International Journal of Pure and Applied Mathematics
Volume 119 No. 12 2018, 13409-13421
ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

ijpam.eu

Implementation of Deduplication on Encrypted Big-data using Signcryption for cloud storage applications

Saravanan Palani*, Sangeetha E, Archana A

School of Computing, SASTRA Deemed University, India

*Corresponding Author

ABSTRACT

As Big Data Cloud storage servers are getting widespread the shortage of disc space within the cloud becomes a major concern. The elimination of duplicate or redundant data, particularly in computer data is named deduplication. Data deduplication is a method to regulate the explosive growth of information within the cloud storage, most of the storage providers are finding more secure and efficient methods for their sensitive method.

Recently, a noteworthy technique referred to as signcryption has been proposed, in which both the properties of signature (ownership) and encryption are simultaneously implemented with better performance

According to deduplication, we introduce a method that can eliminate redundant encrypted data owned by different users. Furthermore, we generate a tag which will be the key component of big data management. We propose a technique called digital signature for ownership verification. Convergent encryption also called for a content hash key cryptosystem. Convergent encryption is an encryption approach that supports deduplication. With this encryption technique, the encryption key is generated out of a hash of plain text. Therefore applying this technique, identical plaintexts would turn out the same ciphertext.

Keywords: Big data, digital signature, cloud, data deduplication, convergent encryption

1. INTRODUCTION

Cloud Storage is an online storage service which provides services such as data maintenance, data management, and data backup. The user is permitted to store their file online and can access the stored files from anywhere. A recent survey states that some business organizations gained an advantage due to cloud adoption is nearly double the previously shown records, and it is estimated that by 2017, the cloud services business may exceed $244 billion.

Some of the most commonly known cloud services are Dropbox, Google drive, etc. Stored files may be accessed from anyplace via net association. Each cloud service allows a specific bandwidth. If a corporation exceeds the given bandwidth, extra charges would be applied. Some suppliers permit unlimited information storage. It is one among the issue that corporations ought to take into account when looking at a cloud storage provider [15-24].

Privacy is a major concerned factor one has to look for. Cloud encryption is the conversion of a cloud user's data into ciphertext [25-32]. The Cloud storage providers provide services like cloud encryption which is an encryption of the user's data before stored in the cloud [33-39]. Encryption is the scrambling of user's data into a form such that it is impossible to decrypt the data without the acknowledgment of the cryptography key.

Data security can be effectively achieved by encryption. Encryption and secure encryption key management allow only authorized users to access the uploaded data. The encrypted data is meaningless without its respective key.

2. RELATED WORK

Hybrid data deduplication in cloud environment [1].They propose a mechanism called hybrid data deduplication. The solution for the mechanism mentioned above is provided with partial semantics security. In this system, CSP knows the encryption key of the data. This system model is not applicable where the Cloud storage supplier cannot be trusted. The main advantage of this technique is that it supports deduplication on plaintext and ciphertext. The main disadvantage of this technique is that it does not support encrypted data deduplication. In this scheme, they use client-end de-duplication, and the encryption is done at the client. In this way, the computation pressure can be shared by all the clients.

Encrypted Data Deduplication in Cloud Storage [2]. The proposed model had cloud server eliminates duplicated ciphertexts which ultimately improves the privacy protection. In this model, the security of the ciphertext is improved in a better way compared with the existing encryption models. Usually, the encryption keys are hash values of plain text, but here the encryption keys are randomly chosen. Here the cloud server has no information except the hash value.

DupLESS: Server aided encryption for deduplicated Storage [3]. They have proposed an architecture system called DupLESS in such a way that the proposed architecture can resist the brute force attacks through secure deduplicated storage. The messages were encrypted via PRF protocol using the key which was obtained from the server.The duplication check was done by an existing service on client behalf.The main advantage of these techniques is that brute force attacks are avoided, and clients can encrypt their data with a key server which is different from the separate storage server. The main drawback of this technique is that flexibility to other data users cannot be provided

Policy-based de-duplication model was proposed based [4] on trust relationship enabled on cloud storage components, de-duplication related components, and different security requirements. The proposed system model has a key management mechanism based on proxy re-encryption algorithm for decryption of deduplicated data and data access. Even if the malicious party tries to access the data store in the cloud, it can't gain knowledge about Map box and the lockbox decryption

private key. In case the chunk key is known, it may result in information leak.

