Integrating Okapi BM25 and Jaccard Algorithms in Thesis Search Engine

Fema Rose Bronda-Ecraela [1], Remia L. Doctora [2]

Abstract: The integration of technology into society has brought transformative changes, enhancing efficiency and accessibility across diverse domains. In the academic realm, the role of documents, particularly in versatile soft copy formats, is pivotal. This article introduces a groundbreaking Thesis Search Engine developed in response to challenges faced by the existing thesis document repository at the College of Computer Studies. Focused on Information Technology theses, the innovative tool leverages advanced algorithms like Okapi BM25 and Jaccard to systematically organize and manage documents. The study's objectives include designing search engine modules, integrating these algorithms to uncover trends and similarities in content and providing insights into the evolution of research themes. Rigorous evaluations assess data relevancy, irrelevancy, and computational efficiency. Findings reveal the efficient use of Okapi BM25 for document ranking, showcasing a user-friendly design. While Jaccard exhibits versatility, its inclination to return all similar documents raises considerations. Computational efficiency tests favor Okapi BM25, establishing its effectiveness in delivering prompt and relevant search results. The study's insights contribute to optimizing the search engine and offer valuable considerations for future developments in Information Technology research, emphasizing the importance of aligning algorithms with research goals and user expectations.

Keywords: Search Engine, Document Management, Okapi BM25, Jaccard algorithm

1. Introduction

The current state of knowledge and research on the topic of thesis document management within academic institutions reflects a growing need for efficient and streamlined systems, particularly in the face of expanding document repositories and increasing reliance on digital formats. With the pervasive influence of technology in society, the management of thesis documents has become increasingly

[1] College of Computer Studies, University of Antique, Sibalom, Antique, Philippines
Email: femarose.ecraela@antiquespride.edu.ph
[2] College of Arts and Sciences, Computer Department, Iloilo Science and Technology University, Iloilo City, Philippines
Email: lopez_remia_05@yahoo.com

Acknowledgement: This research is supported and funded by University of Antique and Iloilo Science and Technology University.

Received [January 31, 2024]; Revised [March 10, 2024]; Accepted [April 4, 2024]

Published by InnoCon Publishing
ISSN 2704-4440

complex, requiring sophisticated computational strategies to organize and retrieve information effectively [1]. Previous research in the field as highlighted by Manning *et al.* [2] highlighted challenges such as inefficient retrieval systems, cumbersome manual processes, and the need for improved search accuracy and relevance [2].

In the academic context, the study addresses challenges in the existing repository of thesis documents by introducing a prototype Thesis Search Engine. Focused on BS Information Technology theses, this prototype aims to enhance thesis document management efficiency through a sophisticated ranking algorithm tailored for content comparison [3]. The study's overarching goal is to develop a Thesis Search Engine for Document Content Analysis, with specific objectives including the integration of Okapi BM25 (Best Match 25) and the Jaccard algorithms, evaluation of data relevancy and irrelevancy in keyword, phrase, and paragraph searches, and assessment of computational efficiency. The study's outcomes seek to validate the prototype's efficacy, offering a valuable tool for faculty, students, and administrators at the University of Antique College of Computer Studies.

The challenges surrounding thesis document management within academic institutions highlight the critical need for innovative solutions capable of navigating the complexities of digital repositories and evolving research landscapes. As the volume of thesis documents continues to expand and reliance on digital formats deepens, traditional management approaches are proving inadequate in meeting the demands of modern academia. In response to these pressing concerns, this study endeavors to address the deficiencies inherent in existing document repositories by introducing a prototype Thesis Search Engine tailored specifically for BS Information Technology theses [4]. By harnessing ranking algorithms and advanced computational strategies, the prototype seeks to streamline document management processes and enhance search accuracy and relevance.

Motivated by the imperative to improve efficiency and accessibility in thesis document management, the study sets out to develop a Thesis Search Engine for Document Content Analysis. This initiative aims to address the deficiencies identified in the study regarding e-theses management in India [5]. Through the integration of Okapi BM25 and Jaccard algorithms, coupled with evaluations of data relevancy, irrelevancy, and computational efficiency, the research aims to offer a transformative solution that empowers faculty, students, and administrators at the University of Antique College of Computer Studies to navigate the academic landscape with greater ease and efficacy. This approach draws inspiration from information retrieval systems, such as the development of hybrid similarity measures using fuzzy logic, as proposed by Gupta *et al.* [6], which addresses the limitations of traditional similarity measures and enhances the performance of information retrieval systems.

