multilingual knowledge graph (KG) into the CLIR task and is supplied with a hierarchical information fusion mechanism to gain all advantage of the KG information. This model can integrate the knowledge level with the KG information in each language, also the language-level fusion joins the information from both source and destination languages.

The study presented by Kumari & Ahlawat [40] regarding a model for IIR that might successfully retrieve valuable information from datasets related to breast cancer and use it to create a classification model. To aid medical professionals in establishing accurate diagnosis when there is uncertainty or insufficient data in one system, the suggested support system's hybrid model can analyze both structured (Fine Needle Aspiration) FNA data and unstructured mammography observations. Educated and examined by a K-NN algorithm using K as the ideal value, and semi-structured mammography pattern records in a single system are trained and verified using an SGD classifier by fine-tuning the parameter using the grid-search method. The model aids medical professionals by providing prompt, accurate, and trustworthy advice that may be used during the crucial decision-making stage and enhance patient survival rates and life quality.

Erbacher et al. [41] focused in their article on a new area of IIR called conversational information retrieval (CIR) crossing reacting IIR with dialogue systems to meet open-domain information demands, and most importantly establishing a practice and assessment framework for user-centered methodologies. Though human evaluation is ideal, models based on reinforcement or deep learning cannot handle it since it takes too much time. The goal of this work is to compile the most recent user modeling and simulation techniques for a variety of information access jobs in order to provide opportunities for the adoption of CIR systems.

B. Clustering Algorithms

Arora et al. [42] presented the outcomes of both K-means and K-Medoids clustering algorithms with regard to the number of clusters created and distance metric. The results of the comparison showed that K-Medoids perform much better than K-Means in terms of cluster head selection time and spatial complexity of cluster overlapping. The dataset's results also demonstrate that K-Medoids are superior to K-Means in all respects, including execution speed, no sensitivity to outliers, and noise reduction, while on the downside that they are more complicated.

One of the most useful algorithms in data mining is the K-means algorithm, the result of this clustering method depends excessively on its initial cluster centers. Xiong et al. [43] proposes an enhanced K-means text clustering technique through the optimization of the first cluster centers, this algorithm's core idea is to choose the first objects cluster centers based on the intensity parameter of the data, which guarantees the initial cluster centers' rationalism. The suggested approach can produce improved text clustering results by largely eliminating the K-means algorithm's sensitivity to the center of the primary cluster.

Due to the text's high dimensionality, conventional clustering techniques may not produce satisfactory results. Also, text clustering written in the Arabic language is a taxing task because of a lot of reasons. Consequently, in order to improve the accuracy of clusters for Arabic text documents Alhawarat & Hegazi [44] suggest a study using a hybrid approach for Arabic text employing topic modeling and clustering technologies. For the purpose of comparing the topic modeling/k-means combination technique with the k-means clustering algorithm, this research utilized a news textual dataset with five different versions. The results of this research indicate that, in comparison to other studies of a similar kind, adding normalization to the VSM improves the outcomes of the straightforward K-means algorithm with the straightforward Euclidean measurement.

Passarat & Shedge [45] they were touched on how to increase the effectiveness and reduce the complexity of the IIR process by employing the process of clustering documents in classifying the actual data. The study indicated the possibility of extracting the most prominent features through the approach of modelling the subject and named entity recognition, this process improves the time complexity of the clustering algorithms as it takes into account the features of the term rather than the entire document. K-prototype, which focuses on the clustering features and takes into account the frequency of mismatches, is more successful than fuzzy clustering.

Yuan & Yang [46] clearly tested the K-means clustering algorithm that is used widely due to it being a plain algorithm and rapid convergence. Although the value of k is determined from the beginning of the algorithm, it greatly affects the final convergence result, to bypass this, researchers Analyse and apply four K-value algorithms namely (Silhouette Coefficient, Elbow Method, Gap Statistic, and Canopy) these four algorithms capable of meet the request for big and complex data sets. The experiment confirmed that the Canopy algorithm is the best choice.

