Proceedings of the 10th World Congress on Electrical Engineering and Computer Systems and Sciences (EECSS'24)
Barcelona, Spain - August 19 - 21, 2024
Paper No. CIST 162
DOI: 10.11159/cist24.162

A BoW-BoC Indexing Method to Enhance Business-Related Document Representation and Retrieval

Sara Bouzid
Complex Systems Modeling Laboratory
Cadi Ayyad University
Marrakesh, Morocco
sara.bouzid@uca.ac.ma;

Abstract – Document indexing is crucial for efficient information retrieval systems. However, when documents are business-related and contains extensive figures and domain-specific terms, retrieving such documents poses challenges due to their lack of semantic context. Traditional Bag-of-Words (BoW) representation relying on word lists extracted from documents has limitations in such cases since it is based on a low-context approach. To address this issue, a BoW-BoC indexing method is proposed. This method utilizes a lexicon tailored to document contexts to associate BoW representations with Bag-of-Concepts (BoC), providing the necessary semantics for improved retrieval. The LSA model is used in conjunction with cosine similarity to automatically identify associations between document representations and lexicon concepts grouped in topics within a low-dimensional space. This BoW-BoC association is leveraged during document retrieval, supported by a weighted scheme intended to balance the contributions of both representations. Initial experiments conducted on an open document collection have shown promising results, demonstrating the potential effectiveness of the BoW-BoC indexing method for business-related documents.

***Keywords*:** Document indexing; Bag of words; Bag of concepts; Semantic context; LSA; Information retrieval

1. Introduction

To facilitate effective analysis and decision making, many companies use documents containing figures, charts and business-related terms including acronyms and abbreviations. However, searching for such documents presents a significant challenge due to their lack of semantic context. Yet, business-related documents are essential in supporting employees in their daily tasks. This challenge stems from real needs identified during a previous work in the industry, where documents pertaining to the core business predominantly consist of figures, charts, and numerous domain-specific terms [1], [2] . Unlike web documents and lengthy textual materials, these documents have information-poor features that are insufficient for indexing and retrieval purposes.

One main issue on which relies the efficiency of an information retrieval system (IRS) is document indexing. Indexing involves creating document representations using relevant terms extracted from documents for query matching. The list of extracted terms is known as Bag of Words (BoW) [3]. However, the limitations of BoW become apparent when dealing with business-related documents, which often contain limited text and lack the contextual information necessary for effective indexing and retrieval. Recent techniques used by researchers to improve document representation rely on external knowledge resources, like ontologies and thesauri, to create Bag-of-Concepts (BoC)[4] representation. In this paper, we focus on improving document indexing by leveraging both BoW and BoC representations. While lexical databases like WordNet [5] can be used to build BoC representations, they suffer from limited coverage, particularly in specialized domains such as industry, biology, and medicine [6], [7]. Domain ontologies offer a potential solution to the limitations of general lexical resources like WordNet. However, building and maintaining ontologies can be challenging due to the complexity of conceptualizing the business environment and the significant time investment required for these tasks.

To address these limitations in document indexing, we propose associating the BoW representation of documents with a controlled vocabulary using a domain-specific lexicon. This lexicon, organized into topics, offers a simple and sustainable knowledge structure within a given business domain, avoiding the complexity and time-consuming nature of ontology construction and updates. The lexicon serves as the foundation for constructing a semantic network around each document, enriching its BoW representation with contextual information and facilitating the creation of a BoC representation. In the

proposed indexing method, terms extracted from documents are matched with lexicon concepts using Latent Semantic Analysis [8], [9] and cosine similarity. This process enables the selection of the topic that best represents the document's context. Consequently, each BoW representation of a document is associated with a corresponding BoC representation. This BoW-BoC representation is leveraged during document retrieval using a weighted scheme.

The remainder of this paper is organized as follows. Section 2 reviews recent related literature. Section 3 introduces the BoW-BoC indexing method. Section 4 presents the initial findings from applying the proposed method on an open document collection, and finally the conclusion and future work are presented in the last section.

