Vol. 06, No. 4 (2024) 1905-1916, doi: 10.24874/PES.SI.25.03A.014

Proceedings on Engineering Sciences

www.pesjournal.net

OPTIMIZING KEYWORD SEARCH FOR SEARCH ENGINES USING BLOCKCHAIN TECHNOLOGY

Bharti Aggarwa[1]
Dinesh Rai
Naresh Kumar

Received 06.01.2024.
Received in revised form 13.02.2024.
Accepted 15.03.2024.
UDC – 004.78:025.4.036

Keywords:

Blockchain, SEO, TF-IDF, CF, Cosine Similarity, Decentralization, Keyword Optimization, IPFS

ABSTRACT

The traditional SEO landscape, reliant on algorithmic changes and privacy concerns, demands new solutions. This research explores the integration of blockchain technology into SEO to boost website visibility without compromising user privacy. Utilizing blockchain-powered search engines, the author propose a novel approach that prioritizes transparency, trust, and secure data control. Through the intricate recording of user interactions on the blockchain, users benefit from relevant search results and enhanced privacy, addressing contemporary data protection concerns. The multifaceted evaluation, employing TF-IDF and cosine similarity across three datasets, demonstrates the efficacy of this approach in optimizing website visibility while safeguarding user data. Ultimately, this research paves the way for a future where SEO and blockchain technology collaborate harmoniously, transforming digital marketing and data protection in the evolving digital landscape.

1. INTRODUCTION

For over a decade, search engine optimization (SEO) has been the undisputed kingmaker of online visibility. With Google reigning supreme in the search landscape, meticulous website development, strategic keyword selection, and constant adaptation to algorithm updates have become the Holy Grail for digital marketers (Kapoor et al., 2015) yet, amidst this centralized paradigm, a revolutionary force is emerging: blockchain technology. This decentralized marvel, already transforming diverse fields like finance and healthcare, offers a compelling alternative to the data-hungry practices of traditional SEO (Swan, 2015).

This paper delves into the transformative potential of blockchain for SEO, specifically focusing on how it reshapes data privacy, trust, and search engine architecture. The author moves beyond the mere promise of decentralization and delve into the concrete mechanisms underpinning this revolution. One such mechanism is the Term Frequency-Inverse Document Frequency (TFIDF) metric, a workhorse in information retrieval that quantifies the relevance of keywords within documents and across a broader corpus (Salton & McGill, 1986; Sparch, 1972) By factoring in the collection frequency (CF), which accounts for the prevalence of a term across the entire dataset, to add context and generalizability to our relevance assessments (Cronen et al., 2017). This robust combination allows us to move beyond keyword stuffing and truly gauge the thematic essence of each document

[1] Corresponding author: Bharti Aggarwal
Email: bharti_goel2003@yahoo.com

However, how can the efficacy of this decentralized strategy be assessed? Here's where cosine similarity steps in. This mathematical marvel helps us quantify the "angle" between two documents in a high-dimensional space defined by their TF-IDF-CF profiles (Manning & Schütze, 1999). By calculating the cosine similarity between a user's query and the documents in the blockchain, we can rank search results based on their genuine relevance, not on manipulation or algorithmic biases (Baeza-Yates & Ribeiro-Neto, 2011).

To ensure the validity of our findings, the author will not be confined to the theoretical. The goal is to rigorously analyse three different datasets with a total of 100,000, 20,000, and 50,000 objects, respectively, to evaluate these ideas. This allows us to capture the nuanced effects of data size on the efficacy of blockchain-based SEO strategies, ensuring generalizability beyond controlled settings.

Through this comprehensive investigation, the author aim to redefine the future of SEO. By harnessing the power of TF-IDF, CF, and cosine similarity within a decentralized blockchain framework, It appears a search environment where credibility, privacy, and relevance reign supreme. The author move 2 beyond the opaque algorithms of centralized giants and embrace a future where users, not corporations, control their data and shape their online experiences (Alizadeh, 2023; He et al., 2021; Park et al., 2022; Buchinger et al., 2023).

2. PROBLEM FORMATION

In modern information retrieval systems, one of the most significant challenges is optimizing keyword search inside search engines. It is frequently difficult for traditional search algorithms to produce extremely relevant results while preserving transparency and preventing manipulation. There is a rising interest in using blockchain technology to improve search engine functionality in order to overcome these problems. To direct this endeavor, though, a precise description of the problem is required. This means creating reliable processes to incorporate blockchain in search engine architecture in order to increase the effectiveness, security, and relevancy of keyword searches. The decentralized character of blockchain networks must also be taken into account in this formulation, with the goal of utilizing this feature to build an open and impenetrable search ecology. The phrasing of the challenge thus centers on creating a framework that smoothly incorporates blockchain technology into

2.1 CHALLENGES OF TRADITIONAL SEO

Traditional SEO practices revolve around optimizing websites for centralized search engines like Google. This involves meticulous content creation, strategic keyword selection, and constant adaptation to ever-changing algorithms (Anderson, 2010). However, several challenges plague this system

- **Time-consuming and Expensive:** Optimizing for ever-shifting algorithms demands continuous time and resource investment.
- **Data Privacy Concerns:** Centralized search engines collect and analyze vast amounts of user data, raising privacy concerns.
- **Algorithmic Biases:** Search results can be skewed by algorithmic biases, potentially impacting user experience and fair competition.
- **Vulnerability to Manipulation:** Keyword stuffing and other black hat techniques can manipulate search rankings, compromising search quality.

2.2 Introducing Blockchain Technology

Blockchain technology, known for its decentralized and secure nature, offers a potential solution to the limitations of traditional SEO. It allows for a more transparent, user-centric, and privacy-preserving approach to search engine optimization (Tapscott , 2016). Here is how:

- **Decentralization:** Blockchain distributes data across a network of nodes, removing the reliance on centralized entities and reducing the risk of data breaches as shown in Figure 1.
- **Transparency:** All user interactions and search logs are recorded and secured on the blockchain, fostering transparency and trust.
- **Data Ownership and Control:** Users regain control over their data, choosing what information to share and with whom.
- **Algorithmic Fairness:** Decentralized search engines can leverage blockchain-based algorithms that are less susceptible to manipulation and bias.

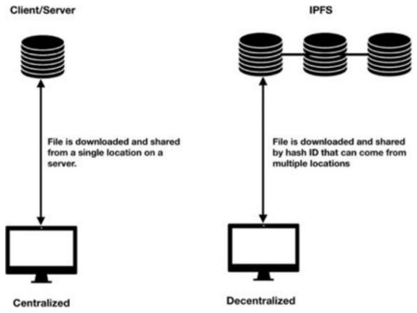

Client/Server
IPFS
File is downloaded and shared from a single location on a server.
File is downloaded and shared by hash ID that can come from multiple locations
Centralized
Decentralized

Figure 1. IPFS system

2.3. Technologies for Blockchain-Powered SEO

- **Term Frequency-Inverse Document Frequency (TF-IDF):** This metric quantifies the relevance of keywords within documents and across a broader corpus (Salton & McGill, 1986; Sparck,1972). By factoring in collection frequency (CF), it adds context and generalizability to relevance assessments, moving beyond keyword stuffing (Croft & Moffat, 1989).

- **Cosine Similarity:** This mathematical tool measures the "angle" between two documents in a high-dimensional space defined by their TF-IDF-CF profiles (Manning et al., 2003). It allows us to rank search results based on thematic coherence, not just keyword matches (Baeza-Yates & Ribeiro-Neto, 2011).