Li et al [47] suggest a creative algorithm for quickly detecting the number of clusters K and the initial cluster centers determination method. Cluster centers are those data points with smaller radius thresholds, and far away from each other, with higher density. This algorithm improved the clustering process effectively by employing MNN (M nearest neighbors), distance, and density to identify the initial cluster centers. The algorithm was validated because it proved highly effective during experiments, in spite of the fact that the process of creating the distance matrix is very time consuming and has the randomness to choose the MNN distance these costs are practically affordable. Besides, no matter how much M is obtained, the final initial center points do not change much, and more stable and high-quality assembly results can be reached.

Gocken & Yaktubay [48] investigate the impact of utilizing various clustering methods in the genetic algorithm's (GA) in the stage of generating the primary population and organizing a multi-objective genetic algorithm approach for the Vehicle Routing Problem with Time Windows (VRPTW) resolution. Customers grouped into feasible clusters By using K-means, Centroid-based heuristic, DBSCAN, and SNN clustering algorithms. Then feasible paths are created for each cluster, which will be taken as the initial population and GA

used for the optimization. To find out the extent of benefit from the use of clustering in the process of creating the initial population of the GA algorithm, and for comparison, five algorithms are used; i.e. CBased, SNN, K-means, RN, and, DBSCAN. The result shows that cluster partitioning techniques are better suited for VRPTW solution method than density-based clustering algorithms.

The quantum clustering technique that described by Bhagawati [49] might aid in improved information administration and retrieval. The development of different sentence clusters for certain acquaintances may be accomplished using this innovative clustering method, together with the coordination of content analysis, as one of the fundamental strategies in quantum computation. Additionally, it would aid in the construction of a quicker browsing structure and a faster digitizing process in the quick-paced world of today.

Alonso et al. [50] proposed three hierarchical clustering techniques that can capture various time series properties. These depend on a collection of "dissimilarity" measurements that were calculated over several characteristics, including quantile auto-covariance, and simple and partial autocorrelations. This model is useful for real-world implementation with huge datasets of time series, such as those received from smart meters, thousands of power usage time series are used to assess how well these clustering methods function. The findings demonstrate the capability of obtaining highly representative clusters that capture a variety of power load exhaustion patterns and establish the degree of efficacy of each model component.

Sinaga & Yang [51] proposed a new Unsupervised k-means (U-k-means) that finds the optimal number of clusters automatically with no need for parameter estimation or initialization, the initial number of clusters in the proposed algorithm is determined by the number of data points. According to the structure of data, the U-k-means method discards further clusters during iterations, and after that, a suitable number of clusters may be discovered automatically. The results show the superiority of the U-k-means clustering approach when comparing the U-K means algorithm with most existing algorithms actually demonstrate on many artificial and real data sets.

For the purpose of helping researchers to obtain finding compelling research articles relevant to their area of expertise, Jalal & Ali [52] presented a categorization way for clustering scientific publications, to enhance and automate the task of organizing and classifying scientific papers according topics by using web data mining techniques. After conducting the experiments, the cosine similarity technique produced the finding that more than 96% of scientific publications could be categorized within the same range; this finding was supported by recall metrics and precision metrics.

To overcome the difficulty of determining clustering amounts, and the accuracy of initializing the clustering center for the K-mean clustering procedure effectively. Ran et al. [53] applied The noise-based K-means clustering algorithm has been to catch metropolitan hotspots in large cities from all over the world. The results were evaluated after the completion of clustering by using DB index, PBM index, SC, and SSE, and the investigational marks of each assessment standard were statistically analyzed by Wilcoxon rank sum testing to calculate the noteworthy variances for the urban hotspot distribution for each clustering algorithm, the results show that the proposed clustering approach achieved superior optimal results for metropolitan hotspots over the FCM, K-means, and K-means plus procedures, also some shortcomings.

