International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. , No. (Mon-20..)

http://dx.doi.org/10.12785/ijcds/XXXXXX

Advancing Information Retrieval: Addressing the Challenges of Clustering and Pattern Mining

Vaishali Patel[1], Dilendra Hiran[2] and Kruti Dangarwala[3]

[1,2]*Department of Computer Engineering, PAHER University, Udaipur, Rajasthan, India*
[3]*Department of Computer Science Engineering, SVM Institute of Technology, Bharuch, India*

Received Mon. 20, Revised Mon. 20, Accepted Mon. 20, Published Mon. 20

Abstract: Information retrieval (IR) is booming because any application dealing with knowledge must retrieve relevant information from a huge data collection. The clustering mechanism plays a vital role in efficiently mining data from massive datasets. During a search, the items that have similar characteristics are grouped together using this strategy so that they may be found and retrieved more quickly. Traditional clustering methods are not capable of producing the required results in an efficient manner. When used in conjunction with a pattern mining technique, clustering can significantly boost the effectiveness of a search. The pattern mining method improves the quality of the clusters produced by exploring the dataset for patterns comparable to one another. The primary emphasis of this study is placed on more recent breakthroughs in information retrieval methods, including clustering and pattern mining. The article examines the present state of the art in information retrieval by dividing it into a few different categories and discussing its implications. This paper provides an overview of the most recent developments in the information retrieval field. The comparative analysis outlines the benefits and limitations of many different retrieval algorithms utilized to obtain the information. Open questions, challenges, and emerging trends are studied thoroughly. We have implemented a k-Means clustering algorithm for document clustering. Performance is evaluated in terms of the number of clusters, SSE, and execution time for the 20Newsgroup document dataset, which works well for small-scale datasets. The research community can develop more efficient data retrieval techniques by focusing on this article's challenges and future dimensions.

Keywords: Searching, Clustering, Pattern Mining, Information Retrieval, Knowledge Discovery

1. Introduction

In recent years, 'Information Retrieval (IR)' term is trendy in the field and applications of Information technology. The IR process retrieves information from a massive collection of data from various data sources, including text, web-based databases, audio, and video. IR presents extract nature of most suitable information for a query from a database. The main objective of IR is to extract useful and more relevant information to match with user queries submitted in web searches quickly and efficiently [1], [2], [3]. Essential phases of Information Retrieval systems (IRS) for information search are: Gathering documents, Indexing, Query processing, Retrieval, Ranking, and Visualization of information as shown in Fig. 1.

Cluster-based IR is one of the essential approaches to retrieving information from the vast collection of data from any source. This approach clusters similar documents based on the nearest data points (distance based) against the user queries, which can be helpful to the researcher or system to analyze the data effectively. This objective can be fulfilled by applying the steps to the retrieved documents against the user query: pre-processing, feature extraction, cluster generation, representation, indexing, query processing, and result presentation. Most cluster-based IR approaches organize and retrieve information from massive data collections by generating clusters. The main issue of this approach is that irrelevant, redundant, or useless information is usually retrieved with the important clusters. Another problem with this approach is that clusters depend on extracted features of the dataset and the type of clustering technique used [4], [5].

Pattern mining in IR generate unknown and frequently used item sets with relationships from large textual/transaction dataset to retrieve important information for various task of IR. Pattern mining applies the following steps to retrieve data for numerous applications where knowledge representation is critical [6], [7].

- Data preprocessing
- Pattern discovery
- Frequent item set mining
- Sequential pattern mining
- Graph mining

E-mail address: author 1, author 2, author 3

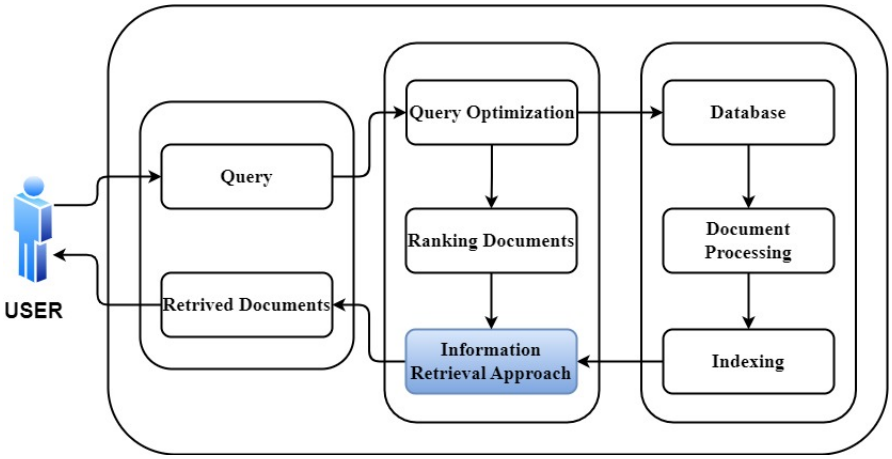

Figure 1. Process diagram of Information Retrieval System

- Evaluation
- Application of generated pattern

Cluster-based IR with a pattern mining approach organizes, analyzes, and retrieves information from a huge collection of documents using clustering and pattern mining algorithms. This approach applies these steps to retrieve important information: Data preparation, apply clustering algorithm, use of pattern mining algorithm, performance measures, and evaluation. This approach can be applied in various areas/domains, including academics, organizations, shopping, online applications, clinical, recommendation, textual data mining, and many more. In this approach, clustering and pattern mining algorithms are used, which extract essential and useful information by generating clusters of similar items and then searching for important or most relevant associations among items within a cluster. The results generated by this approach are most appropriate, suitable, and effective by finding the most frequent item set per cluster [8], [9], [10].

The article provides a thorough analysis of the IR discipline. We've utilized three classification areas to encompass the domain's breadth. The contribution of cited articles is discussed in depth, and comparative analysis is presented to provide a deeper understanding. Issues such as open queries, challenges, and future directions are presented to comprehend retrieval techniques better. The second section offers significant contributions to the IR procedure. The third section provides a comparative analysis of each cited source. The fifth section contains open questions, challenges, and prospective directions. The article concludes with Section 7.

2. State-of-the-Art: Information Retrieval

This section examines the literature published within the last 15 years in the field of IR. As keywords, clustering, pattern mining, knowledge discovery, information retrieval, IR algorithms, efficiency, and optimal information retrieval were used to search the relevant databases. We predominantly utilized Google Scholar, DBLP, IEEE, Springer, Elsevier, and Web of Science to discover suitable articles. Initially, more than 200 documents were selected to determine the bleeding edge of IR. At a later stage of the research, only the most pertinent articles were included, while others were excluded. The following sub-sections classify the included articles into three major categories: cluster-based IR, IR based on pattern mining, and a hybrid approach, and present the contemporary trends in each class. The taxonomy employed for the classification is shown in Fig. 2 where the box with dotted line indicates the major classification techniques.

A. *Cluster-based Information Retrieval*

In traditional IR systems, documents are usually indexed one at a time, and queries are matched against these individual documents to find appropriate results. In cluster-based IR, on the other hand, documents are first grouped into clusters. Then, queries are matched against these clusters to find relevant clusters, which could reduce the search space and speed up retrieval [11], [12], [13], [14].

The study by Naini et al. [15] focuses on a diversification approach where optimization and heuristics methods are integrated with low complexity and effectively in IR. The work discussed by Lan et al. [16] used classification methods has provided a term weighting approach that offers performance improvement. Levi et al. [17] observed that existing cluster-based IR or standard document retrieval methods often retrieves different documents. An approach offered by Bhopale & Tiwari [18] integrates swarm intelligence and data mining techniques to provide adequate IR. The approach proposed by Sheetrit & Kurland [19] ranks the retrieved documents based on their relevancy to a user query motivated by cluster-based IR. The proposed approach generates knowledge (retrieve information) in specific Dialog system technology challenges 7 (DSTC7) [20]. This work trains a model first popular as a generator model to execute an action. Bascur et al. [21] suggested a

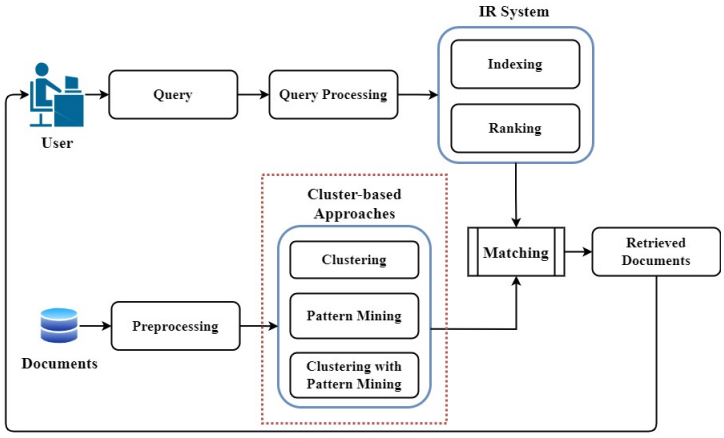

Figure 2. Taxonomy of IR classification

citation-based cluster generation for IR, which retrieves the information by combining a tree hierarchy approach with a cluster generation algorithm.