2. Related Works

In the rapidly evolving field of information technology (IT), this study introduces a groundbreaking Thesis Search Engine, leveraging Okapi BM25 and Jaccard algorithms to enhance data retrieval. Rooted in probabilistic relevance frameworks and adaptive similarity learning, these algorithms demonstrate prowess in diverse information retrieval domains. The synthesis of related studies provides a comprehensive context, guiding the rationale for integrating Okapi BM25 and the Jaccard algorithm and setting the stage for the exploration of how these cutting-edge algorithms converge to revolutionize academic document management.

Several studies contribute to enhancing or developing search engines using the Okapi BM25 and Jaccard algorithms. One study focuses on the theoretical underpinnings of text-retrieval algorithms like BM25, demonstrating its adaptability to modern search algorithms. The evolution of BM25, such as BM25F, showcases the framework's ongoing relevance in incorporating various document metadata for enhanced search algorithms [7]. Another study contributes insights into the challenges faced by digital libraries in information retrieval and proposes Okapi BM25 as a solution, emphasizing its probability-based approach. This aligns with the integration of Okapi BM25 in the proposed thesis search engine, showcasing its effectiveness in enhancing relevance ranking and retrieval [8]. Additionally, other scholars introduce an innovative approach to near-duplicate document detection, emphasizing adaptability and optimization of similarity functions. While not directly related to the search engine, the emphasis on adaptability and refined similarity measures aligns with the study's goal of integrating Okapi BM25 and Jaccard algorithms for improved document retrieval. Similarly, another study addresses the challenges in recommender systems by proposing new similarity models, including Jaccard similarity. Although focused on recommendations, the relevance of Jaccard similarity and the consideration of all rating vectors align with the proposed study's goal of incorporating the Jaccard algorithm for document content analysis [9].

In summary, the synthesis of related studies provides a rich context, showcasing the evolution of probabilistic relevance frameworks, the effectiveness of Okapi BM25 in information retrieval, the importance of adaptability in near-duplicate detection, and the relevance of Jaccard similarity in recommendation systems [10]. These insights collectively inform and support the rationale behind integrating Okapi BM25 and the Jaccard algorithm in the development of an advanced thesis search engine.

3. Methodology

The methodology employed in this study involved a meticulous examination of undergraduate thesis papers within the Bachelor of Science in Information Technology program, with a primary focus on comparing title and content trends against ongoing studies in the graduate program of Information Technology.

To achieve this objective, a specially crafted prototype algorithm was developed, aiming to determine whether proposed documents had been previously conducted or shared similarities with existing ones. The algorithm, integrated into search engine modules leveraging both Okapi BM25 and Jaccard algorithms, generated match scores indicating the level of similarity. The study included a comprehensive evaluation of the search engine's performance, employing statistical analyses such as the T-Test to compare the effectiveness of the Okapi BM25 and Jaccard algorithms in retrieving relevant documents. Additionally, the efficiency of the search engine, particularly concerning response times, was scrutinized using the T-Test. The application of an evolutionary prototype model in the software development process ensured iterative refinement based on user feedback, enhancing the prototype's functionality until its ultimate approval [11]. This innovative methodology aimed to contribute valuable insights to the field of IT research by analyzing and comparing content trends within undergraduate thesis papers.

Figure 1 shows the integrated flowchart for the thesis search engine, encompassing three major operations: upload, search, and view directory. Users could seamlessly upload thesis documents, and during the search operation, the system provided the option to choose between Okapi BM25 and Jaccard algorithms for matching. This dynamic feature allowed users to tailor their search approach based on their preferences. The chosen algorithm then identified relevant documents, offering a personalized and efficient search experience. Additionally, the view directory operation facilitated a structured exploration of the document organization. This strategic integration of algorithms enhanced the overall functionality, providing users with a versatile and effective tool for thesis document management.