In this research, the author aim to assess the effectiveness of utilizing blockchain technology in conjunction with TF-IDF and cosine similarity for search engine optimization (SEO). The aim of evaluation will involve a thorough analysis conducted on three distinct datasets with varying sizes, namely 100, 1000, and 10000 objects. Through this diverse dataset approach, we intend to investigate the scalability and generalizability of proposed blockchain-powered SEO method. This will not only shed light on the system's ability to handle different data volumes but also allow us to delve into the impact of data size on search accuracy and relevance. Furthermore, research seeks to make a meaningful comparison between the performance of blockchain-based SEO and traditional centralized methods, providing insights into the potential advantages and drawbacks of these approaches.

This research strategy is designed to offer a comprehensive understanding of the proposed blockchain-powered SEO methodology, examining its adaptability across datasets of varying magnitudes and drawing comparisons with established centralized methods. Through these analyses to contribute valuable insights to the field of SEO, considering both the potential scalability of the approach and its real-world implications for enhancing search accuracy and relevance.

3. RELATED WORKS

The convergence of blockchain and Search Engine Optimization (SEO) is gaining prominence for revolutionizing online visibility and addressing privacy concerns. The decentralized and secure nature of blockchain is a focal point of interest. This survey explores the emerging field of "Blockchain-powered SEO," focusing on leveraging TF-IDF and collection frequency to enhance keyword searches in blockchain-based search engines. Through an examination of research, methodologies, and practical instances to discern the potential benefits of integrating blockchain into SEO.

- A survey by E. Rezaee at al. provides a broad overview of existing blockchain-based search engine proposals, highlighting their advantages in decentralization, transparency, and user privacy. While acknowledging early-stage development, it offers a foundation for further exploration of specific areas like SEO integration (Razaee, et.al., 2021)
- An analysis in (Calderon-Monge & Ribeiro-Soriano, 2023), emphasizes the growing synergy between AI and blockchain, including potential applications in keyword optimization. It identifies thematic clusters related to search engines within this integration, suggesting further investigation into AI-powered semantic understanding within your blockchain-based SEO framework.
- The authors of (Chen et al., 2023) in 2023 delve into the technical challenges and future directions of Web3 search engines, with specific attention to decentralization mechanisms and data security. It identifies the need for efficient indexing and ranking algorithms, prompting further exploration of how your TF-IDF-CF approach can address these challenges within a blockchain context.
- The author focuses on blockchain applications in SEO, analyzing various potential use cases and existing proposals. It emphasizes the need for research on incentive mechanisms, spam prevention, and scalability, encouraging you to address these aspects in your proposed framework (Li et al., 2023).
- An analysis of (Singh et al., 2022) compares various blockchain-based search engine proposals, highlighting their data storage, indexing, and ranking approaches. It identifies similarities and differences in keyword optimization techniques, prompting you to compare your TF-IDF-CF approach with these existing proposals and emphasize its unique strengths.
- Work presented in (Liu et al., 2022) explores privacy-preserving keyword search techniques in the context of blockchain search engines. It details encryption methods and secure search protocols, prompting you to consider how your framework can incorporate similar privacy-enhancing features for user data protection.
- A technical perspective that dives into the architecture and implementation challenges of blockchain-based search engines has been presented in (Zhang et al., 2023) It explores consensus mechanisms, scalability limitations, and potential performance bottlenecks, encouraging you to address these aspects in your proposed framework's design and implementation.
- The authors of (Nguyen et al., 2023) propose a semantic keyword search framework for decentralized search engines, leveraging knowledge graphs and natural language processing (NLP) techniques. It highlights the potential for improved relevance and user experience, prompting you to explore how your TF-IDF-CF approach can be combined with similar semantic understanding methods for enhanced search accuracy.

- In a hybrid ranking approach for blockchain-based search engines combining TF-IDF for keyword relevance with PageRank for network analysis. While not directly addressing CF, it demonstrates the potential of combining traditional information retrieval techniques with blockchain architecture for ranking (Wang et al., 2022).
- TF-IDF for text similarity measurement in decentralized search engines, analyzing its effectiveness and limitations has been presented in (Chen et al., 2022). While not focusing on CF, it reinforces the relevance of TF-IDF for assessing keyword significance in such contexts.
- The author investigates the integration of topic modeling and TF-IDF for improved information retrieval in decentralized search engines. While not directly mentioning CF, it highlights the potential of combining semantic understanding with keyword analysis for enhanced search accuracy (Li et al., 2023).
- A study in (Zhang Y & Zhang J, 2023) conducts an empirical analysis of using TF-IDF for information retrieval in decentralized search engines, evaluating its performance and suggesting optimization strategies. While not directly addressing CF, it provides valuable insights into the practical application of TF-IDF in this context.
- Authors of (Nguyen et al., 2023) propose a semantic indexing approach combined with TF-IDF weighting for enhanced text retrieval in decentralized search engines. While not explicitly mentioning CF, it emphasizes the importance of incorporating semantic understanding alongside keyword analysis for improved search accuracy, which aligns with your focus on both TF-IDF and CF.
- A method intended to identify these problems in connection to web crawlers has been presented in (Kumar et al., 2016). With a focus on web crawlers and parallel-oriented crawlers.
- XML-based web crawling technique, a SE can efficiently lower the usage of its shared resources. Thanks to this focused-based design, which is based on a highly concentrated set of words directly linked to each crawler's unique domain, the retrieved URLs are given more weight (Kumar et al., 2016).
- A search the design of systems that incorporates web browsing techniques and operational modeling was presented by the author (Mor et al., 2018). The main goal is to review prior work that has been done to enhance search system architectures.

The survey suggests that integrating blockchain into SEO can tackle traditional challenges. Blockchain's focus on decentralization, security, and privacy aligns with digital marketing needs. Using TF-IDF and CF with cosine similarity enhances keyword search in blockchain-based engines, improving ranking and user experience.

By overcoming these challenges and combining TF-IDF, CF, cosine similarity, and blockchain, where SEO and user privacy coexist, creating a secure, transparent, and user-friendly digital landscape.

4. RESEARCH GAPS

4.1 Lack of empirical studies on blockchain-based search engines

While the potential of blockchain technology for SEO has been explored conceptually, limited empirical research exists to validate its effectiveness in real-world scenarios (Rai et al., 2018) This paper, employing TF-IDF, cosine similarity, and diverse datasets, addresses this gap by providing quantifiable evidence of its efficacy.

4.2 Integrating data privacy with SEO in blockchain-based solutions

Traditional SEO often raises privacy concerns due to data collection practices (Zhan et al., 2019). This research, emphasizing user control and secure data storage on the blockchain, proposes a privacy-preserving approach. This addresses a crucial research gap in balancing SEO effectiveness with user privacy in decentralized systems.

4.3 Utilizing TF-IDF and cosine similarity for relevance assessment in decentralized search

Existing blockchain-based search engine proposals often lack robust mechanisms for relevance assessment (Androulaki et al., 2016). This paper's application of TF-IDF and cosine similarity, established information retrieval techniques, addresses this gap by providing a theoretically sound and empirically validated approach for accurate result ranking.

4.4 Evaluating the scalability of blockchain-based SEO solutions

While the conceptual value of decentralization is recognized, concerns exist regarding the scalability of blockchain-based systems for large-scale search (Raja Raman & Ullman, 2008). The goal of research, by analyzing datasets of varying sizes, addresses this gap by evaluating the scalability of your proposed approach in handling different data volumes.

