Information Retrieval

Blockchain Search Engine Abstract

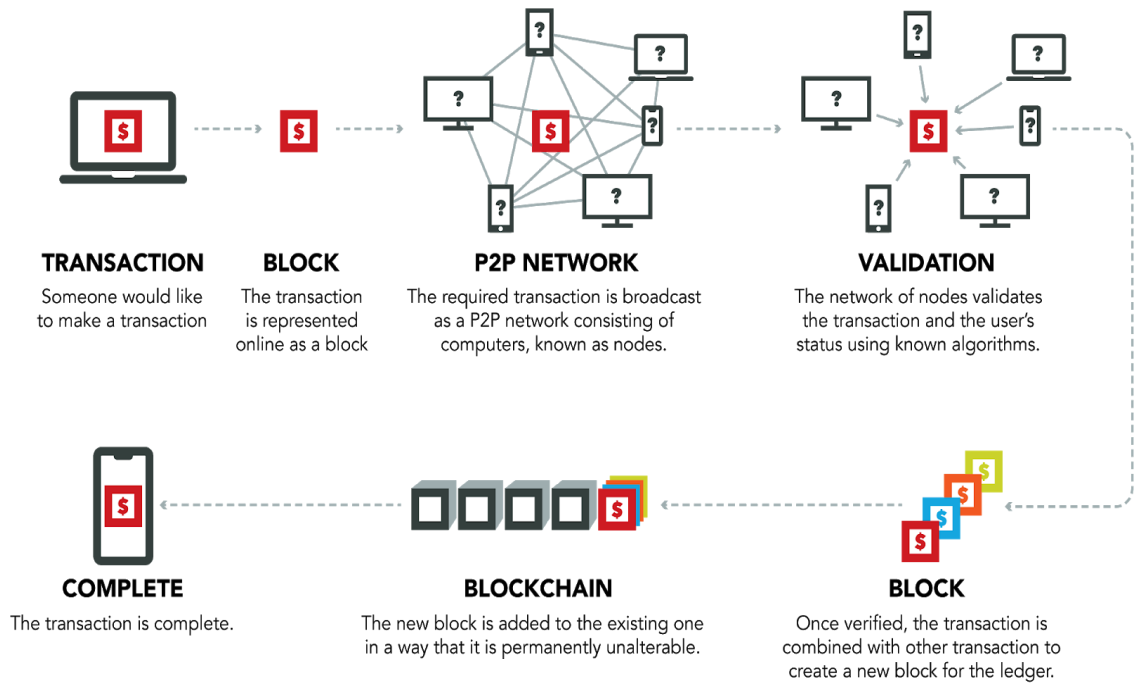
Team

Nethish R (16PT21)
Nithish S (16PT24)

Blockchain

Blockchain, sometimes referred to as Distributed Ledger Technology (DLT), makes the history of any digital asset unalterable and transparent through the use of decentralization and cryptographic hashing. It is a revolutionary technology that reduces risk and keeps track of all the transactions in an organization.

A blockchain is an ever-growing list of records called blocks which are linked using cryptography. Cryptography is a process which encrypts and secures data communication to prevent third-parties from reading private messages. Blockchain technology is most commonly used by cryptocurrencies. Once the data has been recorded in a place, it will not be changed. It works just like a digital notary with timestamps to avoid tampering of information.



