Industrial Engineering Journal
ISSN: 0970-2555
Volume : 54, Issue 9, No.1, September : 2025

A HYBRID APPROACH FOR EFFICIENT DOCUMENT INFORMATION RETRIEVAL USING HIERARCHICAL K-MEANS CLUSTERING AND FP-GROWTH PATTERN MINING

Vaishali Patel and Dilendra Hiran Department of Computer Engineering, PAHER University, Udaipur, India : vaishalirajpatel@gmail.com, sigmapawan72@gmail.com
Kruti Dangarwala Department of Computer Science and Engineering, SVM Institute of Technology, Bharuch, : krutidangarwala@gmail.com

Abstract: The rapid increase in digital information presents several significant challenges in finding accurate and relevant documents. Traditional keyword-based methods and many research approaches for retrieving information, including cluster-based, pattern-based, and the combination of clustering with pattern mining, often fail to provide relevant answers to user queries. This leads to information overload and lower retrieval accuracy. To address the limitations, this research introduces a hybrid combined approach hierarchical document clustering (k-Means) with frequent pattern growth (FP-Growth) mining to enhance the effectiveness of information retrieval. This method has two phases: the first phase consists of offline processing to create an index, while the second phase involves online processing to retrieve queries. In the first phase, documents are preprocessed and transformed into vector representations using the vector space model. They are then organized into a hierarchical structure through recursive k-Means clustering. Within each leaf-level cluster, FP-Growth generates frequent simultaneous terms and semantic patterns. In the second phase, the query is preprocessed and expanded using patterns extracted from the most relevant cluster. This enables a more refined search and better relevance ranking. Experiments conducted on the 20Newsgroups document dataset for user query “NASA space exploration missions” in the sci.space category show that the proposed hybrid approach significantly improves precision, recall, and F1-score compared to traditional keyword-based methods, cluster-only methods, pattern-only methods, and k-Means with pattern mining models. The results indicate that the proposed approach consistently outperforms baseline methods across various newsgroup categories, achieving higher retrieval accuracy while maintaining good runtime efficiency. This validates the effectiveness of combining hierarchical clustering with pattern mining to close the semantic gap and provide more context-aware, intelligent information retrieval systems.

Keywords— Information Retrieval, Hierarchical Document Clustering, k-Means, Frequent Pattern Mining, FP-Growth, Query Expansion, Performance Measures, 20 Newsgroups

Section I: Introduction

The large amount of unstructured text data comes from many sources, including news articles, social media, research publications, and e-commerce applications. This rapid growth of digital content creates both opportunities and challenges in the field of information retrieval. Finding relevant information has become increasingly important in today's information age. Information retrieval systems vary from simple search engines to complex semantic search approaches and an important step to navigate the large amounts of data generated[1]. Traditional IR approaches[2], such as Boolean retrieval and vector space models that use TF-IDF, have been the foundation of retrieval systems for decades. However, these methods have significant limitations. Different words can express the same concept, leading to incomplete documents, which is known as a lexical gap. These measures treat each word independently, ignoring the semantic connections between words. This can create misunderstandings of the same concepts. A single term may have multiple meanings based on the context, resulting in uncertainty in the retrieval results. These limitations make it hard to provide information retrieval that is both accurate and aware of context, especially in large and diverse

document collections that need to meet user queries. While these approaches are computationally efficient, they often do not capture the semantic relationships between terms or the user's intent, leading to information overload, low precision, and decreased user satisfaction.

The difficulties of modified information retrieval approaches have not been adequately addressed by relying solely on conventional keyword-matching techniques[3]. Although keyword-based models are computationally straightforward, they frequently fall short in capturing the contextual relationships between terms in documents or the semantic meaning of user queries. This results in information overload and decreased retrieval precision as these systems often return large sets of documents, many of which may be irrelevant. Recent research has moved toward more intelligent retrieval techniques that reveal hidden patterns between documents and queries in order to overcome these drawbacks.

Integration of document clustering with frequent pattern mining, retrieval systems are able to leverage the benefits of both techniques. Clustering[4] reduces the search space to only the most relevant subset of documents and frequent pattern mining[5] can enhance queries using contextual knowledge. As not only does this hybrid approach yield a more meaningful and robust retrieval system as it is informed semantically, but it also results in an improved retrieval system that overcomes the limitations of traditional IR approaches. Although studies for both methods have been done separately, the hybrid work for accurate and effective information retrieval has not been extensively researched. Document clustering can be quite challenging and complex nowadays due to the high dimensionality and vast size document data generation. Thus, the focus of researchers is to provide enhanced and hybrid document based clustering algorithms, often cited as new directions. One approach advocates generating optimized clusters for improved information retrieval[6]. The literature review analyzed the need for better generation model directions[7]. This research demonstrates a hybrid approach to frequent pattern mining with document clustering with hierarchical k-Means algorithm, and FP Growth for frequent pattern mining by breaking the approach into two distinct phases for improvement and relevant information retrieval. The hybrid technique of clustering with pattern mining can enable collaboration in:

