2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC) | 978-1-6654-2463-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/COMPSAC51774.2021.00202

Lightweight Privacy-Preserving Similar Documents Retrieval over Encrypted Data

1st Zaid Ameen Abduljabbar [1,2]
[1] *College of Education for Pure Science*
University of Basrah
Basrah, Iraq
zaid.ameen@uobasrah.edu.iq
[2] *Technical Computer Engineering Department*
Al-Kunooze University College
Basrah, Iraq

2nd Ayad Ibrahim [1]
[1] *Computer Science Department*
College of Education for Pure Science
University of Basrah
Basrah, Iraq
mraiadibraheem@gmail.com

3rd Mustafa A. Al Sibahee[1,2]
[1] *College of Big Data*
and Internet
Shenzhen Technology University
Shenzhen, China
mustafa@sztu.edu.cn
[2] *Iraq University College*
Basrah, Iraq

4th Songfeng Lu [1,2,*]
[1] *Hubei Engineering Research Center on Big Data Security*
School of Cyber Science and Engineering
Huazhong University of Science and Technology
Wuhan, China
[2] *Shenzhen Institute of Huazhong*
University of Science and Technology
Shenzhen, China
[*]Correspondence: lusongfeng@hust.edu.cn

5th Samir M. Umran [1]
[1] *Hubei Engineering Research Center on Big Data*
Security, School of Cyber Science and Engineering
Huazhong University of Science and Technology
Wuhan, China
samirmuh@yahoo.com

***Abstract*—Document Similarity Detection (DSD) is significant in our real life applications. However, the existing methods ignore the privacy of what is contained in the documents uploaded on remote servers, thus reducing the applicability of these methods. The proposed scheme allows documents to be compared without revealing to those remote servers. For each document, the fingerprint set is calculated. The inverted index is constructed on the basis of the whole fingerprint set. The inverted index is widely used for efficient retrieval. This index is under protection by Paillier cryptosystem before it gets uploaded to the server.**

***Index Terms*—privacy-preserving DSD, secure inverted index, fingerprint set, Paillier cryptosystem**

I. Introduction

Considering the vast amount of content that can be accessed through the World Wide Web [1], it is easy to fit another ideas, manuscripts, or theories as your own without citing the original source. It has become a critical problem both academically and non-academically [2]. Thus, this issue needs a radical solution.

The DSD is a common practice in various applications. DSD is applicable to establish whether the newly submitted article to a journal carries plagiarized contents. However, there is no attention paid to the privacy of documents needing to be matched since the document database is assumed as public for many the current solutions of DSD. Such limitation reduces the utilization of these methods in our real life applications.

In various scenarios, it is necessary to find the similar documentations to a given query document while ensuring privacy. For example, a large proportion of journals avoid submitting the same article twice. Therefore, in order to ascertain whether the same paper is submitted for multiple journals simultaneously without infringing their privacy, a secure DSD is required to protect privacy.

In fact, encryption is most effective in privacy protection for the documentation in storage. However, it is difficult to apply the traditional DSD methods due to encryption, which makes it a necessity to work out the solution to measuring how similar two documents are from the perspective of encryption.

Essentially, the methods currently used can be classified into two directions: hashing [3], [4] and vector space [5]. In hashing approach, the document contains a set of substrings with fixed length that can be extracted. To generate a compressed yet descriptive fingerprint from the hash values, the hash values for all substring are calculated. In case that two documents show more shared fingerprint terms than a predetermined threshold, they are deemed similar. This is more suited to identifying the local similarities.

According to our proposed scheme, the hashing policy is taken to generate the fingerprint set for the document database with fingerprint as a representative yet compressed set of numbers. We build the inverted index based on the fingerprint set of the document database [6]. Inverted index is widely applied for efficient and fast document retrieval [6]. We build a reliable inverted index and build a privacy-preserving DSD scheme based on this index to make use of the attractive features of the inverted index in secure data, where a secret key is employed to encrypt the index in such a way that

978-1-6654-2463-9/21/$31.00 ©2021 IEEE
DOI 10.1109/COMPSAC51774.2021.00202