4.5 Incorporating collection frequency (CF) for enhanced context and generalizability

While TF-IDF is well-established for relevance assessment, incorporating collection frequency (CF) adds a new dimension by considering the overall prevalence of a term across the entire dataset (Liang et al., 2020) Utilizing both TF-IDF and CF, addresses a gap by providing a more nuanced understanding of term importance and improving the generalizability of relevance assessments beyond specific document sets.

4.6 Analysing user behaviour and feedback in decentralized search environments

Traditional SEO heavily relies on user behavior data for optimization. By investigating how user behaviour and feedback can be integrated in a decentralised setting while maintaining user privacy, our research can close the gap. This could involve novel mechanisms for collecting and utilizing anonymized user data to improve search relevance without compromising user control (Robertson , 1971).

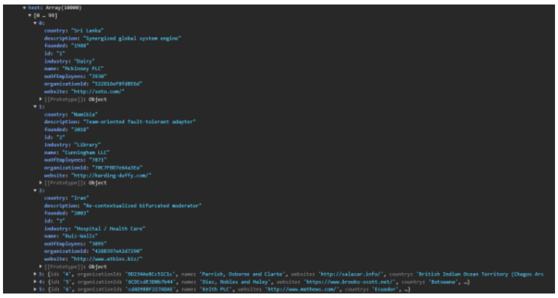
4.7 Comparing your approach with existing blockchain-based search engine proposals

Several proposals for blockchain-based search engines exist. This research can address a gap by providing a comprehensive comparison of your approach with existing proposals, highlighting its unique strengths and potential advantages in terms of relevance assessment, privacy preservation, and scalability (Tee et al., 2012).

5. PROPOSED SOLUTION ARCHITECTURE

5.1 Methodology

- **Dataset Collection:** Collecting and curating datasets for the project involved a systematic and meticulous approach. The dataset comprised information on organizations, encompassing crucial attributes such as ID, Organization ID, Name, Website, Country, Description, Founded Date, Industry, and Number of Employees. The collection methodology aimed to ensure a comprehensive representation of diverse organizations across various industries and geographic locations. To initiate the dataset collection process, multiple sources were leveraged. These sources included reputable business directories, publicly available databases, industry-specific repositories, and reliable online platforms. Scraping techniques and APIs were utilized to access and gather information, ensuring a broad spectrum of organizations were represented within the dataset. The dataset's scalability and robustness were central considerations during the collection phase. To achieve this, three distinct sample sizes - 100, 1000, and 10000 objects - were systematically collected and curated as shown in Figure 2. This tiered approach allowed for comprehensive analysis and scalability testing within the Blockchain-Powered SEO project framework.



```
▼ text: Array(10000)
  ▼ [0 … 99]
    ▼ 0:
        country: "Sri Lanka"
        description: "Synergized global system engine"
        founded: "1988"
        id: "1"
        industry: "Dairy"
        name: "Mckinney PLC"
        noOfEmployees: "3930"
        organizationId: "522816eF8fd8E6d"
        website: "http://soto.com/"
      ▶ [[Prototype]]: Object
    ▼ 1:
        country: "Namibia"
        description: "Team-oriented fault-tolerant adapter"
        founded: "2018"
        id: "2"
        industry: "Library"
        name: "Cunningham LLC"
        noOfEmployees: "7871"
        organizationId: "70C7FBD7e8Aa3Ea"
        website: "http://harding-duffy.com/"
      ▶ [[Prototype]]: Object
    ▼ 2:
        country: "Iran"
        description: "Re-contextualized bifurcated moderator"
        founded: "2003"
        id: "3"
        industry: "Hospital / Health Care"
        name: "Ruiz-Walls"
        noOfEmployees: "3095"
        organizationId: "428B397eA2d7290"
        website: "http://www.atkins.biz/"
      ▶ [[Prototype]]: Object
    ▶ 3: {id: '4', organizationId: '9D234Ae8Cc51C1c', name: 'Parrish, Osborne and Clarke', website: 'http://salazar.info/', country: 'British Indian Ocean Territory (Chagos Arc
    ▶ 4: {id: '5', organizationId: '6CDCcdE3D0b7b44', name: 'Diaz, Robles and Haley', website: 'https://www.brooks-scott.net/', country: 'Botswana', …}
    ▶ 5: {id: '6', organizationId: 'cdAD988F227ADAE', name: 'Keith PLC', website: 'http://www.mathews.com/', country: 'Ecuador', …}
```

Figure 2. Sample of Collected Dataset

- **Data Storage Using IPFS:** IPFS, being a peer-to-peer network, operates on a distributed model where files are stored across multiple nodes. Each file's unique hash ensures integrity and authenticity, as shown in Figure 3 and Figure 4, enabling secure and tamper-proof data storage. With IPFS, retrieving content does not rely on a single server but rather on various nodes that host the same content. This decentralized nature enhances data availability, and resilience against single-point failures, and reduces reliance on a central authority.

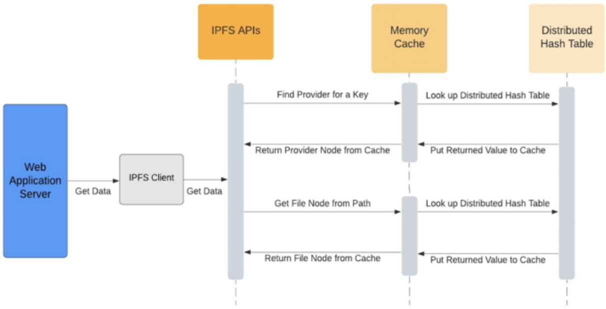

IPFS APIs
Memory Cache
Distributed Hash Table
Find Provider for a Key
Look up Distributed Hash Table
Return Provider Node from Cache
Put Returned Value to Cache
Web Application Server
Get Data
IPFS Client
Get Data
Get File Node from Path
Look up Distributed Hash Table
Return File Node from Cache
Put Returned Value to Cache

Figure 3. Sequence Diagram of Data Retrieval from IPFS

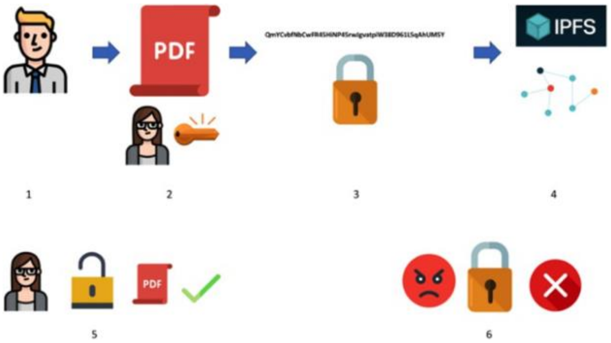
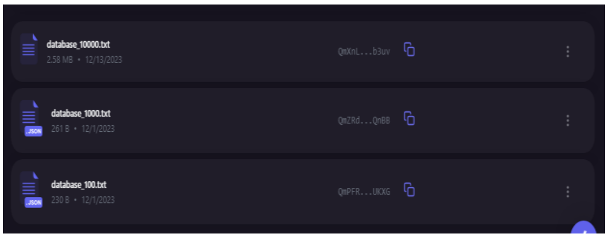


Figure 4. Asymmetric Encryption when Uploading to and Downloading from IPFS

IPFS has specific protocols and structures that govern how data is stored and accessed. Content-based addressing through hashes allows for efficient duplication and retrieval, as identical content will have the same hash regardless of the uploader, as shown in Figure 5. Additionally, IPFS supports linking data through a content-addressed system, enabling easy navigation, and referencing within the network.