- Cluster-based retrieval, where information is retrieved on the ranked[8] query that matches user query
- Improved efficiency of clustering using data-mining patterns that will reduce the documents evaluated for each query.
- Frequent term patterns can also demonstrate the semantic relationship
- Patterns can be helpful for query expansion[9], adding related terms to a query improve recall while not damaging the precision quality.

Motivation: The limitations of keyword-matching approaches have driven research toward intelligent retrieval techniques that can uncover and unseen relationships between documents and queries. Two promising directions are:

- Document Clustering groups similar documents together. This reduces the search space and improves retrieval efficiency. Among clustering techniques, k-Means and its hierarchical versions perform well with high-dimensional text data.
- Frequent Pattern Mining finds term or item sets that appear often in documents with contextual links. These patterns can be used to enhance user queries through query expansion, which improves recall and precision**.**

Contribution: As per our study and analysis, this research is proposing a hybrid approach to overcome the limitations of traditional information retrieval methods. The key contributions can be summarized as follows:

- Hybrid IR Approach: A hybrid information retrieval approach is presented, which includes hierarchical k-means document clustering and frequent pattern mining works in two phases: offline processing (index construction) and online processing (query retrieval).
- Hierarchical Document Clustering: This research developed a hierarchical clustering approach using recursive k-means document clustering, in which documents are put into a multi-layer classification in that documents are put into a hierarchy that captures both productive and local semantic structures of the data.
- Pattern-Based Query Expansion: This research also imposed a method for the user query expansion using the pattering mining FP-Growth algorithm to improve the recall.
- Experimental evaluation: This approach is evaluated by experiments on the 20Newsgroups dataset, and with our chosen cluster feature sets and retailed very good comparative results between proposed hybrid approach and traditional keyword-based approach, cluster approach, and base line hybrid approach.
- Effectiveness and efficiency: Results analysis showed that this proposed approach generates improved evaluation measures.

Outline: The remainder of this paper is organized as follows: Section 2 summarizes related work including traditional approaches for information retrieval, cluster based retrieval, and pattern mining approaches, and hybrid approaches of clustering with pattern mining. Section 3 introduces the proposed hybrid approach with its workflow, phases and algorithm in detail Section 4 presents the experimental set up and performance measures. Section 5 presents discussion and result analysis to demonstrate the effectiveness of the work. At the end, Section 6 discusses conclusion and future enhancement.

Section 2: Related Work

IR using traditional models: Gudivada et al.[10] discusses three Information Retrieval (IR) models: Vector Space, which provides a vector representation of documents and queries; Probabilistic, which provides an estimate of the probability of relevance; and Language models, which focus on generating text. In this case, the paper also mentions term weighting schemes for effective retrieval, such as TF-IDF. Al-rassam et al.[11] investigated information retrieval from the vector space model by comparing features like tf.idf and term frequency. The results indicated that tf.idf, stop words, and stemming improve system effectiveness when compared to traditional tf which does not incorporate these features. Authors[12] aims to improve information retrieval by using a combination of Word2Vec and TFIDF modelling in order to address issue in feature dimension and increase semantic understanding for better article text classification in comparison to traditional models. The Word2vec model does not have essentiality variability in the different texts. Traditional text classification models also have a curse of feature dimension. Jabri et al.[13] illustrates information retrieval in the vector space model using TF-IDF weighting to obtain ranked documents. It emphasizes the limitations of conventional models and presents a new ranking measure that combines association rules to improve the relevance of documents retrieved. Saha et al.[14] hybridized a traditional method of keyword matching, with document retrieval bag-of-embedding model shows that this method is aggressive with large transformer models, whilst allowing for improved performance on relevant tasks across a range of information retrieval tasks.

IR using document clustering: Patel V et al[15] presents the approach of document clustering to facilitate information retrieval by gathering similar documents together, which allows the user to retrieve relevant information more effectively based on their query. Yuan et al.[16] raises an interesting point regarding the possibility of no meaningful outcomes arising from clustering owing to increased data dimensionality, and if, most often, clustering returns unrelated outputs, one might reasonably question the statistical soundness of claims about the effectiveness of clustering.

