Retrieving E-learning document using encrypted semantic words

Farah Amer
Tikrit University, College of Computers Science and mathematics, Computer Science Department, Iraq
farah_amer@gmail.com

Qasim Mohammed Hussein
Department of Control system engineering, College of petroleum process engineering ,Tikrit university،Iraq
kasimalshamry@tu.edu.iq

Mohammed Sahib Mahdi Altaei
Al-Nahrain University, College of Science, Computer Science Department, Iraq
mohammed.sahibmahdi@nahrainuniv.edu.iq

Abstract

The process of retrieving document images in digital media depending on their content is difficult, especially when the number of images is very large. Therefore, there is a need to use an automated computer system to retrieve the documents based on semantic words. This paper presents an automation computer system that uses to retrieve documents images based on semantic words. Semantic words are chosen depending on the content of the document. To keep the security of the document semantic words and prevent unauthorized persons to access these documents, they are encrypted along with the document issue number using hash function and store them in index file. The importance of this work lies in the possibility of using it in the institutions archive departments to retrieve digital documents based on their content instead of their number. The implementation of this system result shows that this system can retrieve the stored documents with high accuracy and short time in easy manner.

Keywords: document retrieval, Encryption process, SHA function, semantic word, Achieve.

استرجاع مستند التعلم الإلكتروني باستخدام كلمات دلالية مشفرة

فرح عامر قاسم محمد حسين محمد صاحب مهدي الطائي
جامعة تكريت ، كلية علوم الحاسب والرياضيات ، قسم علوم الحاسب ، العراق
جامعة النهرين ، كلية العلوم ، قسم علوم الحاسوب ، العراق

الخلاصة

تعتبر عملية استرجاع صور المستند في الوسائط الرقمية حسب محتواها أمرًا صعبًا ، خاصة عندما يكون عدد الصور كبيرًا جدًا. لذلك ، لذا تبرز الحاجة إلى استخدام نظام حاسوبي لاسترجاع المستندات بناءً على الكلمات الدلالية اليا. تقدم هذه الورقة نظام حاسوبي آلي يستخدم لاسترجاع صور المستندات بناءً على الكلمات الدلالية التي يتضمنها المستند. يتم اختيار الكلمات الدلالية اعتمادًا على محتوى المستند. وللحفاظ على أمنية الكلمات الدلالية للمستند ومنع الأشخاص غير المصرح لهم من الوصول إلى هذه المستندات ، يتم تشفيرها مع رقم إصدار المستند باستخدام دالة هاشية وتخزينها في ملف مفهرس. تكمن أهمية هذا العمل في إمكانية استخدامه في أقسام أرشيف المؤسسات لاسترجاع الوثائق الرقمية بناءً على محتواها بدلاً من رقم اصدارها. بينت نتائج تنفيذ النظام المقترح امكانية استرجاع المستندات المخزونة بدقة عالية ووقت قصير بطريقة سهلة.

الكلمات المفتاحية: استرجاع المستندات ، عملية التشفير ، دالة SHA ،الكلمات الدلالية ، الارشيف.

1. Introduction

Institutions have huge documents that are stored as paper or digital format. The main requirements for storing documents are retrieval them easily while maintaining the confidentiality of their information [1]. The process of searching or retrieving these documents is difficult, especially when the storage is done through using arbitrary name or number. To retrieve a stored document in the archives departments of institutions, the employee searches for it sequentially and check depending on the document's issue number, which it waste a large amount of time for the retrieving process, since when the number of documents is large. This method of retrieval makes the effort of intruder in searching for a document close to the employee effort for searching and accessing this document. Therefore, it is necessary to use computer applications to speed up the retrieval process, in addition to, ensuring the security of documents by providing access only to authorized persons [2].

The paper aims to design and implement a computer system that facilitates a secure retrieval of the stored digital documents automatically by the authorized person. It based on creating encrypted ID record for each document image. The ID record includes document issue number and semantic words that existing within the stored document.

Related work

There is no published method that uses the same approach, which is proposed for archival images retrieval. Different approaches are used to retrieve images depend on the image contents. Some approaches depend on processing the image to retrieve the image. In general, these approaches can be classify in two main catagories (i) Text based image retrieval , which utilized for giving keywords, remarks, or descriptions to the images in the database repository, (2) Content based image retrieval which based extracted significant image features like color, texture, and shape features for using in matching of image .[3,4]

In [5] present a solution by transferring the semantics directly from a large word image databases in English andhindi language. They try to exploit the linguistic resources such as WordNetThe accuracy of the proposed framework was 90%.

