Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 8, July 2021: 3222-3230

Content -Based data sharing in Block-IPFS Network

G. Subathra[1], Dr.A. Antonidoss[2]

[1]Department of Computer Science and Engineering, Hindustan Institute of Technology and Sciences, Chennai
[2]Department of Computer science and Engineering, Hindustan Institute of Technology and Sciences, Chennai
rs.sg1018@hindustanuniv.ac.in[1], aro.antoni@gmail.com [2]

Abstract

Interplanetary file system is a peer-peer hypermedia network runs in a distributed environment that heads the content-based data. the sharing of information content in IPFS network enrol a pointer which inhibits from a cryptographic hash function, SHA-256. In an existing system, handling of pointer is far difficult for the user since its unique access. Here in this paper method of creating the hash for finding out the similarities among the content is being proposed. This approach helps the user to identify the content in an IPFS, which enhances the usability.

key terms: IPFS, Hashing functions, Blockchain network.

I. Introduction

The content-based data / information is handled by the Interplanetary File System by inhibiting a pointer that is generated using a cryptographic hash function, SHA-256[1]. This results in a unique value. In an existing system, the content is easily managed by the users. Integrity is being taken into account in order to remedy a problem in an existing file system as well as IPFS by improving the hash function.

Assuming hash function is not limited for its similarity content. Hence, it’s easy for the users to find the similarity content in IPFS and handling them in an existing file system. Here, proposing a hash algorithm for encoding the similarity of the content. [2] perceptual hash and fuzzy hashing [3] are the existing hash algorithm which exhibits the similarity content. Perceptual hash algorithm mainly specialised for images whereas fuzzy hash algorithm computes the content of data. Also, fizzy algorithm inhibits the length of a hash when IPFS is fixed [4].

But still the approach in traditional system is not being compromised therefore here comes Rolling Hash Algorithm where it exposes the similarity in byte sequence with that of the content given. This algorithm mainly focuses on resulting the unique value that depends on byte sequence [5]. The data sharing in Interplanetary file system shown in figure 1.

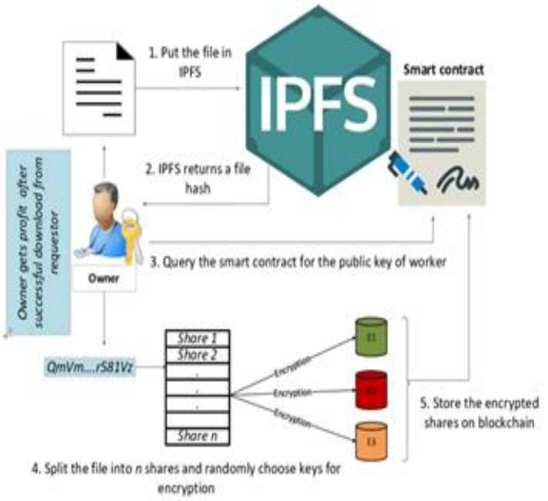

1. Put the file in IPFS
IPFS
Smart contract
2. IPFS returns a file hash
Owner gets profit after successful download from requestor
Owner
3. Query the smart contract for the public key of worker
QmVm....rS81Vz
Share 1
Share 2
Share n
Encryption
Encryption
Encryption
5. Store the encrypted shares on blockchain
4. Split the file into n shares and randomly choose keys for encryption

Figure 1. Storing information using IPFS

The Interplanetary file sharing concept utilized in the web archive to sort the one gigabyte information of HTTPS information [6] [7]. It sorted the 21,994 mementos. It creates hashes for the https header and payload for an URL. Once the user requested an address, the IPFS fetch information based on these hashes. Web archiver takes 66.6 minutes to store information to IPFS and IPFS process 570 files in a minute.

[8] Take advantage of the speed and redundancy of the new interplanetary file system IPFS. In a completely decentralized manner, opus can deliver more than 1000 songs. In IPFS, the encrypted music tracks are stored instead of OPUS directly. While in opus smart contract, the key for the encrypted music and hashes are stored in the Ethereum smart contract. Due to this, the OPUS provides an opportunity to the artist to sale their records to user for a monetary benefit. It also uses additional governance mechanisms, such as Opus DAO (decentralized Autonomous organizations) are funded by a small portion of everyone Sales and artist bounty system, allowing artists to pay a small fee to share their footprints.