2. Related Work

Numerous techniques exist in the literature aimed at enhancing IRSs [10], [11], [12]. Recent approaches in the field have focused on improving document representation by leveraging external knowledge resources [13], [14]. For Instance, Boukhari and Omri [15], [16] proposed a method combining the Vector Space Model (VSM) and Description Logic (DL) to improve biomedical document representation. They utilized the Medical Subject Headings (MeSH) thesaurus to identify morphological variants and most relevant concepts within documents. Similarly, Gabsi et al. [17], introduced a semantic weighting scheme to disambiguate biomedical terms in document representation. The approach identifies the importance of relevant MeSH concepts in biomedical documents through term frequency and semantic similarities with unambiguous MeSH concepts. In [18], [19], the authors focused on improving document representation for classification purposes. Li et al. [18] presented the Bag-of-Concept-Clusters (BoCCl) model which utilizes a probabilistic knowledge base (ProBase) to identify semantically similar concepts in documents and cluster them. The BoC representation is initially constructed using a new concept score-inverse soft document frequency weighting scheme. Then, entity sense disambiguation techniques are applied to BoC to create BoCCI. Lee et al. [19] demonstrated the extension of feature-based document representation using domain-specific ontological concepts to produce an enhanced concept space with reduced dimensionality.

While existing approaches have made significant strides in improving document representation, they were primarily tested on web and lengthy documents [15], [16], [17]. There remains a notable gap in approaches addressing documents predominantly composed of figures with limited textual data including business-specific terms. This paper aims to fill this gap by proposing a hybrid approach that combines traditional BoW representation of documents with BoC representation.

3. The BoW-BoC Indexing Method

To address the challenges of business-related document retrieval, a hybrid method is proposed combining BoW and BoC representations for document indexing. This method offers a suitable combination scheme between document terms and a controlled vocabulary describing the document context. This controlled vocabulary is organized within a domain-specific lexicon defined as follows:

Let $c_i$be a concept in the lexicon L, where each c_i belongs to a Topic T_j :

$L = \{T_1, \dots, T_m\}$ where for each $j \in [1, \dots, m]$, $T_j = \{c_1, \dots, c_k\}$

The choice of a lexicon is motivated by its simplistic structure allowing rapid construction of a knowledge resource and easy adaptation to changes. Indeed, in business contexts where creating knowledge structures requires domain expertise and where information needs can evolve rapidly, using a lexicon simplifies the implementation process of an IRS in such cases.

3.1. Document Indexing

To establish the link between business-related documents and domain-specific lexicon concepts, we use the LSA model [8], also known as Latent Semantic Indexing (LSI). LSA extends the VSM [20] by applying dimension reduction techniques to identify the latent semantic structure between document terms and query terms. The underlying principle is that terms occurring in similar contexts tend to have similar meanings [21]. In the LSA model, a n x m matrix in constructed where n corresponds to query terms and m corresponds to documents. The linear algebra technique Single

Value Decomposition (SVD) [22] is then applied to reduce the number of rows in the matrix while preserving the number of columns[23]. This operation enables to find latent semantic relationship between terms by identifying and grouping together terms that often co-occur in similar contexts across documents.

In our method, LSA is used to find inherent similarities between the BoW representation of documents and the BoC representation of lexicon topics. In this case, *n* corresponds to document terms, and *m* corresponds to lexicon topics. Fig. 1 illustrates the proposed BoW-BoC indexing method. Initially, business-related documents undergo preprocessing to extract *n* relevant terms with their frequencies for the BoW representation. This process includes tokenisation, lowercasing and stopword removal. On the other hand, each lexicon topic, characterized by its set of concepts servers as a BoC representation. Thus, for *m* topics, *m* BoC representations are created. Subsequently, an *n* x *m* matrix is constructed using the BoW representation of each document and the sets of BoC of the lexicon topics. After dimensionality reduction, cosine similarity (equation (1)) is applied in the reduced space to compute the similarity between the BoW and BoC representations of the resulting matrix. The highest similarity score obtained between the BoW_i of a document d_i and the BoC_j of a topic T_j indicates that BoW_i must be associated with BoC_j. In this way, by combining LSA and cosine similarity, we leverage the benefits of both LSA (semantic comparison) and cosine similarity (vector space comparison) to produce relevant term correlations [24].