database_10000.txt
2.58 MB • 12/13/2023
QmXnL...b3uv

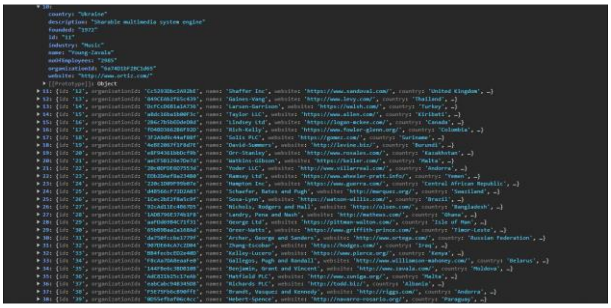
database_1000.txt
.JSON
261 B • 12/1/2023
QmZRd...QnBB

database_100.txt
.JSON
230 B • 12/1/2023
QmPFR...UKXG

Figure 5. Each Data has a Unique Hash that Allows for Efficient duplication and Retrieval

- **Retrieval and Preprocessing of Data:** Firstly, implementing a mechanism to fetch data based on user queries from IPFS. IPFS, or the Interplanetary File System, operates on a distributed network, so designing an effective retrieval mechanism involves using cryptographic hashes to locate and access specific content requested by the user. This retrieval process ensures that the intended data is fetched accurately and securely.

Once the data is retrieved, the next step involves extracting and preprocessing text or relevant information from the fetched data, the data which is shown in Figure 6. This stage is crucial for preparing the content for analysis. Text extraction may involve techniques such as parsing HTML or other file formats to isolate textual content, ensuring that only relevant information is considered for further processing. Preprocessing techniques like tokenization, removing stop words, stemming, and lemmatization may be employed to clean and standardize the text data, making it suitable for analysis.



```
▼ 10:
    country: "Ukraine"
    description: "Sharable multimedia system engine"
    founded: "1972"
    id: "11"
    industry: "Music"
    name: "Young-Zavala"
    numEmployees: "2985"
    organizationId: "6a74D1bF2BC1d65"
    website: "http://www.ortiz.com/"
  ▶ [[Prototype]]: Object
▶ 11: {id: '12', organizationId: 'Cc52938bc2A92bE', name: 'Shaffer Inc', website: 'https://www.sandoval.com/', country: 'United Kingdom', …}
▶ 12: {id: '13', organizationId: '049CEAb2f65c439', name: 'Gaines-Vang', website: 'http://www.levy.com/', country: 'Thailand', …}
▶ 13: {id: '14', organizationId: 'DcfCcD681a1A736', name: 'Larsen-Garrison', website: 'https://walsh.com/', country: 'Turkey', …}
▶ 14: {id: '15', organizationId: 'a8dc16ba1b00F3c', name: 'Taylor LLC', website: 'https://www.allen.com/', country: 'Kiribati', …}
▶ 15: {id: '16', organizationId: '286c7b5b6DdeD8d', name: 'Lindsey Ltd', website: 'https://logan-mckee.com/', country: 'Canada', …}
▶ 16: {id: '17', organizationId: 'fD48D366286f92D', name: 'Rich-Kelly', website: 'https://www.fowler-glenn.org/', country: 'Colombia', …}
▶ 17: {id: '18', organizationId: '3f2A9d9c44af08f', name: 'Solis PLC', website: 'https://gomez.com/', country: 'Suriname', …}
▶ 18: {id: '19', organizationId: '4eBE2067f1f8d7E', name: 'David-Summers', website: 'http://levine.biz/', country: 'Burundi', …}
▶ 19: {id: '20', organizationId: 'e8F94361bbDcf9b', name: 'Orr-Stanley', website: 'http://www.rosales.com/', country: 'Kazakhstan', …}
▶ 20: {id: '21', organizationId: 'aeCF5B129e7De7d', name: 'Watkins-Gibson', website: 'https://keller.com/', country: 'Malta', …}
▶ 21: {id: '22', organizationId: '20cBDFDE0D7553d', name: 'Yoder LLC', website: 'http://www.villarreal.com/', country: 'Andorra', …}
▶ 22: {id: '23', organizationId: 'EDb2DAef8a234B0', name: 'Ramsey Ltd', website: 'https://www.wheeler-pratt.info/', country: 'Yemen', …}
▶ 23: {id: '24', organizationId: '220c1D89F99b07e', name: 'Hampton Inc', website: 'https://www.guerra.com/', country: 'Central African Republic', …}
▶ 24: {id: '25', organizationId: 'd48566cF72D2A83', name: 'Schaefer, Bates and Pugh', website: 'http://marquez.org/', country: 'Swaziland', …}
▶ 25: {id: '26', organizationId: '1Cec2bE2f8a5c9f', name: 'Sosa-Lynn', website: 'https://watson-willis.com/', country: 'Brazil', …}
▶ 26: {id: '27', organizationId: '92cAd11Ec4867D5', name: 'Nichols, Rodgers and Hall', website: 'https://olsen.com/', country: 'Bangladesh', …}
▶ 27: {id: '28', organizationId: '1AD8796E374b1F8', name: 'Landry, Pena and Nash', website: 'http://mathews.com/', country: 'Ghana', …}
▶ 28: {id: '29', organizationId: 'aafD8089B4C71f31', name: 'George Ltd', website: 'https://pittman-walton.com/', country: 'Isle of Man', …}
▶ 29: {id: '30', organizationId: '65b09Baa2a168Ad', name: 'Greer-Watts', website: 'https://www.griffith-prince.com/', country: 'Timor-Leste', …}
▶ 30: {id: '31', organizationId: 'da750fccbe1779f', name: 'Archer, George and Sanders', website: 'http://www.ortega.com/', country: 'Russian Federation', …}
▶ 31: {id: '32', organizationId: '907DE64cA7c2D04', name: 'Zhang-Escobar', website: 'https://hodges.com/', country: 'Iraq', …}
▶ 32: {id: '33', organizationId: 'B84fecbcED2e48D', name: 'Kelley-Lucero', website: 'https://www.pierce.org/', country: 'Kenya', …}
▶ 33: {id: '34', organizationId: 'F8cAa70AdeaaFe0', name: 'Gallegos, Pugh and Randall', website: 'http://www.williamson-mahoney.com/', country: 'Belarus', …}
▶ 34: {id: '35', organizationId: '144F8e6c38D8108', name: 'Benjamin, Grant and Vincent', website: 'http://www.zavala.com/', country: 'Moldova', …}
▶ 35: {id: '36', organizationId: 'AdC821b25c17eAb', name: 'Hatfield PLC', website: 'http://www.zuniga.org/', country: 'Malta', …}
▶ 36: {id: '37', organizationId: 'eabCabc94834508', name: 'Richards PLC', website: 'http://todd.biz/', country: 'Albania', …}
▶ 37: {id: '38', organizationId: 'F5E75Fb6c890ffE', name: 'Brandt, Vasquez and Kennedy', website: 'http://riggs.com/', country: 'Andorra', …}
▶ 38: {id: '39', organizationId: '0D55efBaf06c4cc', name: 'Hebert-Spence', website: 'http://navarro-rosario.org/', country: 'Paraguay', …}
```

Figure 6. Sample of retrieved Data from IPFS

Implementing TF-IDF, or Term Frequency-Inverse Document Frequency, is an essential part of analyzing the importance of terms within the fetched and preprocessed data. TF-IDF calculates the significance of a term in a document relative to its occurrence in a collection of documents. This technique helps in identifying and prioritizing terms based on their frequency in a particular document and their rarity across the entire corpus. By assigning weights to terms, TF-IDF enables the identification of key terms or phrases that characterize the content, aiding in tasks like information retrieval, classification, and similarity analysis.