Efficient Deduplication On encrypted Big Data in Cloud [5]. They have presented an efficient scheme, where the data was encrypted using proxy re-encryption technique and possession undertaking, then stored in the cloud.The outcomes display scheme effectiveness and good performance in deduplication of cloud storage data.

Deduplication based on Hadoop [6].In this paper, an integrated deduplication approach is proposed by taking the features of Hadoop into account and leveraging parallelism based on Map Reduce and HBase to speed up the deduplication procedure. In our proposed approach, a new small-file aggregation scheme is proposed, and the new standard of Secure Hash Algorithm-3, Keccak is employed. The overall deduplication procedure is implemented based on Map Reduce and HBase to provide scalability to cater to the challenges brought by the big data Era.

When there are too many files to process, the tar-to-seq tool's performance may sharply downgrade due to performance bottleneck arising from memory or I/O in a single node.

Attribute-Based Storage Supporting Secure Deduplication of Encrypted Data in Cloud [7]. An attribute-based storage scheme was proposed in a hybrid cloud setting for secure deduplication. The public cloud is responsible for storage. The private cloud manages the duplication check. Confidential data sharing was done efficiently by specifying access policies. Semantic security's standard notion for data confidentiality was achieved. Key management is another major challenge in cloud storage. In [8], security matrix and magic cards are used to identify right binary pattern from the input stream.

A Hybrid Cloud Approach for Secure Authorized Deduplication [9]. The authorized data deduplication issue was first addressed in this paper. Other than the data, user's different privileges were considered for duplication check. The proposed model was a hybrid cloud architecture which supported authorized duplication check [10]. The proposed scheme had minimal overhead, reduced storage space and saved network bandwidth. Key management overhead issue was not addressed in the paper.

A Secure Client Side Deduplication Scheme in Cloud Storage Environments [11]. OpenStack Swift was proposed for secure storage and data sharing via the public cloud. It is client-side deduplication. Unauthorized user confidentiality was better ensured. That is the user itself computed the key required for the respective data encryption [12]. The data owner managed the data access. An authorized user can decrypt a file in its encrypted form with the private key only after integrating access rights. The proposed model was successful only for hash key encryption. Deduplication was also used in cloud storage auditing [13] and improving the access mechanism with higher efficiency [14].

The proposed model in the paper is done by using convergent encryption and digital signature verification for deduplication on

big encrypted data. Confidentiality of data was better ensured. The proposed model provided good performance in achieving deduplication on encrypted big data.

3. PROPOSED APPROACH

Popular service of the cloud is storage of data. Respecting the privacy of the user, the data will be encrypted first and then stored in the cloud. New challenges arise for data deduplication in the cloud due to the encrypted data storage which leads to difficulties in big data storage.

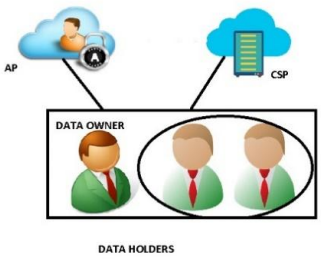

Fig 1. Proposed model

In this paper, the scheme is proposed by applying CE based on digital signature verification. In the proposed scheme, it is not necessary for the data holders to be in online for the duplication check.

The scheme has three entities. They are

1) Data holder who would upload the data which will be saved in the cloud. In this scheme, it's possible that the data holders will try to upload the same data which is already stored in the cloud storage in its encrypted form by a data holder.

2) Cloud Storage Provider is used to store the encrypted data. In this scheme, CSP is only used for storage as it cannot be trusted fully. The trust issue is due to the curiosity of the cloud to know about the stored data.

3) The authorized party is used for data ownership verification and duplication checking. The data holders completely trust AP.

