Application of Fuzzy Search Algorithms and Neural Networks in Fingerprint Document Analysis

A. A. Yuzhakov, A. N. Kokoulin, A. I. Tur
AT Department
Perm National Research Polytechnic University
Perm, Russia
a.n.kokoulin@ieee.org

***Abstract*—Fingerprint is a technology that helps one to find information similar to the reference data by some criteria (features). It is used for tracking the fact of illegal copying of multimedia or electronic documents. The basic algorithm of a «fingerprint» is the statistical evaluation of different features in document: the presence of characteristic elements, and frequently encountered character combinations. The main disadvantage of the technology is the simplicity of "masking" of textual information in comparison with sound and video. Digital fingerprints are very sensitive even to the trivial replacement of symbols in text. Authors propose the fuzzy search algorithm application in digital fingerprinting technology to fix some of the Fingerprint shortcomings. It is demonstrated that the use of fuzzy search algorithms based on Levenshtein distance method and Soundex in DLP systems is efficient. Also we can implement the considered approaches with the help of a neural net-work as it is suggested in paper.**

Keywords—fingerprint; Levenshtein distance; soundex; neural network

I. INTRODUCTION

Fingerprint is a technology that helps one to find information similar to the reference data by some criterion (features). It is used for tracking the fact of illegal copying of music, video and electronic documents, as well as identifying the computer for specific settings and program modules. The basic algorithms is to create a "fingerprint" of the original information by statistical evaluation of several features – a color gradient and contrast indicators for video and images, the frequency of vector change and the top of the sound diagrams, text formatting, the presence of characteristic elements, and frequently encountered character combinations in electronic documents.

The main disadvantage of the technology is the simplicity of "masking" textual information in comparison with sound and video. Digital fingerprints are very sensitive even to the trivial replacement of one of the used symbols with others.

The authors proposed the fuzzy search algorithms application in digital fingerprinting technology to fix some of the Fingerprint shortcomings. It is demonstrated that the use of fuzzy search algorithms based on Levenshtein distance method and Soundex in DLP systems is efficient. It preserve the main advantages of Fingerprint technology and reduce the work on the specialist supervising the system. Also we can implement the considered approaches with the help of a neural network as it is suggested in paper.

II. THE APPLICATION OF FUZZY SEARCH IN DIGITAL FINGERPRINTS

A. DLP security systems

Data loss prevention software detects potential data breaches/data ex-filtration transmissions and prevents them by monitoring, detecting and blocking sensitive data while in-use (endpoint actions), in-motion (network traffic), and at-rest (data storage). In data leakage incidents, sensitive data is disclosed to unauthorized parties by either malicious intent or an inadvertent mistake. Sensitive data includes private or company information, intellectual property (IP), financial or patient information, credit-card data and other information.

The terms "data loss" and "data leak" are related and are often used interchangeably. Data loss incidents turn into data leak incidents in cases where media containing sensitive information is lost and subsequently acquired by an unauthorized party. However, a data leak is possible without losing the data on the originating side. Other terms associated with data leakage prevention are information leak detection and prevention (ILDP), information leak prevention (ILP), content monitoring and filtering (CMF), information protection and control (IPC) and extrusion prevention system (EPS), as opposed to intrusion prevention system.

The basic principle of DLP systems is sniffing every document or network packet in order to detect the critical information leakage. Text analyses is usually based on fingerprint technology.

B. Fingerprints for "masked" documents

Fingerprints are generally considered as high-performance hash functions used to uniquely identify blocks of data, but this is not applicable to analyzing "disguised" texts. The ideal cryptographic hash function has five main properties:

- it is deterministic so the same message always results in the same hash;
- it is quick to compute the hash value for any given message;

978-1-5386-1810-3/17/$31.00 ©2017 IEEE

- it is infeasible to generate a message from its hash value except by trying all possible messages;
- a small change to a message should change the hash value so extensively that the new hash value appears uncorrelated with the old hash value;
- it is infeasible to find two different messages with the same hash value

To overcome these difficulties, it makes sense to use fuzzy search algorithms. There are several algorithms of this kind: the Levenshtein distance, the Dahmerau-Lowenstein distance, the Bitap algorithm, the sampling algorithm, the N-gram method; Hashing by signature, BK-trees, etc. [1, 2, 3, 4]. They are characterized by a metric – a function of the distance between two words, allowing to assess the degree of their similarity in this context. [5] In the Levenshtein method, "distance" is the minimum number of edits of one line (three possible operations are meant for editing: erasing a character, replacing a symbol, and inserting a symbol), which makes it possible to turn it into a second one. Obviously, for the case when two matching words are compared, the distance will be zero. Consider the results of calculations for variants of "masking" by the example of the word "test" (Fig. 1).