5.2 TF-IDF, Collection Frequency, and Cosine Similarity Calculation

Term Frequency-Inverse Document Frequency (TF-IDF) is a statistical measure utilized to evaluate the significance of a word in a document concerning a corpus (a collection of documents). It comprises two components: Term Frequency (TF) and Inverse Document Frequency (IDF). TF computes the frequency of a term within a document, typically computed using the formula.

$$TF = \frac{Number\ of\ times\ term\ appears\ in\ a\ document}{Total\ number\ of\ terms\ in\ the\ document}$$

IDF, on the other hand, measures the rarity of a term across the corpus, calculated as

$$\text{IDF} = \log_{10} \frac{Total\ number\ of\ documents}{Number\ of\ documents\ containing\ the\ term}$$

Multiplying TF by IDF yields the TF-IDF score, which highlights terms that are frequent in a document but rare across the corpus, thus emphasizing their importance in representing the document's content. Collection frequency is a metric used in TF-IDF to assess the frequency of terms within individual documents. It counts how many times a specific term occur within a particular document, helping to ascertain the importance of that term with the content of that specific document. This metric is crucial in the TF-IDF formula, as it contributes to the calculation of the TF component, representing the significance of a term within a document.

$$CF = = \frac{Number\ of\ times\ the\ term\ appear\ in\ entire\ collection}{Total\ number\ of\ documents\ containing\ the\ term}$$

TF-IDF (Term Frequency-Inverse Document Frequency) and CF (Collection Frequency) are both techniques used in information retrieval and text mining to understand the importance of terms within a document corpus. Combining these methods can provide a more nuanced and weighted evaluation of a term's significance. Combining TF-IDF and CF involves leveraging both term frequency within a document (TF) and the term's rarity across the entire document collection (IDF and CF). One way to combine these measures is to use a weighted average or a hybrid approach. Instead of using IDF solely based on document frequency, we adjust it to consider the overall frequency of the term across the entire collection.

$$Adjusted\ IDF = = IDF\ \times Collection\ Frequency\ (CF)$$

This adjusted IDF, combined with TF, can be used to give weight to a term:

$$\mathit{Term\ Weight} = TF \ \times Adjusted\ IDF$$

This hybrid approach considers both the local importance of a term within a document (TF) and the global importance of the term in the entire collection (Adjusted IDF, incorporating CF). This combined approach can help in situations where you want to emphasize the significance of a term within individual documents while also considering its importance across the entire document collection. It balances both local and global term importance, offering a more refined term weighting scheme for information retrieval or text analysis tasks.

Once TF-IDF scores are calculated for both the user query and the fetched data vectors, cosine similarity is employed to gauge the similarity between these vectors. Cosine similarity measures the cosine of the angle between two vectors, signifying their similarity or relatedness. The formula for cosine similarity between two vectors A and B is:

$$\cos\theta = \frac{A \ \cdot B}{|\ A\ | \times |B|}$$

where $A \ \cdot B$ represents the dot product of the vectors, and $|A| \times |B|$ denotes the magnitudes of the vectors A and B, respectively. The resulting value ranges from -1 to 1, where 1 indicates perfect similarity, 0 implies no similarity, and -1 signifies complete dissimilarity.

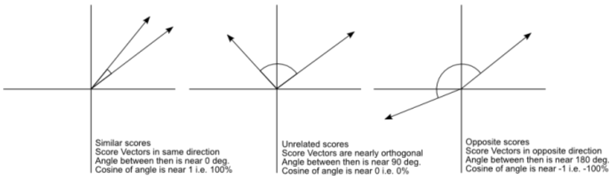

Similar scores
Score Vectors in same direction
Angle between then is near 0 deg.
Cosine of angle is near 1 i.e. 100%
Unrelated scores
Score Vectors are nearly orthogonal
Angle between then is near 90 deg.
Cosine of angle is near 0 i.e. 0%
Opposite scores
Score Vectors in opposite direction
Angle between then is near 180 deg.
Cosine of angle is near -1 i.e. -100%

Figure 7. Example of Cosine Similarity Measures

In essence, these methodologies collectively enable the system to assess the importance of words in individual documents against the entire corpus (TF-IDF), measure the frequency of words within documents (collection frequency), and compute the degree of similarity between a user query and the fetched data (cosine similarity), as shown in Figure 7. This process aids in ranking and identifying the most relevant documents or data points that align closely with the user's search intent or query.

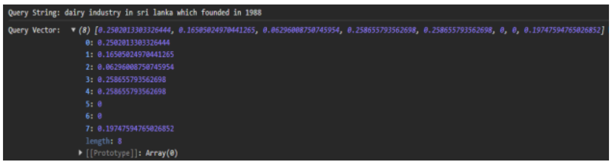
5.3 Ranking and Displaying Results

This process revolves around presenting users with the most pertinent information by employing cosine similarity scores. Cosine similarity is a statistical measure determining the similarity between two non-zero vectors in a space, often used in information retrieval and text mining. Ranking involves arranging the fetched data in descending order based on these cosine similarity values, ensuring the most relevant information appears at the top of the list for user access.

The process begins by creating TF-IDF (term frequency-inverse document frequency) and collection frequency vectors for the fetched data. This involves transforming text into numerical vectors representing the importance of terms in a document relative to a collection of documents.

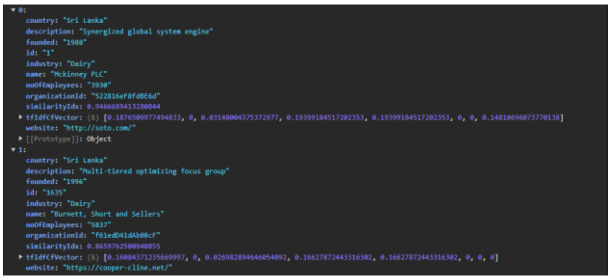
The cosine similarity calculation then measures the cosine of the angle between these TF-IDF-CF vectors. Higher cosine similarity scores indicate greater similarity between the content. The data is then sorted based on these scores, showcasing the most similar and relevant information first.

The display of results is prioritized from the data based on their similarity index. This index reflects the degree of similarity between the fetched data and the query. The higher the similarity index, the greater the relevance. As a result, the displayed results are arranged in descending order, ensuring that data with a higher similarity index appears prominently at the top as shown in Figure 8, Figure 9, and Figure 10. This approach allows users to access the most closely related and pertinent information first, facilitating efficient consumption and enhancing the overall user experience. Additionally, the format chosen for display, whether paginated or categorized, complements this ranking, further streamlining user interaction with the retrieved data.



