The Management Of Security For Text Document Protection Using Various Algorithms

Prof.K.S. Warke, Pratibha Singh, Mokshi Pandita, Maclina Biswas

Abstract:
In the recent digital world, information security has become an important issue. Due to the advanced technology different ways are used to copy, redistribute and store the digital contents easily. The authentication of the original owner and copyright protection of data is a challenging task. Digital watermarking provides a solution for digital contents copyright protection and ownership verification. A secret message is placed inside a digital content without compromising valuable data. This secret information is used later for ownership identification, Secure Hash Algorithms, also known as SHA, are a family of cryptographic functions designed to keep data secured. It works by transforming the data using a hash function: an algorithm that consists of bitwise operations, modular additions, and compression functions, in our project we are going use the SHA for file downloading from Cloud by using Key. We can consider key as a security purpose, if user key gets match than only download functionality works otherwise use cannot download the available file from cloud.

Keyword: Information Security, Digital Watermarking, Copyright Protection.

Introduction:

In the recent digital world, information security has become an important issue. Due to the advanced technology different ways are used to copy, redistribute and store the digital contents easily. There are unlimited benefits of digital media and advanced technology such as cloud computing and Internet of Things. On the other hand, they have created a problem for original data owners against illegal usage. Internet of Things and cloud has received significant support from governments and research institutes around the world . Data is shifted on cloud computing in the form of audio, video, image and text. The authentication. of the original owner and copyright protection of data is a challenging task. Data is the crucial element in smart cities which sustains the infrastructure of data and helps people to gain access to digital contents. The architecture of the smart city is presented in Fig. 1, where data is store, process and analyze the central location. The privacy and ownership of the original data are significant. Digital watermarking provides a solution for digital contents copyright protection and ownership verification. A secret message is placed inside a digital content without compromising valuable data. This secret information is used later for ownership identification. Digital text watermarking is an active area of research: the individuals, government officials, and military facing data security problems which also affect the smart cities. Digital publishers have rights but facing many threats, such as illegal use of copyrights,

data manipulation, and redistribution of information. Text documents are part of almost every organization or company such as audit firms, banks, or any large private or public corporation. These documents are in the form of financial statements, legal notes, birth certificates,

soft degrees, classified reports and declarations.

Related work

Methodology :-Text digital watermarking is a critical area of research.

1) LINGUISTIC-BASED APPROACH

The linguistic-based approach consists of semantic and syntactic techniques which emphasize the semantic that is used for embedding the watermark and does not change the meaning of the text. Using a synonym substitution technique semantic approach is developed, in which specific words are exchanged with their synonyms for data hiding. In this technique, grammatical alternations are used for watermark embedding without affecting the original text meaning of the text. The verb, adverb, noun, pronoun, adjective, preposition, acronyms, and conjunction are language parts which are used for the watermarking. Digital Watermarking is often used to discourage illegal copying, and it is also used to stop the distribution of digital assets. the watermarking contains two phrases, embedding, and detection of the watermark. The peace of secret information which is embedded into the original document is called watermark. The watermark embedding process includes three steps, first, watermark generation that includes the information about the owner, e.g., author name and other information like a publisher. Second, watermark securing, where the watermark is transformed into a binary string or groups. The last one is inserting a watermark, where the watermark is inserted without affecting the whole document. where "SM" denotes the secret message, "T" represents the original document, "WD" is a watermarked document and "K" denotes Key. The watermarked document is shared via communication channels such as e-mail, website, and social media. The reverse process of watermark embedding is called The reverse process of watermark embedding is called extraction or verifying.

Algorithm Used :

NLP :- In that project NLP algorithm use for extracting noun from PDF document which is uploaded by the sender. For extracting noun in PDF document we use the Maxine Tagger API **.** A Part-Of-Speech Tagger (POS Tagger) is a piece of software that reads text in some language and assigns parts of speech to each word (and other token), such as noun, verb, adjective.

AES :- This algorithm is use for encryption of extracted noun .

- The Advanced Encryption Standard **(AES)** is a fast and secure form of encryption that keeps prying eyes away from our data.
- AES is based on ‘substitution–permutation network’. It comprises of a series of linked operations, some of which involve replacing inputs by specific outputs (substitutions) and others involve shuffling bits around (permutations).

- Interestingly, AES performs all its computations on bytes rather than bits. Hence, AES treats the 128 bits of a plaintext block as 16 bytes. These 16 bytes are arranged in four columns and four rows for processing as a matrix

SHA: Hash functions are extremely useful and appear in almost all information security applications.

A hash function is a mathematical function that converts a numerical input value into another compressed numerical value. The input to the hash function is of arbitrary length but output is always of fixed length.