Cozzolino et al.[4] enhances information retrieval and provides better precision, recall, and retrieval efficiency by clustering similar documents. On the downside, prototype-based clustering is sensitive to choices made by the user and model-based clustering. Authors[17] organize text into topics, assisting users in finding information quickly. But computationally expensive and sensitive to the choice of cluster. Haji[18] enhances information retrieval because it groups documents based on semantic similarities rather than just keywords, but don't always represent the full richness of the text collections used for clustering.

IR using patterning mining: Authos[19] extracts meaningful patterns from structured and unstructured data to improve information retrieval and address challenges in navigating the vast web. He et al.[20] introduces a category pattern mining strategy for image retrieval, enhancing traditional methods by extending individual query images to entire categories, thereby improving retrieval effectiveness. Zhou et al.[21] discusses pattern matching information retrieval approach, specifically in question-answer system for computer textbook. Authors[22] proposed and data mining application to prove the effectiveness of retrieval and analyzes its application value in internet information retrieval. Singh et al.[23] proposed a technique within text mining that identifies trends and relationships in large datasets, enhancing information retrieval by uncovering latent topics, sentiment, and key terms.

IR using Document clustering with pattern mining: Authors[24] analyzes frequent item sets and association rules to enhance library management, but does not specifically address document clustering or information retrieval techniques. Ren et al.[25] proposed deep document clustering model to improve the document clustering performance by adaptively learning a hybrid representation. Authors[26] focuses on a clustering-based frequent pattern mining approach for recommender systems to alleviate the cold-start problem. Bascur et al.[27] focuses on citation-based clusters for information retrieval tasks, evaluating their performance in finding relevant documents for systematic reviews without mentioning FP growth techniques. Authors[28] focuses on a clustering approach for semantic information retrieval in long documents,.

Research GAP: Based on the study and analysis of literature survey for information retrieval, this work finds the following research gap,

- Flat clustering methods (e.g., K-Means) reduce search space but fail to capture hierarchical topic structures and fine-grained semantics. FP-Growth identifies frequent term associations but is inefficient for large-scale retrieval when used alone.
- Few studies combine clustering with pattern mining; most ignore hierarchical structures and rarely use mined patterns for query expansion. hierarchical clustering and pattern mining
- A hybrid approach combining can overcome these gaps, improving retrieval efficiency and effectiveness.

Section 3: Proposed Approach

This section introduces the hybrid information retrieval approach which integrates hierarchical document clustering and frequent pattern mining to provide the solutions against of issues of traditional methods and effectiveness in retrieval over large document datasets. A hybrid information retrieval approach consists of two phases: first, offline processing for index construction and that consists of: data preprocessing and feature extraction: tokenization, stop-word removal, stemming, and document victimization using TF-IDF, document clustering and frequent pattern mining. Document clustering is generated using hierarchical k-Means with cosine similarity to generate documents into semantically meaningful groups. Frequent pattern within clusters are generated using FP-Growth to find frequent co-occurrences of each term in the generated. The second phase is online processing for query retrieval which includes: query processing, cluster selection, query expansion and document retrieval-ranking. The expansion of user queries will be derived from the generated

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 54, Issue 9, No.1, September : 2025

frequent patterns. TF-IDF relevance score and pattern-support measures are combined to produce an overall document ranking. Figure 1 depicts the process of the hybrid approach of hierarchical document clustering with frequent pattern mining for effective Information retrieval.

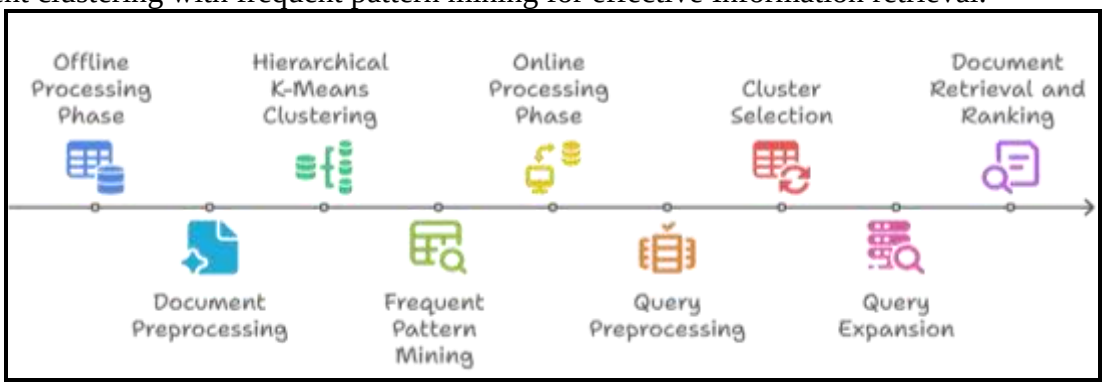

Fig 1: Overview of Hybrid Approach: Hierarchical Document Clustering with Pattern Mining for IR

Phase 1-Offline Processing (Index Creation):

In this step, the document collection is organized for retrieval. This includes the domain-wise document collection and pre-processing, followed by document clustering with hierarchical k-Means, generating frequent patterns with FP growth using a pattern-based cluster construction for relevant information retrieval on a user query.