3.1. PRELIMINARY AND NOTATION

3.1.1PRELIMINARIES

1)Convergent Encryption

Here the convergent encryption technique is represented as (KG; E; D)Where KG is the key generation, E is the encryption, D is the decryption. On the input of the hash key K and the message M, the encryption algorithm E generates a ciphertext C where C=E (M, K). On the input of hash key K and the ciphertext C, the decryption algorithm generates a plain text M where M=D(C, K)

2).Symmetric Encryption

The encryption algorithm takes two parameters namely the plaintext M and the secret key K. Using these parameters, it encrypts the input M with the secret key S and generates the ciphertext CT (I.e.) E (M, K). This process is conducted at the respective data owner U; then the encrypted data is uploaded to the cloud storage provider CSP

The decryption algorithm takes two parameters namely the ciphertext CT and the secret key K. Using these parameters, it decrypts the ciphertext CT with the secret

key S and generates the plaintext M (I.e.) D (CT, K). This process is conducted to obtain the encrypted data stored in the cloud in the form of plain text

3.1.2 NOTATION

Table 1 represents the list of parameters used in the proposed model

Key	Description
M	The plaintext
CT	The ciphertext
H()	The hash function
KG	The Key generation algorithm
K	The hash key
T()	The tag generation function
T	The tag used for data duplication check
S	The secret key
(PK1, Pk2)	The public key and private key for ownership verification
E	The encryption
D	The decryption
SIG	The digital signature
HD	The hashed document
Encrypt(M,S)	The encryption function on M using secret key S
Decrypt(CT.S)	The decryption function on CT
CS(HD,PK2)	The signature creation function
VS(HD,SIG,PK1)	The signature verification function

4. SCHEME

The proposed model has the following aspects

1) Data upload- AP does the duplication check. If the check results in negative, then the data holder encrypts the data with the secret key S given by AP. The encrypted data is then stored at CSP

2) Data duplication- when a data holder tries to store a data which is already stored in the cloud storage, duplication occurs. AP does the duplication check using tag comparison. If the comparison results in positive, AP contacts the data holder and informs about the duplication occurred

3) Data download- If the data holder wants to decrypt a particular data stored at CSP, it will request AP to provide the key needed to decrypt the respective data. AP goes for data holder eligibility check. If the check results in positive, then AP will issue the secret key that will be used to decrypt the encrypted data by the respective data holder

4.1DATA OWNERSHIP VERIFICATION SCHEME

In this paper, the data ownership verification is done based on Rivest-Shamir-Adleman scheme. AP checks data holder's eligibility to make sure that it is a real party which will upload the data. Here H (M) cannot be disclosed by CSP.

4.2 PROCEDURE

4.2.1 Data deduplication

The procedure for data deduplication is given by the following steps:

Step 1) user U1 selects the data M it wants to upload

Step 2) Tag generation for the particular data is done by user U1. It will generate a tag for the particular data M, t1= H (E (H (M), M))

Step 3) AP will do the duplication check. It will check whether the tag t1 already exist

If the check results in negative, then the data holder will encrypt the data M with the secret key S given by AP. The encrypted data is then stored at CSP. If the check results in positive, then AP contacts the data holder and informs about the duplication occurred

Step 4) When User U2 request to upload the same data M encrypted by the user U1 following step 1 and 2. In step 3, duplication happens because t2 exists so that U2 will be informed about the situation.

4.2.2 Ownership verification

Step 1) the public key and the private key (PK1, PK2) is generated for each message which has to be uploaded by the data holder.

Step 2) Signature creation: The signature is created on the input of hashed message (HD) using private key PK2

Step 3) AP verifies the signature on the input of hashed message HD, signature SIG using public key PK1

Step 4) if the verification is positive, AP proceeds to check for duplication. If the verification is negative AP will conclude it as an attempt by a malicious party.

4.2.3 Data ownership verification

The data owner logs into its account. It will select the document that it wants to upload to the cloud. Then, the document is hashed. Public and private key pair is generated for the selected document.

