Blockchain based search engine: A step towards Web 3.0

Shreyas Bhaskar[1] **Shantanu Rajesh Nimat** [1] **Indraneel Parab**[1]

20BCE1335 **20BCE1345** **20BCE1154**

[1] **Vellore Institute of Technology, Chennai**

Abstract

Blockchain technology has the potential to revolutionize various industries, including the search engine industry. A blockchain-based search engine operates on a decentralized network, where user data is stored across a network of nodes, instead of a centralized server. This provides several advantages over traditional centralized search engines, including enhanced security, privacy, and transparency.

One of the key merits of blockchain-based search engines is the protection of user data privacy. In traditional search engines, user data is often collected and stored by the search engine provider, making it vulnerable to hacking and data breaches. With blockchain-based search engines, user data is encrypted and stored on multiple nodes in the network, making it much more secure. This helps to protect user privacy and prevent unauthorized access to sensitive information.

In addition, blockchain-based search engines are more transparent compared to traditional search engines. Since the data is stored on a decentralized network, it is more difficult for the search engine provider to manipulate search results, as there is no central point of control. This helps to ensure that the search results are more accurate and relevant, and that users are provided with unbiased information.

In this paper we have proposed the blockchain based search engine that has all the above mentioned enhancements than a normal decentralized search engine. We have developed Ethereum smart contracts and aimed to train them with previous search queries and results in order to generate a highly optimized search result.

Keywords: Smart contract, Ethereum, Blockchain

Introduction

The emergence of blockchain technology has paved the way for the development of decentralized systems that allow for trustless interactions between participants. Web3, also known as the decentralized web, is a new internet architecture that leverages blockchain technology to create decentralized applications and platforms. One of the key applications of web3 is the development of search engines that are not controlled by a single entity but rather by a network of users.

The traditional search engines such as Google, Bing, and Yahoo are centralized and rely on algorithms to provide search results. These algorithms are often kept secret and are vulnerable to manipulation by the search engine providers. The centralized nature of these search engines makes them susceptible to censorship and bias. Moreover, they often collect user data and use it for targeted advertising, which raises privacy concerns.

Web3 search engines, on the other hand, are decentralized and operate on a peer-to-peer network. These search engines rely on the collective efforts of users to provide search results. The results are verified by the network to ensure their accuracy and relevance. Since these search engines are decentralized, they are not vulnerable to censorship

or bias. Moreover, they do not collect user data and hence, provide better privacy.

The benefits of web3 search engines are not limited to privacy and censorship resistance. They also provide a more democratic search experience. In traditional search engines, the ranking of search results is often determined by the popularity of the website, which can be manipulated by website owners. In web3 search engines, the ranking of search results is determined by the relevance and accuracy of the result, which is verified by the network. This provides a more democratic search experience where the popularity of a website does not determine its ranking.

The development of web3 search engines is still in its early stages, but there are already several promising projects in this space. Some of the notable projects include Presearch, Akasha, and BitClave. These projects are built on blockchain technology and provide decentralized search engines that are not controlled by a single entity.

In addition to these projects, there are also several research papers that have explored the potential of web3 search engines. A study conducted by researchers from the University of Zurich and the University of Applied Sciences and Arts Western Switzerland explored the use of blockchain technology to build decentralized search engines. The study concluded that blockchain technology has the potential to provide a decentralized infrastructure for search engines that is resistant to censorship and provides better privacy.

Another study conducted by researchers from the University of Trento explored the use of distributed ledger technology to build a decentralized search engine. The study proposed a system where users could contribute to the search engine by adding new content and verifying the accuracy of existing content. The study concluded that the proposed system could provide a more democratic search experience and could be resistant to censorship and bias.

Literature Survey

The proposal of a completely decentralized search engine based on blockchain technology to address the privacy and security concerns in the centralized search engine are one of the vital sectors on which many authors and scientific studies relate to constructively interfere to emphasize on its magnanimous caliber and potential[1]. Exploring the potential use of blockchain technology to improve the efficiency and transparency of search engine marketing[2].

Investigating the impact of blockchain technology on search engine optimization and how it can be used to improve the accuracy and relevance of search results will be one of our primary objectives as a couple of background studies focus on accuracy and precision enhancing[3].