Values returned by a hash function are called message digest or simply hash values. The following picture illustrated hash function –

Motivation:

The motivation of the project is in the digital world quite easy to produce an illegal copy of the text document. The verification of digital content is one of the major issues because digital content is generated daily and shared via the internet. There is limited techniques are available for document copyright protection.

System Architecture:

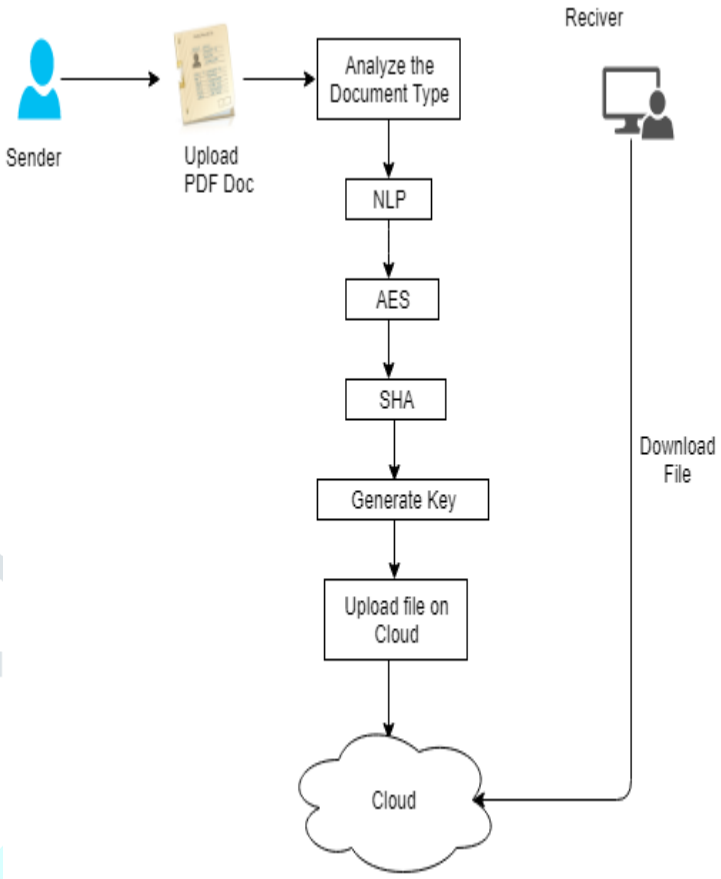

Fig 1. System Overview

In above diagram, user first upload the PDF file and using NLP algorithm we will extract the five nous from file, after that we will apply AES on file, after that we will apply SHA algorithm and generated file we can upload into cloud, if user wants to download the file from cloud they need KEY to download the file.

Conclusion:

We conclude how to more secure text documents by applying various algorithms for text documents copyright protection. To provide more security to the document by using the AES algorithm , NLP algorithm and SHA algorithm,.

Reference:

[1] Seyoung Huh,Sangrae Cho, Soohyung Kim

, "Managing IoT Devices using Blockchain Platform ," 2017.

[2] Reem A. Alotaibi Lamiaa A. Elrefaei "Improved capacity Arabic text watermarking 4 methods based on open word space " 2014.

[3] Stefano Giovanni Rizzo, Flavio Bertini, Danilo Montesi "Content-preserving Text Watermarking through Unicode Homoglyph Substitution " 2016.

[4] Jack T. Brassil ,Steven Low, Nicholas F. Maxemchuk ,Lawrence O'Gorman "Electronic Marking and Identification Techniques to Discourage Document Copying " 1995.

[5] Ooi Wei Liang, Vahab Iranmanesh "Information Hiding using Whitespace Technique in Microsoft Word. " 2016.

[6] Y. M. Alginahi, M. N. Kabir, and O. Tayan, "An enhanced Kashidabased watermarking approach for Arabic text-documents," in Electronics, Computer and Computation (ICECCO), 2013 International Conference on, 2013, pp. 301-304: IEEE.

[7] Y. Meng, T. Guo, Z. Guo, and L. Gao, "Chinese text zero-watermark based on sentence's entropy," in Multimedia Technology (ICMT), 2010 International Conference on, 2010, pp. 1-4: IEEE.

[8] Z. Jalil and A. M. Mirza, "Text watermarking using combined image-plustext watermark," in Education Technology and Computer Science (ETCS),2010 Second International Workshop on, 2010, vol. 1, pp. 11-14: IEEE.

[9] R. J. Jaiswal and N. N. Patil, "Implementation of a new technique for web document protection using unicode," in Information Communication and Embedded Systems (ICICES), 2013 International Conference on, 2013, pp.69-72:IEEE.