Document Collection and Preprocessing: This step collects the 20Newsgroup[29] document dataset from kaggle which is an open dataset. This dataset is the most popular dataset for document based machine learning and information retrieval applications. It is around 18,000 newsgroup documents with 20 different news categories in roughly similar amounts. The data collection step applies two methods to perform the cleaning and filtering of the data. Pre-processing step includes the steps of splitting the text into tokens, removing stop words, stemming, and vector representation.

For example, Given set of document D={d1,d2,…,dn} each document is processed as follows:

- Tokenization: The process of separating a document into separate tokens (words/phrases). Example: "Document clustering pattern mining important" → [Document, clustering, pattern, mining, important].
- Stop word Removal: The removal of semantically irrelevant words (e.g. "are", "and", "is").
- Stemming/Lemmatization: The act of reverting words back to their base form. Example: "clustering", "clustered" → "cluster"
- Vector Representation: the transformation of unstructured text into structured, where each document is mapped to vector space and every dimension mapped to a feature (derived from the text. e.g., words). For Example, TF–IDF assigns weight to words by its important and included by common terms. Result is a generated preprocessed document vectors.

Document clustering using Hierarchical k-Means: Dataset is divided into clusters C={C1,C2,…,Cn}, which are minimizing respective elements of different clusters. Clustering is chosen to minimize the common repeated terms among the created clusters in that each cluster shares a non-common term. Such clustering algorithms that are applied including k-Means[30], spectral clustering[31], and DBSCAN[32]. This work applies the multi level hierarchical k-Means document clustering using cosine similarity as a distance measure to cluster documents based on similarity in terms present in each document also minimize the variation within clusters, with respect to centroids. Cosine similarity works by computing the angle to find the directional difference between the vectors for the two documents. In the process, mean of centroid position will be recalculated every time a document is assigned for every iteration of all the documents in the clusters until the clusters have minimal overlap and optimal internal cohesion for the clusters. In this particular step, the hierarchical k-Means clustering is applied to the processed document vectors.

Frequent Pattern using FP-Growth: FP growth algorithm is applied to each generated cluster to find valuable associations between terms in documents. In Boolean and weighted transformation when creating the transactional database from the clusters, each cluster is turned into a Boolean transactional database and transformed in a transactional database for the frequent and high utility patterns, respectively. In Boolean transformation, a transactional database is created from the cluster where each object is taken from the cluster and mapped to a transaction with Boolean values for each term where 1 = present and 0= absent. On each cluster, FP-growth algorithm [33] is applied to the documents in each leaf cluster getting all patterns generated using the a minimum support value and confirm frequency of terms in the clusters. In weighted transformation each term is taken as an item and it is given weight, where the internal utility of each item is term frequency and external utility indicates the equality of clusters having value of 0. Sometimes, UP growth algorithm [34] can apply to retrieve high utility patterns.

Phase 2: Online Processing (Query Retrieval):

This phase handles user queries and retrieves relevant documents. It involves the steps to fulfill the objective of relevant terms retrieval: query preprocessing, cluster selection, query expansion, and document Retrieval and Ranking [35].

Query Preprocessing: Query processing [36] is used to analyze, modify and interpret the user query to enable the retrieval system to effectively match the query against indexed documents. It ensures that the query accurately describes the user intent and is able to rapidly fetch relevant documents. It starts with query input, which is typically in natural language. Queries can take many forms, from short keyword queries such as " NASA missions" to longer descriptive expressions such as "space exploration missions by NASA." Once the query has been submitted, query preprocessing (like document preprocessing) clean and prepare the query terms so they can be effectively retrieved. First, tokenization is performed, where the query is divided into individual terms or tokens; for instance, the input “NASA space missions” is transformed into the set {NASA, space, missions}. Then, stop words removal and stemming are performed, in which the word(s) are reduced to their root.

C**luster Selection:** The preprocessed query vector is considered as an input and compared against the hierarchical clustering structure. The similarity between the preprocessed query vector and the centroid is computed at the level of the hierarchy. Once the similarity computation has been performed, the most similar cluster is selected for further processing based on the highest similarity value.