In the proposed hybrid method by Chawla [22], the fuzzy c-Means document clustering algorithm and ontology are used in query session mining to make intelligent IR clusters that improve the performance and quality of the clusters they make. The suggested model gets information from social media based on what the user asks for [23]. The result shows a more promising performance than other social media data retrieval methods. The work by Mbate et al. [24] is an alternative to deep neural networks for semantic IR from long documents. Toman et al. [25] have proposed an HPGA algorithm integrating the k-Means algorithm with hybrid master/slave algorithms.

In recent years, spectral clustering has become a popular clustering technique in machine learning. The study by Janani & Vijayarani [26] offers a novel spectral clustering algorithm with particle swarm optimization (SCPSO) to enhance the clustering of text documents. Results demonstrate that the proposed SCPSO algorithm produces more accurate clustering than other clustering techniques. The article by Zubair et al. [27] proposed a method for effectively locating optimal initial centroids to reduce iterations and execution time. Xie et al. [28] proposed IIEFA and CIEFA models to focus on the significant issue of initializing several clusters in the k-Means algorithm.

B. Pattern Mining Based Information Retrieval

Pattern mining-based information retrieval is a method that uses data mining to find meaningful patterns in big sets of textual data. These patterns can then be used to make information retrieval systems more efficient and effective. The idea is to get helpful information out of unstructured text and improve the retrieval process so that users get more accurate and useful search results. Pattern mining-based information retrieval can be especially helpful when standard keyword-based methods are insufficient [29], [30], [31], [32].

The work by Thirugnanasambandam et al. [33] proposed Document Information Retrieval (DIR) using ant colony optimization, which provides an advanced analysis in the field of data mining with the evolutionary concept of exploring document space to retrieve information. The study by Gan et al. [34] proposed an efficient utility mining approach that is popular as a non-redundant Correlated high-Utility Pattern Miner (CoUPM) to address the issues of inheritance correlation of frequent itemset by considering the positive correlation and profitable value. Yun et al. [35] proposed a High Average Utility Pattern Mining approach, which generates valuable patterns with relatively more meaning using a novel utility measure.

Zinga et al. [36] have introduced the HQE model, which focuses on integrating external resources with association rule mining to improve the retrieval process. PM-HR is proposed by Belhadi et al. [37], which has used a pattern mining algorithm to improve the accuracy and processing speed of information retrieval. Babashzadeh et al. [38] proposed a model for medical query contexts based on the mining semantic-based association rule mining and semantic relatedness measures, which is exploited for re-rank to improve the IR of a medical dataset. A new framework is proposed by Cai & Li [39], which generates clusters based on the sentence to improve overall sentence-based ranking performance.

C. Cluster-based IR with Pattern Mining

Clustering with pattern mining-based information retrieval is a hybrid approach that combines clustering with pattern mining. Though selecting these hybrid algorithms requires careful tuning, we can use multimodel datasets, get faster results and find hidden relationships [40], [41], [42], [43].

Intelligent Cluster-based Information Retrieval is proposed by Djenouri et al. [44] to address critical challenges

like relevant information retrieval, performance improvement, and quality clusters generation of cluster-based IR approaches. The method discussed in [45] uses frequent and high-utility pattern mining to find relevant patterns for each cluster from the pre-processed collection. Bokhabrine et al. [46] implemented an "IDETEX" platform that generates item sets from textual data using various clustering techniques. The study of Saili et al. [47] handles the main challenge of retrieving information from large-scale datasets, such as those with high dimensionality. The work by Yarlagadda et al. [48] introduced a Modsup-based frequent item set and Rider Optimization-based Moth Search Algorithm (Rn-MSA). Yarlagadda et al. [48] proposed a framework where pattern mining is based on Modsup and Rider Optimization which generates document clustering and overcome the challenge to generate informative and efficient clusters.

Rouane et al. [49] proposed a novel biomedical text summarization method that integrates data mining approaches. The method by Sato et al. [50] employs a data mining technique to discover knowledge using Pareto-optimal solutions to surmount multi-objective topology optimization problems. The method proposed by Kim & Chung [51] generates valuable associative feature data from health big data using text mining.

The work of Waghere et al. [52] focuses on the significant issues of frequent item generation algorithms in a distributed computing environment using MapReduce framework. The work by Sovia et al. [53] aims to improve the quality of these expert people in the company, which automatically improves the quality of human resources. The work by Gayathri & Arunodhaya [54] proposed combined RFM and k-Means clustering approach for segmenting the correct consumers with a recommendation feature for commercial web application data. The work by Kusak et al. [55] analyzed landslide data and evaluated the pre-landslide conditions of the region using conventional statistical and spatial data mining techniques.

Scells et al. [56] explained the latest trends in green information retrieval. A new IR framework is proposed by Lin [57] for a symbolic approach. Legal issues regarding IR are discussed in the work of Sansone & Sperlí [58]. Improvements in query expansion for web-based information retrieval are discussed in [59], [60]. Single server private IR is addressed in [61], [62]. Tavares et al. [63] explain the relation between cyberspace and IR. The use of modern machine learning algorithms for IR is presented in Zamani et al. [64]. Research and Development in performance improvement for IR are discussed by Rao et al. [65].

3. Comparative Analysis: Research Contributions, Strengths and Limitations

The previous section discusses the methodology adopted by the particular article in detail. The findings of the cited article are crucial in advancing scientific knowledge and promoting overall development. Table 1 shows the research contribution, article strength, and presented approaches' limitations. The tabular representation also highlights the recent developments, the algorithms, the outcomes, and the constraints. Comparative analysis of present approaches for information retrieval is based on three main categories: cluster-based information retrieval, pattern-based information retrieval, and clustering with pattern-based information retrieval. Analysis shows the main contribution, strengths, and limitations of said research. Most research in clustering and pattern mining uses a combined approach; clustering with pattern mining needs to be generalized and should perform efficiently for large document datasets.

TABLE I. Comparative Analysis of Cited Research Articles

Research	Contributions	Strength	Limitations
	Cluster-based Information Retrieval		
[66]	Fusion based method generates number of clusters using chameleon clustering algorithms with improved experiments results	Optimized performance	Need to be generalized
[67]	Selective search to reduce time complexity and greater effectiveness	Faster response time, higher throughput, save storage cost	Useful for low resource environment
[68]	Clustering algorithms are used with weight function for effective information retrieval	Provide optimized and effective IR	Need high dimensions datasets
[69]	Hybrid indexing method with cluster based IR technique to improve the complexity and cost	Reduce time complexity and less expansive	Disjunctive query is not considered to reduce costing time and space
[70]	Clusters are generated with ranking using cluster based algorithms and re-rank methods	Apply re-rank method to generate more precise clusters	Different nature of datasets are not re-ranked.
[71]	Optimized using ranking to clicked URLs using genetic algorithm	Computation time is reduced with more relevant and effective IR	Appropriate for personalized web search only. More performance measures needed
[15]	Provides detailed analysis for large set of documents effectively and efficiently	Quality and efficient information is retrieved in distributed environment in less time	Extended for generalized analysis
[16]	KNN and SVM are integrated with term weighting method to overcome the identification of different terms of document	Experiment results shows an improvement	Performance is not uniform for large datasets
[17]	Applied cluster based, query expansion and term proximity methods for effective result generation	Performance is improved in terms of efficiency and effectiveness	Extended for generalized approach
[18]	Data mining technique is integrated with swarm intelligence to generate fuzzy clusters frequent patterns respectively	Problem of local minima, complexity and effectiveness are improved	Approach can be extended for more datasets
[19]	Ranking method is integrated to assign rank to each generated cluster	Result of various data sets shows an improvement	Extended for different ranking methods
[20]	Conversational history modeling is relevant and interesting	Results shows an improvement like more diverse and informative retrieval	Can be extended using large and high dimensional dataset
[21]	Approach retrieve information from systematic literature reviews specifically	Evaluation shows an improvement	Not cover all possible citation in review using corpus and sensitive to noise
[72]	Boolean queries is integrated with cluster based IR approach	an effective workflow for adaptive visual search of complex information	Needed more high dimentional datasets
[73]	Measure the quality of document clustering for large corpora	Results shows an effective information retrieval	Sensitive of k value