Shuja et al. [54] provided a hierarchical clustering for efficient use of the Social Internet of Things (SIoT). In two steps the proposed framework can clusters geo-textual data. first, all data points in the dataset are clustered depending on geographical positions only by utilizing Euclidean distance. second, measuring the similarity of the text based on unsupervised neural network training of Word-to-Vec to build the geo-textual clusters in the geographical clusters. The two-step hierarchical clustering reduced dimensionality by almost 54% memory and 52% time consumption when compared to hybrid clustering.

Almazroi & Atwa [55] presented a semi-supervised clustering algorithm for clustering multi-density data (SSMD). By examining the statistical properties for their variance with respect to density, the proposed algorithm divides the dataset into groups of distinct intensity levels and then grows the groups using defined active pairwise limitations. By the use of real datasets of distinct dimensions and volumes Experiments were conducted to evaluate the clustering Performance and execution time on and the results showed a better clustering performance achieved by SSMD compared to The current state-of-the-art, SSMD pointedly Reduced execution time.

Awad & Hamad [56] discussed in their paper Many related clustering techniques to give a general understanding of the clustering methodology to implement the suggested technique the k-means base clustering algorithm has been run on a machine-learning-based processor to enhance the clustering's effectiveness, k-means performs better with numeric values than with categorical in terms of efficiency.

Sohrabi et al. [57] discussed the microarray technique, which enables the analysis of genes that determined

breast cancer. Five bi-clustering methodologies including (Fabia, Xmotif, Cheng & Church (CC) , PLAID (PL), and Bimax) were assessed in order to uncover diverse gene subsets connected to distinct forms of breast cancer. This was done through the use of inferential and descriptive statistical analysis. The Results showed the success of all methodologies, excluding CC, finding in the data good bi-clusters, and assured that the PL model was much preferable to that of other bi-clusters structure models and it is convenient for mining in the gene expression data.

3. Conclusion

It has become vital to create and add new mechanisms in the science of IIR due to the expansion of the volume of data, the growing need for its retrieval in a timely manner, and its entry into all spheres of life. In order to assist us in developing an IIR system based on clustering algorithms, we have collated a prior study pertaining to these schemes in this review article. This research seeks to increase our knowledge of the IIR system's efficiency. Many academics interested in the various clustering strategies used in document clusters have examined and highlighted the advantages and disadvantages of each methodology in order to retrieve the current clustering relevant to the query provided by the user. The study came to the conclusion that clustering is used in IIR for many different purposes, such as document grouping, query expansion, document indexing, and search result visualization, and that text clustering in document browsing serves the primary purpose of describing and summarizing a larger set of documents. In order to achieve better results, our proposal for future work is to perform more studies relating to utilizing hybrid clustering algorithms in the IIR system.

References

[1] V. K. Singh and V. K. Singh, "Vector space model: an information retrieval system," *Int. J. Adv. Engg. Res. Studies/IV/II/Jan.-March*, vol. 141, no. 143, 2015.

[2] Y. Wang, M. Rastegar-Mojarad, R. Komandur-Elayavilli, and H. Liu, "Leveraging word embeddings and medical entity extraction for biomedical dataset retrieval using unstructured texts," *Database*, vol. 2017, 2017.

[3] M. Arora, U. Kanjilal, and D. Varshney, "Evaluation of information retrieval: precision and recall," *International Journal of Indian Culture and Business Management*, vol. 12, no. 2, pp. 224-236, 2016.

[4] N. Rastogi, P. Verma, and P. Kumar, "Evaluation of information retrieval performance metrics using real estate ontology," in *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 2020: IEEE, pp. 102-106.

[5] N. Mehra, "A Novel Kernelized Fuzzy Clustering Algorithm for Data Classification," *International Journal*, vol. 9, no. 8, 2021.

