

Low-Complexity and Secure Clustering-Based Similarity Detection for Private Files

Duaa Fadhel Najem¹, Nagham Abdulrasool Taha², Zaid Ameen Abduljabbar^{2,3,4},
Vincent Omollo Nyangaresi^{5,6}, Junchao Ma³, Dhafer G. Honi^{2,7}

¹ Department of Cyber Security, College of Computer Science and Information Technology,
University of Basrah, Basrah 61004, Iraq

² Department of Computer Science, College of Education for Pure Sciences,
University of Basrah, Basrah, 61004, Iraq

³ College of Big Data and Internet, Shenzhen Technology University, Shenzhen, 518118, China

⁴ Shenzhen Institute, Huazhong University of Science and Technology, Shenzhen 518000, China

⁵ Department of Computer Science and Software Engineering, Jaramogi Oginga Odinga
University of Science & Technology, Bondo 40601, Kenya;

⁶ Department of Applied Electronics, Saveetha School of Engineering, SIMATS,
Chennai, Tamil Nadu 600124, India

⁷ Department of IT, University of Debrecen, Debrecen, 4002, Hungary

Abstract – Detection of the similarity between files is a requirement for many practical applications, such as copyright protection, file management, plagiarism detection, and detecting duplicate submissions of scientific articles to multiple journals or conferences. Existing methods have not taken into consideration file privacy, which prevents their use in many delicate situations, for example when comparing two intellectual agencies' files where files are meant to be secured, to find file similarities. Over the last few years, encryption protocols have been developed with the aim of detecting similar files without compromising privacy. However, existing protocols tend to leak important data, and do not have low complexity costs. This paper addresses the issue of computing the similarity between two file collections belonging to two entities who desire to keep their contents private.

We propose a clustering-based approach that achieves 90% accuracy while significantly reducing the execution time. The protocols presented in this study are much more efficient than other secure protocols, and the alternatives are slower in terms of similarity detection for large file sets. Our system achieves a high level of security by using a vector space model to convert the files into vectors and by applying Paillier encryption to encrypt the elements of the vector separately, to protect privacy. The study uses the application of the Porter algorithm to the vocabulary set. Using a secure cosine similarity approach, a score for similar files was identified and the index of the similarity scores is returned to the other party, rather than the similar files themselves. The system is strengthened by using clustering for files, based on the k-means clustering technique, which makes it more efficient for large file sets.

DOI: 10.18421/TEM133-61

<https://doi.org/10.18421/TEM133-61>


Corresponding author: Zaid Ameen Abduljabbar,
Department of Computer Science, College of Education for
Pure Sciences, University of Basrah, Basrah, 61004, Iraq
Email: zaid.ameen@uobasrah.edu.iq

Received: 23 January 2024.

Revised: 11 May 2024.

Accepted: 18 May 2024.

Published: 27 August 2024.

 © 2024 Duaa Fadhel Najem et al;
published by UIKTEN. This work is licensed under the
Creative Commons Attribution-NonCommercial-NoDeriv
4.0 License.

The article is published with Open Access at
<https://www.temjournal.com/>

Keywords – File similarity, privacy, similarity
detection.

1. Introduction

File similarity detection techniques have begun to be used in many important applications since the first research in this field began in 1993 [1]. For example, this approach is used in a file management system, which can work more efficiently if similar files are identified. It is also used to improve the function of web crawlers in terms of detecting similar pages [2], [3], [4]. Finally, this method is used in applications related to plagiarism detection and copyright protection [5], [6].

The problem of security is considered very important in the process of data matching.