[74]	User query is accepted in form of phrases where SVM classifier and ranking methods are applied to identify and assign rank	Performance is improved	Integrated algorithms required
[75]	Information is retrieved for topic modeling of news to discover more precise topics	Integration of classification and clustering retrieves better information	Need to apply more algorithms
[76]	k-Means with hierarchical technique of clustering using cosine measures	Overcome each other's limitations (partition and hierarchy approaches)	Can be improved the retrieval using huge datasets
[22]	Fuzzy c-Means with the integration of semantic ontology retrieve intelligent information	Evaluation shows a significant improvement	Approach can be generalized
[23]	Social media contents are integrated with the expansion of word queries which provides more reliable and effective IR	Results shows an improvement in performance	Approach cab be generalized
[24]	Combined approach of clustering algorithm	Performance shows a significant improvement	More datasets can be tested for better analysis
[25]	Hierarchical parallel genetic algorithm retrieve more relevant information	Quality and performance is improved	more datasets can be tested
[26]	Generates automatic clusters for unstructured text documents	Improve the accuracy and provides optimal solution	Enhancements for more text documents
[27]	Generates optimal number of iterations which reduce the overall execution time using high dimensional healthcare datasets	Reduce the computational power and work faster	can solve many problems of real-world application areas including smart city and IOT
[28]	Overcome the problems associated with initialization sensitivity and local optima traps of the conventional KM clustering algorithm	Increases search diversification and efficiency.	Objective functions for inter and intra cluster measurement can be employed
	Pattern mining based Information Retrieval		
[33]	Fuzzy c-Means is integrated with ant colony optimization	Information is retrieved effectively with improved accuracy	Data Mining approaches can be consider to reduce the complexity
[34]	Correlated significance is used with high utility frequent pattern mining algorithm	Result proves work's effectiveness and efficiency	More pruning strategies could be explored
[35]	A damped window model with pruning strategies consider time factors to search significant and recent pattern generation	Performance is increased	More reliable patterns can be generated
[36]	Candidate term generation and selection phases are developed using local as well as global methods	Results show improvement in quality	Weight and embedded vectors can be integrated to discard redundancy and filter irrelevant terms
[37]	Information is retrieved from large number of tweets, hashtags and real time user queries to generate high utility patterns	Solve the issue of hash-tag information retrieval. Results shows outperform in terms of execution time and accuracy	Advanced computing tools
[38]	Designed for medical query contexts using semantic association rule mining algorithm	Result shows an improvement	Timer series and sequential pattern mining algorithms can be integrated

[39]	Integrate ranking with sentence clusters to refine the results which generates direct clusters uniquely	Ranking and clustering by mutually and simultaneously updating each other to improve the performance	Need to improve the effective speech summarization to minimize execution time
[77]	Improved ARM with swarm optimization using different strategies to generate effective and efficient association rules	Effectively generates association rules for large datasets by improving fitness criteria and processing time	Need to extend for high dimensional transactional datasets
[78]	Deep learning with implicit user feedback to represent item's preferences based on the observations	Address the challenges like item cold-start and recommendation based on user's implicit preference feedback	Explore with other deep learning architectures for high dimensional and large datasets
[79]	Designed using Galois connection, granular computing and connection function of smallest frequent closed granule to reduce the coasted I/O	Performance is improved	Many smallest frequent closed item-sets are generated
[80]	Discovery technique to process the deploying and evolving of patterns to find relevant information	Effective patterns in text mining and overcome low frequency and misinterpretation problems of text mining	More datasets can be explored for more analysis and evaluation
	Cluster-based IR with Pattern Mining		
[44]	Clustering with closed frequent item-set mining algorithm to extract rich knowledge to answer user query	Outperforms in terms of quality and run-time on large datasets	Enhance for large and other data types such as images and videos
[45]	Transformation and ranking approaches are integrated to assign weight, rank and generate high quality clusters	Improvement in the IR in terms of relevancy and quality	Deep learning algorithms and high performance computing tools are suggested to improve the performance
[47]	Generate frequent patterns in minimum execution time with improvement in IR	Increased efficiency and reduced computation time of high dimensional data-set	Need more experiment analysis
[48]	Combined Rider Optimization Algorithm and Moth Search Algorithm approach applied pre-processing to remove stop word, stemming, feature extraction and selection	Performance evaluation shows higher accuracy	Analysis is extended for highly advanced features datasets
[49]	Biomedical text summarization based on clustering and frequent patterns to enhance the quality	Result shows an improvement	More semantic analysis on biomedical texts required.
[50]	Topology optimization based clustering with association rules to discover important knowledge and effective solution	Discover effective design solutions	No guarantee of meaningful clusters retrieval
[51]	Extracts useful associative feature information using text mining	Generate efficient associative feature based information for large healthcare data	Apply and analyze for other domain datasets
[52]	Address major issue of frequent item generation in short time	Effectively extract frequent patterns and IR from huge dataset in reduced time	More Data mining approaches can be integrated and evaluated using large and high dimensional datasets

[53]	k-Means with Apriori approach search most frequent expert persons handling	Quality of human resources can be improved	Other domains can be explored
[54]	k-Means and RFM approach segment the right customers with recommendation feature and provides a solution to company's marketing problem	Provides services to customer and marketing segments	Enhancement could be done using dynamic datasets
[55]	Evaluate the pre-landslide conditions in Karahacılı District using traditional statistics and data mining approaches	Evaluation of landslides in small areas that can be detected	As data grow, more advanced methods of machine learning cab be applied

4. Performance Analysis

Based on the literature review on various approaches to clustering, pattern mining, and combining clustering with pattern mining, we have implemented a k-Means document clustering algorithm using the public 20Newsgroups dataset. This dataset contains approximately 18,000 documents in a variety of 20 categories. Out of these 20 categories, we have implemented categories 3, 4, and 5 to reduce computational time. Performance evaluation shows that for more number of clusters, less Sum of Squared Error (SSE) is generated, which is shown in Table II and Fig. 3.

TABLE II. Performance Analysis of k-Means Clustering Algorithm(SSE)

			Clusters	SSE
k-Means	**20Newsgroups**	**kaggle**	3	2
			4	1
			5	0

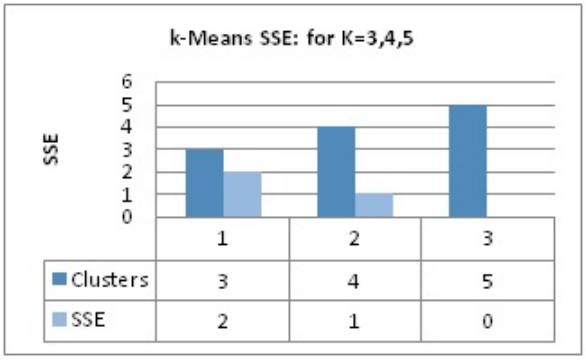

Figure 3. Performance Analysis: SSE of k-Means

As per the evaluation, k-Means clustering algorithm performs well for small scale dataset. But still k-Means and variation of k-Means for large size and dimensional document datasets is a challenging issue in the field of Information Retrieval. Performance analysis of k-Means clustering algorithm based on execution time for number of iterations and clusters is shown in table III and Fig. 4.

5. Open Issues, Challenges and Future Directions

Information Retrieval retrieves information from a massive collection of databases based on the user query on the web. Due to the documents' structured, semi-structured, or unstructured nature, information retrieval is challenging in IR. Various works are proposed to overcome this document retrieval using traditional approaches/IR models. These approaches often retrieve irrelevant information and need to work better in the case of large datasets (time-consuming). Based on the analysis, we compiled the following twelve critical challenges (Fig. 5) required to be addressed first for future-proof IR solutions [81].