[9] The content of this paper describes the tutorial sessions provided on ICWE 2016, with a focus on building distributed web applications using IPFS. IPFS, the Interplanetary File System, is a distributed permanent network, a protocol that makes the network faster, safer and more open. It gives information regarding the library files and data structures used in IPFS to store the web applications through Core applications, HTTP applications and CLI.

[10] Linked Open Data (LOD) is a method for releasing the machine-readable, open data, so it will be more effective through a combination of semantic queries. LOD's first prototype using interplanetary file system (IPFS) built, P2P system based on Merkel DAGs, and content-addressed block storage.

[11] They offered a new zigzag based storage model to enhance the volume storage model provided by IPFS. Also, they combined the Blockchain to connect IPFS with this storage model. The analysis showed that this project can effectively solve the High-performance issue for individual users in IPFS by introducing the role of content service providers. It also achieved data reliability and availability, and reduced storage overhead and other issues for service providers.

[12] Semantic web suffers from the problem of reuse and sharing of information due to its linked data nature. This problem can overcome by the prototype format. Based on this, the author combined this linked data and prototype format in interplanetary file system that helps to share the knowledge between the systems.

[13] Stated that the Internet of Things suffered from the data privacy and sharing problem due to its centralized access. Hence, this problem is overcome by utilizing the blockchain technology and Interplanetary file system in Internet of Things. In block chain model, the IoT communication and pee-peer data sharing at the top layer and it incorporate the IPFS for preserve the information.

[14] Stated that the cloud computing cannot be directly applied to the Internet of Things due to its high latency problem. Because the data centres located at a far distance. Due to this high latency occurs. To overcome this, a storage device is located in the fog/edge computing system and it termed as Scale-out Network Attached Storage systems and it mitigated the data access problems. Then, it also incorporated the interplanetary file system that helps to reduce the storage space and provide privacy to the data. By using this approach, the cloud computing can be incorporate with the Internet of Things.

This paper proposes a hash algorithm that outlets the similar values for similar content in IPFS.

II. Literature survey

Ohashi et al., (2019) suggested that it's possible that a new approach to data management in digital assets may develop [15]. Data is now managed individually since storing a large amount of data in a block chain can cause the ledger to grow enormously. The data associated with the token is saved using the specified way in a distributed content-addressable file system.

Vimal and Srivatsava (2019) utilized the IPFS and block chain, to offer a new way for improving the efficiency of P2P file sharing networks, which incorporates file transfer credibility and proximity awareness [16]. It also talked about the issues that block chain technology faces and how IPFS might help overcome them. The Interplanetary File System (IPFS) is a peer-to-peer (P2P) technology that uses hash file clusters in each node to share resources or data in a distributed system. Any user who wants to get a copy of one of these hashed files can simply call the file's hash. Issues with high throughput. Because the role of an IPFS single user is to construct a content service provider, IPFS employs the Filecoin idea, which stores data on a distributed network of local providers. Filecoin miners, particularly the transfer of resources for its successful collaboration service, have been awarded.

Benet (2014) IPFS is a high-throughput content-addressable block storage format with hyperlinks. Versioned file systems, blockchains, and even eternal networks can all be created with it. There is no single point of failure with IPFS, and nodes do not need to trust one another [17].

Sicilia et al., (2016) LOD (Linked Open Data) is a way to publish machine-readable open data that can be linked together. The current decentralized design of LOD clouds relies on location-specific services, which are known to cause availability and disconnection issues [18]. It describes the first prototype concept of an IPFS-based LOD, a Merkel DAG-based peer-to-peer (P2P) system, and a content-addressed block storage paradigm.

Naz et al., (2019) proposed the solution that achieves transparency, security, access control, owner reliability, and data quality [19]. Smart contracts are reliably written and applied to the local Ethereum test network. The user first uses the RSA signature for identity verification, and then submits the requested amount as the price of the digital content. After successfully submitting the data, users are encouraged to post comments about the data. Watson Analyzer will verify the comments and filter out fake comments. Customers who sign up for appropriate reviews will be rewarded.

Steichen et al., (2018) proposed the Interplanetary File System (IPFS) is a modified version of the Interplanetary File System (IPFS), which uses Ethereum smart contracts to provide access-controlled file sharing [20]. The access control list is maintained by a smart contract and enforced by the updated IPFS software. When uploading, downloading or transferring files, it will interact with smart contracts.