Q**uery Expansion:** It combines selected cluster from the above step (preprocessed query vector, frequent patterns) to improve the expressiveness s of the query. The frequent patterns in the cluster are examined to find those that share overlapping terms with the original query. From this cluster, new terms are added into the query to improve richness. The new terms are weighted based on the support of the pattern and the semantic relevance to the query terms. An expanded query vector results in a better recall, by encompassing related concepts that may not have explicitly appeared in the original query. Query expansion refines and enriches the user query for better retrieval performance. For example, “space exploration” may be expanded to include semantically related terms such as NASA, mission, planets, rover, and discovery. This step enhances recall by including documents that may use different but related terminology.

Document Retrieval and Ranking: Lastly, this step makes use of the expanded query vector and compares it to the document vectors that fall within the specific cluster being searched. Using similarity scores calculated for each document, the document vectors representing their respective topics were ranked in declining order of similarity[35]. As a result of this step, a set of ranked documents is produced: although irrelevant results may not have perfect accuracy, they will be both

relevant and comprehensive. The search process[37] begins with the top cluster ranked on the user query and can extended to using other machine learning[38] or clustering techniques.
Proposed algorithm for hierarchical k-Means document clustering FP growth pattern mining is as below with two phases: Index construction and query retrieval.

Algorithm: Hybrid Approach of Document Hierarchical Clustering with Pattern Mining for IR Hybrid_HKMeans_FP(D, Q, k, d, min_supp):	
1	**Input:** Document set D={d_1,d_2,…,d_n} , User query Q, Number of clusters k, hierarchy depth d, Minimum support threshold minsup for frequent pattern mining **Output:** Ranked list of top N retrieved documents relevant to Q
PHASE I: Offline Processing (Query Construction)	
2	Document Preprocessing: • For each document di∈D: - Remove stop-words and punctuation, Convert to lowercase - Apply stemming/lemmatization - Represent each document di as TF–IDF vectors • Store all vectors in set V
3	Document Clustering (Hierarchical k-Means) • Apply recursive k-Means clustering on V up to depth d. • Construct hierarchical cluster structure H
4	Frequent Pattern Mining (FP-Growth) For each leaf-level cluster in H: • Apply FP-Growth with threshold min_supp. • Store frequent patterns FP
PHASE 2: Online Processing (query Retrieval)	
5	Query Preprocessing Preprocess the query Q (same steps as document processing in step 2). Convert Q into a TF–IDF vector. • Tokenize query Q, remove stop-words, apply stemming • Represent as TF–IDF vector Qv
6	Cluster Selection • f or each leaf-level cluster in H Compute cosine similarity between Q and each cluster centroid • Select cluster $C_{selected}$ with most simiar
7	Query Expansion Using Frequent Patterns • Retrieve frequent patterns FP from $C_{selected.}$ • For each pattern containing query terms: - Identify patterns containing one or more query terms - Add the associated terms from these patterns to Q to form the expanded query Q′ • Construct expanded query vector Qe.
8	Document Retrieval and Ranking

- For each document in cluster $C_{selected}$ using Q:
  - Compute cosine similarity between Qe and document vector
- Rank documents in descending order of similarity
- Return ranked list R

9 Output
Return ranked list of top-N retrieved documents to the user

These algorithms provide a detailed overview of the hybrid document retrieval system. The combination of hierarchical clustering, frequent pattern mining, and query expansion aims to improve both the relevance and efficiency of document retrieval.

Section 4: Experimental Setup: Dataset and Performance Measures

This section explains the datasets, implementation details, evaluation metrics, baseline systems, and experimental evaluation parameters for proposed hybrid information retrieval (IR) approach using 20Newgroups document dataset which contains approximately 18,846 newsgroup posts divided into 20 different topics in subjects like politics, sports, technology, religion and many with approximately 900 posts per category.

Performance Measures: Proposed hybrid information retrieval approach is evaluated using three traditional evaluation measures of effectiveness - Precision, Recall, and F1-score. These measures are fundamental in assessing the quality of retrieved documents, as they offer a balanced evaluation of system performance. Precision is defined as the number of relevant documents retrieved divided by the total number of retrieved documents, meaning precision represents the ratio of relevant documents out of all retrieved documents. In the context of information retrieval[39], high precision means the system is effectively filtering unwanted results and providing a user with relevant information. Recall is measured by dividing retrieved number of relevant documents divided by total number of existing relevant documents meaning recall is the ratio of relevant documents covered by the system out of all relevant documents found in an entire collection of documents. Higher recall is desirable as it means the system has a relatively high likelihood of covering most of the relevant information available and minimizing chances of important documents being left unfiltered by the retrieving system.F1-score is the harmonic mean of precision and recall that offers to achieve a balance between the two measures in assessing retrieval performance. F1-score becomes important whenever there is some trade-off between precision and recall, as it provides.