1) **Framework (A):** Most researchers suggested a framework specific to some application. There is a need for a generalized framework that can be employed in most applications of IR. [82], [83], [84]
2) **Multidimensionality (B):** Few solutions include multi-dimensional data for generating clusters or finding patterns. The inclusion of multifaceted data is a need for today's IR algorithms. [85], [86].
3) **Dynamic Data (C):** Most real-world situations generate dynamic data. The IR technique must be able to process the dynamic nature of the information effectively. [87], [88].
4) **Scalability (D):** Scalability is a significant challenge due to the enormous volume of data generated in today's Internet world. Future algorithms must be able to handle the complexity generated by this information. [89], [90].
5) **Pattern Selection (E):** Pattern quality dramatically affects the IR system's performance. Appropriate pattern selection methods generate quality clustering results. [91], [92].
6) **Data Pre-processing (F):** Real-world data may be ambiguous and generate noise or irrelevant clustering or patterns. To achieve optimum results, data must be pre-processed effectively, including the proper feature selection. [93], [94].
7) **Enhanced Data Mining approaches (G):** Effective and enhanced clustering/classification/pattern generation techniques must be applied to address the quality of the IR approach. [95], [96].
8) **Rank (H):** Few methods have integrated ranking methods with IR approaches. It is required to consider Information retrieval for better and more effective results. [97], [98].
9) **Weight (I):** Approaches with weighting methods: Weight assignment to a user query is crucial to prune redundancy in information retrieval. [99], [100].
10) **Semantic (J):** It is helpful to improve the quality of clustering results by integrating semantic information. This leads to complexity in the clustering process, so it takes time to generate clustering results effectively. [101], [102].
11) **Pruning Strategy (K):** Numerous patterns are generated using pattern mining algorithms. Pruning strategies are required to prune and filter relevant and informative patterns, a critical step in IR. [103], [104].
12) **Advanced Techniques (L):** Approaches and advancements (machine learning techniques including deep learning) can be integrated with cluster-based IR using pattern mining for effective IR and to overcome significant challenges. [105], [106], [107].

These challenges are compared to the cited articles; the graphical representation is given in Fig. 6. It shows the contribution of each research article to a specific challenge. In cluster-based Information Retrieval with pattern-mining

TABLE III. Performance Analysis of k-Means Clustering Algorithm (Execution Time)

Clustering Algorithm	Dataset	Data Source	Number of Iterations	Execution Time for No of Clusters 3	4	5
k-Means	20Newsgroups	kaggle	15	4.482	7.631	11.398
			30	10.208	8.767	16.416
			42	4.127	10.379	14.557
			50	4.735	8.049	11.622
			72	3.084	11.296	12.1

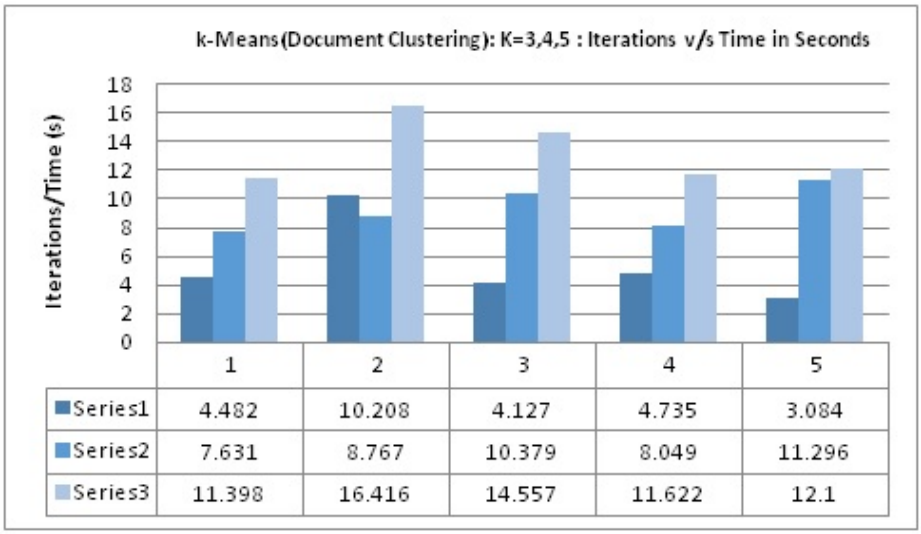

Figure 4. Performance Analysis:Execution Time of k-Means

approach, weight and rank are proposed for high-quality information retrieval. A major challenge of this approach is generalization and support various form of data. Cluster-based IR approach is also combined with advanced pattern mining technique to identify the patterns, including maximal, rare, and closed for generated clusters for more accurate information retrieval. Other machine learning technique like deep learning model can also be used for more effective and precise information retrieval from a high volume of datasets.

Here are some potential future directions and opportunities for information retrieval:

- User-level personalized information retrieval [108]
- Integration of heterogeneous data like text, image, audio, and video [109]
- Use of advanced machine learning algorithms like deep learning [110]
- Integration of multi-language and cross-language searching [111]
- Instant on-demand, real-time result [112]
- Use of explainable artificial intelligence [113]
- Integration of blockchain for secure information exchange [114]
- Searching based on the semantics of the user's query [115]

6. Discussion

Information retrieval is a searching technique that retrieves relevant information from large and different forms of data. Due to the tremendous use of the web, massive amounts of data are increasing in various forms, such as audio, video, image, etc. Cluster-based information retrieval is an Information retrieval technique that improves searching and retrieval. It overcomes the major IR problem by generating clusters with similar data objects based on the user query. Pattern mining-based information retrieval is an approach that creates informative patterns and improves information retrieval processes. It is suitable when traditional IR methods are not generating relevant searches. Cluster-based Information Retrieval and pattern mining are combined approaches that improve the effectiveness of information retrieval systems and have the benefits of both methods for more accurate, precise, and relevant information retrieval. In this discussion, we have explored the significant findings and challenges addressed by the information retrieval approaches: cluster-based, pattern mining-based, and clustering with pattern mining. These challenges are addressed based on a literature survey, comparative

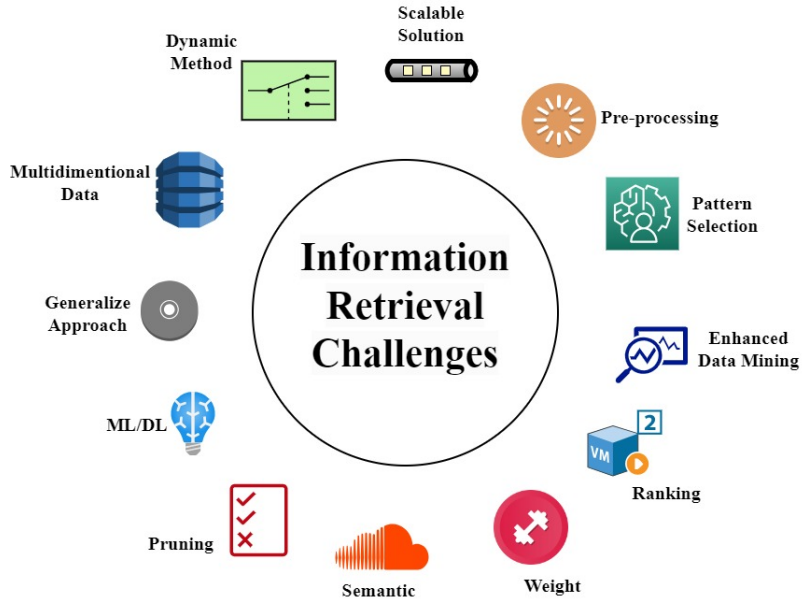

Figure 5. IR Challenges

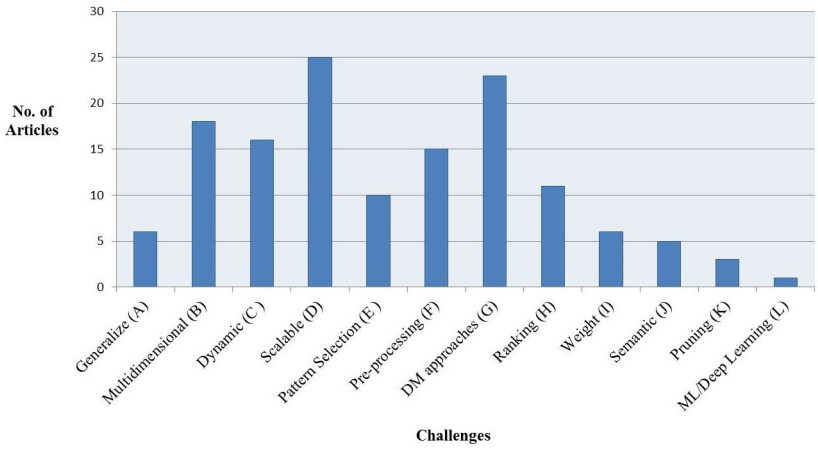

Figure 6. Number of articles contributing for each research challenge of IR

analysis, and challenges discussed earlier section 2. These challenges are related to computational complexity, cluster quality, scalability, and high-dimensional data.