Huang et al., (2020) proposed a file sharing scheme through the usage of IPFS proxy, a secure file sharing system is suggested, which includes distributed access control and group key management [21]. The control technique is implemented using the IPFS agent, which plays an essential part in the design. A secure file sharing system is created by combining an IPFS server with a blockchain network, as well as the use of an IPFS proxy. Members of the system can form new groups or join existing ones based on their preferences. The secure file sharing system manages the access control approach, despite the fact that the IPFS server and blockchain network lack an access control mechanism. Only the groups that members have authorised have access to files.

Kumar et al., (2021) recommended to create a decentralized peer-to-peer image and video sharing platform based on blockchain technology based on IPFS (Interplanetary File System) [22]. To detect copyright infringement in multimedia, we use perceptual hashing technology (pHash). When media is loaded into IPFS, the pHash of the same content is calculated and compared to the existing pHash of the blockchain network. Due to the similarity to the existing pHash value, the media will be marked as modified. Non-involvement of third parties is an advantage of blockchain technology, which eliminates single points of failure.

Khatal et al., (2021) lays forth a framework for distributing digital content in a secure environment that is protected from illegal access and allows users to read, alter, and exchange data with one another [23]. Our application is built on top of IPFS and Ethereum smart contracts. To manage access to digital content and record lineage data, blockchain technology is used. The suggested FileShare application assures that digital content is only available within the app and not on the end user's operating system.

Kumar and Tirpathi (2019) proposed a blockchain storage architecture based on IPFS to solve the problem of storing transactions in blocks and granting transactions access to specific blocks [24]. The miner stores the transaction in IPFS DFS storage, and receives the returned IPFS transaction hash into the blockchain block under the recommended storage paradigm. The IPFS network function and the hashes it generates reduce the size of transactions in the block. By pressing the content address, a solution (IPFS hashing) has been proposed to protect access to transactions stored in designated blocks. This mode is applied to transactions including IPFS image storage and blockchain hash storage.

III. Roll -Hashing Algorithm

This Algorithm comes from the study while proposing spam-sum algorithm [25].

Consider file in whole form consists of bytes B1, B2,….., Bn . The rolling hash is indicated as Rhx for position x in the content.

where, Rhx is determined by the content from Bx-a to Bx and 'a' denotes area where the rolling hash algorithm processed. A simple roll hash algorithm is shown in the figure 2.

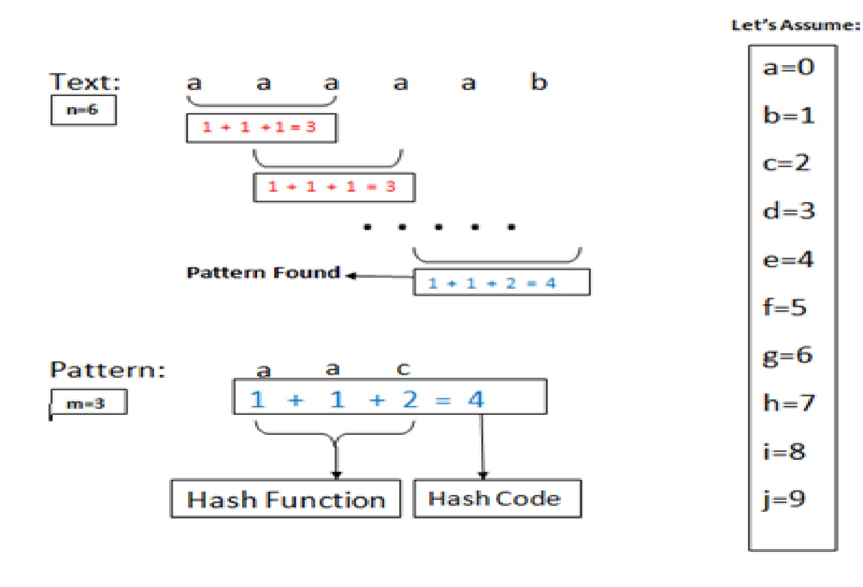


Figure 2. Roll hash algorithm

consider,

a=6

Bi=0 if i<0

Then it defines with function F′as written as

$$Rhx = F'\, Bx, Bx - 1, Bx - 2, \ldots., Bx - n) \text{ ------------ } (1)$$

IV. Algorithm process

Consider the proposed method process in text data.

(A) Definition

The process of Rolling Hash Algorithm is being repeated till the last byte in text data are being processed.

Then, set of Rh is denoted as U' where,

Rh denotes the number of events with the same Rh.

Additionally, the DRh is the compatible with Rh in slope order of Rhe.

In other words,

if DRhi > DRhi+1, condition is being satisfied for any I of same Rhe

where,

DRh_o has a maximum Rhe.