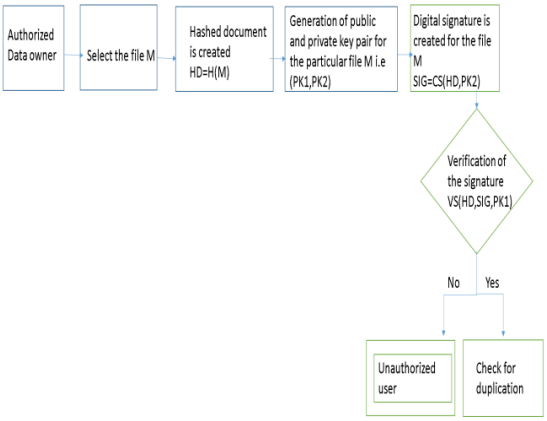

Fig 2. Data ownership verification

Now, the digital signature is created for the file using the hashed document and the private key. AP does the verification of the signature using its respective public key. If the verification is true, the AP proceeds to check for duplication

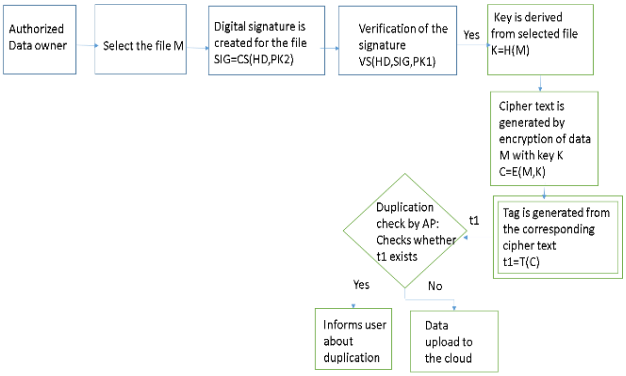

Fig 3. Data deduplication

4.2.4 Data deduplication:

The data owner logs into its account and selects the document that it wants to upload to the cloud. A digital signature is created for the particular document, and AP verifies it. If the verification is true, then AP checks

for duplication. The process is done by 1) deriving a hash key from the selected document. 2) The ciphertext of the particular document is done by encryption of the selected document and the hash key. 3) A tag is generated from the corresponding ciphertext created. 4) The created tag is compared with the other tags present in AP if the comparison is false, then it is uploaded to the cloud

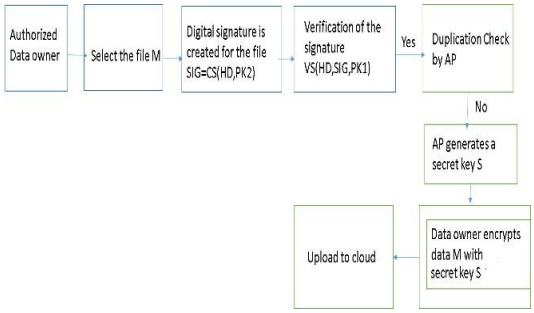

Fig 4. Data Encryption

4.2.5 Data Encryption:

The data owner logs into its account and selects the document that it wants to upload to the cloud. A digital signature is created for the particular document, and AP verifies it. If the verification is true, then AP checks for duplication. If the comparison is false, the AP generates a secret key. The data owner encrypts the selected document with the secret key and then uploads it to the cloud.

4.2.6 Data decryption

A data holder logs into its account, it will select the document that it wants to decrypt.

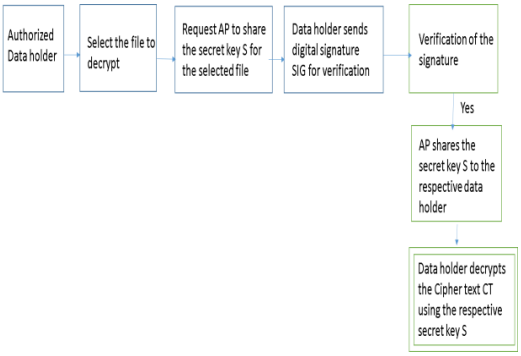

Fig 5. Data Decryption

It will request AP to share the secret key for the selected document by sending its digital signature for verification. AP verifies the digital signature. If the verification is true, then AP shares the secret key to the respective data holder. Now the data holder can decrypt the selected document using the respective secret key.