Clustering techniques are introduced to generate clusters to retrieve information effectively and overcome these problems of traditional approaches. Quality, quick and correct information is not retrieved using cluster based information retrieval approaches against of the user queries. As per the literature survey of various research articles of cluster based, pattern mining based and clustering with pattern mining for information retrieval, twelve major unique challenges are addressed. These challenges are mentioned using alphabets A to L which represents framework, multi dimensionality, dynamic data, scalability, pattern selection, data pre-processing, data mining (ML), ranking, weighting, semantic, pruning strategy and advanced techniques respectively. These challenges are summarized in Table IV. Major merits of this cluster based information retrieval are relevant information generation by creating similar groups or clusters, often scalable, provides visualization of the retrieval, reduce redundancy in the generated results. Major demerits of cluster based information retrieval are selection of clustering algorithm based on the domain, loss of granularity, more complexity in case of scalable. According to Table V, the areas with the least amount of research are the general framework, pattern selection, weighting, semantic, pruning strategy and advanced techniques (enhanced/hybrid machine learning or deep learning) for cluster based information retrieval approach. Compared to other issues these issues have received less attention; therefore, more efficient work is required to address them.

TABLE IV. Challenges addressed in research article (Cluster-based)

Research	A	B	C	D	E	F	G	H	I	J	K	L
[66]	-	-	✓	✓	-	✓	-	-	-	-	-	-
[67]	-	-	✓	✓	-	-	-	✓	-	-	-	-
[68]	✓	-	-	✓	-	✓	-	✓	✓	-	-	-
[69]	-	-	✓	-	-	-	-	-	-	-	-	-
[70]	-	✓	-	✓	-	-	-	✓	-	-	✓	-
[71]	-	✓	✓	✓	-	-	-	✓	-	-	-	-
[15]	-	✓	-	✓	-	-	-	-	-	-	-	-
[16]	-	-	-	-	-	✓	-	-	✓	-	-	-
[17]	-	✓	-	✓	-	-	-	-	-	-	-	-
[18]	-	✓	-	-	✓	-	✓	-	-	-	-	-
[19]	-	-	-	-	-	✓	-	✓	-	-	-	-
[20]	-	-	-	-	-	✓	-	-	-	-	-	-
[21]	-	-	-	-	-	-	✓	-	-	-	-	-
[72]	-	✓	-	-	-	✓	-	-	-	-	-	-
[73]	-	-	-	-	-	✓	-	-	-	-	-	-
[74]	✓	-	✓	-	-	✓	-	-	-	-	-	-
[75]	-	✓	✓	-	-	-	-	-	-	-	-	-
[76]	-	-	-	-	-	✓	✓	-	-	-	-	-
[22]	-	-	-	-	-	-	✓	-	-	-	-	-
[23]	-	-	✓	✓	-	✓	-	-	✓	-	-	-
[24]	-	-	✓	✓	-	-	✓	-	-	-	-	-
[25]	-	✓	-	✓	-	-	✓	-	-	-	-	-
[26]	✓	-	-	✓	-	-	✓	-	-	-	-	-
[27]	-	-	-	✓	-	-	✓	-	-	-	-	-
[28]	-	✓	-	-	-	✓	-	-	-	-	-	-

TABLE V. Challenges addressed in each research article (Pattern mining based)

Research	A	B	C	D	E	F	G	H	I	J	K	L
[33]	-	-	✓	✓	-	✓	✓	-	-	-	-	-
[34]	✓	-	-	✓	-	-	-	-	-	-	✓	-
[35]	-	-	✓	-	-	-	-	-	-	-	✓	-
[36]	-	✓	-	-	✓	-	-	-	-	-	-	-
[37]	-	✓	✓	✓	✓	-	-	-	-	-	-	-
[38]	-	✓	✓	✓	-	-	✓	✓	-	✓	-	-
[39]	-	-	-	-	-	-	✓	✓	-	-	-	-
[77]	-	-	-	✓	-	-	✓	-	-	-	-	-
[78]	-	-	-	-	-	-	-	✓	✓	-	-	✓
[79]	✓	-	-	-	✓	-	-	-	-	-	-	-
[80]	-	-	-	✓	-	-	✓	-	-	-	-	-

According to the literature survey Researchers have proposed cluster-based information retrieval with the integration of pattern mining approaches to overcome these issues, where various clustering algorithms are integrated with pattern mining algorithms. Many existing approaches need to be revised for massive, high-dimensional datasets. Works are also proposed to overcome these challenges by mutually integrating ranking techniques with clustering algorithms to improve the quality of generated clusters. In this case, generated clusters are getting more accurate and informative results. Many times a term weighting is a significant and important task in the field of information retrieval. The main objective of term weighting in IR with cluster-based approaches with pattern mining is to improve the performance evaluation regarding precision and recall. A weighted rank also assigns the rank to some retrieval from the datasets using various weighting and ranking functions. Based on the user query, high/more weight is assigned to repeated terms in IR. The problems for pattern mining-based IR is shown in Table V. According to Table V, the areas with the least amount of research are the general framework, multi dimensionality, pattern selection, data

pre-processing, weighting, semantic, pruning strategy and advanced techniques (enhanced/hybrid machine learning or deep learning) for pattern mining based approach. More work is required to address these challenges in pattern mining based information retrieval.

In short, weight and rank are proposed to integrate with cluster-based IR using pattern-mining approaches for high-quality information retrieval. Still, these proposed approaches must be generalized and should work for the data types, including image, audio or video. Other advanced frequent mining techniques can be integrated with a cluster-based IR approach to identify the patterns, including maximal, rare, and closed for generated clusters in IR. Also, different cluster-based techniques or other machine learning techniques, including deep learning models, can be applied for more effective and precise information retrieval from a huge collection of datasets of various domains. The challenges associated with this hybrid approach are summarized in Table VI. According to Table 4, the areas with the least amount of research are the general framework, ranking, weighting, semantic, pruning strategy and advanced techniques (enhanced/hybrid machine learning or deep learning) for cluster based information retrieval with pattern mining approach. Great research and work is required to address these challenges in cluster based information retrieval with pattern mining.

In essence, more and more work is required to address significant challenges for the three information retrieval approaches discussed here. The focus should be on generalized framework, ranking, weighting, semantics, pruning strategy, and advanced techniques (enhanced/hybrid machine learning or deep learning). In the future, our work will concentrate on information retrieval challenges for effective and more precise searches.

7. Conclusion

The review article provides a comprehensive overview of the IR domain by highlighting its basics, methodologies, categories, and current trends. The work presented here mainly covers three main categories of IR: cluster-based information retrieval, pattern mining-based information retrieval, and hybrid approach. Moreover, this survey provides valuable insights by exploring the recent research articles and finding the strengths and limitations of each article. This provides a platform to discover significant challenges faced in the modern era of information retrieval. The investigation of evaluation criteria and bench-marking techniques has greatly helped accurate performance assessment of IR systems. We have also discussed the importance of machine learning and deep learning techniques, which may play a significant role in shaping the future of IR. Finally, the open issues, challenges, and future directions are discussed. This study clearly shows that Information Retrieval is a dynamic field that changes quickly and has many real-world uses, such as web search, digital libraries, and recommendation systems. The work presented here gives newcomers and researchers in the field a solid base. It also shows how important it is to keep researching and coming up with new ideas to deal with new problems and use Information Retrieval to its fullest. Performance evaluation shows it works well for small-scale document datasets with fewer dimensions. Collaboration between science and enterprises will help push the limits of IR and build next-generation systems that meet the different needs of information seekers worldwide.

References

[1] H.-J. Choi and C. H. Park, "Emerging topic detection in twitter stream based on high utility pattern mining," *Expert systems with applications*, vol. 115, pp. 27–36, 2019.

[2] D. V. Suma, "A novel information retrieval system for distributed cloud using hybrid deep fuzzy hashing algorithm," *Journal of Information Technology and Digital World*, vol. 2, no. 3, pp. 151–160, 2020.

[3] F. Gulnashin, I. Sharma, and H. Sharma, "A new deterministic method of initializing spherical k-means for document clustering," in *Progress in Advanced Computing and Intelligent Engineering: Proceedings of ICACIE 2017, Volume 1*. Springer, 2019, pp. 149–155.

[4] R. V. Pamba, E. Sherly, and K. Mohan, "Self-adaptive frequent pattern growth-based dynamic fuzzy particle swarm optimization for web document clustering," in *Computational Intelligence: Theories, Applications and Future Directions-Volume II: ICCI-2017*. Springer, 2019, pp. 15–25.

[5] A. K. Sangaiah, A. E. Fakhry, M. Abdel-Basset, and I. El-henawy, "Arabic text clustering using improved clustering algorithms with dimensionality reduction," *Cluster Computing*, vol. 22, pp. 4535–4549, 2019.

[6] I. Sharma, A. Jain, and H. Sharma, "Stream and online clustering for text documents," in *International Conference on Advanced Computing Networking and Informatics: ICANI-2018*. Springer, 2019, pp. 469–475.