$$\cos(\vec{d_i}, \vec{T_j}) = \frac{\sum_{r=1}^{n} d_{ir} T_{jr}}{\sqrt{\sum_{r=1}^{n} d_{ir}^2} \cdot \sqrt{\sum_{r=1}^{n} T_{jr}^2}} \quad (1)$$

At the end of this process, two structures are created: (i) an inverted index structure where document BoW representations are stored along with their term frequencies and (ii) a BoW-T structure where the linkage between the BoW of each document and a lexicon topic is stored.

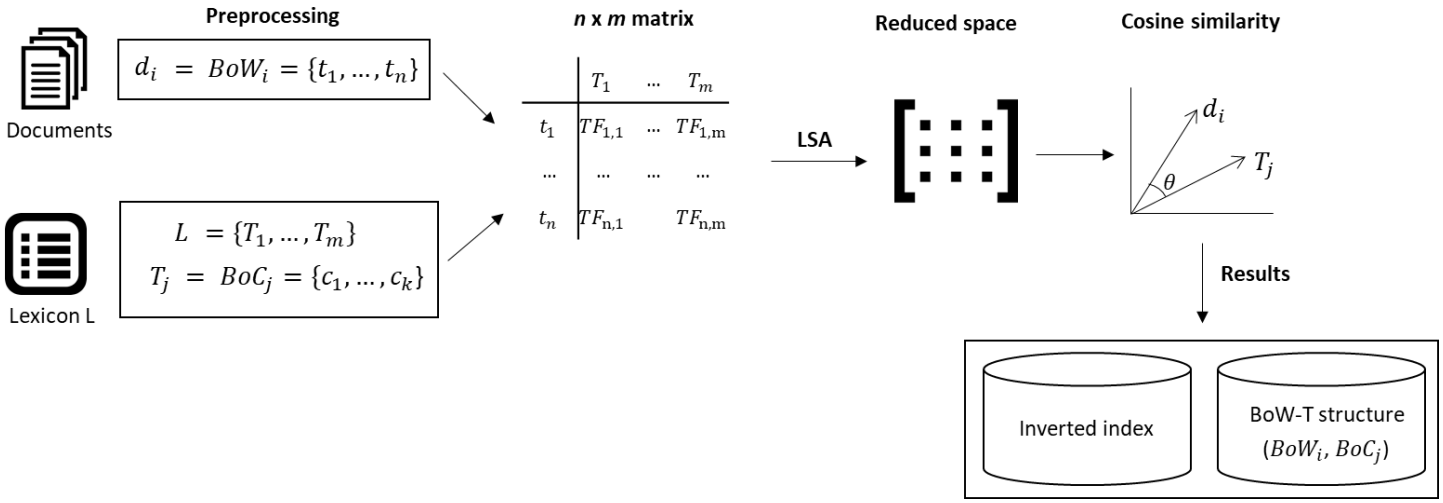

Fig. 1 : The BoW-BoC indexing method.

3.2. Document Retrieval Using the BoW-BoC Indexing Method

In traditional IRS, document retrieval typically involves matching query terms with document terms (BoW representation) using an information retrieval model such as the Boolean model [25] or the VSM. In our approach, the retrieval process utilizes both the BoW and the BoC representations of documents with a weighted scheme to compute similarity with query terms, adapted from the Weighted Sum Model (WSM) [26], a multi-criteria decision-making method.

