Journal of Office Administration: Education and Practice

Volume 4 Issue 3, pp.193-202 (2024)

Hompage: https://ejournal.unesa.ac.id/index.php/joa

Application of Inverted Index Technique to Improve Document Search Effectiveness in Office Administration System

Septa[a], Dirmansyah[b], Fithria Rizka S[c], Le Minh Tu[d]

[a,b,c]Universitas Sumatera Utara, Medan, Indonesia
[d]Thai Nguyen University, Thai Nguyen City, Vietnam

ARTICLE INFO

Keywords:

inverted index, information retrieval, office administration, document similarity, search system

Article History:

Received November 2, 2024
Revised November 19, 2024
Accepted November 20, 2024
Available online November 30, 2024

Correspondence:

Septa, Department of Library and Information Science, Faculty of Humanities, Universitas Sumatera Utara, Medan, Indonesia. Email: septa@usu.ac.id

ABSTRACT:

This study aims to analyze the use of the inverted index technique in improving the effectiveness of information retrieval in office administration systems. As a core structure in modern search engines, the inverted index allows fast and relevant document access. This research uses a descriptive qualitative method, focusing on literature review as the primary data collection approach. The literature review was conducted in four stages: determining relevant keywords, accessing indexed academic sources, selecting articles based on relevance, and analyzing theoretical contributions. Articles were obtained from databases such as Google Scholar, IEEE Xplore, SpringerLink, and ScienceDirect. A total of 13 relevant articles were reviewed. Findings indicate that systems applying inverted index techniques significantly enhance search performance, especially in managing large volumes of office documents. The inverted index enables systems to locate document content more precisely, improving both recall and precision rates. By indexing every term along with its document location, the system can better identify documents that match user queries. This leads to faster, more accurate search processes, essential for office environments where timely and relevant information retrieval is critical.

INTRODUCTION

Scientific work is the culmination of an in-depth learning and research process, typically manifested in the form of research papers, journal articles, or theses. These works are composed of structured information derived from both primary analysis and supporting literature. In office administration systems, managing scientific work involves organizing, storing, and

ensuring effective accessibility of documents. (Swales, J. M., & Feak, C. B. (2016)., n.d.) in "Academic Writing for Graduate Students" explain that managing this information requires an efficient system to ensure that scientific works can be accessed and utilized effectively.

A major problem is the difficulty in accessing relevant documents quickly and accurately, especially when using conventional keyword-based search systems. These systems often produce irrelevant results due to poor indexing and lack of semantic understanding. Moreover, ensuring the originality of documents and detecting content similarity or plagiarism remain pressing concerns in academic and administrative settings. Existing tools such as Turnitin are widely used; however, they lack effectiveness in local language contexts like Indonesian, limiting their utility in many educational and administrative institutions.

Prior studies have focused on general plagiarism detection or indexing methods, yet few have addressed the application of inverted index techniques specifically within office administration systems. This study proposes that using an inverted index—a data structure that stores term-to-document mappings—can significantly improve the efficiency and accuracy of document searches and similarity detection.

The novelty of this research lies in its emphasis on integrating the inverted index model into administrative systems to manage unstructured academic data more effectively. This contributes to the development of smarter, localized, and scalable search infrastructures. The study aims to fill the gap in existing literature by offering both a conceptual framework and a practical foundation for improving document retrieval systems in educational institutions.

However, one of the main challenges in office administration is ensuring the originality of scientific work and avoiding plagiarism. Bergstrom and West (2020) explain that plagiarism, whether intentional or unintentional, is often a significant issue in academic writing. In office administration systems, it is crucial to have mechanisms that enable the verification of the originality of scientific work to ensure that stored or used documents do not violate copyright or academic ethics.

Similarity between scientific works is often a complex issue in office administration, especially when the works are in similar fields of study. Boehme (2019) indicates that the process of measuring similarity between documents can be very intricate. It requires appropriate tools and techniques to effectively compare documents. The limited availability of comparative sources, especially in certain languages like Indonesian, adds to this difficulty.

In the context of office administration, applications such as Turnitin, developed by iParadigms LLC, provide solutions for detecting similarities and plagiarism in academic

documents. According to iParadigms LLC 2017, this application has been widely used globally to check the originality of scientific work. However, in office administration systems, this application has limitations in certain languages due to the limited availability of document data. To enhance the effectiveness of office administration systems in addressing these issues, the inverted index technique can be applied. According to Manning et al. (2019) in "Introduction to Information Retrieval," the inverted index technique is a highly effective method for managing and searching information in large documents. By using an inverted index, the system can store each word along with its occurrence location in the document, enabling fast and accurate information retrieval while ensuring that the assessed scientific work is genuinely original.