[7] M. S. Rani and G. C. Babu, "Efficient query clustering technique and context well-informed document clustering," in *Soft Computing and Signal Processing: Proceedings of ICSCSP 2018, Volume 1*. Springer, 2019, pp. 261–271.

[8] J. M.-T. Wu, C. W. Lin, P. Fournier-Viger, Y. Djenouri, C.-H. Chen, and Z. Li, "The density-based clustering method for privacy-preserving data mining," 2019.

[9] Y. Wan, X. Liu, Y. Wu, L. Guo, Q. Chen, and M. Wang, "Icgt: A novel incremental clustering approach based on gmm tree," *Data & Knowledge Engineering*, vol. 117, pp. 71–86, 2018.

[10] C. Vidyadhari, N. Sandhya, and P. Premchand, "Automatic incremental clustering using bat-grey wolf optimizer-based mapreduce framework for effective management of high-dimensional data," *International Journal of Ambient Computing and Intelligence (IJACI)*, vol. 11, no. 4, pp. 72–92, 2020.

[11] X. Liu and W. B. Croft, "Cluster-based retrieval using language models," in *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, 2004, pp. 186–193.

TABLE VI. Challenges addressed in each research article (Cluster-based with Pattern Mining)

Research	A	B	C	D	E	F	G	H	I	J	K	L
[44]	-	✓	-	✓	-	-	✓	✓	✓	-	-	-
[45]	-	-	-	-	-	✓	✓	✓	✓	-	-	-
[46]	-	-	-	-	✓	-	✓	-	-	-	-	-
[47]	✓	-	-	-	-	-	✓	-	-	-	-	-
[48]	-	-	-	✓	-	✓	✓	-	-	-	-	-
[49]	-	✓	✓	✓	✓	-	-	-	-	✓	-	-
[50]	-	-	-	-	✓	-	✓	-	-	-	-	-
[51]	-	✓	✓	✓	-	✓	✓	✓	-	-	-	-
[52]	-	-	-	-	✓	-	✓	-	-	-	-	-
[53]	-	✓	✓	✓	-	-	✓	-	-	-	-	-
[54]	-	✓	-	✓	✓	-	✓	-	-	-	-	-
[55]	-	✓	✓	✓	✓	-	-	-	-	-	-	-

[12] B.-Y. Kang, D.-W. Kim, and S.-J. Lee, "Exploiting concept clusters for content-based information retrieval," *Information sciences*, vol. 170, no. 2-4, pp. 443–462, 2005.

[13] M. K. Abbasi and I. Frommholz, "Cluster-based polyrepresentation as science modelling approach for information retrieval," *Scientometrics*, vol. 102, pp. 2301–2322, 2015.

[14] H.-S. Oh and Y. Jung, "Cluster-based query expansion using external collections in medical information retrieval," *Journal of biomedical informatics*, vol. 58, pp. 70–79, 2015.

[15] K. D. Naini, I. S. Altingovde, and W. Siberski, "Scalable and efficient web search result diversification," *ACM Transactions on the Web (TWEB)*, vol. 10, no. 3, pp. 1–30, 2016.

[16] M. Lan, C. L. Tan, J. Su, and Y. Lu, "Supervised and traditional term weighting methods for automatic text categorization," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 4, pp. 721–735, 2008.

[17] O. Levi, F. Raiber, O. Kurland, and I. Guy, "Selective cluster-based document retrieval," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, 2016, pp. 1473–1482.

[18] A. P. Bhopale and A. Tiwari, "Swarm optimized cluster based framework for information retrieval," *Expert Systems with Applications*, vol. 154, p. 113441, 2020.

[19] E. Sheetrit and O. Kurland, "Cluster-based focused retrieval," in *Proceedings of the 28th ACM international conference on information and knowledge management*, 2019, pp. 2305–2308.

[20] Y.-C. Tam, "Cluster-based beam search for pointer-generator chatbot grounded by knowledge," *Computer Speech & Language*, vol. 64, p. 101094, 2020.

[21] J. P. Bascur, S. Verberne, N. J. van Eck, and L. Waltman, "Academic information retrieval using citation clusters: in-depth evaluation based on systematic reviews," *Scientometrics*, vol. 128, no. 5, pp. 2895–2921, 2023.

[22] S. Chawla, "Application of fuzzy c-means clustering and semantic ontology in web query session mining for intelligent information retrieval," *International Journal of Fuzzy System Applications (IJFSA)*, vol. 10, no. 1, pp. 1–19, 2021.

[23] H. Khalifi, S. Dahir, A. El Qadi, and Y. Ghanou, "Enhancing information retrieval performance by using social analysis," *Social Network Analysis and Mining*, vol. 10, pp. 1–7, 2020.

[24] P. Mbate Mekontchou, A. Fotsoh, B. Batchakui, and E. Ella, "Information retrieval in long documents: Word clustering approach for improving semantics," *arXiv e-prints*, pp. arXiv–2302, 2023.

[25] S. H. Toman, M. H. Abed, and Z. H. Toman, "Cluster-based information retrieval by using (k-means)-hierarchical parallel genetic algorithms approach," *arXiv preprint arXiv:2008.00150*, 2020.

[26] 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.

[27] M. Zubair, M. A. Iqbal, A. Shil, M. Chowdhury, M. A. Moni, and I. H. Sarker, "An improved k-means clustering algorithm towards an efficient data-driven modeling," *Annals of Data Science*, pp. 1–20, 2022.

[28] H. Xie, L. Zhang, C. P. Lim, Y. Yu, C. Liu, H. Liu, and J. Walters, "Improving k-means clustering with enhanced firefly algorithms," *Applied Soft Computing*, vol. 84, p. 105763, 2019.

[29] R. Sagayam, S. Srinivasan, and S. Roshni, "A survey of text mining: Retrieval, extraction and indexing techniques," *International Journal of Computational Engineering Research*, vol. 2, no. 5, pp. 1443–1446, 2012.

[30] S. V. Gaikwad, A. Chaugule, and P. Patil, "Text mining methods and techniques," *International Journal of Computer Applications*, vol. 85, no. 17, 2014.

[31] J. K. Tarus, Z. Niu, and A. Yousif, "A hybrid knowledge-based recommender system for e-learning based on ontology and sequential pattern mining," *Future Generation Computer Systems*, vol. 72, pp. 37–48, 2017.

[32] J. Liu, X. Kong, X. Zhou, L. Wang, D. Zhang, I. Lee, B. Xu, and F. Xia, "Data mining and information retrieval in the 21st century: A bibliographic review," *Computer science review*, vol. 34, p. 100193, 2019.

[33] K. Thirugnanasambandam, R. Anitha, V. Enireddy, R. Raghav, D. K. Anguraj, and A. Arivunambi, "Pattern mining technique derived ant colony optimization for document information retrieval,"

Journal of Ambient Intelligence and Humanized Computing, pp. 1–13, 2021.

[34] W. Gan, J. C.-W. Lin, H.-C. Chao, H. Fujita, and S. Y. Philip, "Correlated utility-based pattern mining," *Information Sciences*, vol. 504, pp. 470–486, 2019.

[35] U. Yun, D. Kim, E. Yoon, and H. Fujita, "Damped window based high average utility pattern mining over data streams," *Knowledge-Based Systems*, vol. 144, pp. 188–205, 2018.

[36] M. A. Zingla, C. Latiri, P. Mulhem, C. Berrut, and Y. Slimani, "Hybrid query expansion model for text and microblog information retrieval," *Information Retrieval Journal*, vol. 21, pp. 337–367, 2018.

[37] A. Belhadi, Y. Djenouri, J. C.-W. Lin, C. Zhang, and A. Cano, "Exploring pattern mining algorithms for hashtag retrieval problem," *IEEE Access*, vol. 8, pp. 10 569–10 583, 2020.

[38] A. Babashzadeh, M. Daoud, and J. Huang, "Using semantic-based association rule mining for improving clinical text retrieval," in *Health Information Science: Second International Conference, HIS 2013, London, UK, March 25-27, 2013. Proceedings 2*. Springer, 2013, pp. 186–197.

[39] X. Cai and W. Li, "Ranking through clustering: An integrated approach to multi-document summarization," *IEEE transactions on audio, speech, and language processing*, vol. 21, no. 7, pp. 1424–1433, 2013.

[40] M. Sukanya and S. Biruntha, "Techniques on text mining," in *2012 ieee international conference on advanced communication control and computing technologies (icaccct)*. IEEE, 2012, pp. 269–271.

[41] D. Agnihotri, K. Verma, and P. Tripathi, "Pattern and cluster mining on text data," in *2014 fourth international conference on communication systems and network technologies*. IEEE, 2014, pp. 428–432.