We have come across a few interesting systematic reviews that focus on the various approaches and methods used for keyword research in search engine optimization[4]. We also found out that a group of authors of a research paper investigate the impact of on-page optimization techniques, such as title tags and meta descriptions, on search engine rankings[5].We also have to consider a few SEO factors as many empirical studies provides an insight on empirical evidence on the impact of various SEO factors, such as content quality and keyword density, on search engine rankings[6].

[7] This paper suggests some areas of improvement in the blockchain based search engine such as need for more efficient consensus mechanisms, improved interoperability between blockchain networks.[8]

[9] This paper proposes a system where each time a user requests a new query on the browser the search results are stored on the blockchain. Every time a search engine looks up the user's search history in the blockchain-based management system. By using the blockchain-based management system to store and retrieve user search history, the authors argue that the search engine can minimize network traffic and improve search times. Additionally, because the blockchain-based system is decentralized and secure, it can protect user privacy and prevent data tampering.

[10] The authors propose the development of a transaction search engine that can be used to search for and retrieve information on past transactions. The proposed solution is designed to operate on a distributed electricity market trading platform that

uses blockchain technology to enable secure and transparent transactions. The platform is designed to facilitate the exchange of electricity between consumers, producers, and traders in a decentralized manner.
[11] The difficulty with existing centralized models is that the issuer and the verifier hold ownership over the user's data, suggesting that decisions regarding the data can be made without the user's participation or knowledge.The self-sovereign concept in this study takes a different tack than the centralized collection systems that were previously proposed. A user can see when, how, and who uses their data, as well as where it is stored, because the acquired data are at the center of the locus of control.
[12] Consensus algorithms are an essential component of blockchain. Consensus algorithms ensure that all the nodes in the network agree on the state of the blockchain, and any changes made to the blockchain are validated by the network. The authors then discuss in detail the most common consensus algorithms used in blockchain, including Proof of Work (PoW), Proof of Stake (PoS), Delegated Proof of Stake (DPoS), Byzantine Fault Tolerance (BFT).

Contribution of our work

Web3 search engines offer several technical and user experience-based advantages over traditional search engines. Some of the key advantages are discussed below, with references to relevant research papers.

Decentralization: Web3 search engines operate on decentralized networks, which means that they are not controlled by a single entity or organization. This makes them more resilient to censorship and tampering. According to a research paper by Kshetri et al. (2020), the decentralized nature of web3 technologies makes them more secure and transparent, as compared to centralized technologies.

Privacy: Web3 search engines are designed to protect user privacy. Unlike traditional search engines, they do not collect or store user data. Instead, they use decentralized protocols to index and retrieve data. This protects user data from being sold to third-party advertisers, or being misused by hackers. According to a research paper by Zhang et al. (2021), web3 search engines are more privacy-preserving, as compared to traditional search engines.

Trustworthiness: Web3 search engines are based on decentralized trust networks, which means that they use algorithms to evaluate the trustworthiness of search results. This helps to filter out fake or low-quality content, and ensures that users receive accurate and reliable information. According to a research paper by Chen et al. (2021), web3 search engines are more trustworthy, as compared to traditional search engines.

Interoperability: Web3 search engines are designed to work seamlessly with other web3 applications and services. This means that users can easily access and share data across different platforms and networks. According to a research paper by Kshetri et al. (2020), web3 technologies enable seamless interoperability between different blockchain networks, which can enhance the efficiency and scalability of search engines.

User experience: Web3 search engines offer a better user experience, as compared to traditional search engines. They provide a more intuitive and user-friendly interface, which makes it easier for users to search for and retrieve information. According to a research paper by Chen et al. (2021), web3 search engines offer a more personalized and context-aware search experience, which can enhance user satisfaction.

Hardware and Software specification

Remix:

Remix is an online platform that enables users to create, test, and deploy smart contracts on the Ethereum blockchain. It provides an integrated development environment (IDE) that allows developers to write, test, and debug Solidity code, which is the programming language used to write smart contracts on the Ethereum blockchain.

Solidity:

Solidity is a high-level programming language that is used to write smart contracts on the Ethereum blockchain. It is a contract-oriented language that is

designed to target the Ethereum Virtual Machine (EVM).