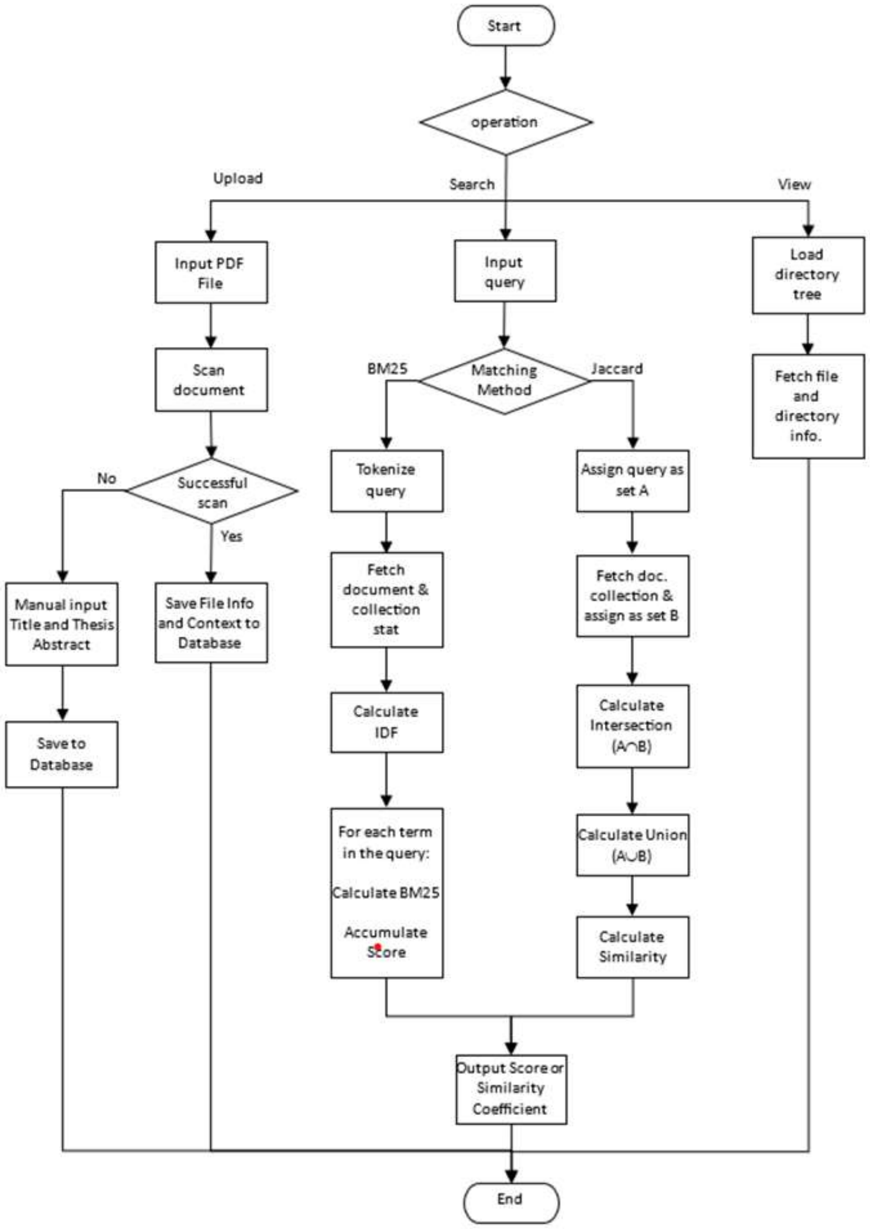
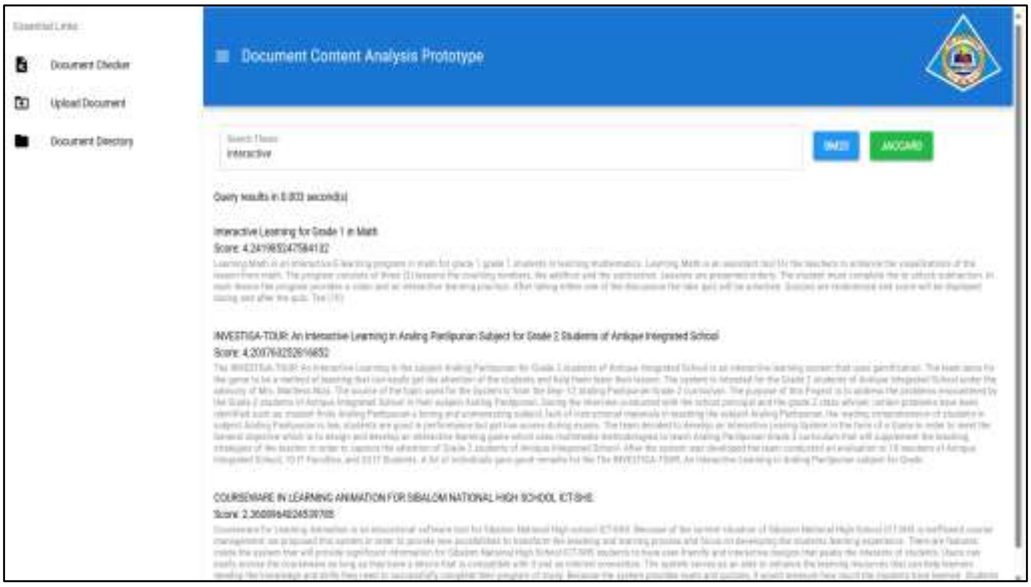