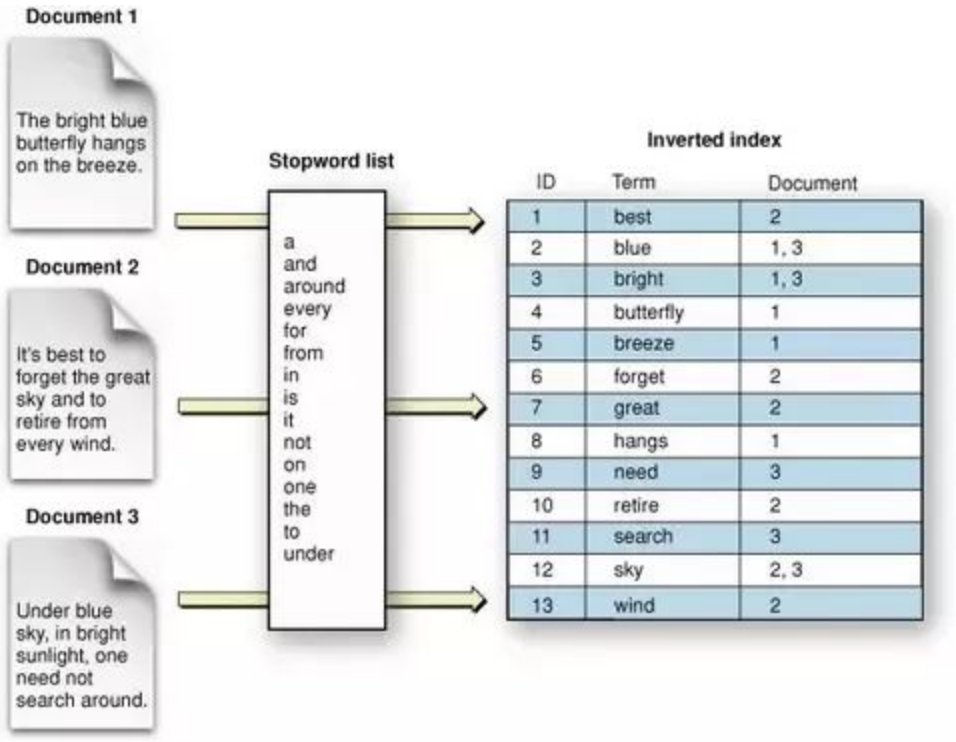
Blockchain Search Engine

There are numerous applications of blockchain and it's still being evolved. The aim is to build a search engine to investigate the current state of blockchain technology and its applications and how this technology can revolutionize modern business.

There are many theories and research papers published on IEEE(https://www.ieee.org/), Springer(https://www.springer.com/in) and in many more websites. If one wants to build an application or a model that uses blockchain technology they would have to research this area a bit to enhance the use of this technology. With its wide range of applications in the field of Cryptocurrencies, Financial services, Supply chain, Video games, Energy trading etc, many research papers are also being published in other fields too. It is very likely that the same topic has been researched and published on the internet. So building a blockchain search engine dedicated to search for the papers of the area of interest would be helpful to organizations that want to use this technology in their own way.

Indexing - Inverted Index

An inverted index (also referred to as a postings file or inverted file) is a database index storing a mapping from content, such as words or numbers, to its locations in a table. The inverted index is mainly used in search engines. The goal is to find all documents that have a particular word in it. The inverted index helps to find these kinds of queries. Building requires scraping the document, removing stop words and indexing. Exposing the data to preprocessing techniques like stemming would also increase accuracy.



TF-IDF

The term frequency of a word in a document. There are several ways of calculating this frequency, with the simplest being a raw count of instances a word appears in a document. Then, there are ways to adjust the frequency, by length of a document, or by the raw frequency of the most frequent word in a document.

The inverse document frequency of the word across a set of documents. This means, how common or rare a word is in the entire document set. The closer it is to 0, the more common a word is. This metric can be calculated by taking the total number of documents, dividing it by the number of documents that contain a word, and calculating the logarithm.

So, if the word is very common and appears in many documents, this number will approach 0. Otherwise, it will approach 1.

$$idf\,(t, D) = log\,(\ \frac{N}{count\ (d \in D{:}t \in d)}\ )$$

Querying and Ranking

The documents are first used to construct a document term matrix. These values are normalized and tf-idf scores are calculated. Now each of the documents is represented as vectors with normalized tf-idf values for terms.

The structure of a query is a simple string consisting of space separated words. This query undergoes a construction and restructuring process corresponding to the tf-idf matrix. Then this restructured query is used for Ranking the documents. The words that are not present in the copus is completely ignored.

Cosine similarity metric is used for finding the similar documents and ranking.

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

The lowest similarity is 0 which says there are no similarities between documents and highest is 1 which indicates both the documents are the same. Basically higher the value more similar the documents are. A simple search returns most similar 5 links.

```
>>> Search: blockchain technology
https://dl.acm.org/doi/abs/10.5555/2093889.2093944
https://link.springer.com/chapter/10.1007/978-3-662-44774-1_12
https://www.tandfonline.com/doi/abs/10.1080/13504851.2014.916379
https://dl.acm.org/doi/abs/10.1145/2810103.2813686
https://link.springer.com/chapter/10.1007/978-3-319-07536-5_6
>>> Search:
```