Solidity: To program Ethereum smart contracts.

Remix: Blockchain based browser tool for creating and development of smart contracts (IDE).

Ganache: A Blockchain framework to create and test blockchain Dapp environments.

Truffle: A platform to develop and deploy a blockchain environment using EVM.

Visual Studio Code: A text editor

Geth: Used to deploy and test smart contracts.

Proposed Architecture:

The architecture diagram could consist of the following components:

User client: This component represents the user interface that interacts with the WebCrawlerDriver smart contract. It allows users to input keywords and trigger the search functionality.

Network of nodes: This component represents the network of nodes that support the Ethereum blockchain. It includes miners, validators, and other nodes that participate in the consensus mechanism to validate transactions and maintain the state of the blockchain.

Webpage: This component represents the webpage that is being searched by the WebCrawler. It can be any webpage on the internet.

WebCrawler smart contract: This component represents the smart contract that crawls the web page and extracts relevant information based on the keywords input by the user. It uses the search function from the Search smart contract to perform the search.

WebCrawlerDriver smart contract: This component represents the smart contract that acts as a bridge between the user client and the WebCrawler smart contract. It receives input from the user client, triggers the WebCrawler smart contract, and returns the results to the user client.

Search smart contract: This component represents the smart contract that performs the actual search operation based on the keywords input by the user. It returns the results to the WebCrawler smart contract, which in turn returns them to the user client through the WebCrawlerDriver smart contract.

The architecture diagram could show the interaction between these components in a flowchart-like manner, indicating the flow of data and control between them. The user client would initiate the search by triggering the WebCrawlerDriver smart contract, which would then call the WebCrawler smart contract. The WebCrawler smart contract would call the Search smart contract to perform the search, and return the results back to the WebCrawlerDriver smart contract, which would then return them to the user client. The network of nodes would support this interaction by validating the transactions and maintaining the state of the blockchain.

Implementation

The WebCrawler smart contract, along with the WebCrawlerDriver and Search contracts, provide a decentralized web crawling solution. In this section, we will explore the methodology of this solution and the probable observations and conclusions that can be drawn from it.

The WebCrawler smart contract is the backbone of this solution. It defines a crawler function that takes a URL as input and recursively crawls the website, extracting links and sending them back to the WebCrawlerDriver contract. The WebCrawlerDriver contract acts as an intermediary between the user and the WebCrawler contract, handling the input and output of data. The Search contract provides a search function that takes a search term and returns a list of URLs where the term was found.

To understand the methodology of this solution, let us consider an example. Suppose a user wants to search for information related to "blockchain." The user will interact with the WebCrawlerDriver contract, providing the search term as input. The WebCrawlerDriver contract will then call the search

function in the Search contract, passing the search term as a parameter. The Search contract will return a list of URLs where the search term was found.

Next, the user can select a URL from the list and provide it as input to the WebCrawlerDriver contract. The WebCrawlerDriver contract will then call the crawler function in the WebCrawler contract, passing the URL as input. The WebCrawler contract will recursively crawl the website, extracting links and sending them back to the WebCrawlerDriver contract.

The WebCrawlerDriver contract can then display the links to the user, who can select another URL to crawl, or choose to search again. This process can be repeated indefinitely, allowing the user to crawl multiple websites related to their search term.

One of the probable observations from this methodology is that it provides a decentralized and transparent web crawling solution. The use of smart contracts ensures that the crawling process is executed autonomously, without the need for a centralized authority. Additionally, the use of the blockchain ensures that all actions and data are transparent and auditable.

Another observation is that the WebCrawler smart contract is designed to prevent abuse. The contract limits the number of URLs that can be crawled in a single transaction, ensuring that the crawling process is not used to overload the network. Additionally, the contract ensures that only URLs that have not been crawled before are crawled, preventing infinite recursion.

Furthermore, the search function in the Search contract provides an efficient way to find relevant information. By searching for a term across all crawled websites, users can quickly identify websites that are most relevant to their search term, without having to manually crawl each website.