5. PERFORMANCE EVALUATION

The proposed model was implemented and tested for performance evaluation. The time taken for selection of data to be uploaded and downloaded were not taken into consideration. Testing was mainly focused on algorithms designed and deduplication procedure performance

5.1 Data Encryption and Decryption:

Data encryption and decryption were done by using AES. The operation time was estimated using AES key size 256 bits for different data size (range from 80MB to 150MB). It was observed that the time taken for the data which was as big as 150MB to encrypt and decrypt were 64 milliseconds and 60 milliseconds respectively.

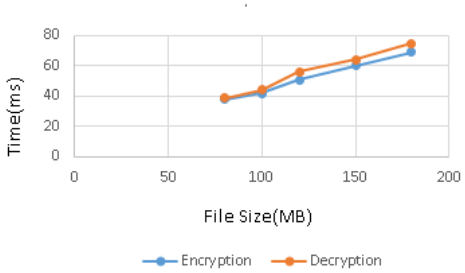

Fig 6. Data Encryption and decryption

As the size of the data increased, the time is taken for data encryption and decryption also increased. Data encryption and decryption were efficiently done by using AES.

5.2 Convergent Encryption:

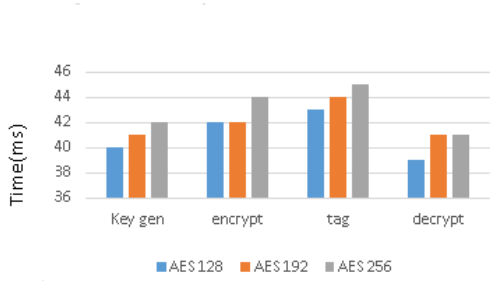

Fig 7. Convergent Encryption

Each operation of convergent encryption was tested under different AES key size 128bits, 196bits, and 256bits. For a data of 100MB, the time is taken for key generation, encryption, tag generation, decryption were observed. For the different AES key size, the encryption time was less than 44 milliseconds. The decryption time was less than 41 milliseconds. For each operation, the computation time did not have much difference. The proposed model was efficiently applicable in various scenarios.

5.3 Data ownership verification:

Fig 8. Data ownership verification

For each step in data ownership verification, the operation time was estimated using 2048 bit RSA. For a data of 100MB, the time is taken for key generation, signature creation and signature verification were observed. The time taken for a key generation was 11milli second. The time taken for creation of signature was 16 milliseconds. The time taken for verification of signature was 15 milliseconds. In the proposed model, the data holder didn't need to proceed for data encryption if that data was already uploaded by another data holder

References

[1]. C. Fan, S. Y. Huang, and. C. Hsu, Hybrid data deduplication in a cloud environment, in Proc. Int. Conf. Inf. Secure. Intell. Control, 2012, pp. 174177

[2]. Encrypted Data Deduplication in Cloud Storage, Information Security (AsiaJCIS), 2015 10th Asia Joint Conference on, Issue Date: 24-26 May 2015

[3]. M. Bellare, S. Keelveedhi, and T. Ristenpart, "DupLESS: Server aided

encryption for deduplicated Storage," in Proc. 22nd USENIX Conf. Secur., 2013, pp. 179–194.

[4]. Chuanyi Liu, Xiaojian Liu and Lei Wan "Policy-based deduplication in secure cloud storage," in Proc. Trustworthy Compute. Serv., 2013, pp. 250–262, doi: 10.1007/978-3-642-35795-4_32.

[5]. Y.Saritha, A. Vineela, "Efficient Deduplication On encrypted Big Data in Cloud", International Journal of emerging trends and technology in computer science, vol.6, no.4, 2017.

[6]. Zhang, D., Liao, C., Yan, W., Tao, R., & Zheng, W. (2017, August). Data Deduplication Based on Hadoop. In Advanced Cloud and Big Data (CBD), 2017 Fifth International Conference on (pp. 147-152). IEEE.