The proposed formula, denoted WScore in equation (2), calculates the similarity between a query q and a document d by computing the inner product between q and the BoW and BoC associated with d, and by assigning two weights, α and β, to each resulting scores. From the initial experiments, α and β weights were set to 0.7 and 0.3 respectively. This weighted scheme ensures a well-adjusted contribution of both document terms and additional concepts that delimit document contexts, while also mitigating the influence of noisy data through the appropriate weights. It's worth noting that the proposed formula can be adapted to accommodate any string similarity measure (levenshtein, n-gram, …) [27] to compute the syntactic comparison between query terms and document representations.

$$WScore\ (q,d) = \alpha\ \times \sum_{i=1}^{n} q_i\ BoW_i + \ \beta \times \sum_{j=1}^{m} q_j\ BoC_j \tag{2}$$

4. First Experiments

4.1. Implementation Details

A primary experimentation of the BoW-BoC indexing method was conducted using Apache Lucene (version 9.9), specifically leveraging the Lucene Core Java library[1]. Apache Lucene offers a wide range of indexing and search features including tokenization, lowercasing, spellchecking, high-dimensionality vectors and pluggable ranking models like VSM and BM25. The experimentation was conducted on a machine equipped with a Core i7 CPU, 16GB of RAM and a 64-bit Windows operating system. The business-related document collection was extracted from the World Development Indicators[2] dataset of the open data of the World Bank Group. This collection contains documents encompassing figures and business-related descriptions across six data themes. Specifically, 784 documents were extracted in XML format from three specific themes: *States and Markets*, *Economy*, and *Environment*. For the lexicon used in the experimentation, the classification structure associated with these themes, along with their underlying descriptions, were utilized as topics and concepts, respectively. Table 1 provides examples of the extracted topics and concepts used in the lexicon, illustrating the diversity of business-related topics covered in the dataset.

Table 1: Examples of concepts used in the domain-specific lexicon.

Topic ID	Topic	Concepts
1	Business environment	business, firm, start, create, day, time, calendar, cost, connection, procedure, legal, operate, tax, official, speed, fast, CEO, female, top manager, manage, high position
2	Financial access and stability	depositor, commercial, bank, borrower, non-performing, loan, gross value, amount, overdue, retail, financial, service, subsidiary, balance sheet, resident, household, corporation
3	Stock markets	stock, market, capitalization, value, business, company, trade, turnover, annual, price, investment, funds, unit trust, domestic, hold, share
4	Government finance and taxes	revenue, grant, current, LCU, net, expense, lending, borrow, compensation, employee, tax, good, service, commercial, profit, rate, cash, receipt, fee, income, investment, value, asset, finance, transaction, business, authority, benefit
…	…	…

[1] https://lucene.apache.org/core/

[2] https://datatopics.worldbank.org/world-development-indicators/

The documents extracted for the experimentation were approximately the same size (± 4MB). Lucene uses index file formats to store document-related indexes following an inverted index structure. Typically, an index encompasses a set of documents, each referred with an integer document number. Both the lexicon and the BoW-T structure were implemented in XML format. An excerpt from the BoW-T structure that establishes the linkage between the BoW representations of documents and the lexicon topics is provided in Fig. 2. The algorithm of the LSA Model used in the experimentation was adapted from [28].

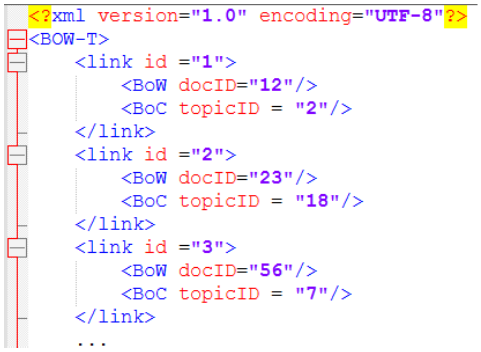

Fig. 2 : Extract from the BoW-T structure.