Figure 1. Flowchart Integrating Okapi BM25 and Jaccard Algorithm in Thesis Search Engine

The software's main interface, illustrated in Figure 2, offers three user functions: document checker, upload document, and document directory. The “Document Checker” button is central to the GUI, providing a search bar for users to input queries seamlessly. Two additional buttons, “OKAPI BM25” and “Jaccard,” allow users to select their preferred search algorithm. Clicking on “OKAPI BM25” initiates a search using the OKAPI BM25 algorithm, while “Jaccard” utilizes the Jaccard algorithm for an alternative method of finding relevant documents.

The “Upload Document” function enables users to expand the document database by effortlessly uploading PDF files or manually entering data, including title and description, in case of unsuccessful uploads. Lastly, the “Document Directory” function serves as a user-friendly window into the repository, allowing easy access to the list of neatly stored documents for transparent and efficient document management.



interactive

Query results in 0.003 second(s)

Interactive Learning for Grade 1 in Math
Score: 4.241985247584132
[illegible]

INVESTIGA-TOUR: An Interactive Learning in Araling Panlipunan Subject for Grade 2 Students of Antique Integrated School
Score: 4.200763252616852
[illegible]

COURSEWARE IN LEARNING ANIMATION FOR SIBALOM NATIONAL HIGH SCHOOL ICT-SHS
Score: 2.360896482450765
[illegible]

Figure 2. Interface of the Thesis Search Engine Application

4. Results and Discussion

A comprehensive performance test was conducted to assess the efficiency, accuracy, and responsiveness of the thesis search engine, utilizing both the Okapi BM25 and Jaccard algorithms. The evaluation aimed to provide a holistic examination of the search engine's capabilities, considering the effectiveness of the Okapi BM25 algorithm in document ranking and the Jaccard algorithm in measuring content similarity. Systematic testing and analysis were employed to reveal insights into the relative strengths and weaknesses of each algorithm within the context of the thesis search engine. The performance test not only sought to optimize the overall functionality of the search engine but also aimed to contribute valuable data and considerations for further refinement and enhancement of the system's search capabilities.