[6] Used a set of computable image-text metrics and translated into intuitive, distinct semantic image-text classes. They derived image-text categories using metrics cross-modal mutual information, semantic correlation, and the status relation metrics. The accuracy was between 83% and 90% for the aforementioned metrics.

In our proposed method, instead of processing the entire image, we create ID record to represent the document image. the processing treatments deal with the ID image only.

2. SHA-256 Function

Cryptography plays an important role in data security by converting the data to unreadable text[6]. SHA-256 is one of six secure hash cryptographic algorithms from SHA-2 family designed by the National Security Agency (NSA), in 2001, and published by the National Institute of Standards and Technology (NIST). SHA-256 can used in many applications in cyber security areas such as information encryption, message authentication and digital signature. It uses relatively simple nonlinear logic operations and considers a strong hash functions available. SHA-256 generates unique 256-bit (32-byte) data signature for any size of original text. [9].

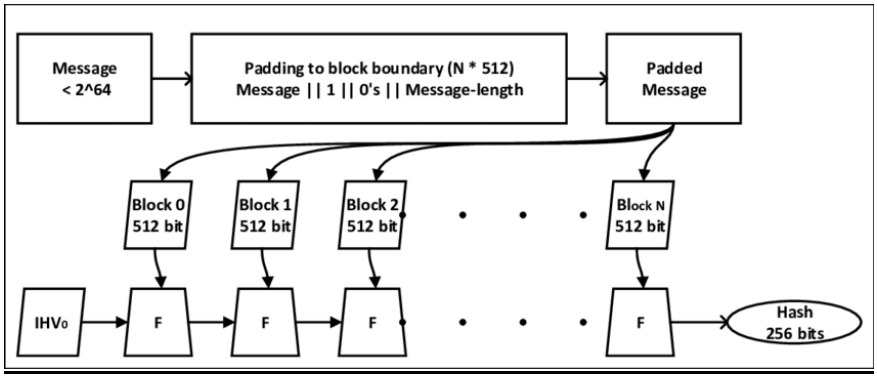

Figure (1) Hash function process

3. Proposed retrieving image document

The long-distance exchange of information on a wide network needs encryption to secure the information from unauthorized access. The reliability of the network plays an important role in protecting data on an insecure network. The researchers also established many cryptographic strategies for protecting and efficiently transmitting knowledge on an insecure network. Image encryption techniques are commonly used by all cryptographic techniques to transmit photos on an insecure network. The purpose of this paper is to demonstrate the few encryption techniques that are used on an insecure network to encrypt the image. To encrypt the image, this paper proposes a new encryption technique.

The proposed method depends on create a record for each document image that contain semantic words of this document and document issue number. The record is encrypted using one SHA-256 hash function. The encrypted record are store in index file that contain two fields: the serial number and the encrypted record.

4.1. Creation and encryption of image document ID

At first, ID record for each document image is created. It contains 256 bits as maximum. The details of each ID record and the size of each field are illustrated in table (1) .

Table (1) ID document contains

Details	Semantic words No.	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	Document issue
Size (in bit)	4	36	36	36	36	36	36	36

characters as a maximum. Each character will represent in 6 bits. The character may be Arabic letters, numbers, comma and "/ ", which is used in writing the date and other characters, as the user like. Table 2 shows the characters coding.

The number of selected semantic words in that entered in ID record must not exceed six words, which it depends on the content of the document and the employee's desire to choose the number that expresses the document. Each word may contain 6

Table (2) Characters coding

Character	Code	Character	code	Character	Code	Character	Code
أ	000000	ع	010001	5	000010		
ب	000001	غ	010010	6	100011		
ت	000010	ف	010011	7	100100		
ث	000011	ق	010100	8	100101		
ج	000100	ك	010101	9	100110		
ح	000101	ل	010110	,	100111		

خ	000110	م	010111	.	101000		
د	0001110	ن	011000	"	101001		
ذ	001000	ه	011001	/	101010		
ر	001001	ك	011010	-	101011		
ز	001010	و	011011	:	101100		
س	001011	ي	011100		101101		
ش	001100	0	011101		101110		
ص	001101	1	011110		101111		
ض	001110	2	011111		110000		
ط	001111	3	100000				
ظ	010000	4	100001				

Algorithm 1 presents the creation process of ID record and encrypted the created ID for each document

.