[6] I. Osuna-Galán, Y. Pérez-Pimentel, and C. Aviles-Cruz, "A Novel 2D Clustering Algorithm Based on Recursive Topological Data Structure," *Symmetry*, vol. 14, no. 4, p. 781, 2022.

[7] M. Bendechache, M.-T. Kechadi, and N.-A. Le-Khac, "Efficient large scale clustering based on data partitioning," in *2016 IEEE international conference on Data science and advanced analytics (DSAA)*, 2016: IEEE, pp. 612-621.

[8] H. Gulati and P. Singh, "Clustering techniques in data mining: A comparison," in *2015 2nd international conference on computing for sustainable global development (INDIACom)*, 2015: IEEE, pp. 410-415.

[9] I. Osuna-Galán, Y. Pérez-Pimentel, C. Avilés-Cruz, and J. Villegas-Cortez, "Topology: A theory of a pseudometric-based clustering model and its application in content-based image retrieval," *Mathematical Problems in Engineering*, vol. 2019, pp. 1-14, 2019, doi: https://doi.org/10.1155/2019/4540731.

[10] P. Praveen and B. Rama, "An empirical comparison of Clustering using hierarchical methods and K-means," in *2016 2nd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*, 2016: IEEE, pp. 445-449.

[11] P. S. Nishant, S. Mehrotra, P. R. Sree, and P. Srikanth, "Hierarchical clustering based intelligent information retrieval approach," in *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 2020: IEEE, pp. 862-866.

[12] J. Lu and Q. Zhu, "An effective algorithm based on density clustering framework," *Ieee Access*, vol. 5, pp. 4991-5000, 2017.

[13] X. Chen, "Clustering Algorithms In Data Mining," 2017.

[14] Y. Yang and Z. Zhu, "A fast and efficient grid-based K-means++ clustering algorithm for large-scale datasets," in *Proceedings of the Fifth Euro-China Conference on Intelligent Data Analysis and Applications 5*, 2019: Springer, pp. 508-515.

[15] R. Janani and S. Vijayarani, "Text document clustering using spectral clustering algorithm with particle swarm optimization," *Expert Systems with Applications*, vol. 134, pp. 192-200, 2019, doi: https://doi.org/10.1016/j.eswa.2019.05.030.

[16] D. Huang, C.-D. Wang, J.-S. Wu, J.-H. Lai, and C.-K. Kwoh, "Ultra-scalable spectral clustering and ensemble clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 6, pp. 1212-1226, 2019, doi: 10.1109/TKDE.2019.2903410.

[17] Y. Djenouri, A. Belhadi, P. Fournier-Viger, and J. C.-W. Lin, "Fast and effective cluster-based information retrieval using frequent closed itemsets," *Information Sciences*, vol. 453, pp. 154-167, 2018.

[18] K. K. Agbele, E. F. Ayetiran, K. D. Aruleba, and D. O. Ekong, "Algorithm for information retrieval optimization," in *2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, 2016: IEEE, pp. 1-8.

[19] M. Ahmed, "Semantic Based Intelligent Information Retrieval through Data Mining and Ontology," *International Journal of Computer Sciences and Engineering*, pp. 210-217, 2017.

[20] A. M. Siregar and A. Puspabhuana, "Improvement of term weight result in the information retrieval systems," in *2017 4th International Conference on New Media Studies (CONMEDIA)*, 2017: IEEE, pp. 108-112.

[21] T. Gahlawat, P. K. Bhatia, and D. Mehrotra, "The relationship between user preferences in interactive information retrieval evaluation," in *2017 6th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*, 2017: IEEE, pp. 423-426.

[22] N. T. W. Khin and N. N. Yee, "Query classification based information retrieval system," in *2018 international conference on intelligent informatics and biomedical sciences (ICIIBMS)*, 2018, vol. 3: IEEE, pp. 151-156.

[23] S. H. Alahmadi, "Information retrieval of distributed databases a case study: search engines systems," in *2018 1st International Conference on Computer Applications & Information Security (ICCAIS)*, 2018: IEEE, pp. 1-5.