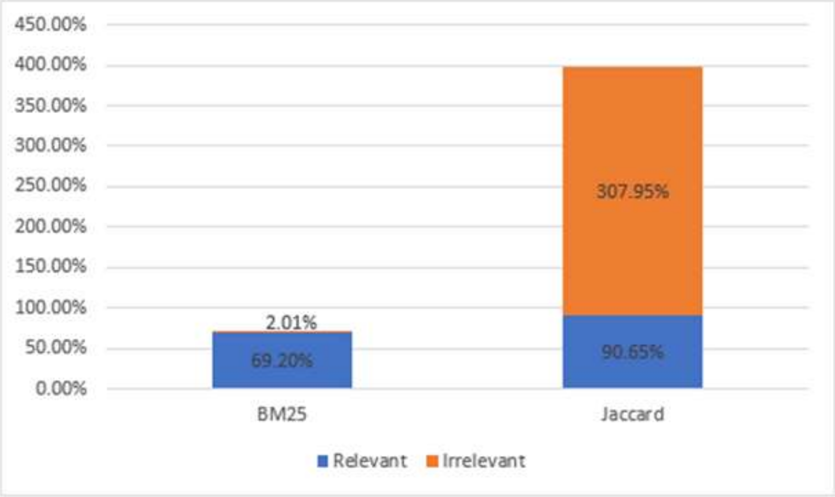


Figure 3. Bar Graph of Relevant and Irrelevant Search Result thru Keyword

The bar graph in Figure 3 presents the performance metrics for the Okapi BM25 and Jaccard algorithms in document retrieval. BM25 achieves a relevance rate of 69.20% and an impressively low 2.01% rate of incorrectly identified irrelevant documents. On the other hand, Jaccard outperforms with a higher relevance rate of 90.65%, but this improvement is accompanied by a significantly elevated 307.95% rate of incorrectly identified irrelevant documents. This comparison underscores the nuanced trade-off between precision and recall in the context of these two retrieval algorithms.

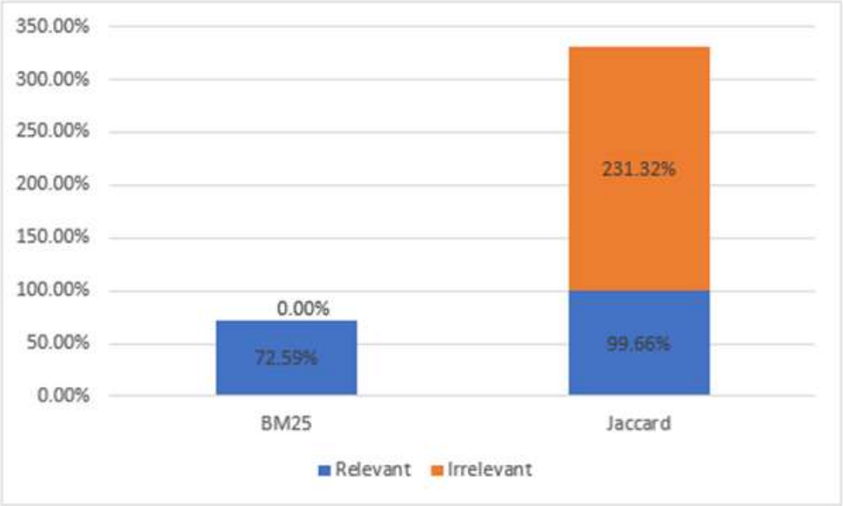


Figure 4. Bar Graph of Relevant and Irrelevant Search Result thru Phrases

The bar graph illustrated in Figure 4 shows a striking contrast in the performance of the BM25 and Jaccard algorithms in terms of relevance and irrelevance scores. For relevant documents, BM25 achieves a moderate 72.59%, while Jaccard excels with an impressive 99.66%. Notably, the graph reveals that Jaccard reports a 0.00% irrelevance score, showcasing its proficiency in filtering out irrelevant documents. In contrast, BM25 shows a 231.32% irrelevance score, emphasizing a substantial difference and underscoring Jaccard's superiority in handling both relevant and irrelevant documents in this comparison.