[7]. Hui Cui, Robert H. Deng, Yingjiu Li, and Guowei Wu "Attribute Based Storage supporting secure deduplication of encrypted data in cloud" IEEE Transactions on Big Data, January 2017

[8]. Sekar, K. R., Saravanan, P., & Sethuraman, J. (2006). Discovery of right binary pattern in key management using security matrix and magic cards. ARPN Journal of Engineering and Applied Sciences, 11(15), pp. 9187-9194

[9]. Jin Li, Yan Kit Li, Xiaofeng Chen, Patrick P.C. Lee, and Wenjing Lou "A Hybrid Cloud Approach for Secure Authorized Deduplication" IEEE Transactions On Parallel And Distributed Systems, Vol. 26, No. 5, May 2015

[10]. C. Fan, S. Y. Huang, and W. C. Hsu, "Hybrid data deduplication in cloud environment," in Proc. Int. Conf. Inf. Secur. Intell. Control, 2012, pp. 174–177, doi:10.1109/ISIC.2012.6449734.

[11]. Nesrine Kaaniche, Maryline Laurent "A Secure Client Side Deduplication Scheme in Cloud Storage Environments" IEEE Transactions on Mobility and Security (NTMS) in Cloud Computing, Issue Date:April.2.2014

[12]. C. Yang, J. Ren, and J. F. Ma, "Provable ownership of file in deduplication cloud storage," in Proc. IEEE Global Commun. Conf, 2013,pp.695700,doi:10.1109/ GLOCOM. 2013.6831153.

[13]. J. W. Yuan and S. C. Yu, "Secure and constant cost public cloud storage auditing with deduplication," in Proc. IEEE Int. Conf. Communic. Netw. Secur., 2013, pp. 145–153, doi:10.1109/CNS.2013. 6682702.

[14]. T. Y. Wu, J. S. Pan, and C. F. Lin, "Improving accessing efficiency of cloud storage using de-duplication and feedback schemes,"IEEE Syst. J., vol. 8, no. 1, pp. 208–218, Mar. 2014, doi:10.1109/ JSYST.2013.2256715.

[15] Indragandhi, V., Logesh, R., Subramaniyaswamy, V., Vijayakumar, V., Siarry, P., & Uden, L. (2018). Multi-objective optimization and

energy management in renewable based AC/DC microgrid. Computers & Electrical Engineering.

[16] Subramaniyaswamy, V., Manogaran, G., Logesh, R., Vijayakumar, V., Chilamkurti, N., Malathi, D., & Senthilselvan, N. (2018). An ontology-driven personalized food recommendation in IoT-based healthcare system. The Journal of Supercomputing, 1-33.

[17] Arunkumar, S., Subramaniyaswamy, V., & Logesh, R. (2018). Hybrid Transform based Adaptive Steganography Scheme using Support Vector Machine for Cloud Storage. Cluster Computing.

[18] Indragandhi, V., Subramaniyaswamy, V., & Logesh, R. (2017). Resources, configurations, and soft computing techniques for power management and control of PV/wind hybrid system. Renewable and Sustainable Energy Reviews, 69, 129-143.

[19] Ravi, L., & Vairavasundaram, S. (2016). A collaborative location based travel recommendation system through enhanced rating prediction for the group of users. Computational intelligence and neuroscience, 2016, Article ID: 1291358.

[20] Logesh, R., Subramaniyaswamy, V., Malathi, D., Senthilselvan, N., Sasikumar, A., & Saravanan, P. (2017). Dynamic particle swarm optimization for personalized recommender system based on electroencephalography feedback. Biomedical Research, 28(13), 5646-5650.

[21] Arunkumar, S., Subramaniyaswamy, V., Karthikeyan, B., Saravanan, P., & Logesh, R. (2018). Meta-data based secret image sharing application for different sized biomedical images. Biomedical Research,29.

[22] Vairavasundaram, S., Varadharajan, V., Vairavasundaram, I., & Ravi, L. (2015). Data mining-based tag recommendation system: an overview. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 5(3), 87-112.