Algorithm (1): Creation of ID image of document

Input	Empty record
Output	Encrypted ID Document Image
Process	Do for each document: 1. Enter serial number of the document. 2. Determining the number of document semantic words 3. Enter the selected semantic words of the document 4. Enter Issue number of the document 5. Coding the characters, each character has (6) bits starting from $(000000)_2$ to $(111111)_2$. 6. Create a record that represents the ID of the document image. 7. Encrypt the ID record using SHA -256 function 8. Save the serial number , encrypted image ID in the index file

The semantic words can be selected manually by the employees or automatically by the system depending on the most frequently used words in the document. The index file can access by authorized persons only.

4.2 Retrieval the document image

Algorithm (1.3) Retrieval Algorithm

Input	Document image , index
Output	ID of document
Process	Do for each document retrieving 1. Enter the number semantic words that use to searchfor the document 2. Enters the semantic keywords or issue number 3. Search sequential in index file until the reaches the desire image. - Read ID document - Decrypt the ID record - Recover the semantic words - compare the recover and input semantic words - IF there is matching, get the serial number and retrieve the document image.

	- Else read next ID 4. Need other document; if yes goto step 1 5. End.

It is not necessary to enter all the semantic words of the retrieval image when searching for a document, the user can search by any number of semantic words.

5. Discussion the implementation

The proposed method is implemented using Python programming language, which is executed under windows 10 operating system of 64 bit type. After the image, the hash library was called to encrypt the semantic word have been entered. In each ID, The length of each word is (6) letters as a maximum, and use six bits for the document issue number. The semantic words and the document number are converting to the binary representation, and then the ID is encrypted using SHA-256. The encrypted IDs are store sequentially in index file. The documents are stores depending on their serial number. Figures (2 - 4) illustrate an example of the implementation the proposed method for one of the document.

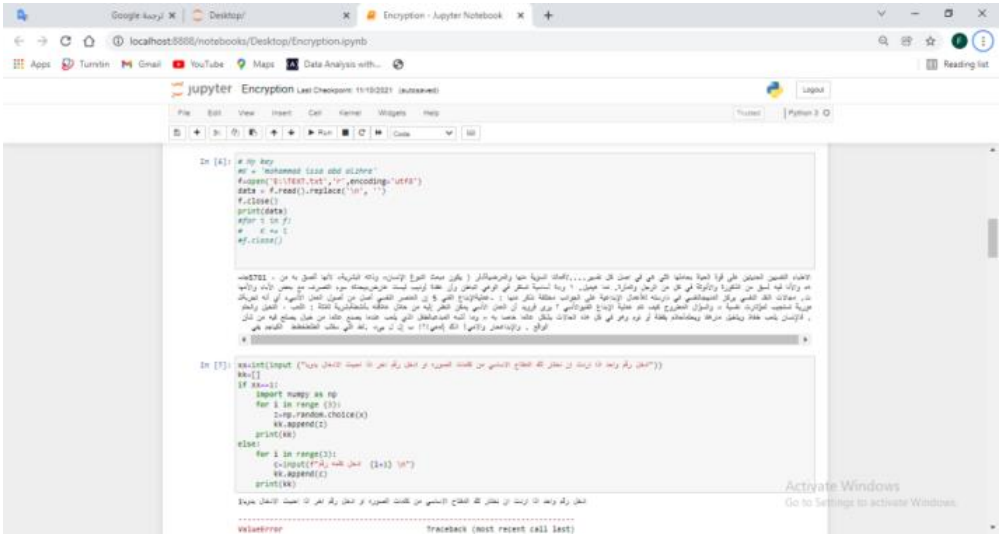

Figure (2) The document image

```
K1 = "".join(K1) # list into string
kk=[]
for i in range (3):
    c=input(f"ادخل كلمه رقم  {i+1} \n")
    kk.append(c)
kk
```

```
ادخل كلمه رقم  1
فرح
ادخل كلمه رقم  2
فرحج
ادخل كلمه رقم  3
فرحجح
['فرح', 'فرحج', 'فرحجح']
```

```
K1=str(kk)
```

Figure (3) Input the semantic words

```
import hashlib
SK1 = hashlib.sha256(K1.encode())

print("The hexadecimal equivalent of SHA256 is : ")
print(SK1.hexdigest())
```

```
The hexadecimal equivalent of SHA256 is :
6e8c2e3fca440e74fabb6eaeb1868c90d4a2e0b8a260442a77a6bd80f26bd3bd
```

Figure (4) the encrypted ID

The importance of this paper is the possibility to use the proposed method to facilitate and speed up the searching or retrieving processes of the achieve document. This ability to retrieve images will reduce the time and efforts even there are the huge number of documents stored.