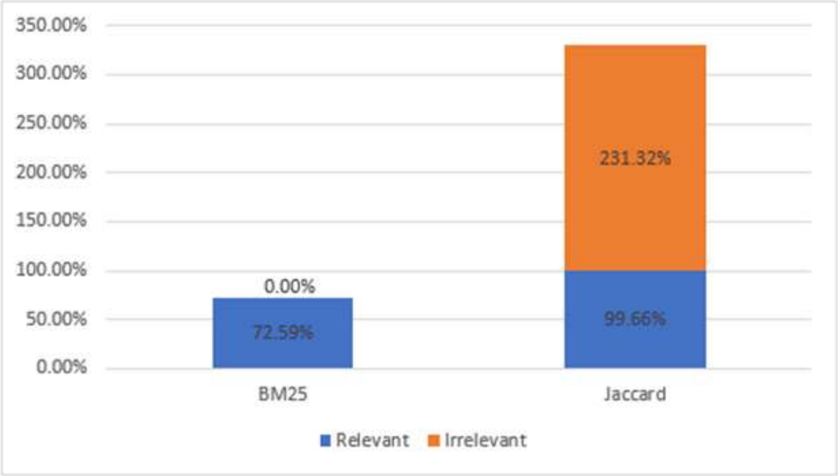


Figure 5. Bar Graph of Relevance and Irrelevant Search Result thru Paragraph

The presented bar graph in Figure 5 vividly illustrates the considerable disparity in performance between the BM25 and Jaccard algorithms concerning relevant and irrelevant documents. In terms of relevant documents, BM25 achieves a respectable 72.59%, whereas Jaccard outperforms significantly with an impressive 99.66%. A noteworthy observation is the stark contrast in handling irrelevant documents, where BM25 reports a 0.00% score, indicating effective filtration, while Jaccard, with a score of 231.32%, showcases a notable difference, emphasizing its superior capability in efficiently distinguishing between relevant and irrelevant content.

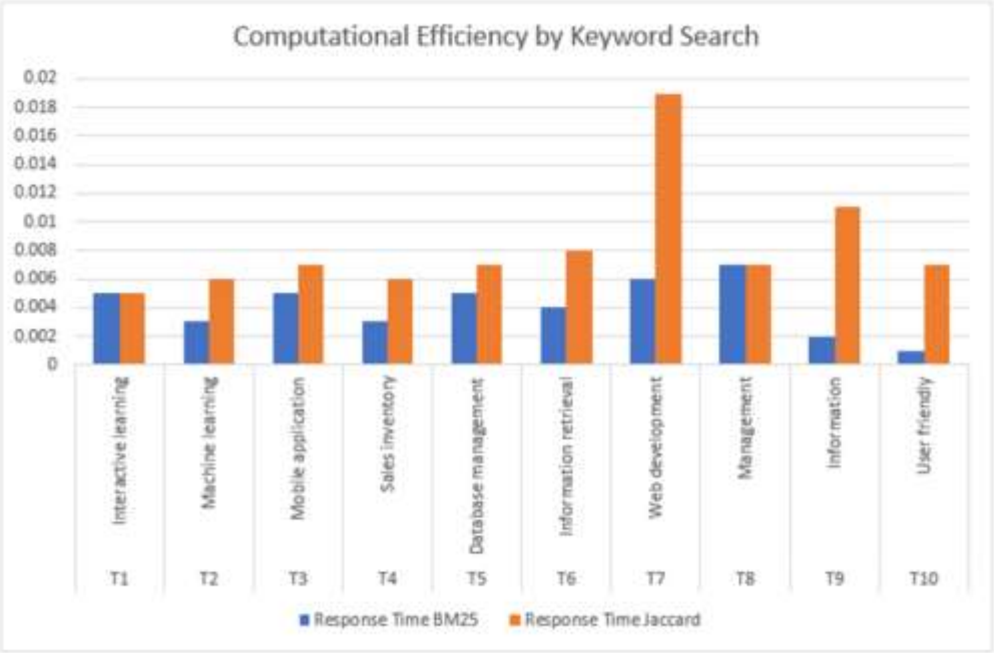


Figure 6. Bar Graph of Search Results of Response Time thru Keyword

The Figure 6 visually illustrates the response times of the search engine using Okapi BM25 and Jaccard algorithms for ten keyword-based test cases. Both algorithms generally exhibit low response times, with the graph indicating slight variations, notably in cases like "Web Development," where Jaccard shows a slightly higher response time of 0.019 seconds. It highlights the efficiency and speed of the search engine for each keyword search scenario, with lower bars indicating faster response times and, conversely, higher bars suggesting slightly longer processing times.