By setting the maximum allowed distance between words equal to 2 (relevant for the case with "misspellings" in two letters), we will be able to detect most of the protected information typed by the user with misspellings. [6]

Not so long ago in social networks, users used this warped spelling of words, which is the pass of almost all the vowels. It allows you to read a word to a person, but when analyzing it, the system can have a number of problems. If we consider an example with a longer word "testing" (in the case of substituting it with the word "tank"), then the Levenshtein distance method gives a sufficiently large distance - 4 corrections. This approach will produce too many false positives.

This problem can be solved using phonetic algorithms. They compare two words with a similar pronunciation to the same codes, which makes it possible to compare and index many of these words on the basis of their phonetic similarity. There are quite a number of such algorithms - Soundex, Daitch-Mokotoff Soundex, NYSIIS, Metaphone, Double Metaphone, Caverphone.

Soundex is a phonetic algorithm for indexing names by sound, as pronounced in English. The goal is for homophones to be encoded to the same representation so that they can be matched despite minor differences in spelling. The algorithm mainly encodes consonants; a vowel will not be encoded unless it is the first letter. Soundex is the most widely known of all phonetic algorithms (in part because it is a standard feature of popular database software) and is often used as a synonym for "phonetic algorithm". Improvements to Soundex are the basis for many modern phonetic algorithms.

The first problem is the Russian language and alphabet, since this algorithm is designed for the Latin alphabet. As a solution, you can first translate Russian words into transliteration, but we need to perform some analysis of symbols statistics. For example instead of JA transliteration for "Я" it is preferably to use YA, since J for the algorithm is considered an essential letter.

		T	E	C	T
	0	1	2	3	4
T	1	0	1	2	3
Q	2	1	1	2	3
C	3	2	2	1	2
T	4	3	3	2	1

а

		T	E	C	T
	0	1	2	3	4
T	1	0	1	2	3
E	2	1	0	2	3
E	3	2	1	1	2
C	4	3	2	1	2
T	5	4	3	2	1

б

Fig. 1. The results of the calculation (a - for the case of replacing one of the letters with a symbol not used in the text, b - for the case of doubling one of the letters of the text)

The next problem is the accident letters of European languages (for example, Hôtel La Défense Kléber). We have to replace these characters with analogs of the Latin alphabet.

Let's consider in detail the algorithm Daitch-Mokotoff Soundex[7] which was developed in 1985 by two genealogies – Harry Mokotoff and Randy Day, aiming to achieve the best results when working with Eastern European (including Russian) languages. This algorithm has little to do with the original Soundex, except that the result is still a sequence of numbers, but now the first letter is also encoded. It has much more complex conversion rules – now in the formation of the resulting code, not only single characters, but also sequences of several characters are involved. In addition, the result of type 023689 provides about 600,000 different variations of code, which, combined with complicated rules, reduces the number of "superfluous", i.e. "False positive" words in the resulting set.

Calculation of the proposed problematic example based on the rules of the transliteration version of the algorithm is shown in Fig. 2.

B, P	1
F, V	2
C, K, S	3
G, J	4
Q, X, Z	5
D, T	6
L	7
M, N	8
R	9

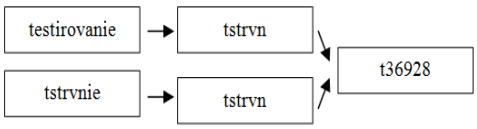

Fig. 2. Soundex phonetic algorithm testing

Thus, the "testing" and the "test" get a common code - "t36928". In the analysis, these two words will be perceived as one - the desired "testing". Of course, because of the method used, one code corresponds to several words. This can be critical for consonant surnames.

When applying Soundex, the division of the text into fragments can be represented as separate words (each new word starts with a letter and ends with a digit preceding the next letter) or more arbitrary constructions. [7]

The examples clearly show that in many cases, fuzzy search algorithms successfully cope with the recognition of "disguised" text. With the correct choice of the algorithm and proper tuning, you can get quite high results of searching for information leaks.