A number of experiments were carried out to find out the efficiency of the proposed method. Tens ID document images records are created encrypted and stored them in index file, some of them in table 3. Then, attempt to retrieve these documents were made. The results were successful at rate 100%.

Table (3) Creation some of IDs document images

Image No.	Words No.	Selected Semantic words	Encrypted record
1	4	يصنع ، يغني ، العمل ، النظامي	2fd4a3ed8e86f10cb77b7a1b64977cc18b05b726ae64c066decaa0d4e0dadd1
2	4	المنهج ، للطفل .النفسي، يركز	372afc3b6c579721d3dff802e097ae817ca842dfeef14231e46d2409990adce3
3	5	ادارة ، البينات ، الضخمة ،المعلومات ، تعريفها	da1340a204bf2408df12f86a09d5bc76975dcadf18877a6735e1864eac6bd96b
4	4	بناء ، المناهج ، الادبية ، اللغوية	3563e406db5d13de05c050d54a80fb558e4f2c9cd88f266980ff5524d1fe87e8
5	3	شكر .تقدير .دورة	0eff4047039accb22e83e048f45faaf497f7dbeeb9b7b263ee57e80115c15f20
6	5	امر ، اداري ، نقل ، عباس ، جعفر	c7997fe8d2e446c4b6261c9db255d82718e105f04b5ba1734fcc99d25e6a9805
7	5	خصائص، المنهج، البحث ، لغة ، ادب	46fe2033a12e990027a88f47c232111305bd60b51c2850b5971722e31d955207
8	3	الاطباء، العلماء ، النفسيين	8f32b5205be54023cf437a9fa2ec51a8bf849fa67fcd996accc9b7136fed2337
9	4	امر ، اداري ، تقييم ، اداء ، موظفين	f0dbe263d80fd0ce2d127772748980d68c0f1117b1f0098015990bf508fd7534

6. Conclusion & Recommendations

This paper presents a secure and fast way to retrieve documents images based on semantic words within the document. The proposed method recovers images without going into the details of the image. The experiments proved that this method has high accurate, the rate is 100%, and the time depend on the number of the ID that will be check. In addition to, using this method can retrieve all document that include the selected semantic words. The institutions archive departments can make use of to retrieve digital documents based on their content

instead of their number with less efforts and time

7. References

[1] Farah Amer, Qasim Mohamed Hussein and Mohammed Sahib Mahdi, Design Engineering Arabic Word Recognition Using SURF Descriptor, Design Engineering , Issue 9, 2021, pp 8380 – 8393.

[2] William Stallings, Cryptography and Network Security Principles and practice, seventh edition, Global edition. Pearson Education Limited 2017 .

[3] Narjis Mezaal Shati , Noor khalid Ibrahim and Taha Mohammed Hasan , A Review of Image Retrieval Based on Ontology model, Journal of Al-Qadisiyah for Computer Science and Mathematics Vol.12 (1) 2020 , pp 10–14.

[4] Gong, Y.; Cosma, G.; Fang, H. On the Limitations of Visual-Semantic Embedding Networks for Image-to-Text Information Retrieval. J. Imaging 2021, 7, 125

[5] Praveen Krishnan and C. V. Jawahar, Bringing Semantics in Word Image Retrieval, 2013 12th International Conference on Document Analysis and Recognition,IEEE, pp733-737.

[6] Christian Otto, Matthias Springstein, Avishek Anand and· Ralph Ewerth, Characterization and classification of semantic image-text relations, International Journal of Multimedia Information Retrieval (2020) 9. pp:31–45.

[7] Nada Qasim Mohammed,Qasim Mohammed Hussein, Ahmed M. Sana, Layth A. Khalil, A Hybrid Approach to Design Key Generator of Cryptosystem, Journal of Computational and Theoretical Nanoscience, Volume 16, Number 3, March 2019, pp. 971-977(7).

[8] Mellila Bouam, Charles Bouillaguet, Claire Delaplace, Camille Noûs. Computational Records with Aging Hardware: Controlling Half the Output of SHA-256. 2021. ffhal-02306904v2

[9] Mellila Bouam, Charles Bouillaguet, Claire Delaplace, Camille Noûs. Computational Records with Aging Hardware: Controlling Half the Output of SHA-256. 2021.