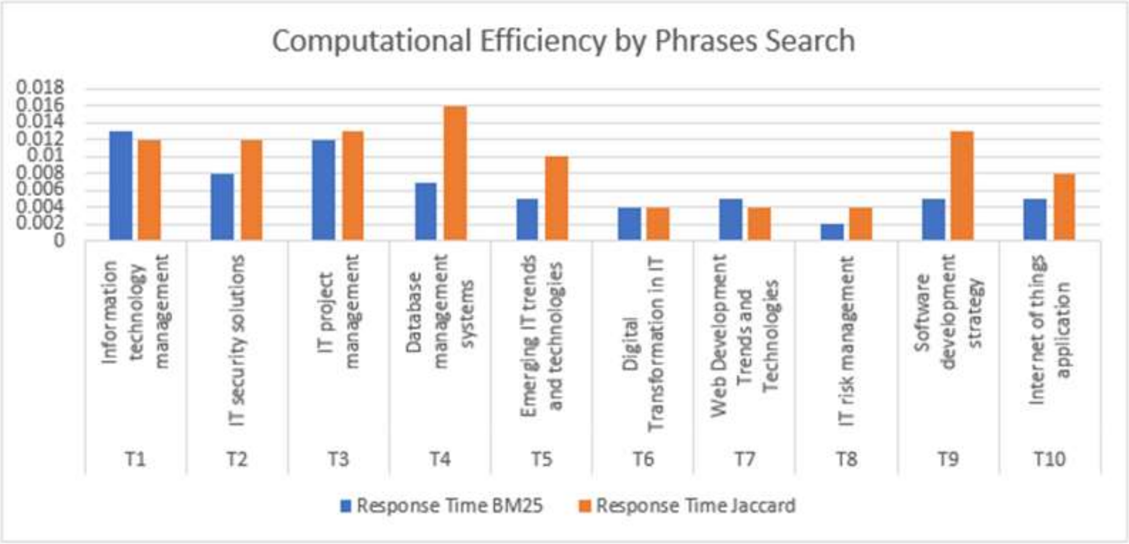


Figure 7. Bar Graph of Search Results of Response Time thru Phrases

The Figure 7 visually shows how quickly the search engine responds to different search phrases using the Okapi BM25 and Jaccard algorithms. Both algorithms perform quickly, with response times ranging from 0.002 to 0.016 seconds. Notably, phrases like “Digital Transformation in IT” and “Web Development Trends and Technologies” result in very fast responses of 0.004 seconds for both algorithms, suggesting efficient processing for these specific search queries. The graph simplifies the comparison of how fast Okapi BM25 and Jaccard handle different phrase searches.

Table 1. T-Test Results Comparing OKAPI BM25 and Jaccard Algorithm on Response Time thru Phrases Search

t-Test: Paired Two Sample for Means		
	Okapi BM25	*Jaccard*
Mean	0.0066	46.7
Variance	1.22667E-05	218.6778
Observations	10	10
Hypothesized Mean Difference	0	
Df	9	
T Stat	-9.984035561	
P(T<=t) one-tail	1.81307E-06	
t Critical one-tail	1.833112933	
P(T<=t) two-tail	3.62614E-06	
t Critical two-tail	2.262157163	

The t-test results in Table 1 shows a highly significant difference in response time between the Okapi BM25 and Jaccard algorithms for phrase searches. The mean response time for Okapi BM25 is 0.0066 seconds, while Jaccard demonstrates a substantially higher mean of 46.7 seconds. The negative Pearson correlation of -0.457382015 suggests an inverse relationship, indicating that as one algorithm's response time increases, the other tends to decrease. The t-statistic of -9.984035561, along with very low p-values (both one-tail and two-tail) significantly below 0.05, strongly supports rejecting the null hypothesis. This provides robust evidence that Jaccard, on average, has a significantly longer response time compared to OKAPI BM25 for phrase searches. The negative correlation further suggests an inverse relationship in response times between the two algorithms.

5. Conclusion and Recommendations

The development and evaluation of the thesis search engine, incorporating Okapi BM25 and Jaccard algorithms, highlight its user-friendly design and functional efficiency. Okapi BM25 serves as a crucial ranking algorithm, optimizing document relevance in large-scale information retrieval, while Jaccard focuses on similarity assessment, enhancing text analysis for clustering and classification. Despite BM25's competitive accuracy, Jaccard stands out with higher mean relevance scores, emphasizing the need for refined evaluation, especially in phrases related to information technology management. The comparative analysis of response times favors Okapi BM25, contributing to a streamlined, user-friendly search experience in the thesis search engine.

Looking ahead, future innovations in the thesis search engine could explore advancements in algorithmic refinement, ensuring a balanced approach that combines the strengths of Okapi BM25 and Jaccard. Continuous optimization of algorithms and periodic updates can keep pace with evolving academic needs and information retrieval challenges. Moreover, incorporating machine learning and artificial intelligence technologies may introduce adaptive features, allowing the search engine to learn and adapt to user preferences over time.