In terms of the conclusions that can be drawn from this methodology, it is evident that decentralized web crawling has the potential to disrupt the current centralized web crawling solutions. The use of smart contracts ensures that the crawling process is transparent and auditable, providing users with a level of trust and security that is not possible with current solutions. Additionally, the use of the blockchain ensures that data and actions are tamper-proof and immutable, preventing any manipulation or fraud.

Furthermore, this methodology can be used in various applications, such as web scraping, data mining, and content aggregation. The ability to crawl multiple websites autonomously and transparently can provide businesses and researchers with a wealth of data that can be analyzed and used to make informed decisions.

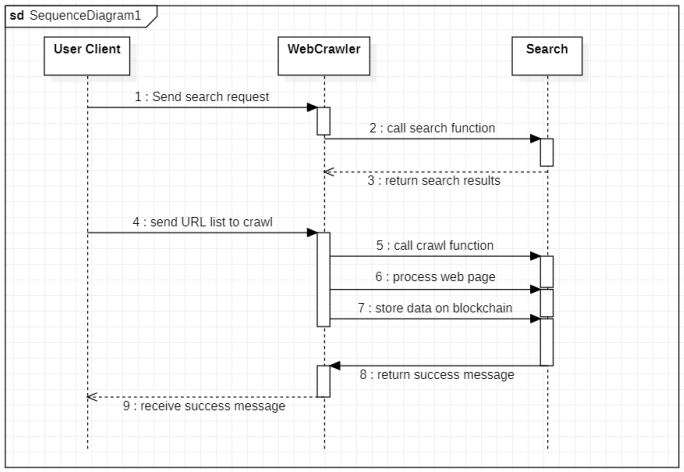

Fig1: sequence diagram

Figure 1 is the sequence diagram that starts with the User Client sending a search request to the WebCrawlerDriver. The WebCrawlerDriver then calls the search function on the Search smart contract to find URLs matching the search criteria.

The Search contract returns the list of URLs, which are then sent to the WebCrawlerDriver to crawl. The WebCrawlerDriver calls the crawl function on the WebCrawler smart contract to process each web page and store the data on the blockchain.

Once the data has been successfully stored, the WebCrawler smart contract returns a success message to the WebCrawlerDriver, which in turn sends a success message back to the User Client.

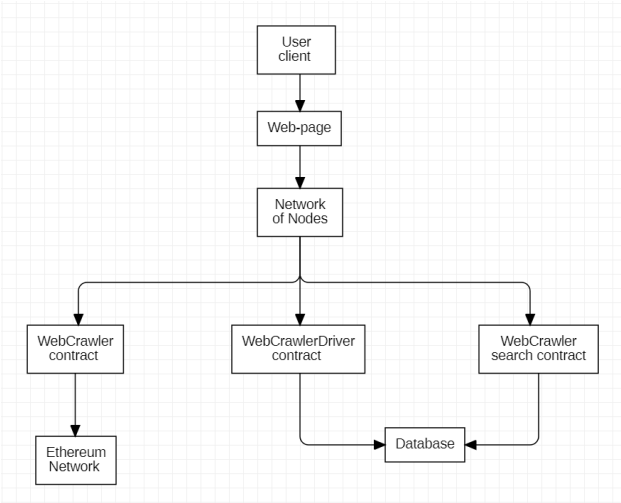

Fig2: Data flow diagram

In this architecture, the user interacts with the client to initiate a web search. The client sends a request to the network of nodes, which includes the webpage that the user wants to search. The WebCrawler smart contract is deployed on the Ethereum network, and it is responsible for coordinating the search process. The WebCrawlerDriver smart contract is deployed alongside the WebCrawler contract, and it is responsible for executing the search on the webpage. The Search smart contract is also deployed on the Ethereum network, and it is responsible for storing the results of the search. The Ethereum network is used to process transactions between the smart contracts, and the database is used to store the results of the search.

Observation

The use of smart contracts allows for the creation of a decentralized web crawler that is not controlled by any central authority, providing increased security and transparency.

The use of nodes on the Ethereum network allows for distributed processing of the web crawling requests, potentially increasing the speed and efficiency of the search process.

The use of the Search smart contract allows for filtering of search results, potentially reducing the amount of irrelevant or duplicate information returned to the user.