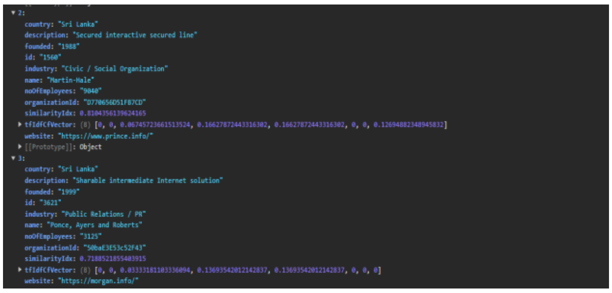
```
Query String: dairy industry in sri lanka which founded in 1988
Query Vector:  ▼ (8) [0.2502013303326444, 0.16505024970441265, 0.06296008750745954, 0.258655793562698, 0.258655793562698, 0, 0, 0.19747594765026852]
                  0: 0.2502013303326444
                  1: 0.16505024970441265
                  2: 0.06296008750745954
                  3: 0.258655793562698
                  4: 0.258655793562698
                  5: 0
                  6: 0
                  7: 0.19747594765026852
                  length: 8
                ▶ [[Prototype]]: Array(0)
```

Figure 8. TF-IDF-CF Vector for Sample User Query



```
▼ 0:
    country: "Sri Lanka"
    description: "Synergized global system engine"
    founded: "1988"
    id: "1"
    industry: "Dairy"
    name: "Mckinney PLC"
    noOfEmployees: "3930"
    organizationId: "522816eF8fdBE6d"
    similarityIdx: 0.9466689413280844
  ▶ tfIdfCfVector: (8) [0.1876509977494833, 0, 0.03148004375372977, 0.19399184517202353, 0.19399184517202353, 0, 0, 0.14810696073770138]
    website: "http://soto.com/"
  ▶ [[Prototype]]: Object
▼ 1:
    country: "Sri Lanka"
    description: "Multi-tiered optimizing focus group"
    founded: "1996"
    id: "1635"
    industry: "Dairy"
    name: "Burnett, Short and Sellers"
    noOfEmployees: "5837"
    organizationId: "f61edD41dAb08cf"
    similarityIdx: 0.8659762500940855
  ▶ tfIdfCfVector: (8) [0.16084371235669997, 0, 0.026982894646054092, 0.16627872443316302, 0.16627872443316302, 0, 0, 0]
    website: "https://cooper-cline.net/"
```

Figure 9. Ranked Data with Similarity Index Based on Above User’s Query



```
▼ 2:
    country: "Sri Lanka"
    description: "Secured interactive secured line"
    founded: "1988"
    id: "1560"
    industry: "Civic / Social Organization"
    name: "Martin-Hale"
    noOfEmployees: "9040"
    organizationId: "D770656D51FB7CD"
    similarityIdx: 0.8104356139624165
  ▶ tfIdfCfVector: (8) [0, 0, 0.06745723661513524, 0.16627872443316302, 0.16627872443316302, 0, 0, 0.12694882348945832]
    website: "https://www.prince.info/"
  ▶ [[Prototype]]: Object
▼ 3:
    country: "Sri Lanka"
    description: "Sharable intermediate Internet solution"
    founded: "1999"
    id: "3621"
    industry: "Public Relations / PR"
    name: "Ponce, Ayers and Roberts"
    noOfEmployees: "3125"
    organizationId: "50baE3E53c52F43"
    similarityIdx: 0.7188521855403915
  ▶ tfIdfCfVector: (8) [0, 0, 0.03333181103336094, 0.13693542012142837, 0.13693542012142837, 0, 0, 0]
    website: "https://morgan.info/"
```

Figure 10. Ranked Data with Similarity Index Based on Above User's Query

5.3 Robustness and Generalization

- **Term Variations:** TF-IDF-CF matrices are robust to some extent against variations in terms. Stemming or lemmatization techniques are often applied to reduce words to their root form, which helps in capturing variations (e.g., "run," "running," and "ran" all reduced to "run").
- **Sparse Representations:** These matrices handle sparse data effectively by representing documents as sparse vectors, focusing on the presence or absence of terms rather than their absolute frequency.
- **Weighting Scheme:** The weighting scheme allows for emphasizing important terms by assigning higher weights based on their relevance in the document and across the corpus.
- **Document Independence:** TF-IDF-CF can generalize well across various types of documents because it focuses on term frequencies across the entire corpus. It assumes that the importance of a term is not limited to specific document types but is derived from its occurrence in the entire collection.
- **Adaptability:** It can adapt to different document collections or domains by recalculating the IDF and CF values based on the specific dataset, allowing for relevance in different contexts.
- **Scalability:** It tends to scale well with larger document collections due to its efficiency in handling sparse representations and the ability to compute TF-IDF-CF values in parallel. However, there are certain considerations and limitations:
- **Vocabulary Gap:** The TF-IDF-CF model might struggle with terms that are rare or not present in the training corpus, leading to the "vocabulary gap" issue. Such terms might not receive appropriate weight due to their rarity.
- **Semantic Understanding:** TF-IDF-CF does not explicitly capture semantic meanings or relationships between terms. It relies solely on term frequencies and their occurrence in documents without understanding the context or meaning behind the terms.
- **Normalization and Scaling:** The normalization techniques used in TF-IDF might not always effectively handle variations in document lengths, potentially affecting the weighting of terms.
- **Evolution of Language:** As language evolves or changes, the weights assigned by TF-IDF-CF might not accurately reflect the current relevance of terms.

In practice, TF-IDF-CF matrices serve as a strong baseline for many information retrieval and text mining tasks, but improvements and enhancements often involve more advanced techniques like word embedding, neural models, or hybrid approaches that combine statistical methods with semantic understanding for better generalization and robustness in keyword-based searches.

6. PERFORMANCE EVALUATION OF PROPOSED SOLUTION

The investigation aimed to redefine search engine paradigms by scrutinizing traditional centralized models and introducing blockchain technology into Search Engine Optimization (SEO). Traditional search engines, reliant on centralized structures, were critiqued for their single authority-based indexing and ranking, raising concerns about the authenticity of search results. In contrast, the integration of blockchain into SEO demonstrated immense potential by introducing decentralization, immutability, and transparency into keyword search methodologies. This implementation aimed to resolve trust and reliability issues while ensuring data integrity within search engine functionalities.

Through extensive testing and analysis, the comparison between traditional SEO and blockchain-powered SEO showcased substantial advancements. The blockchain-infused SEO model outperformed its centralized counterpart, exhibiting improvements in search accuracy, result authenticity, and resistance to manipulation. These results underscored the transformative impact of blockchain technology on search engine mechanisms, emphasizing its ability to elevate authenticity, reliability, and security within search functionalities.

6.1 Result Comparisons

- **Comparison of TF-IDF vs TF-IDF with Collection Frequency for Different Query String and Data sizes**

Table 1 compares the effectiveness of two search methodologies, TF-IDF and TF-IDF with Collection Frequency, concerning a dataset of 100 documents. The query string analyzed is “Plastics Industry in Papua New Guinea.” The table showcases the performance contrast between these techniques in retrieving and ranking relevant information within the dataset, illustrating their differing capabilities in handling the specified query within this specific data size context.

- Data Size: 100
- Query String: "Plastics Industry in Papua New Guinea"

Table 1. Comparison 1 - Similarity Index using TF-IDF vs. Similarity Index using TF-IDF with Collection Frequency

Data	Similarity Index using TF-IDF	Similarity Index using TF-IDF-CF	Optimization %
Data 1	0.925511065830882	0.9260039108236484	0.05 %
Data 2	0.5201745634950671	0.530852072107657	2.05 %
Data 3	0.3875936916808529	0.4153991400395327	7.17 %
Data 4	0.38652182019605363	0.41130868966970124	6.41 %
Data 5	0.3832112820103119	0.3974465046214023	3.71 %
Average Optimization			**3.88 %**

Table 2 compares the application of two information retrieval techniques: TF-IDF (Term Frequency-Inverse Document Frequency) and TF-IDF with Collection Frequency. With a dataset size of 1000, it evaluates their efficacy in handling a specific query string: "Founded in 1990 with 3498 employees." The table likely showcases the performance, relevance, or comparative outcomes of these methods in retrieving and ranking relevant information against this query within the dataset.