The robust features of the thesis search engine led to several recommendations for further enhancement. Prioritizing continuous optimization of the Okapi BM25 algorithm is essential for its effectiveness in diverse academic contexts, with regular updates to meet evolving information retrieval needs. Educational resources for both the Okapi BM25 and Jaccard algorithms are crucial for informed user choices. Integration with external academic databases can expand the engine's scope, while incorporating user feedback mechanisms and regular system updates fosters a collaborative academic community. A strategic approach, leveraging Jaccard for comprehensive document exploration and Okapi BM25 for precision in keyword searches, enhances the search engine's versatility. Based on consistently superior response times, prioritizing the implementation and further optimization of Okapi BM25 is recommended. This multi-pronged strategy aims to ensure the thesis search engine remains a reliable and efficient tool for information retrieval in academic contexts, with a forward-looking perspective on integrating emerging technologies for continuous innovation.

References

[1] G. Cabaleiro-Cerviño, C. Vera, "*The Impact of Educational Technologies in Higher Education*", GiST Education and Learning Research Journal, vol. 20, June 2020, pp. 155-169, doi: 10.26817/16925777.711.

[2] C. D. Manning, P. Raghavan, H. Schütze, "*Introduction to Information Retrieval*", Cambridge University Press, Cambridge, United Kingdom, 2008, doi: 10.1017/CBO9780511809071.

[3] T. N. Tran, "*Enhanced Retrieval and Discovery of Desktop Documents*", MS Thesis, Vrije Universiteit Brussel, Ixelles, Belgium, 2015, www.wise.vub.ac.be/sites/default/files/theses/Thesis TrungNgocTran_0.pdf (Accessed January 20, 2024).

[4] N. G. Momoti, "*Records Management for an Intelligent University: The Case of University of the Western Cape*", MS Thesis, University of the Western Cape, Cape Town, South Africa, 2017, www.etd.uwc.ac.za/bitstream/handle/11394/6191/Momoti_MLIS_ARTS_2017.pdf?sequen (Accessed January 20, 2024).

[5] K. V Satyanarayana, B. R. Babu, "*Trends in the Development of E-Theses in India: Issues, Constraints and Solutions*", www.epc.ub.uu.se/ETD2007/files/papers/paper-17.pdf (Accessed January 20, 2024).

[6] Y. Gupta, A. Saini, A. K. Saxena, "*Fuzzy Logic-based Approach to Develop Hybrid Similarity Measure for Efficient Information Retrieval*", Journal of Information Science, vol. 40, no. 6, December 2014, pp. 846-857, doi: 10.1177/0165551514548989.

[7] S. Robertson, H. Zaragoza, "*The Probabilistic Relevance Framework: BM25 and Beyond*", Foundations and Trends in Information Retrieval, vol. 3, no. 4, April 2009, pp. 333-389, doi: 10.1561/1500000019.

[8] L. Bonetti, "*Design and Implementation of a Realworld Search Engine Based on Okapi Bm25 and Sentencebert*", MS Thesis, Department of Computer Science and Engineering, Alma Mater Studiorum

- Università di Bologna, Bologna, Italy, 2021, www.amslaurea.unibo.it/24774/1/Thesis_Bonetti.pdf (Accessed January 20, 2024).

[9] S. Bag, S. K. Kumar, & M.K Tiwari, "*An Efficient Recommendation Generation Using Relevant Jaccard Similarity*", Information Sciences, vol. 483, May 2019, pp. 53-64, doi: 10.1016/j.ins.2019.01.023.

[10] H. Hajishirzi, W. T. Yih, A. Kolcz, "*Adaptive Near-Duplicate Detection via Similarity Learning*", in Proc. 33rd international ACM SIGIR conference on Research and development in information retrieval, July 2010, pp. 419-426, doi: 10.1145/1835449.1835520.

[11] W. Agustino, "*Applying Evolutionary Prototyping in Developing LMIS: A Spatial Web-Based System for Land ManagementI*", Journal of Physics: Conference Series, vol. 953, no. 1, January 2018, doi: 10.1088/1742-6596/953/1/012147.