[23] Logesh, R., Subramaniyaswamy, V., & Vijayakumar, V. (2018). A personalised travel recommender system utilising social network profile and accurate GPS data. Electronic Government, an International Journal, 14(1), 90-113.

[24] Vijayakumar, V., Subramaniyaswamy, V., Logesh, R., & Sivapathi, A. (2018). Effective Knowledge Based Recommeder System for Tailored Multiple Point of Interest Recommendation. International Journal of Web Portals.

[25] Subramaniyaswamy, V., Logesh, R., & Indragandhi, V. (2018). Intelligent sports commentary recommendation system for individual cricket players. International Journal of Advanced Intelligence Paradigms, 10(1-2), 103-117.

[26] Indragandhi, V., Subramaniyaswamy, V., & Logesh, R. (2017). Topological review and analysis of DC-DC boost converters. Journal of Engineering Science and Technology, 12 (6), 1541–1567.

[27] Saravanan, P., Arunkumar, S., Subramaniyaswamy, V., & Logesh, R. (2017). Enhanced web caching using bloom filter for local area networks. International Journal of Mechanical Engineering and Technology, 8(8), 211-217.

[28] Arunkumar, S., Subramaniyaswamy, V., Devika, R., & Logesh, R. (2017). Generating visually meaningful encrypted image using image splitting technique. International Journal of Mechanical Engineering and Technology, 8(8), 361–368.

[29] Subramaniyaswamy, V., Logesh, R., Chandrashekhar, M., Challa, A., & Vijayakumar, V. (2017). A personalised movie recommendation system based on collaborative filtering. International Journal of High Performance Computing and Networking, 10(1-2), 54-63.

[30] Senthilselvan, N., Udaya Sree, N., Medini, T., Subhakari Mounika, G., Subramaniyaswamy, V., Sivaramakrishnan, N., & Logesh, R. (2017). Keyword-aware recommender system based on user demographic attributes. International Journal of Mechanical Engineering and Technology, 8(8), 1466-1476.

[31] Subramaniyaswamy, V., Logesh, R., Vijayakumar, V., & Indragandhi, V. (2015). Automated Message Filtering System in Online Social Network. Procedia Computer Science, 50, 466-475.

[32] Subramaniyaswamy, V., Vijayakumar, V., Logesh, R., & Indragandhi, V. (2015). Unstructured data analysis on big data using map reduce. Procedia Computer Science, 50, 456-465.

[33] Subramaniyaswamy, V., Vijayakumar, V., Logesh, R., & Indragandhi, V. (2015). Intelligent travel recommendation system by mining attributes from community contributed photos. Procedia Computer Science, 50, 447-455.

[34] Vairavasundaram, S., & Logesh, R. (2017). Applying Semantic Relations for Automatic Topic Ontology Construction. Developments and Trends in Intelligent Technologies and Smart Systems, 48.

[35] Logesh, R., Subramaniyaswamy, V., Vijayakumar, V., Gao, X. Z., & Indragandhi, V. (2017). A hybrid quantum-induced swarm intelligence clustering for the urban trip recommendation in smart city. Future Generation Computer Systems, 83, 653-673.

[36] Subramaniyaswamy, V., & Logesh, R. (2017). Adaptive KNN based Recommender System through Mining of User Preferences. Wireless Personal Communications, 97(2), 2229-2247.

[37] Logesh, R., & Subramaniyaswamy, V. (2017). A Reliable Point of Interest Recommendation based on Trust Relevancy between Users. Wireless Personal Communications, 97(2), 2751-2780.

[38] Logesh, R., & Subramaniyaswamy, V. (2017). Learning Recency and Inferring Associations in Location Based Social Network for Emotion Induced Point-of-Interest Recommendation. Journal of Information Science & Engineering, 33(6), 1629–1647.

[39] Subramaniyaswamy, V., Logesh, R., Abejith, M., Umasankar, S., & Umamakeswari, A. (2017). Sentiment Analysis of Tweets for Estimating Criticality and Security of Events. Journal of Organizational and End User Computing (JOEUC), 29(4), 51-71.