III. FINGERPRINTS BASED ON NEURAL NETWORKS

The task of processing the results of preliminary processing of the source text and compiling a digital print can be solved by means of a neural network. However, for this you have to solve the task of entering text into the neural network.

To enter text into the system, we can break the analyzed information into separate words and create a special bond structure – a convolution filter (this will preserve the importance of word order without turning the search into a sorting of "familiar" words), as shown in Fig. 3.

The idea of such a neural network is simple - one neuron is taken, which receives two (or more) words. For the next - the input is "shifted" by one word and the operation is repeated. Thus, we get an idea of the whole sentence.

To reduce the complexity of the neural network, it makes sense to pre-process the texts with fuzzy search algorithms (for example, Soundex). The essence of this approach is that the input of the neural network will be fed not by text, but by some result of the hash function. To do this, all texts from the training database of original documents and information "dummies" are broken into words and processed by the chosen algorithm. Otherwise, the training of the system will take into account the complexity and variability of the natural language, as well as the ability of an attacker to "mask" information.

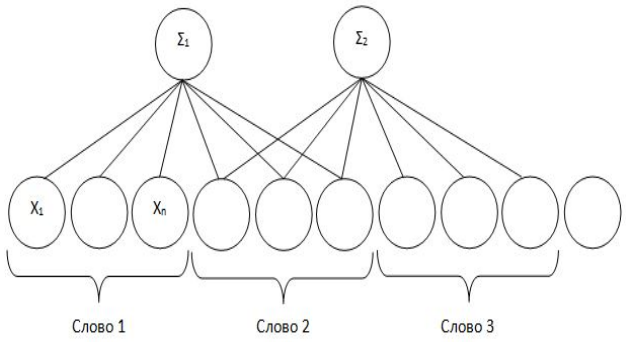

Fig. 3. Example of a neural network using Soundex

The disadvantages of these approaches to solving the problem of searching for protected information are all the main disadvantages of neural networks: the probabilistic result (the answer will be true only with a certain degree of probability), a long learning time, the need to have a sufficient sample of test texts.

IV. CONCLUSION

The authors proposed the use of fuzzy search algorithms in digital fingerprint technology. Algorithms were considered on the basis of the Levenshtein distance method and Soundex. It is demonstrated that the use of fuzzy search algorithms in DLP systems to search for protected information in verified electronic text documents is quite possible and may allow to refuse additional modules of text analysis (in comparison with the classical approach implemented in digital prints). In addition, it allows you to preserve the main advantages of Fingerprint technology and reduce the burden on the specialist supervising the system. However, when implementing such an approach, one has to think about optimization of the algorithm and the possibility of more efficient use of system resources. The possibility of implementing an analogue of the approaches considered by means of a neural network is proposed.

REFERENCES

[1] Alneyadi S., Sithirasenan E., Muthukkumarasamy V. A survey on data leakage prevention systems // Journal of Network and Computer Applications. 2016. № 62. P.137–152.

[2] Collins L. Chapter 59 – System Security // Computer and Information Security Handbook (Second Edition). 2013. P.993–1000.

[3] Smetanin N.. Fuzzy string search [Online] // Nikita's blog Search algorithms, software devel-opment and so on. 2011. URL: http://ntz-develop.blogspot.ru/2011/03/fuzzy-string-search.html

[4] Made Edwin Wira Putra, Iping Supriana Suwardi. Structural Off-line Handwriting Character Recognition Using Approximate Subgraph Matching and Levenshtein Distance // Procedia Computer Science – 2015. № 59. P. 340–349.

[5] Smetanin N. Phonetic algorithms [Online] // Nikita's blog Search algorithms, software devel-opment and so on. 2011. http://ntz-develop.blogspot.ru/2011/03/phonetic-algorithms.html.

[6] Beider A., Stephen P. Morse. Phonetic Matching: A Better Soundex // Association of Professional Gene-alogists Quarterly. 2010.

[7] Soundex System The Soundex Indexing System [Online] // National Archives. 2007. URL: .https://www.archives.gov/research/census/soundex.html

[8] High Performance Content Defined Chunking [Online] // The Pseudo Random Bit Bucket. URL: https://moinakg.wordpress.com/tag/rabin-fingerprint/.