(B) **Text pattern in byte sequence**:

The Lord of the Rings text contains 10 chapters and about 966 KB in total, to the same text with chapter 6 removed which is 135 KB. The graph shows the majority matches of Rhe to Rh from the corresponded text by J.R.R Tolkien [26].

The graph shows the frequency of similar texts in byte pattern sequence, only if Rhe nearer to Rh is associated with hash. If that case frequency of Rh being used within text is higher than, it is considered a hash and not a word.

(C) Proposed Hashing sequence:

Here , Let us consider the proposed hash function 'H' and end result are computed as follows

H'(u) = [DRh1, DRh2,……..,DRh24, SHA 256($\sum_{i=1}^{|U'|}$ ui € $U')mod60\ for\ each$

= [Hb1, Hb2,…..Hb34]

and the output as calculated as

output = $\sum_{i=0}^{34}$ f(Hbi)

where,

U 'denotes the text that is removed

i.e., DRh1, DRh2, …….,DRh24 from U

|U '| indicates number of texts.

H 'outputs a value.

here Hb1, Hb2,……,Hb34 are replaced by DRh1, DRh2,….,DRh24 with defined SHA 256 hash function

f ' denotes function changes the end result being calculated by H.

According to Base 60 symbol chart [30] and the sum operation is concatenating to the strings. Data accessibility model is reviewed in [31].

V. Evaluation on performance

(A) Evaluating Similarity:

To determine the degree of similarity between texts, we employed the Edit distance technique. Let's look at t1 and t2. The minimum number of insertions, deletions, and substitutions required to convert t1 and t2. The highest similarity between two texts is shown by smaller values. The following equation, which spans from 0 to 200, calculates the distance between two sentences using the edit distance technique.

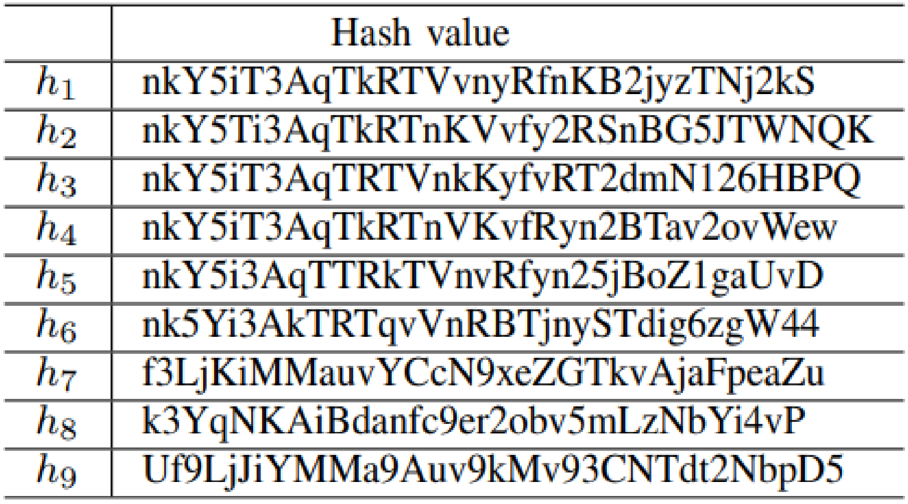
$$\text{similarity} = (1 - \frac{ED(t1,t2)}{34})100 \text{ ------(4)}$$

(B) Evaluating Objective:

Evaluating target involves 11 texts in total. five texts in a sentence as s1, s2,….,s5 have the original texts from “The Lord of the Rings” also same text with chapter 6 to 10 removed. The remaining sentence s6, s7,….,s10 respectively are “wuthering Heights” by Emily Bronte,”Middlemarch” by George Eliot. All the above texts are encoded in UTF-8 where it retrieves from “Oxford Royale Academy”.

(C) Results and discussions

Table 1 shows the end result of applying sentences s1,s2,….,s10 in equation (2) and (3).