In this context, web-based applications developed by iParadigms LLC, such as Turnitin, offer solutions to these problems. Since 1997, iParadigms LLC has developed technology used by over 30 million students from 15,000 institutions in 150 countries to detect similarities and plagiarism in scientific work (iParadigms LLC, 2017). Although this application is highly useful, it has limitations in specific languages, such as Indonesian, due to the limited availability of document data.

To address these limitations, the inverted index technique can be integrated to improve the process of text searching and comparison. According to Manning et al. (2019) in their book "Introduction to Information Retrieval," the inverted index is an effective method for managing and searching information in large texts. With this technique, each word in a document is indexed along with its occurrence location, allowing for fast and efficient searches.

The use of inverted index can be further developed by linking words to their source sentences and documents containing those words. This enables deeper searches to find similarities between documents, making it a valuable tool for detecting content similarity and assisting writers or researchers in measuring the similarity of their scientific works.

Based on the growing challenges in managing academic and administrative documents, this research explores the use of the inverted index technique in information retrieval systems within office administration. In administrative environments, the rapid increase in unstructured data—such as reports, memos, and proposals—demands a system that can perform fast and accurate document searches. One of the most pressing problems is the inefficiency of conventional search methods, which often fail to retrieve relevant results due to keyword mismatches and the inability to measure content similarity effectively.

Another urgent issue is the risk of plagiarism and content duplication, particularly in

academic works. With limited tools that support local languages like Indonesian, existing systems such as Turnitin face limitations. As a result, there is a clear gap in the current technology for document originality verification and similarity detection in office administration systems.

This research is motivated by the need for an advanced, language-independent search mechanism that enhances both retrieval precision and performance. The inverted index, which maps every term in a document to its occurrence, offers a solution by enabling quick and relevant access to matching documents. Integrating this technique into administrative systems can significantly improve the management of large document collections and address similarity detection more effectively. Therefore, this study is essential to provide theoretical insights and practical implications for building a more intelligent and efficient document search infrastructure in office administration.

METHOD

This research falls into the category of descriptive qualitative research, aimed at understanding and describing phenomena in depth. According to Sugiyono (2016), qualitative research methods are used to study the condition of objects in their natural settings, with the researcher playing a key role as the primary instrument in data collection and analysis. This study relies on contextual understanding and deep interpretation of the phenomenon under investigation, which is the use of inverted index techniques in information retrieval processes.

The data collection method employed in this research is a literature review, conducted to establish a solid and comprehensive theoretical foundation related to the inverted index technique and information retrieval processes. The literature review was carried out in several systematic stages. First, the researcher identified relevant keywords such as *inverted index*, *information retrieval*, and *document similarity* to narrow the search scope. Second, academic sources were accessed through reputable indexed databases, including Google Scholar, IEEE Xplore, SpringerLink, and ScienceDirect, ensuring the credibility and quality of the information obtained. Third, articles were selected based on their direct relevance to the research topic, theoretical contributions, and recency. A total of 13 relevant articles were reviewed and analyzed in depth to support the conceptual framework of this study.

These steps ensure that the method used aligns with the rational foundation of the study and addresses the research objective of enhancing document search effectiveness in office administration systems through the application of the inverted index technique.In this context,

the literature review focuses on theories and concepts related to inverted index and information retrieval. Manning et al. (2019) in "Introduction to Information Retrieval" explain that an inverted index is a data structure that allows efficient searching within large text documents by indexing each word and its occurrence location.

RESULTS AND DISCUSSIONS

The growth and increase in data every year for documents or scholarly publications demand a faster and more accurate information retrieval system. In the context of office administration, the ever-expanding volume of documents requires tools that can efficiently manage and search for information. With the growing amount of stored documents and information, having a quick and accurate search system is crucial for improving productivity and work efficiency.

Search results using keywords often lack relevance to the author or the title of the documents, and are more frequently based on the actual content of the documents. To address this issue, inverted index technology is employed by libraries and information systems. In office environments, keyword-based document searches often fail to produce satisfactory results due to mismatches between keywords and document content. An inverted index, with its ability to index every word in a document and its location, can enhance search result relevance by more accurately matching keywords with document content.

According to Darojad (2015), an information system can be implemented using an inverted index data structure based on HashTable and Ordered LinkedList, along with TFIDF weighting. This system was tested with 281 documents, 3,278 root words, 395 stopwords, and generated 25,737 terms. This data structure enables document searches based on 1 to 4 keywords with an average search time of 0.1 seconds.

In office administration, applying a data structure like an inverted index allows for faster and more relevant searches. By indexing keywords and their locations within documents, the system can speed up the search process and improve the accuracy of finding relevant documents.