[24] R. Aygun and W. Benesova, "Multimedia retrieval that works," in *2018 IEEE conference on multimedia information processing and retrieval (MIPR)*, 2018: IEEE, pp. 63-68.

[25] L. Azzopardi, P. Thomas, and A. Moffat, "cwl_eval: An evaluation tool for information retrieval," in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2019, pp. 1321-1324.

[26] R. Gao and C. Shah, "How fair can we go: Detecting the boundaries of fairness optimization in information retrieval," in *Proceedings of the 2019 ACM SIGIR international conference on theory of information retrieval*, 2019, pp. 229-236.

[27] E. Wahyudi, S. Sfenrianto, M. J. Hakim, R. Subandi, O. R. Sulaeman, and R. Setiyawan, "Information retrieval system for searching JSON files with vector space model method," in *2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT)*, 2019: IEEE, pp. 260-265.

[28] X. Li, K. Li, D. Qiao, Y. Ding, and D. Wei, "Application research of machine learning method based on distributed cluster in information retrieval," in *2019 International Conference on Communications, Information System and Computer Engineering (CISCE)*, 2019: IEEE, pp. 411-414.

[29] H. Khalifi, W. Cherif, A. E. Qadi, and Y. Ghanou, "Query expansion based on clustering and personalized information retrieval," *Progress in Artificial Intelligence*, vol. 8, pp. 241-251, 2019.

[30] M. Eminagaoglu, "A new similarity measure for vector space models in text classification and information retrieval," *Journal of Information Science*, vol. 48, no. 4, pp. 463-476, 2022.

[31] A. Alnaied, M. Elbendak, and A. Bulbul, "An intelligent use of stemmer and morphology analysis for Arabic information retrieval," *Egyptian Informatics Journal*, vol. 21, no. 4, pp. 209-217, 2020.

[32] A. Hammache and M. Boughanem, "Term position-based language model for information retrieval," *Journal of the Association for Information Science and Technology*, vol. 72, no. 5, pp. 627-642, 2021, doi: https://doi.org/10.1002/asi.24431.

[33] P. Joby, "Expedient information retrieval system for web pages using the natural language modeling," *Journal of Artificial Intelligence*, vol. 2, no. 02, pp. 100-110, 2020.

[34] S. Neji, T. Chenaina, A. M. Shoeb, and L. B. Ayed, "HIR: a hybrid IR ranking model," in *2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC)*, 2021: IEEE, pp. 1717-1722.

[35] S. Jain, K. Seeja, and R. Jindal, "A fuzzy ontology framework in information retrieval using semantic query expansion," *International Journal of Information Management Data Insights*, vol. 1, no. 1, p. 100009, 2021.

[36] J. Joe, "Information Retrieval based on Cluster Analysis Approach," *Available at SSRN*, 2021.

[37] Q. Ye, "RETRACTED ARTICLE: Situational English Language Information Intelligent Retrieval Algorithm Based on Wireless Sensor Network," *International Journal of Wireless Information Networks*, vol. 28, no. 3, pp. 287-296, 2021.

[38] F. Lechtenberg, J. Farreres, A.-L. Galvan-Cara, A. Somoza-Tornos, A. Espuña, and M. Graells, "Information retrieval from scientific abstract and citation databases: A query-by-documents approach based on Monte-Carlo sampling," *Expert Systems with Applications*, vol. 199, p. 116967, 2022.

[39] F. Zhang *et al.*, "Mind the gap: Cross-lingual information retrieval with hierarchical knowledge enhancement," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, vol. 36, no. 4, pp. 4345-4353.

[40] M. Kumari and P. Ahlawat, "Intelligent Information Retrieval for Reducing Missed Cancer and Improving the Healthcare System," *International Journal of Information Retrieval Research (IJIRR)*, vol. 12, no. 1, pp. 1-25, 2022.