Section 5: Result Analysis and Discussion

The experimental result analysis is demonstrated the effectiveness of the proposed hybrid approach, hierarchical k-Means clustering with FP-Growth pattern mining, to specify information retrieval which is shown in Table 1. The proposed hybrid method addressed two critical issues in information retrieval. These issues include reducing irrelevant document retrieval or precision improvement, and increasing coverage of semantically related documents or recall improvement. This work clustered in a multi-level semantic structure and reduced the search space at query time and retrieved more contextual documents as semantically relevant candidates. As a result, proposed the retrieval system had a greater precision and F-score than baseline TF–IDF, k-Means. FP-Growth, and k-Means with. For the query *"*NASA space exploration missions" in the *sci.space* category, TF–IDF retrieved many documents but included unrelated "space" contexts, while k-Means and FP-Growth alone performed poor, k-Means combining with FP-Growth generates more balanced and improved relevance through query expansion. The proposed Hierarchical k-Means with FP-Growth approach achieved the best performance, retrieving top 10 documents specially related to term NASA missions such as Mars exploration, shuttle programs, and satellite launches and it can be be useful for large, heterogeneous document collections where semantic clustering and vocabulary expansion.

Table 1: Comparison of Performance Analysis of IR approaches with Proposed Hybrid Approach

IR Approaches	Precision	Recall	F-Measure
Baseline TF-IDF Traditional	0.4514	0.566	0.569
k-Means clustering	0.388	0.314	0.311
FP Growth pattern mining	0474	0.438	0.437
k-means with FP-Growth	0.558	0.550	0.550
Hierarchical kMeans and FP Growth	0.688	0.607	0.604

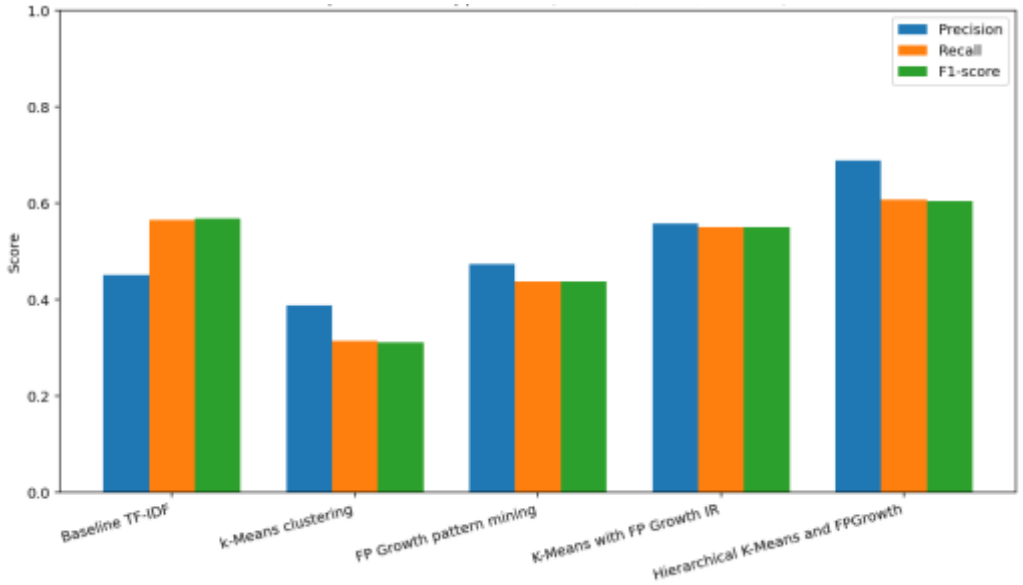

Fig 2: Comparison of Performance Analysis of IR approaches with Proposed Hybrid Approach

Section 6: Conclusion and Future Work

The study illustrated a hybrid information retrieval approach that combines hierarchical k-Means document clustering with FP-Growth pattern mining, using the 20Newsgroups dataset for evaluation. The hierarchical clustering aspect grouped documents in a semantically meaningful way, reducing the search space, and provided filtering of non-relevant documents thereby increasing the precision. The FP-growth aspect of the experiment mined frequent term associations in the clusters and those associations were then used for query expansion, addressing the vocabulary mismatch issue and significantly improving recall. The results from the experiments on 20Newsgroup dataset for user query "NASA space exploration missions" in the *sci.space* category show that while TF–IDF retrieval was poor at addressing the semantic gaps, retrieval with clustering only had very poor recall, and the hybrid approach is able to consistently achieve the highest F1-score, demonstrating a better balance of precision and recall. In conclusion, the study showed that clustering and pattern mining techniques can complement each other and establish a combined synergy, thereby making the hybrid or combination model more robust and effective than each technique when used alone. It is especially effective for large heterogeneous text corpuses data when overlaps in topics and synonymy create possibilities for conventional retrieval methods to fail.