4.2. First Results

In this first experimentation, the VSM model was employed for query matching with the document collection using *WScore* weighted scheme (equation (2)). The VSM is a widely used information retrieval model that represents documents and queries as vectors in a high-dimensional space, where the similarity between vectors indicates the relevance of documents to queries. A total of six queries were tested on the 784 extracted documents. Additionally, a traditional document representation method, i.e. without the application of the BoW-BoC indexing method, was conducted using the same queries for comparison purposes. Table 2 provides examples of the used queries in the experimentation.

The experimental results are presented in Table 3 and Table 4, using standard information retrieval metrics [29], namely precision, recall and F1-Score. Precision (equation (3)) represents the proportion of correct results among the total retrieved results. Recall (equation (4)) measures the proportion of correct results among both the retrieved and non-retrieved results. F1-Score (equation (5)) represents the harmonic mean between precision and recall, providing a balanced measure of retrieval performance.

Table 2: Queries used for the experimentation.

Queries	Content
Q1	figures about starting a business in emerging countries, facilities and average costs
Q2	figures about start-ups and firms with female top managers, type of positions and salaries
Q3	cellular mobile subscription rates all over the world, rates per gender and income
Q4	proportion of cash payments in Middle East in current LCU
Q5	proportion of countries with high GDP per capita with their unemployment rate
Q6	countries in Europe and America with less CO2 emissions, number of industries in each country

$$precison = \frac{\#correctResults}{\#totalFound} \tag{3}$$

$$recall = \frac{\#correctResults}{\#correctResults + \#missedResults} \tag{4}$$

$$F1 - Score = 2 \times \frac{precision \times recall}{precision + recall} \tag{5}$$

The findings presented in Table 3 and Table 4 highlight important improvements across all queries with the BoW-BoC indexing method compared to the traditional method. In Table 3, we can notice that the lack of contextual information in document representation adversely affects both precision and recall. Indeed, users do not necessarily know which terms must be used in queries to retrieve relevant documents meeting their needs. Contextual information in document representation is crucial for addressing this issue, but it must be incorporated in a balanced manner to avoid introducing excessive noise in the results. The BoW-BoC indexing method demonstrates this balanced approach by contextualizing document representation, as evidenced in Table 4. By leveraging both BoW and BoC representations, the method effectively improves retrieval performance while mitigating the impact of noisy data.

Ultimately, the proposed method underscores the importance of introducing additional knowledge in IRS with rational use to enable effective semantic retrieval of business-related documents.

Table 3: Query processing results with traditional document indexing.

Queries	Precision	Recall	F1-Score
Q1	0,667	0,77	0,715
Q2	0,6	0,5	0,546
Q3	0,579	0,688	0,629
Q4	0,5	0,637	0,561
Q5	0,5	0,6	0,546
Q6	0,334	0,5	0,401

Table 4: Query processing results with the BoW-BoC indexing method.

Queries	Precision	Recall	F1-Score
Q1	0,819	0,9	0,858
Q2	0,875	1	0,934
Q3	0,917	0,847	0,881
Q4	0,778	0,875	0,824
Q5	0,786	0,847	0,816
Q6	0,706	0,8	0,751

4. Conclusion

This paper has put forward a BoW-BoC indexing method to enhance business-related document representation in IRS. Leveraging a lexicon containing business concepts, the method enhances BoW representation using the LSA model and cosine similarity. A weighted scheme is also proposed to balance BoW and BoC contributions during retrieval, adaptable to various syntactic similarity measures. Initial

experiments conducted with the BoW-BoC indexing method yielded significant results, demonstrating that document representation is much more enhanced with the proposed method in comparison with traditional BoW BoW representation of documents. To our knowledge, no approach in IRS-related studies has dealt with documents documents featuring extensive figures and business terms as experimented in this study. Yet, such documents are are challenging to retrieve in IRS and are integral to fulfilling users' needs in business contexts.

In future work, further experiments are planned with other sets of business-related documents, testing several similarity measures to assess their impact on the weighted scheme during document retrieval. There are also plans to improve the domain-specific lexicon model to leverage its concepts for query expansion with lexical terms like synonyms and variants. This includes developing a specific feature within the lexicon to handle lexical terms.