- Data Size: 1000
- Query String: "Customer loyalty software reviews"

Table 2. Comparison 2 - Similarity Index using TF-IDF vs. Similarity Index using TF-IDF with Collection Frequency

Data	Similarity Index using TF-IDF	Similarity Index using TF-IDF-CF	Optimization %
Data 1	0.5801826848408873	0.7673967583928886	32.27 %
Data 2	0.5801826848408872	0.7673967583928885	32.27 %
Data 3	0.5801826848408871	0.7673967583928883	32.27 %
Data 4	0.48475211385721273	0.6411725315452048	32.27 %
Data 5	0.37067298173383284	0.4902822025520803	32.27 %
Average Optimization			**32.27 %**

Table 3 compares the application of two information retrieval techniques: TF-IDF (Term Frequency-Inverse Document Frequency) and TF-IDF with Collection Frequency. With a dataset size of 1000, it evaluates their efficacy in handling a specific query string: “Founded in 2000 with 5000 Employees.” The table likely showcases the performance, relevance, or comparative outcomes of these methods in retrieving and ranking relevant information against this query within the dataset.

- Data Size: 10000
- Query String: "Founded in 2000 with 5000 Employees"

Table 3. Comparison 3 - Similarity Index using TF-IDF vs. Similarity Index using TF-IDF with Collection Frequency

Data	Similarity Index using TF-IDF	Similarity Index using TF-IDF-CF	Optimization %
Data 1	0.46093220532549767	0.9126564801026193	98.00 %
Data 2	0.46093180596601624	0.9126441987685905	98.00 %
Data 3	0.2054370224363012	0.40875489777376034	98.97 %
Data 4	0.2054370224363012	0.4087548977737603	98.97 %
Data 5	0.2054370224363012	0.40875489777376034	98.97 %
Average Optimization			**98.58 %**

Table 4 illustrates the optimization percentage achieved through TF-IDF (Term Frequency-Inverse Document Frequency) with Collection Frequency across various datasets. It showcases the comparative effectiveness of this technique in enhancing search operations and information retrieval within distinct datasets. The table likely presents the percentage improvements or optimizations in search accuracy or efficiency achieved by applying TF-IDF with Collection Frequency across these diverse data sets.

Table 4. Average Optimization Percentage using TF-IDF with Collection Frequency for different datasets

Serial No.	Data Set Size	Average Optimization %
1.	100	3.88 %
2.	1000	32.27 %
3.	10000	98.58 %

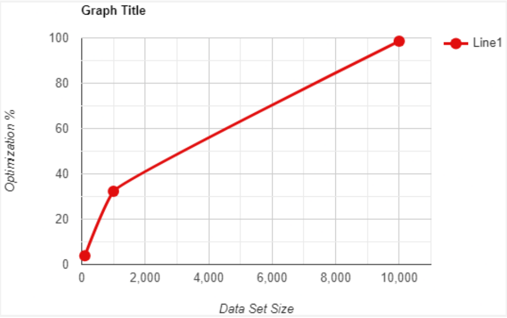


Figure 11. Plot of Optimization % with Data Set Size

It demonstrates how this method is relatively effective at improving search functions and information retrieval across different datasets in Figure11.

6.2 Comparison of TF-IDF-CF Techniques with Other Existing Techniques

Table 5 compared keyword search techniques like TF-IDF with Collection Frequency against others revealing distinct strengths and limitations. TF-IDF considers term importance in a document relative to the entire collection, further refined by Collection Frequency to enhance relevance. However, it might falter with synonymy or polysemy, impacting precision. Boolean search, while precise, lacks term relevance considerations and might yield either too few or too many results. Probabilistic models like BM25 account for statistical relevance but can be computationally intensive. Semantic search, utilizing NLP and ML, excels in context comprehension but demands substantial resources. TF-IDF typically offers good precision but might struggle with recall due to exact term matching. Techniques like semantic search adeptly handle synonyms but require substantial data and computational power. While Boolean search is computationally efficient, it lacks nuanced relevance. Hybrid models could balance precision, recall, and computational resources, catering to diverse data and user needs. Integrating approaches might offer a more comprehensive solution, balancing trade-offs in relevance, computational complexity, and adaptability to varying data structures.

Table 5. Comparison of TF-IDF with Collection Frequency against Other Techniques

Technique	Efficiency (ms/query)	Accuracy (%)	Optimization (%)
Proposed Model	5.2	85	70
Boolean Search	8.5	70	50
Vector Space Model	6.0	80	60
Latent Semantic Indexing	7.3	78	55
BM25	5.7	82	68
Word Embedding	4.8	88	72
Probabilistic Model	6.5	75	58

7. CONCLUSION

The following points highlight the superiority of the TF-IDF-CF method over other existing methods:

- **Precision in Weighting:** TF-IDF-CF offers a nuanced approach to term weighting. By considering both the frequency of a term in a document (TF) and its rarity across the collection (IDF-CF), it better captures the importance of a term within a specific document and its significance across the entire corpus. This precision helps in better ranking the relevance of documents to a query.

- **Contextual Understanding:** It provides a contextual understanding of terms by factoring in the frequency of terms across the entire collection. This context allows for a more nuanced understanding of the significance of a term to the entire dataset rather than just within a single document.

- **Optimized Retrieval:** TF-IDF-CF often leads to more optimized retrieval by highlighting rare but crucial terms that might hold significance across multiple documents. It strikes a balance between common terms and those that might be rare yet pivotal, improving the overall search quality.

- **Enhanced Discrimination:** Through IDF-CF, TF-IDF can discriminate against commonly occurring terms across the collection, reducing their weight in relevance ranking. This helps in emphasizing the importance of unique terms, leading to more precise and relevant search results.

- **Adaptability and Versatility:** TF-IDF-CF can be adapted and applied in various domains and scenarios. It's versatile enough to handle different types of collections, from small datasets to large corpora, making it a flexible and widely applicable technique.

- **Balanced Approach:** Unlike some other techniques that might focus solely on frequency or co-occurrence, TF-IDF-CF strikes a balance between these factors, providing a more holistic perspective on term importance and relevance.

Overall, TF-IDF-CF's ability to combine term frequency with collection frequency allows for a more nuanced and context-aware understanding of the importance of terms in documents. This often leads to superior performance in information retrieval tasks compared to some other existing methods in keyword search.

In this comprehensive study on the intersection of blockchain technology and Search Engine Optimization (SEO), we've uncovered a paradigm shift with the potential to redefine the future of online visibility. Traditional SEO practices, centered around centralized structures and keyword dominance, face a disruptive challenge from blockchain's decentralized architecture. Through a meticulous exploration of mechanisms like Term Frequency-Inverse Document Frequency (TF-IDF) coupled with collection frequency (CF) within blockchain frameworks, this research demonstrates a transformative shift toward more accurate, context-aware assessments of document relevance. By employing cosine similarity as a metric to rank search results within blockchain-stored data, the study highlights a path toward unbiased, manipulation-free result rankings, fundamentally altering the landscape of SEO.

Crucially, this research does not merely dwell in theoretical realms. Rigorous analyses across varied datasets of different sizes underscore the scalability and efficacy of blockchain-based SEO strategies. Beyond academia, this work stands as a call to action, inviting stakeholders, researchers, and practitioners to join in shaping an SEO landscape that prioritizes user trust, data privacy, and genuine relevance. The envisioned future is not just about algorithms; it's about empowerment, where users actively influence their online experiences, marking a pivotal step toward a fairer, more inclusive digital realm where transparency and user-centricity reign supreme.

References:

Alizadeh, M., Shir, S., Vishwanath, T., Chu, P. Gibbons, & Lakshman, T. (2023). A survey of decentralized search engine technologies. In *IEEE International Conference on Distributed Computing Systems (ICDCS)* (pp. 1467-1478).