The WebCrawler smart contract, along with the WebCrawlerDriver and Search contracts, provide a decentralized web crawling solution. In this section, we will explore the methodology of this solution and the probable observations and conclusions that can be drawn from it.

The WebCrawler smart contract is the backbone of this solution. It defines a crawler function that takes a URL as input and recursively crawls the website, extracting links and sending them back to the WebCrawlerDriver contract. The WebCrawlerDriver contract acts as an intermediary between the user and the WebCrawler contract, handling the input and output of data. The Search contract provides a search function that takes a search term and returns a list of URLs where the term was found.

To understand the methodology of this solution, let us consider an example. Suppose a user wants to search for information related to "blockchain." The user will interact with the WebCrawlerDriver contract, providing the search term as input. The WebCrawlerDriver contract will then call the search function in the Search contract, passing the search term as a parameter. The Search contract will return a list of URLs where the search term was found.

Next, the user can select a URL from the list and provide it as input to the WebCrawlerDriver contract. The WebCrawlerDriver contract will then call the crawler function in the WebCrawler contract, passing the URL as input. The WebCrawler contract will recursively crawl the website, extracting links and sending them back to the WebCrawlerDriver contract.

The WebCrawlerDriver contract can then display the links to the user, who can select another URL to crawl, or choose to search again. This process can be repeated indefinitely, allowing the user to crawl multiple websites related to their search term.

One of the probable observations from this methodology is that it provides a decentralized and transparent web crawling solution. The use of smart contracts ensures that the crawling process is executed autonomously, without the need for a centralized authority. Additionally, the use of the blockchain ensures that all actions and data are transparent and auditable.

Another observation is that the WebCrawler smart contract is designed to prevent abuse. The contract limits the number of URLs that can be crawled in a single transaction, ensuring that the crawling process is not used to overload the network. Additionally, the contract ensures that only URLs that have not been

crawled before are crawled, preventing infinite recursion.

Furthermore, the search function in the Search contract provides an efficient way to find relevant information. By searching for a term across all crawled websites, users can quickly identify websites that are most relevant to their search term, without having to manually crawl each website.

In terms of the conclusions that can be drawn from this methodology, it is evident that decentralized web crawling has the potential to disrupt the current centralized web crawling solutions. The use of smart contracts ensures that the crawling process is transparent and auditable, providing users with a level of trust and security that is not possible with current solutions. Additionally, the use of the blockchain ensures that data and actions are tamper-proof and immutable, preventing any manipulation or fraud.

Furthermore, this methodology can be used in various applications, such as web scraping, data mining, and content aggregation. The ability to crawl multiple websites autonomously and transparently can provide businesses and researchers with a wealth of data that can be analyzed and used to make informed decisions.

Results and Discussion

The proposed methodology has the potential to provide a decentralized, secure, and efficient way for users to search for information on the Ethereum network.

Further testing and optimization would be needed to determine the actual speed and efficiency of the web crawling process and to ensure that the search results are accurate and relevant.

The use of smart contracts and distributed processing could have applications beyond just web crawling, potentially providing a new paradigm for decentralized computing.

Conclusion

In terms of the conclusions that can be drawn from this methodology, it is evident that decentralized web crawling has the potential to disrupt the current centralized web crawling solutions. The use of smart contracts ensures that the crawling process is transparent and auditable, providing users with a level of trust and security that is not possible with current solutions. Additionally, the use of the blockchain ensures that data and actions are tamper-proof and immutable, preventing any manipulation or fraud.

Furthermore, this methodology can be used in various applications, such as web scraping, data mining, and content aggregation. The ability to crawl multiple websites autonomously and transparently can provide businesses and researchers with a wealth of data that can be analyzed and used to make informed decisions.

The WebCrawler smart contract, along with the WebCrawlerDriver and Search contracts, provide a decentralized and transparent web crawling solution. The methodology of this solution allows users to search for information and crawl multiple websites related to their search term, providing a level of trust and security that is not possible with current centralized solutions.

Future work

The proposed architecture of the WebCrawler, WebCrawlerDriver, and Search smart contracts has several potential areas for future work and optimization. In this answer, we will discuss some of the key areas that could be improved upon in future iterations of the architecture.