A system using an inverted index also helps in measuring search performance with parameters like precision, recall, and search time. Darojad demonstrates that while indexing can decrease precision by up to 41.88%, search speed can increase up to 3,800 times. In office administration systems, balancing search speed and result accuracy is essential. Even though

precision might slightly decrease, a significant increase in search speed can enhance work efficiency and ease the management of large volumes of documents.

Hidayat (2015) classifies documents or scholarly works as unstructured data, which is challenging to store, manage, secure, and retrieve. Documents in office environments often fall into the category of unstructured data, such as reports, memos, and emails. Using an inverted index helps manage and search these documents in a more systematic and structured way, addressing the challenges of handling unstructured data. The indexing process in an inverted index involves parsing, stopping, stemming, sorting, and merging. This process results in differences in search outcomes compared to conventional methods, with noticeable changes in search time and result relevance. Implementing an inverted index in office administration systems introduces a more efficient method for indexing and searching, reducing search times and improving result relevance. This facilitates more effective and efficient document management.

According to Rachman (2022), information retrieval systems are expected to meet the precision level required by users. Fika adds that a system with an inverted index can achieve a recall rate of 84.7% and a precision rate of 39.7%. By using an inverted index, office administration systems can achieve high recall rates, ensuring that most relevant documents are found while maintaining precision to ensure search results remain relevant. Information retrieval models integrating inverted index, stemming, and vector space can enhance search performance, meeting users' needs for complete and relevant documents. The combination of inverted index with techniques like stemming and vector space provides a comprehensive solution for document searches, improving search results and efficiency in managing documents in office environments. Setiawan & Bunyamin (2018) also state that inverted index techniques can aid in finding similarities in document content and improve search results based on user-entered keywords. In office administration systems, the capability of inverted index to identify document content similarities and enhance keyword searches strengthens the system’s functionality, ensuring users can easily find relevant and similar documents.

The rapid and continuous growth of digital data presents serious challenges for information management, especially in office administration where large volumes of documents—such as reports, proposals, and memos—accumulate daily. One major barrier is the inefficiency of traditional keyword-based search systems, which often produce irrelevant results due to limitations in indexing and the inability to account for context or semantic meaning. These systems struggle to process unstructured data effectively, leading to low

precision and high user frustration in document retrieval tasks. According to the International Data Corporation (IDC), global data creation is expected to reach 175 zettabytes by 2025, driven by the digitization of work processes and the proliferation of electronic records. This dramatic increase underscores the urgent need for modern, scalable information retrieval systems capable of handling complex, high-volume data environments.

Given these findings and the data explosion in administrative environments, the development and adoption of inverted index-based systems are no longer optional—they are essential. Such systems not only resolve the limitations of conventional methods but also ensure timely, accurate, and efficient access to critical information in the digital era.

Overall, the application of inverted index in office administration systems can significantly improve search efficiency and document management, providing faster and more relevant solutions for users' information needs.

DISCUSSIONS

Information Retrieval Process

Information retrieval (IR) in office administration systems involves a multi-stage process essential for accessing relevant documents quickly and accurately. The process is divided into first workflow where it involves the collection and preprocessing of various documents such as memos, emails, reports, and academic papers. These documents, often in PDF or text formats, are cleaned, normalized, and prepared for indexing. According to Swales & Feak (2016), well-structured document management is a foundation of academic communication, and preprocessing is critical to achieving effective document retrieval. Second workflow is conducted by user queries, where is the system must accurately interpret and match these keywords with the document database. Inadequate indexing often leads to irrelevant results, as shown by Boehme (2019), who highlights the complexity of semantic interpretation in IR. The IR process includes indexing and searching, as outlined by Manning et al. (2019). Indexing converts the document corpus into an inverted index—a data structure mapping words to their locations in documents. Searching compares user queries to this index using vector space models and similarity measures to retrieve relevant documents. Implication is the structured process ensures a significant improvement in the accuracy and speed of information retrieval, enhancing administrative decision-making efficiency. The inverted index supports better document discoverability, even within large, unstructured datasets.

Stages of the Inverted Index Process

The implementation of an inverted index comprises three critical stages. The first is Parsing and Frequency Counting, where each document is broken down into tokens, and the frequency of each word is calculated (Robinson, 2014). This helps the system understand the weight or significance of terms. The second is Sorting for Optimization, where the sorted elements enable quicker access, ensuring efficient grouping and comparison. This is vital in reducing search latency, especially in systems handling thousands of documents. The third is Construction of Dictionary and Posting List, where the words and their frequencies are stored in a dictionary, and a posting list identifies the exact document locations. This architecture, also used in systems like Google Search, allows scalable and fast access (Manning et al., 2019). The implication is the structured mapping of terms to document locations significantly reduces search complexity. Future studies like Setiawan & Bunyamin (2018) support this approach for detecting content similarity across multiple documents, which is essential for plagiarism detection and content verification.