[41] P. Erbacher, L. Soulier, and L. Denoyer, "State of the Art of User Simulation approaches for conversational information retrieval," *arXiv preprint arXiv:2201.03435*, 2022.
[42] P. Arora and S. Varshney, "Analysis of k-means and k-medoids algorithm for big data," *Procedia Computer Science*, vol. 78, pp. 507-512, 2016.
[43] C. Xiong, Z. Hua, K. Lv, and X. Li, "An Improved K-means text clustering algorithm By Optimizing initial cluster centers," in *2016 7th International Conference on Cloud Computing and Big Data (CCBD)*, 2016: IEEE, pp. 265-268.
[44] M. Alhawarat and M. Hegazi, "Revisiting k-means and topic modeling, a comparison study to cluster arabic documents," *IEEE Access*, vol. 6, pp. 42740-42749, 2018.
[45] S. Pasarate and R. Shedge, "Concept based document clustering using K prototype Algorithm," in *2018 International Conference on Control, Power, Communication and Computing Technologies (ICCPCCT)*, 2018: IEEE, pp. 579-583.
[46] C. Yuan and H. Yang, "Research on K-value selection method of K-means clustering algorithm," *J*, vol. 2, no. 2, pp. 226-235, 2019.
[47] Y. Li, J. Cai, H. Yang, J. Zhang, and X. Zhao, "A novel algorithm for initial cluster center selection," *ieee Access*, vol. 7, pp. 74683-74693, 2019.
[48] T. Gocken and M. Yaktubay, "Comparison of different clustering algorithms via genetic algorithm for VRPTW," 2019.
[49] R. Bhagawati, "Clusters Analyzer Algorithm for Informative Acquaintances-Quantum Clustering Algorithm," in *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, 2020: IEEE, pp. 38-42.
[50] A. M. Alonso, F. J. Nogales, and C. Ruiz, "Hierarchical clustering for smart meter electricity loads based on quantile autocovariances," *IEEE Transactions on Smart Grid*, vol. 11, no. 5, pp. 4522-4530, 2020.
[51] K. P. Sinaga and M.-S. Yang, "Unsupervised K-means clustering algorithm," *IEEE access*, vol. 8, pp. 80716-80727, 2020, doi: 10.1109/ACCESS.2020.2988796.
[52] A. A. Jalal and B. H. Ali, "Text documents clustering using data mining techniques," *International Journal of Electrical & Computer Engineering (2088-8708)*, vol. 11, no. 1, 2021, doi: 10.11591/ijece.v11i1.pp664-670
[53] X. Ran, X. Zhou, M. Lei, W. Tepsan, and W. Deng, "A novel k-means clustering algorithm with a noise algorithm for capturing urban hotspots," *Applied Sciences*, vol. 11, no. 23, p. 11202, 2021.
[54] J. Shuja, M. A. Humayun, W. Alasmary, H. Sinky, E. Alanazi, and M. K. Khan, "Resource efficient geo-textual hierarchical clustering framework for social IoT applications," *IEEE Sensors Journal*, vol. 21, no. 22, pp. 25114-25122, 2021.
[55] A. A. Almazroi and W. Atwa, "An Improved Clustering Algorithm for Multi-Density Data," *Axioms*, vol. 11, no. 8, p. 411, 2022.
[56] F. H. Awad and M. M. Hamad, "Improved k-means clustering algorithm for big data based on distributed smartphoneneural engine processor," *Electronics*, vol. 11, no. 6, p. 883, 2022.
[57] A. Sohrabi, N. Saraygord-Afshari, and M. Roudbari, "The Application of Bi-clustering and Bayesian Network for Gene Sets Network Construction in Breast Cancer Microarray Data," *Middle East Journal of Cancer*, vol. 13, no. 4, pp. 624-640, 2022, doi: 10.30476/mejc.2022.89998.1557.