For future applications, approach can be experimented on more document datasets with various user queries. Approach can be enhanced by integrating deep learning or enhanced machine learning algorithms[40] like Word2Vec, BERT to capture richer contextual meaning beyond term co-occurrence. This would improve hierarchical clustering, yield more meaningful frequent patterns, and enable smarter query expansion, ultimately addressing issues like synonymy and the semantic gap for more effective information retrieval.

References:

1 Hambarde, K. A., & Proenca, H. (2023). Information retrieval: recent advances and beyond. IEEE Access, 11, 76581-76604.
2 Ai, Q., Bai, T., Cao, Z., Chang, Y., Chen, J., Chen, Z., ... & Zhu, X. (2023). Information retrieval meets large language models: a strategic report from chinese ir community. *AI open*, *4*, 80-90.
3 Bouhini, C., Géry, M., & Largeron, C. (2016, June). Personalized information retrieval models integrating the user's profile. In *2016 IEEE Tenth International Conference on Research Challenges in Information Science (RCIS)* (pp. 1-9). IEEE.
4 Cozzolino, I., & Ferraro, M. B. (2022). Document clustering. Wiley Interdisciplinary Reviews: Computational Statistics, 14(6), e1588.
5 Fournier-Viger, P., Gan, W., Wu, Y., Nouioua, M., Song, W., Truong, T., & Duong, H. (2022, April). Pattern mining: Current challenges and opportunities. In *International Conference on Database Systems for Advanced Applications* (pp. 34-49). Cham: Springer International Publishing.
6 Anand, S. K., & Kumar, S. (2022). Experimental comparisons of clustering approaches for data representation. *ACM Computing Surveys (CSUR)*, *55*(3), 1-33.
7 Kathiria, P., Pandya, V., Arolkar, H., & Patel, U. (2023, May). Performance analysis of document similarity-based dbscan and k-means clustering on text datasets. In *Proceedings of International Conference on Recent Innovations in Computing: ICRIC 2022, Volume 1* (pp. 57-69). Singapore: Springer Nature Singapore
8 Shah, R., Sheng, C., Thankachan, S., & Vitter, J. (2023). Ranked Document Retrieval in External Memory. ACM Transactions on Algorithms, 19(1), 1-12.
9 Afuan, L., Ashari, A., & Suyanto, Y. (2021). A new approach in query expansion methods for improving information retrieval. *JUITA: Jurnal Informatika*, *9*(1), 93-103.
10 Gudivada, V. N., Rao, D. L., & Gudivada, A. R. (2018). Information retrieval: concepts, models, and systems. In Handbook of statistics (Vol. 38, pp. 331-401). Elsevier.
11 Al-rassam, O., Amin, M. H. S. M., & Faeq, Z. S. (2021). Performance evaluation of information retrieval system using vector space model: a comparative analysis. Iraqi journal for computers and informatics, 47(2), 6-9.
12 Wang, R., & Shi, Y. (2022, February). Research on application of article recommendation algorithm based on Word2Vec and Tfidf. In 2022 IEEE International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA) (pp. 454-457). IEEE.
13 Jabri, S., Dahbi, A., Gadi, T., & Bassir, A. (2018, April). Ranking of text documents using TF-IDF weighting and association rules mining. In 2018 4th international conference on optimization and applications (ICOA) (pp. 1-6). IEEE.
14 Saha, A., Hassanzadeh, O., Gittens, A., Ni, J., Srinivas, K., & Yener, B. (2023). Improving Neural Ranking Models with Traditional IR Methods. arXiv preprint arXiv:2308.15027.
15 Patel, V., Hiran, D., & Dangarwala, K. (2024, October). Assessing Document Clustering Algorithms for Effective Information Retrieval. In 2024 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC) (pp. 952-956). IEEE.
16 Yuan, M., Zobel, J., & Lin, P. (2022). Measurement of clustering effectiveness for document collections. Information Retrieval Journal, 25(3), 239-268.
17 Amiri, B., & Karimianghadim, Cozzolino, I., & Ferraro, M. B. (2022). Document clustering. Wiley Interdisciplinary Reviews: Computational Statistics, 14(6), e1588.R. (2024). A novel text clustering model based on topic modelling and social network analysis. Chaos, Solitons & Fractals, 181, 114
18 Haji, S. H., Al-zebari, A., Sengur, A., Fattah, S., & Mahdi, N. (2023). Document clustering in the age of big data: Incorporating semantic information for improved