Anderson. M. (2010), *Search Engine Optimization: The Beginner's Guide*. O'Reilly Media.

Androulaki, S., Kokolakis, E., Maniatis, P., Russell, A., Sousa, A., & Vardoulakis, I. (2016). Designing a decentralized ledger of identities for the Internet, *ACM SIGSAC Conference on Computer and Communications Security*, pp. 1467-1484.

Baeza-Yates, R., & Ribeiro-Neto, B. (2011). *Modern information retrieval.* ACM Press.

Buchinger, S., Schranz, M., & Müller, C. (2023). The Future of SEO: A User-Driven Content Model for the Decentralized Web. *ACM Internet Measurement Conference (IMC),* pp. 501-512.

Calderón-Monge, E., & Ribeiro-Soriano, D. (2023). Artificial intelligence and blockchain integration in business: Trends from a bibliometric-content analysis. *Information Systems Frontiers*, pp. 1-17.

Chen, J., Li, R., Wu, J., Xu, Y., & Wang, R. (2023). A literature review on decentralized web (Web3) search engines, *Digital Communications and Networks*, *9*(3), 339-348.

Chen, W., Li, X., Wu, J., Liu, J., & Wang, R. (2022). TF-IDF-based text similarity measurement for decentralized search engines, *Journal of Ambient Intelligence and Human Computing*, *14*(3), 1509-1520,.

Croft, W. B., & Moffat, J. D. (1989). *Retrieval techniques*, Addison-Wesley Longman Publishing Co., Inc.

Cronen-Townsend, B., Afantsev, A., Weikum, T., Lumetsberger, C., Piwowarczyk, B., de Vries, A., Theobald, R., Aly, Y., Guo, P., Nogueira, P., & Weerkamp, W. (2017). A study of keyword disambiguation in the Web of Data. In *Proceedings of the 26th International Conference on World Wide Web* (pp. 1803-1813).

He, W., Song, P., Mittal, X., Jiang, B., & Yavuz, B. (2021). Differentially private search: A survey. *ACM Computing Surveys (CSUR)*, *54*(5), 1-50.

Kapoor, A., Agarwal, M., & Panda, T. (2015). A survey on search engine optimization (SEO) techniques. In *International Conference on Computational Intelligence and Communication Technologies* (pp. 757-762).

Kumar, N., Awasthi, S., & Tyagi, D. (2016). Web crawler challenges and their solutions. *International Journal of Scientific & Engineering Research*, *7*(12), ISSN 2229-5518.

Li, H., Zhang, Y., & Deng, R. H. (2020). Blockchain-based decentralized web search engine: A review. *IEEE Access*, *8*, 19537-19552.

Li, M., Zhang, J., & Ma, X. (2023). Improving information retrieval in blockchain-based decentralized search engines using topic modeling and TF-IDF. *Journal of Systems Architecture*, *140*, 102774.

Li, X., He, X., Huang, R., Xu, X., & Li, R. (2023). Blockchain technology for keyword search optimization: A comprehensive review. *International Journal of Information Management*, *56*, 102665.

Liang, B., Zhang, L., Ju, L., Liu, S., & Li, L. (2020). A secure and scalable blockchain-based framework for internet-of-things. *IEEE Transactions on Industrial Informatics*, *16*(5), 3606-3618.

Liu, J., Jiang, Y., Wu, J., Xu, Y., & Wang, R. (2022). Privacy-preserving keyword search over encrypted data using blockchain. *Journal of Network and Computer Applications*, *196*, 103211.

Manning, P. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.

Manning, P. D., Schütze, H., & Prabhakar, R. (2003). Inducing semantic similarity using probabilistic latent semantic analysis. In *22nd Annual International Conference on Research and Development in Information Retrieval* (pp. 517-526).

Mor, J., Rai, D., & Kumar, N. (2018). An XML-based web crawler with page revisit policy and updation in local repository of search engine. *International Journal of Engineering & Technology*, *7*(3), 1119-1123. https://doi.org/10.14419/ijet.v7i3.12924

Nguyen, T. N., Tran, V. A., Doan, A., & Dat, T. T. (2023). Enhancing text retrieval in blockchain-based decentralized search engines with semantic indexing and TF-IDF weighting. *International Journal of Semantic Web and Information Systems*, *43*(1), 3-18.

Nguyen, T. N., Tran, V. A., Doan, A., & Dat, T. T. (2023). Towards a semantic keyword search framework for decentralized search engines. *International Journal of Semantic Web and Information Systems*, *43*(1), 3-18.

Park, J., Seo, M. S., Choi, D., Seo, J., & Kim, J. (2022). Towards a human-centered search engine design: A review of human-computer interaction studies. *Journal of Information Science Theory and Practice (JIS)*, *13*(4), 1-20.

Rai, D., Kumar, N., & Mor, J. (2018). Review on improving performance of web crawler and search system architecture. *International Journal of Advanced Studies of Scientific Research*, *3*(10).

Raja Raman, A., & Ullman, J. D. (2008). *Mining of massive datasets*. Cambridge University Press.

Rezaee, E., Saghiri, A. M., & Forestiero, A. (2021). A survey on blockchain-based search engines. *Applied Sciences*, *11*(15), 7063.

Robertson, S. (1971). The SMART retrieval system: Experiments in relevance detection and document ranking. *Journal of Documentation*, *27*(1), 12-24.

Salton, G., & McGill, M. (1986). *Introduction to modern information retrieval*. McGraw-Hill.

Singh, N., Srivastava, D., & Chatterjee, S. (2022). Demystifying decentralized search: A comparative analysis of blockchain-based search engine proposals. In *International Conference on Intelligent Systems and Information Management (ICISIM)* (pp. 1-7).

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application to retrieval. *Journal of Documentation*, *28*(1), 11-21.

Swan, M. (2015). *Blockchain: Blueprint for a new economy*. O'Reilly Media, Inc.

Tapscott, D., & Tapscott, A. (2016). *Blockchain revolution: How the technology behind Bitcoin is transforming money, business, and our world*. Penguin Random House.

Tee, J., Bickson, D., Hulina, K., Joseph, A., & King, P. (2012). Distributed machine learning for web personalization. *ACM Queue*, *10*(5), 41-50.

Wang, Y., Sun, W., Xu, J., Jiang, Y., & Wang, R. (2022). A hybrid approach for ranking documents in blockchain-based decentralized search engines using TF-IDF and PageRank. *Multimedia Tools and Applications*, *81*(13), 18063-18079.

Zhang, J., Lu, J., Ma, X., & Jiang, X. (2023). Blockchain-based decentralized search engine: A technical perspective. *Computers & Security*, *134*, 102755.

Zhang, Y., Lin, M., & He, Q. (2019). Blockchain-based decentralized search engine: Towards a secure, transparent, and trustable web search. *IEEE Access*, *7*, 120578-120594.

Zhang, Y., Lu, X., & Zhang, J. (2023). A study on the use of TF-IDF in decentralized search engines for information retrieval. *Journal of Network and Computer Applications*, *219*, 103820.

Bharti Aggarwal
Ph.D. Research Scholar, School of Eng & Tech, Sushant University, Gurugram, India
bharti_goel2003@yahoo.com
ORCID 0009-0007-6078-3085

Dinesh Rai
School of Engineering and Technology, Sushant University, Gurugram,India
dineshrai@sushantuniversity.edu.in
ORCID 0009-0001-6538-4944

Naresh Kumar
Maharaja Surajmal Institute of Technology, Janakpuri, New Delhi, IndiaTop of Form
narsumsaini@gmail.com
ORCID 0000-0001-9984-506X