	Hash value
h_1	nkY5iT3AqTkRTVvnyRfnKB2jyzTNj2kS
h_2	nkY5Ti3AqTkRTnKVvfy2RSnBG5JTWNQK
h_3	nkY5iT3AqTRTVnkKyfvRT2dmN126HBPQ
h_4	nkY5iT3AqTkRTnVKvfRyn2BTav2ovWew
h_5	nkY5i3AqTTRkTVnvRfyn25jBoZ1gaUvD
h_6	nk5Yi3AkTRTqvVnRBTjnySTdig6zgW44
h_7	f3LjKiMMauvYCcN9xeZGTkvAjaFpeaZu
h_8	k3YqNKAiBdanfc9er2obv5mLzNbYi4vP
h_9	Uf9LjJiYMMa9Auv9kMv93CNTdt2NbpD5

Table 1. *Hash Output*

Table 2 explains the similarity of texts in sentences.

	h1	h2	h3	h4	h5	h6	h7	h8	h9
h1	100.99	47.900	42.850	51.450	43.765	29.235	4.500	6.525	6.525
h2	47.900	100.99	35.720	47.900	40.640	29.235	4.500	6.525	0.000
h3	42.850	35.720	100.99	51.450	35.720	29.235	4.500	9.375	6.525
h4	51.450	47.900	51.450	100.99	42.850	30.250	4.500	4.500	6.525
h5	43.765	42.850	35.720	42.850	100.99	29.250	4.500	12.200	4.500
h6	29.235	29.235	29.250	30.250	29.235	100.99	0.000	6.525	6.525

h7	4.500	4.500	4.500	4.500	4.500	0.000	100.99	12.200	25.250
h8	6.525	6.525	9.500	4.500	12.200	6.525	12.200	100.99	6.525
h9	6.525	0.000	6.525	6.525	4.500	6.525	25.250	6.525	100.99

Table2. *Similarity between the text data*

Assume that in table II, s1,s2,s3...s5 have a relatively high similarity to “”. Also, because their texts are unrelated, there is minimal resemblance in S7......S10. The presence of multibyte sequences of text in UTF-8 may explain the high resemblance across different texts. Using the hash function, the suggested method computes reasonably high similarity for similar text in a sentence.

VI. Conclusion

As a result, offering a hash algorithm approach that recognises the similarity of content in texts/ phrases overcomes the limitations of standard file systems in IPFS networks, resulting in relatively high similarity between the texts in content. As a result of the impact of multi-byte sequences in texts, the calculated similarity isn't up to par. As a result, in the future, the proposed method will be improved by taking into account the multisystem sequence in the IPFS network.

References

1. Announcing the secure hash standard,” https://csrc.nist.gov/csrc/media/publications/fips/180/2/archive/2002-08-01/documents/ fips180-2withchangenotice.pdf, 2002.
2. D. N. Krawetz, http://www.hackerfactor.com/blog/index.php?/archives/ 432-Looks-Like-It.html.
3. ssdeep project ssdeep - fuzzy hashing program,” https://ssdeep-project. github.io/ssdeep/index.html, 2006.
4. J.Batiz-Benet,“go-multihash,”https://github.com/multiformats/go-multihash/blob/master/multihash.go.
5. J. Kornblum, “Identifying almost identical files using context triggered piecewise hashing,” 2006. [Online]. Available: https://dfrws.org/sites/default/files/session-files/paper-identifying almost identical files using context triggered piecewise hashing.pdf
6. Kelly, M., Alam, S., Nelson, M. L., & Weigle, M. C. (2016, September). Interplanetary wayback: Peer-to-peer permanence of web archives. In International Conference on Theory and Practice of Digital Libraries (pp. 411-416). Springer, Cham.
7. Alam, S., Kelly, M., & Nelson, M. L. (2016, June). Interplanetary wayback: The permanent web archive. In Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries (pp. 273-274).
8. Jia, B., Xu, C., Gotla, R., Peeters, S., Abouelnasr, R., & Mach, M. (2016). Opus-Decentralized music distribution using InterPlanetary File Systems (IPFS) on the Ethereum blockchain V0. 8.3. Opus Foundation, 2017.
9. Dias, D., & Benet, J. (2016, June). Distributed web applications with IPFS, tutorial. In International Conference on Web Engineering (pp. 616-619). Springer, Cham.
10. Sicilia, M. A., Sánchez-Alonso, S., & García-Barriocanal, E. (2016, November). Sharing linked open data over peer-to-peer distributed file systems: the case of IPFS. In Research Conference on Metadata and Semantics Research (pp. 3-14). Springer, Cham.
11. Chen, Y., Li, H., Li, K., & Zhang, J. (2017, December). An improved P2P file system scheme based on IPFS and Blockchain. In 2017 IEEE International Conference on Big Data (Big Data) (pp. 2652-2657). IEEE.
12. Cochez, M., Hüser, D., & Decker, S. (2017, June). The future of the semantic web: Prototypes on a global distributed filesystem. In 2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS) (pp. 1997-2006). IEEE.
13. Ali, M. S., Dolui, K., & Antonelli, F. (2017, October). IoT data privacy via blockchains and IPFS. In Proceedings of the seventh international conference on the internet of things (pp. 1-7).
14. Confais, B., Lebre, A., & Parrein, B. (2017, June). An object store for Fog infrastructures based on IPFS and a Scale-Out NAS. In RESCOM 2017 (p. 2).
15. Ohashi, S., Watanabe, H., Ishida, T., Fujimura, S., Nakadaira, A., & Kishigami, J. (2019, July). Token-Based Sharing Control for IPFS. In *2019 IEEE International Conference on Blockchain (Blockchain)* (pp. 361-367). IEEE.