Relevance in Office Administration Systems

The adoption of inverted index techniques delivers various tangible benefits in office administration. The first is Improved Search Efficiency, where the inverted index structures drastically reduce the time needed to access relevant information. Lin & Dyer (2018) explain that efficient indexing systems improve performance in big data environments. The average search time can be reduced to milliseconds, enabling faster decision-making and increasing productivity. The second is Management of Unstructured Data, where office documents are typically unstructured and scattered across formats. Hidayat (2015) emphasized the challenges of organizing such data. An inverted index imposes structure on this unstructured content, allowing for logical, searchable organization. The third is Following Study, where Darojad (2015) demonstrated in his empirical study that a system using inverted index and TF-IDF weighting could manage 25,000+ terms with an average search time of 0.1 seconds across 281 documents. This highlights its scalability in office settings. The fourth is Performance Measurement, where the system's effectiveness is measured using precision and recall. Rachman (2022) reported that inverted index-based systems could achieve recall levels of up to 84.7%, meaning most relevant documents are retrieved. Although precision may drop slightly due to broader matches, the trade-off is often acceptable in high-volume environments.

The implication of this study underscores that inverted index techniques are not only theoretical constructs but have practical and measurable impacts in administrative efficiency. When integrated with stemming, stop-word removal, and vector space modeling, they form a comprehensive and adaptable framework for information retrieval.

CONCLUSION

The application of the inverted index technique in office administration systems significantly improves the effectiveness of document searches by speeding up the search process, increasing the relevance of search results, and managing unstructured data more efficiently. This technique enables the system to handle large volumes of documents more effectively, provide more accurate and relevant search results, and enhance productivity and efficiency in an office environment. Thus, the inverted index empowers the system to deliver more precise and relevant search outcomes, contributing to overall work performance improvement.

Future research could focus on integrating inverted index techniques with artificial intelligence or machine learning technologies to enhance semantic analysis capabilities in document retrieval. Additionally, further studies could explore the implementation of inverted index systems in cloud-based administrative environments and examine their impact on data security and real-time information accessibility.

REFERENCES

Bergstrom, C. T., & West, J. D. (2020). *Calling bullshit: The art of skepticism in a data-driven world*. Penguin Books.

Boehme, M. (2019). *A comprehensive study on similarity and plagiarism detection*. Springer.

Darojad, R. M. (2015). *Sistem temu balik informasi dokumen makalah ilmiah berbahasa Indonesia menggunakan struktur data inverted index berbasis hash table dan linked list* (Tesis, Universitas Sanata Dharma, Yogyakarta).

Fika, A. (2020). Analysis of inverted index in information retrieval systems. *Journal of Information Science, 38*(2), 123-135. https://doi.org/10.1177/0165551519876617

Hidayat, W. (2015). *Indexing and retrieval engine untuk dokumen berbahasa Indonesia dengan menggunakan inverted index*. Semin. Nas. Inform. dan Aplikasi (SNIA), 2015 October

Lin, J., & Dyer, C. (2018). *Data intensive text processing with MapReduce*. Morgan & Claypool Publishers.

Manning, C., Raghavan, P., & Schütze, H. (2019). *Inverted indexing*. Retrieved August 19, 2024, from

http://home.deib.polimi.it/lbondi/data/uploads/irdm116/slides/07_inverted_indexing_v1.pdf

Rachman, F. H. (2022). *Sistem temu kembali informasi dalam mesin pencarian menggunakan model ruang vektor dan inverted index* (Disertasi Doktoral, Universitas Gadjah Mada, Yogyakarta).

Robinson, L. (2014). *Implementasi metode generalized vector space model pada aplikasi information retrieval untuk pencarian informasi pada kumpulan dokumen teknik elektro di UPT BPI LIPI* (Disertasi Doktoral, Universitas Komputer Indonesia).

Setiawan, D. K., & Bunyamin, H. (2018). Pemanfaatan inverted index pada proses penelusuran kesamaan isi file dokumen PDF tugas akhir mahasiswa. In *Seminar Nasional Teknologi Informasi dan Komunikasi* (pp. 1-10). Yogyakarta.

Sugiyono. (2016). *Metode penelitian kuantitatif, kualitatif, dan R&D*. Alfabeta.

Swales, J. M., & Feak, C. B. (2016). *Academic writing for graduate students: Essential tasks and skills* (3rd ed.). University of Michigan Press.