[42] M. J. H. Mughal, "Data mining: Web data mining techniques, tools and algorithms: An overview," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 6, 2018.

[43] S. S. Tandel, A. Jamadar, and S. Dudugu, "A survey on text mining techniques," in *2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS)*. IEEE, 2019, pp. 1022–1026.

[44] 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.

[45] Y. Djenouri, A. Belhadi, D. Djenouri, and J. C.-W. Lin, "Cluster-based information retrieval using pattern mining," *Applied Intelligence*, vol. 51, pp. 1888–1903, 2021.

[46] A. Bokhabrine, I. Biskri, and N. Ghazzali, "New descriptors of textual records: getting help from frequent itemsets," *Vietnam Journal of Computer Science*, vol. 7, no. 04, pp. 355–372, 2020.

[47] C. E. Saili, S. Fatimi, and L. Alaoui, "Frequent itemsets methods for text clustering," in *2020 International Conference on Intelligent Systems and Computer Vision (ISCV)*. IEEE, 2020, pp. 1–5.

[48] M. Yarlagadda, K. G. Rao, and A. Srikrishna, "Frequent itemset-based feature selection and rider moth search algorithm for document clustering," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 4, pp. 1098–1109, 2022.

[49] O. Rouane, H. Belhadef, and M. Bouakkaz, "Combine clustering and frequent itemsets mining to enhance biomedical text summarization," *Expert Systems with Applications*, vol. 135, pp. 362–373, 2019.

[50] Y. Sato, K. Izui, T. Yamada, and S. Nishiwaki, "Data mining based on clustering and association rule analysis for knowledge discovery in multiobjective topology optimization," *Expert systems with applications*, vol. 119, pp. 247–261, 2019.

[51] J.-C. Kim and K. Chung, "Associative feature information extraction using text mining from health big data," *Wireless Personal Communications*, vol. 105, pp. 691–707, 2019.

[52] S. S. Waghere, P. RajaRajeswari, and V. Ganesan, "Retrieval of frequent itemset using improved mining algorithm in hadoop," in *International Conference on Innovative Computing and Communications: Proceedings of ICICC 2020, Volume 2*. Springer, 2021, pp. 787–798.

[53] R. Sovia, S. Defit, and N. Fatimah, "Determination potential experts by application the apriori algorithm and the k-means algorithm," *International Journal of Artificial Intelligence Research*, vol. 6, no. 1, 2022.

[54] K. Gayathri and R. Arunodhaya, "Customer segmentation and personalized marketing using k-means and apriori algorithm," in *Proceedings of the First International Conference on Combinatorial and Optimization, ICCAP 2021, December 7-8 2021, Chennai, India*, 2021.

[55] L. Kusak, F. B. Unel, A. Alptekin, M. O. Celik, and M. Yakar, "Apriori association rule and k-means clustering algorithms for interpretation of pre-event landslide areas and landslide inventory mapping," *Open Geosciences*, vol. 13, no. 1, pp. 1226–1244, 2021.

[56] H. Scells, S. Zhuang, and G. Zuccon, "Reduce, reuse, recycle: Green information retrieval research," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022, pp. 2825–2837.

[57] J. Lin, "A proposed conceptual framework for a representational approach to information retrieval," in *ACM SIGIR Forum*, vol. 55, no. 2. ACM New York, NY, USA, 2022, pp. 1–29.

[58] C. Sansone and G. Sperlí, "Legal information retrieval systems: State-of-the-art and open issues," *Information Systems*, vol. 106, p. 101967, 2022.

[59] H. K. Azad, A. Deepak, C. Chakraborty, and K. Abhishek, "Improving query expansion using pseudo-relevant web knowledge for information retrieval," *Pattern Recognition Letters*, vol. 158, pp. 148–156, 2022.

[60] H. Li, S. Zhuang, A. Mourad, X. Ma, J. Lin, and G. Zuccon, "Improving query representations for dense retrieval with pseudo relevance feedback: A reproducibility study," in *European Conference on Information Retrieval*. Springer, 2022, pp. 599–612.

[61] H. Corrigan-Gibbs, A. Henzinger, and D. Kogan, "Single-server private information retrieval with sublinear amortized time," in *Annual International Conference on the Theory and Applications of Cryptographic Techniques*. Springer, 2022, pp. 3–33.

[62] A. Henzinger, M. M. Hong, H. Corrigan-Gibbs, S. Meiklejohn, and V. Vaikuntanathan, "One server for the price of two: Simple and fast single-server private information retrieval," in *Usenix Security*, vol. 23, 2023.

[63] J. Tavares, P. Dutta, S. Dutta, and D. Samanta, *Cyber Intelligence and Information Retrieval*. Springer, 2022.

[64] H. Zamani, F. Diaz, M. Dehghani, D. Metzler, and M. Bendersky, "Retrieval-enhanced machine learning," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022, pp. 2875–2886.

[65] J. Rao, F. Wang, L. Ding, S. Qi, Y. Zhan, W. Liu, and D. Tao, "Where does the performance improvement come from? -a reproducibility concern about image-text retrieval," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022, pp. 2727–2737.

[66] Q. Xu, Y. Huang, S. Wu, and C. Nugent, "Clustering-based fusion for medical information retrieval," *Journal of Biomedical Informatics*, vol. 135, p. 104213, 2022.

[67] Y. Kim, J. Callan, J. S. Culpepper, and A. Moffat, "Efficient distributed selective search," *Information Retrieval Journal*, vol. 20, pp. 221–252, 2017.

[68] J.-P. Mei and L. Chen, "Proximity-based k-partitions clustering with ranking for document categorization and analysis," *Expert systems with applications*, vol. 41, no. 16, pp. 7095–7105, 2014.

[69] X. Jin, D. Agun, T. Yang, Q. Wu, Y. Shen, and S. Zhao, "Hybrid indexing for versioned document search with cluster-based retrieval," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, 2016, pp. 377–386.

[70] F. Raiber and O. Kurland, "Ranking document clusters using markov random fields," in *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, 2013, pp. 333–342.

[71] S. Chawla, "A novel approach of cluster based optimal ranking of clicked urls using genetic algorithm for effective personalized web search," *Applied Soft Computing*, vol. 46, pp. 90–103, 2016.

[72] J. He, Q. Ping, W. Lou, and C. Chen, "Paperpoles: Facilitating adaptive visual exploration of scientific publications by citation links," *Journal of the Association for Information Science and Technology*, vol. 70, no. 8, pp. 843–857, 2019.

[73] M. Yuan, J. Zobel, and P. Lin, "Measurement of clustering effectiveness for document collections," *Information Retrieval Journal*, vol. 25, no. 3, pp. 239–268, 2022.

[74] S. Amudha and I. Elizabeth Shanthi, "Phrase based information retrieval analysis in various search engines using machine learning algorithms," in *Data Management, Analytics and Innovation: Proceedings of ICDMAI 2019, Volume 2*. Springer, 2020, pp. 281–293.

[75] J. Rashid, S. M. A. Shah, and A. Irtaza, "An efficient topic modeling approach for text mining and information retrieval through k-means clustering," *Mehran University Research Journal of Engineering & Technology*, vol. 39, no. 1, pp. 213–222, 2020.

[76] 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)*. IEEE, 2020, pp. 862–866.

[77] Y. Djenouri, H. Drias, and Z. Habbas, "Bees swarm optimisation using multiple strategies for association rule mining," *International Journal of Bio-Inspired Computation*, vol. 6, no. 4, pp. 239–249, 2014.

[78] T. Ebesu and Y. Fang, "Neural semantic personalized ranking for item cold-start recommendation," *Information Retrieval Journal*, vol. 20, pp. 109–131, 2017.

[79] G. Fang, Y. Wu, M. Li, and J. Chen, "An efficient algorithm for mining frequent closed itemsets," *Informatica*, vol. 39, no. 1, 2015.

[80] N. Zhong, Y. Li, and S.-T. Wu, "Effective pattern discovery for text mining," *IEEE transactions on knowledge and data engineering*, vol. 24, no. 1, pp. 30–44, 2010.

[81] A. Patel and J. Shah, "Sensor-based activity recognition in the context of ambient assisted living systems: A review," *Journal of Ambient Intelligence and Smart Environments*, vol. 11, no. 4, pp. 301–322, 2019.

[82] S. Buttcher, C. L. Clarke, and G. V. Cormack, *Information retrieval: Implementing and evaluating search engines*. Mit Press, 2016.

[83] A. Patel and J. Shah, "Real-time human behaviour monitoring using hybrid ambient assisted living framework," *Journal of Reliable Intelligent Environments*, vol. 6, pp. 95–106, 2020.