References

[1] S. Bouzid, C. Cauvet, C. Frydman, and J. Pinaton, *A semantic mapping approach to retrieve manufacturing information resources: STMicroelectronics' case study*, vol. 46, no. 9. IFAC, 2013. doi: 10.3182/20130619-3-RU-3018.00201.

[2] S. Bouzid, "A bottom-up semantic mapping approach for exploring manufacturing information resources in industry," *Comput. Syst. Sci. Eng.*, vol. 32, no. 3, pp. 243–256, 2017.

[3] M. Carrillo, E. Villatoro-Tello, A. Lopez-Lopez, C. Eliasmith, M. Montes-y-Gomez, and L. Villasenõr-Pineda, "Representing Context Information for Document Retrieval," in *International Conference on Flexible Query Answering Systems*, 2009, pp. 239–250. doi: https://doi.org/10.1007/978-3-642-04957-6_21.

[4] M. Carrillo and A. Lopez-Lopez, "Concept Based Representations as Complement of Bag of Words," in *International Conference on Artificial Intelligence Applications and Innovations*, 2010. [Online]. Available: https://inria.hal.science/hal-01060663

[5] A. Kilgarriff and C. Fellbaum, "WordNet: An Electronic Lexical Database," *Language (Baltim).*, vol. 76, no. 3, 2000, doi: 10.2307/417141.

[6] M. Ben Aouicha and M. A. Hadj Taieb, "Computing semantic similarity between biomedical concepts using new information content approach," *J. Biomed. Inform.*, vol. 59, pp. 258–275, 2016, doi: 10.1016/j.jbi.2015.12.007.

[7] M. Batet, D. Sánchez, A. Valls, and K. Gibert, "Semantic similarity estimation from multiple ontologies," *Appl. Intell.*, vol. 38, no. 1, pp. 29–44, 2013, doi: 10.1007/s10489-012-0355-y.

[8] A. Kontostathis and W. M. Pottenger, "A framework for understanding Latent Semantic Indexing (LSI) performance," *Inf. Process. Manag.*, vol. 42, no. 1 SPEC. ISS, 2006, doi: 10.1016/j.ipm.2004.11.007.

[9] C. H. Papadimitriou, P. Raghavan, H. Tamaki, and S. Vempala, "Latent Semantic Indexing: A Probabilistic Analysis," *J. Comput. Syst. Sci.*, vol. 61, no. 2, pp. 217–235, 2000, doi: 10.1006/jcss.2000.1711.

[10] J. N. Chimah and F. I. Ude, "Current trends in information retrieval systems: review of fuzzy set theory and fuzzy Boolean retrieval models," *J. Libr. Serv. Technol.*, vol. 2, no. 2, 2020, doi: 10.47524/jlst.v2i2.5.

[11] Y. V. B. Reddy, S. N. Reddy, and S. S. S. N. Reddy, "Efficient web-information retrieval systems and web search engines: A survey," *Int. J. Mech. Eng. Technol.*, vol. 8, no. 12, 2017.

[12] E. Karimi, M. Babaei, and M. S. H. Beheshti, "The study of semantic and ontological features of thesaurus and ontology-based information retrieval systems," *Iran. J. Inf. Process. Manag.*, vol. 34, no. 4, 2019.

[13] M. N. Asim, M. Wasim, M. U. G. Khan, N. Mahmood, and W. Mahmood, "The Use of Ontology in Retrieval: A Study on Textual, Multilingual, and Multimedia Retrieval," *IEEE Access*, vol. 7, pp. 21662–21686, 2019, doi: 10.1109/ACCESS.2019.2897849.

[14] J. Guo, Y. Cai, Y. Fan, F. Sun, R. Zhang, and X. Cheng, "Semantic Models for the First-Stage Retrieval: A Comprehensive Review," *ACM Trans. Inf. Syst.*, vol. 40, no. 4, pp. 1–42, 2022, doi: 10.1145/3486250.