results. Journal of Applied Science and Technology Trends, 4(01), 34-53.
19 Sial, A. H. (2024). Web Content Mining: A Review on Concepts, Techniques, and Tools.
20 He, H., Hao, G., & Wen, B. (2022, November). Category pattern mining based image retrieval. In Second International Conference on Optics and Communication Technology (ICOCT 2022) (Vol. 12473, pp. 28-35). SPIE.
21 Zhou, T., Li, Y., Zhang, Y., & Wang, L. (2022). Pattern matching method for q&a information retrieval system. In Advances in Intelligent Information Hiding and Multimedia Signal Processing: Proceeding of the IIH-MSP 2021 & FITAT 2021, Kaohsiung, Taiwan, Volume 2 (pp. 101-112). Singapore: Springer Nature Singapore.
22 Xiao, H., & Wang, J. (2023, December). Application Analysis of Data Mining technology in Internet Information Retrieval. In 2023 9th International Conference on Systems and Informatics (ICSAI) (pp. 1-6). IEEE.
23 Singh, V., Varalakshmi, S., Velusudha, N. T., Nawadkar, A. R., Srivastava, M., & Kalra, H. (2024, June). Text Mining for Automated Data Analysis and Information Retrieval. In 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT) (pp. 1-6). IEEE.
24 Faradilah, R., Harahap, S. Z., & Irmayanti, I. (2024). Application of Apriori and Fp-Growth Methods in Analyzing Book Lending Patterns Penerapan Metode Apriori dan Fp-Growth dalam Analisis Pola Peminjaman Buku. INFORMATIKA, 12(3), 538-545.
25 Ren, L., Qin, Y., Chen, Y., Lin, C., & Huang, R. (2023). Deep document clustering via adaptive hybrid representation learning. Knowledge-Based Systems, 281, 111058.
26 Kannout, E., Grzegorowski, M., Grodzki, M., & Nguyen, H. S. (2024). Clustering-based frequent pattern mining framework for solving cold-start problem in recommender systems. IEEE Access, 12, 13678-13698.
27 Bascur, J. P., Verberne, S., van Eck, N. J., & Waltman, L. (2023). Academic information retrieval using citation clusters: in-depth evaluation based on systematic reviews. Scientometrics, 128(5), 2895-2921.
28 Mekontchou, P. M., Fotsoh, A., Batchakui, B., & Ella, E. (2023). Information Retrieval in long documents: Word clustering approach for improving Semantics. arXiv preprint arXiv:2302.10150.
29 Rajindran, Y., & Salim, H. P. (2025). A Comparative Analysis of Clustering Methods on the 20 Newsgroups Dataset for Analytics.
30 Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, *9*(8), 1295.
31 Berahmand, K., Saberi-Movahed, F., Sheikhpour, R., Li, Y., & Jalili, M. (2025). A comprehensive survey on spectral clustering with graph structure learning. *arXiv preprint arXiv:2501.13597*.
32 Kulkarni, O., & Burhanpurwala, A. (2024, February). A survey of advancements in DBSCAN clustering algorithms for big data. In *2024 3rd International conference on Power Electronics and IoT Applications in Renewable Energy and its Control (PARC)* (pp. 106-111). IEEE.
33 Wu, Z., & Fang, G. (2024). The Research on the Improvement of FP-growth Algorithm. *Artificial Intelligence Technology Research*, *2*(1).
34 Wang, L., & Wang, S. (2021). HUIL-TN & HUI-TN: Mining high utility itemsets based on pattern-growth. *Plos one*, *16*(3), e0248349.
35 Hambarde, K. A., & Proenca, H. (2023). Information retrieval: recent advances and beyond. *IEEE Access*, *11*, 76581-76604.
36 Ahmad, M., Qadir, M. A., Rahman, A., Zagrouba, R., Alhaidari, F., Ali, T., & Zahid, F. (2023). Enhanced query processing over semantic cache for cloud based relational

databases. *Journal of Ambient Intelligence and Humanized Computing*, *14*(5), 5853-5871.

37 Ibrihich, S., Oussous, A., Ibrihich, O., & Esghir, M. (2022). A Review on recent research in information retrieval. *Procedia Computer Science*, *201*, 777-782.

38 Patel, A., & Shah, J. (2020, October). Smart ecosystem to facilitate the elderly in ambient assisted living. In Proceedings of International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications: ICMISC 2020 (pp. 501-510). Singapore: Springer Singapore.

39 Patel, V., Hiran, D., & Dangarwala, K. (2024). Advancing Information Retrieval. International Journal of Computing and Digital Systems, 17(1), 1-18.

40 Patel, A., & Shah, J. (2022). Towards enhancing the health standards of elderly: role of ambient sensors and user perspective. International Journal of Engineering Systems Modelling and Simulation, 13(1), 96-110.