[84] G. González-Sáez, P. Galuscáková, R. Deveaud, L. Goeuriot, and P. Mulhem, "Exploratory visualization tool for the continuous evaluation of information retrieval systems," in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023, pp. 3220–3224.

[85] P. Hu, L. Zhen, D. Peng, and P. Liu, "Scalable deep multimodal learning for cross-modal retrieval," in *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, 2019, pp. 635–644.

[86] R. Zhao, H. Chen, W. Wang, F. Jiao, X. L. Do, C. Qin, B. Ding, X. Guo, M. Li, X. Li *et al.*, "Retrieving multimodal information for augmented generation: A survey," *arXiv preprint arXiv:2303.10868*, 2023.

[87] S. Li, A. Karatzoglou, and C. Gentile, "Collaborative filtering bandits," in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 2016, pp. 539–548.

[88] Y. Fan, X. Xie, Y. Cai, J. Chen, X. Ma, X. Li, R. Zhang, J. Guo *et al.*, "Pre-training methods in information retrieval," *Foundations and Trends® in Information Retrieval*, vol. 16, no. 3, pp. 178–317, 2022.

[89] A. Davidson, G. Pestana, and S. Celi, "Frodopir: Simple, scalable, single-server private information retrieval," *Cryptology ePrint Archive*, 2022.

[90] P. A. Kalyankar, A. O. Mulani, S. P. Thigale, P. G. Chavhan, and M. M. Jadhav, "Scalable face image retrieval using aesc technique," *Journal Of Algebraic Statistics*, vol. 13, no. 3, pp. 173–176, 2022.

[91] P. Yang, H. Fang, and J. Lin, "Anserini: Enabling the use of lucene

for information retrieval research," in *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, 2017, pp. 1253–1256.

[92] P. Fournier-Viger, W. Gan, Y. Wu, M. Nouioua, W. Song, T. Truong, and H. Duong, "Pattern mining: Current challenges and opportunities," in *International Conference on Database Systems for Advanced Applications*. Springer, 2022, pp. 34–49.

[93] M. Anandarajan, C. Hill, T. Nolan, M. Anandarajan, C. Hill, and T. Nolan, "Text preprocessing," *Practical text analytics: Maximizing the value of text data*, pp. 45–59, 2019.

[94] K. Maharana, S. Mondal, and B. Nemade, "A review: Data preprocessing and data augmentation techniques," *Global Transitions Proceedings*, vol. 3, no. 1, pp. 91–99, 2022.

[95] X. Guo, "Wireless network health information retrieval method based on data mining algorithm." *Journal of Information Processing Systems*, vol. 19, no. 2, 2023.

[96] T. Sarwar, S. Seifollahi, J. Chan, X. Zhang, V. Aksakalli, I. Hudson, K. Verspoor, and L. Cavedon, "The secondary use of electronic health records for data mining: Data characteristics and challenges," *ACM Computing Surveys (CSUR)*, vol. 55, no. 2, pp. 1–40, 2022.

[97] J. Guo, Y. Fan, L. Pang, L. Yang, Q. Ai, H. Zamani, C. Wu, W. B. Croft, and X. Cheng, "A deep look into neural ranking models for information retrieval," *Information Processing & Management*, vol. 57, no. 6, p. 102067, 2020.

[98] M. Zehlike, K. Yang, and J. Stoyanovich, "Fairness in ranking, part ii: Learning-to-rank and recommender systems," *ACM Computing Surveys*, vol. 55, no. 6, pp. 1–41, 2022.

[99] Z. Dai and J. Callan, "Context-aware term weighting for first stage passage retrieval," in *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, 2020, pp. 1533–1536.

[100] Y. Yang, C. Huang, L. Xia, and C. Li, "Knowledge graph contrastive learning for recommendation," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022, pp. 1434–1443.

[101] S. Vithana, K. Banawan, and S. Ulukus, "Semantic private information retrieval," *IEEE Transactions on Information Theory*, vol. 68, no. 4, pp. 2635–2652, 2021.

[102] L. Ismail, D. Niyato, S. Sun, D. I. Kim, M. Erol-Kantarci, and C. Miao, "Semantic information market for the metaverse: An auction based approach," in *2022 IEEE Future Networks World Forum (FNWF)*. IEEE, 2022, pp. 628–633.

[103] S. Halder, K. H. Lim, J. Chan, and X. Zhang, "Efficient itinerary recommendation via personalized poi selection and pruning," *Knowledge and Information Systems*, vol. 64, no. 4, pp. 963–993, 2022.

[104] X. Wang, Z. Zheng, Y. He, F. Yan, Z. Zeng, and Y. Yang, "Progressive local filter pruning for image retrieval acceleration," *IEEE Transactions on Multimedia*, 2023.

[105] A. Patel and J. Shah, "Towards enhancing the health standards of elderly: role of ambient sensors and user perspective," *International Journal of Engineering Systems Modelling and Simulation*, vol. 13, no. 1, pp. 96–110, 2022.

[106] A. Chugh, V. K. Sharma, S. Kumar, A. Nayyar, B. Qureshi, M. K. Bhatia, and C. Jain, "Spider monkey crow optimization algorithm with deep learning for sentiment classification and information retrieval," *IEEE Access*, vol. 9, pp. 24 249–24 262, 2021.

[107] A. Breit, L. Waltersdorfer, F. J. Ekaputra, M. Sabou, A. Ekelhart, A. Iana, H. Paulheim, J. Portisch, A. Revenko, A. t. Teije *et al.*, "Combining machine learning and semantic web: A systematic mapping study," *ACM Computing Surveys*, 2023.

[108] Y. Li, H. Chen, S. Xu, Y. Ge, and Y. Zhang, "Towards personalized fairness based on causal notion," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 1054–1063.

[109] S. Heller, V. Gsteiger, W. Bailer, C. Gurrin, B. . Jónsson, J. Lokoč, A. Leibetseder, F. Mejzlík, L. Peška, L. Rossetto *et al.*, "Interactive video retrieval evaluation at a distance: comparing sixteen interactive video search systems in a remote setting at the 10th video browser showdown," *International Journal of Multimedia Information Retrieval*, vol. 11, no. 1, pp. 1–18, 2022.

[110] J. Rodrigues, H. Liu, D. Folgado, D. Belo, T. Schultz, and H. Gamboa, "Feature-based information retrieval of multimodal biosignals with a self-similarity matrix: Focus on automatic segmentation," *Biosensors*, vol. 12, no. 12, p. 1182, 2022.

[111] R. Litschko, I. Vulić, S. P. Ponzetto, and G. Glavaš, "On cross-lingual retrieval with multilingual text encoders," *Information Retrieval Journal*, vol. 25, no. 2, pp. 149–183, 2022.

[112] R. Piening, K. Pfeuffer, A. Esteves, T. Mittermeier, S. Prange, P. Schröder, and F. Alt, "Looking for info: Evaluation of gaze based information retrieval in augmented reality," in *Human-Computer Interaction–INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30–September 3, 2021, Proceedings, Part I 18*. Springer, 2021, pp. 544–565.

[113] D. Minh, H. X. Wang, Y. F. Li, and T. N. Nguyen, "Explainable artificial intelligence: a comprehensive review," *Artificial Intelligence Review*, pp. 1–66, 2022.

[114] A. Ali, M. F. Pasha, O. H. Fang, R. Khan, M. A. Almaiah, and A. K. Al Hwaitat, "Big data based smart blockchain for information retrieval in privacy-preserving healthcare system," in *Big Data Intelligence for Smart Applications*, 2022, pp. 279–296.

[115] A. Sharma and S. Kumar, "Machine learning and ontology-based novel semantic document indexing for information retrieval," *Computers & Industrial Engineering*, vol. 176, p. 108940, 2023.

Vaishali Patel is a Ph.D. Scholar of Computer Engineering at Pacific University, Rajasthan. She has done her Bachelor of Engineering from DDIT, Nadiad, Gujarat and Master of Technology (Research) from SVNIT, Surat. Her area of interest includes Machine Learning, Artificial Intelligence and Information Retrieval.

Dilendra Hiran has completed his Ph.D. in Computer Science and working as a director of Faculty of Computer Application of Pacific Academic Higher Education and Research University, Udaipur, Rajasthan. He is having more than 23 years of working experience and published many research papers.

Kruti Dangarwala Dr kruti dangarwala has completed Ph.D in computer engineering. She is working as a Head of Department of Computer Science Engineering and Information Technology. She is working at SVMIT for 21 years. She has published many research papers on machine learning and image processing area.