[15] K. Boukhari and M. N. Omri, "Approximate matching-based unsupervised document indexing approach: application to biomedical domain," *Scientometrics*, vol. 124, no. 2, pp. 903–924, 2020, doi: 10.1007/s11192-020-03474-w.

[16] K. Boukhari and M. N. Omri, "DL-VSM based document indexing approach for information retrieval," *J. Ambient*

Intell. Humaniz. Comput., vol. 14, no. 5, 2023, doi: 10.1007/s12652-020-01684-x.
[17] I. Gabsi, H. Kammoun, D. Souidi, and I. Amous, "MeSH-Based Semantic Weighting Scheme to Enhance Document Indexing: Application on Biomedical Document Classification," *J. Inf. Knowl. Manag.*, p. 2450035, Mar. 2024, doi: 10.1142/S0219649224500357.
[18] P. Li, K. Mao, Y. Xu, Q. Li, and J. Zhang, "Bag-of-Concepts representation for document classification based on automatic knowledge acquisition from probabilistic knowledge base," *Knowledge-Based Syst.*, vol. 193, p. 105436, 2020, doi: https://doi.org/10.1016/j.knosys.2019.105436.
[19] Y.-H. Lee, P. J.-H. Hu, W.-J. Tsao, and L. Li, "Use of a domain-specific ontology to support automated document categorization at the concept level: Method development and evaluation," *Expert Syst. Appl.*, vol. 174, p. 114681, 2021, doi: https://doi.org/10.1016/j.eswa.2021.114681.
[20] G. Salton, A. Wong, and C. S. Yang, "A Vector Space Model for Automatic Indexing," *Commun. ACM*, vol. 18, no. 11, pp. 613–620, 1975, doi: 10.1145/361219.361220.
[21] G. Hollis and C. Westbury, "The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics," *Psychon. Bull. Rev.*, vol. 23, no. 6, pp. 1744–1756, Dec. 2016, doi: 10.3758/s13423-016-1053-2.
[22] T. G. Kolda and D. P. O'Leary, "A semidiscrete matrix decomposition for latent semantic indexing in information retrieval," *ACM Trans. Inf. Syst.*, vol. 16, no. 4, 1998, doi: 10.1145/291128.291131.
[23] C. Aswani Kumar and S. Srinivas, "Latent Semantic Indexing using eigenvalue analysis for efficient information retrieval," *Int. J. Appl. Math. Comput. Sci.*, vol. 16, no. 4, 2006.
[24] F. Al-Anzi and D. Abuzeina, "Enhanced latent semantic indexing using cosine similarity measures for medical application," *Int. Arab J. Inf. Technol.*, vol. 17, no. 5, 2020, doi: 10.34028/iajit/17/5/7.
[25] G. Salton, E. A. Fox, and H. Wu, "Extended Boolean information retrieval," *Commun. ACM*, vol. 26, no. 11, pp. 1022–1036, Nov. 1983, doi: 10.1145/182.358466.
[26] R. T. Marler and J. S. Arora, "The weighted sum method for multi-objective optimization: new insights," *Struct. Multidiscip. Optim.*, vol. 41, no. 6, pp. 853–862, 2010, doi: 10.1007/s00158-009-0460-7.
[27] D. W. Prakoso, A. Abdi, and C. Amrit, "Short text similarity measurement methods: a review," *Soft Comput.*, vol. 25, no. 6, pp. 4699–4723, 2021, doi: 10.1007/s00500-020-05479-2.
[28] J. Hicklin, C. Moler, and P. Webb, "JAMA : A Java Matrix Package." Accessed: Mar. 26, 2024. [Online]. Available: https://math.nist.gov/javanumerics/jama/
[29] K. Zuva, "Evaluation of Information Retrieval Systems," *Int. J. Comput. Sci. Inf. Technol.*, vol. 4, no. 3, pp. 35–43, 2012, doi: 10.5121/ijcsit.2012.4304.