16. Vimal, S., & Srivatsa, S. K. (2019). A new cluster P2P file sharing system based on IPFS and blockchain technology. *Journal of Ambient Intelligence and Humanized Computing*, 1-7.
17. Benet, J. (2014). Ipfs-content addressed, versioned, p2p file system. *arXiv preprint arXiv:1407.3561*.
18. Sicilia, M. A., Sánchez-Alonso, S., & García-Barriocanal, E. (2016, November). Sharing linked open data over peer-to-peer distributed file systems: the case of IPFS. In *Research Conference on Metadata and Semantics Research* (pp. 3-14). Springer, Cham.
19. Naz, M., Al-zahrani, F. A., Khalid, R., Javaid, N., Qamar, A. M., Afzal, M. K., & Shafiq, M. (2019). A secure data sharing platform using blockchain and interplanetary file system. *Sustainability*, *11*(24), 7054.
20. Steichen, M., Fiz, B., Norvill, R., Shbair, W., & State, R. (2018, July). Blockchain-based, decentralized access control for IPFS. In *2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)* (pp. 1499-1506). IEEE.
21. Huang, H. S., Chang, T. S., & Wu, J. Y. (2020, July). A secure file sharing system based on IPFS and blockchain. In *Proceedings of the 2020 2nd International Electronics Communication Conference* (pp. 96-100).
22. Kumar, R., Tripathi, R., Marchang, N., Srivastava, G., Gadekallu, T. R., & Xiong, N. N. (2021). A secured distributed detection system based on IPFS and blockchain for industrial image and video data security. *Journal of Parallel and Distributed Computing*, *152*, 128-143.
23. Khatal, S., Rane, J., Patel, D., Patel, P., & Busnel, Y. (2021). FileShare: A Blockchain and IPFS Framework for Secure File Sharing and Data Provenance. In *Advances in Machine Learning and Computational Intelligence* (pp. 825-833). Springer, Singapore.
24. Kumar, R., & Tripathi, R. (2019, November). Implementation of distributed file storage and access framework using IPFS and blockchain. In *2019 Fifth International Conference on Image Information Processing (ICIIP)* (pp. 246-251). IEEE.
25. A. Tridgell, https://www.samba.org/ftp/unpacked/junkcode/spamsum/ README, 2002.
26. glaslos (Lukas Rist), https://github.com/glaslos/ssdeep.
27. S. Natsume, “Wagahai wa neko dearu,” http://www.cl.ecei.tohoku.ac.jp/ nlp100/data/neko.txt, 1905.
28. “Base58check encoding,” https://en.bitcoin.it/wiki/Base58Check encoding.
29. V. Levenshtein, “Binary codes capable of correcting deletions, insertions, and reversals,” pp. 707––710, 1965.
30. https://www.aozora.gr.jp/.
31. Dr.A.Rajaram, S Vaithiya lingam, “Distributed Adaptive Clustering Algorithm for Improving Data Accessibility in MANET,” IJCSI International Journal of Computer science, 8(4): 369-373, July 2011.
32. Premanand, R.P., Rajaram, A. Enhanced data accuracy based PATH discovery using backing route selectionalgorithm in MANET. Peer-to-Peer Netw. Appl. 13, 2089–2098 (2020). https://doi.org/10.1007/s12083-019-00824-1
33. Rajaram.A., Dr.S.Palaniswami . Malicious Node Detection System for Mobile Ad hoc Networks. (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 1 (2) , 2010, 77-85
34. Dr.S.Palaniswami, Ayyasamy Rajaram. An Enhanced Distributed Certificate Authority Scheme for Authentication in Mobile Ad hoc Networks. The International Arab Journal of Information Technology (IAJIT).vol.9 (3),291-298.