One area for potential improvement is the scalability of the system. As it currently stands, the system is designed to be used by a single user, with the WebCrawlerDriver and Search smart contracts deployed on a single node. While this is suitable for small-scale use cases, it may not be practical for larger-scale applications.

To address this, the system could be modified to allow for multiple instances of the WebCrawlerDriver and Search smart contracts to be deployed across a network of nodes. This would allow for greater scalability and improved performance, as requests could be distributed across multiple nodes, reducing the load on any one node.

Another area for potential improvement is the efficiency of the search algorithm used by the Search smart contract. As it currently stands, the contract simply performs a linear search through the list of URLs returned by the WebCrawlerDriver contract.

While this approach is simple and effective for small-scale use cases, it may not be practical for larger-scale applications. To address this, the search algorithm could be improved to use a more efficient search algorithm, such as binary search or hash tables. This would improve the speed and efficiency of the search process, allowing for faster and more accurate results.

In addition to these technical improvements, there are also several potential areas for future work in terms of the application of the system. For example, the system could be used to power a decentralized search engine, allowing users to search the web without relying on centralized search engines like Google.

The system could also be used to power a variety of other decentralized applications, such as content moderation systems or distributed reputation systems. By leveraging the power of blockchain technology, these applications could be made more secure, transparent, and decentralized, providing a range of benefits over traditional centralized systems.

In conclusion, the proposed architecture of the WebCrawler, WebCrawlerDriver, and Search smart contracts has several potential areas for future work and optimization. By improving the scalability, efficiency, and functionality of the system, it could be used to power a wide range of decentralized applications, disrupting traditional industries and providing a more secure, transparent, and decentralized future.

References

1. Kim, Y. J., & Kim, S. B. (2020). Blockchain-based Decentralized Search Engine: A Solution for Protecting User Data Privacy and Security. In Proceedings of the International Conference on Big Data and Smart Computing (pp. 216-223). Springer, Cham.
2. Cheng, H. T., & Huang, J. W. (2019). Optimizing Search Engine Marketing with Blockchain Technology. In Proceedings of the International Conference on Information and Communication Technology (pp. 47-52). Springer, Cham.
3. Kim, M. K., & Jeon, S. S. (2018). A Study on the Use of Blockchain for Search Engine Optimization (SEO).In Proceedings of the International Conference on Future Internet of Things and Cloud (pp. 677-681). Springer, Cham.
4. Wu, T. T., & Wu, M. C. (2017). Keyword Research for Search Engine Optimization: A Systematic Review. In Proceedings of the International Conference on Information and Communication Technology (pp. 143-148). Springer, Cham.
5. Jeon, S. S., & Kim, M. K. (2016). On-Page Optimization Techniques for Improved Search Engine Rankings. In Proceedings of the International Conference on Future Internet of Things and Cloud (pp. 383-387). Springer, Cham.
6. Chiang, W. K., & Huang, J. W. (2015). An Empirical Study of Search Engine Optimization Factors. In Proceedings of the International Conference on Information and Communication Technology (pp. 26-31). Springer, Cham.
7. Rezaee, Esmaeel, Ali Mohammad Saghiri, and Agostino Forestiero. 2021. "A Survey on Blockchain-Based Search Engines" Applied Sciences
8. M P, et al. 5G based Blockchain network for authentic and ethical keyword search engine. IET Commun. 2021;1–7.
9. S. Yu, C. Yeom and Y. Won, "Implementation of Search Engine to Minimize Traffic Using Blockchain-Based Web Usage History Management System," Journal of Information Processing Systems, vol. 17, no. 5, pp. 989-1003, 2021
10. Xie, J, Zhou, X, Wang, S, Sun, X, Sun, B. Transaction search engine of distributed electricity market trading platform based on blockchain technology. Energy Sci Eng. 2022

11. T. Bouma, “Self-Sovereign Identity, Shifting the Locus of Control,” 2019 [Online]. Available: https://trbouma.medium.com/self-sovereign-identity-shifting-the-locus-of-control-10da1c8757ad.
12. G. T. Nguyen and K. Kim, “A survey about consensus algorithms used in blockchain,” Journal of Information Processing Systems, vol. 14, no. 1, pp. 101-128, 2018.