International Journal of Electrical and Computer Engineering (IJECE)
Vol. 9, No. 5, October 2019, pp. 3798~3803
ISSN: 2088-8708, DOI: 10.11591/ijece.v9i5.pp3798-3803

Identification of individualization techniques for criminal records in sanction lists

Gonzalo M. Arias[1], Pablo A. Peláez[2], Fredy E. Hoyos[3]
[1,2]Tecnológico de Antioquia, Institución Universitaria, Facultad de Ingeniería, Colombia.
[3]Universidad Nacional de Colombia - Sede Medellín, Facultad de Ciencias, Escuela de Física, Colombia

Article Info

Article history:

ReceiveDec 31, 2018
Revised Apr 9, 2019
Accepted Apr 20, 2019

Keywords:

Criminal records second
False positives
Filters
Sanctions list
Verification methods

ABSTRACT

Using efficient searching techniques on sanctions lists and press articles allows a better filtering on individuals and entities to establish a commercial relationship with, including those who are going to have access to confidential information belonging to the company, in order to minimize the risk of leakage or information mismanagement. That process of filtering on individuals or entities could be automated by using individualization algorithms, searching techniques based on string comparisons, artificial intelligence, and facial recognition. Diverse methods were examined to be applied on each mentioned technique in order to identify which ones are ideal to its application on individualization due to their characteristics, in order to obtain agile and reliable results; taking into account that different methods are complementary and not exclusive, and that their combination allows to minimize human interaction in the classification of information, avoiding analysis of irrelevant data for that particular search.

Corresponding Author:

Fredy E. Hoyos,
Facultad de Ciencias - Escuela de Física,
Universidad Nacional de Colombia - Sede Medellín,
Carrera 65 No. 59A-110, Medellín, Colombia.
Email: fehoyosve@unal.edu.co

1. INTRODUCTION

Currently, relevant information for any subject of study is in a large number of media formats, which can be parameterized and indexed for specialized use in public or private databases [1]. They can be in natural language, captured in images, or in any medium required to facilitate its use and disclosure. Updating the information went from being in hands of a few companies and media, to be available to any person who has an electronic device with access to the Internet, causing sources of information to proliferate, both reliable and of dubious origin.

Taking into account this vast amount of information of all kinds, companies have found the need to classify and individualize [2] it, in order to comply with the regulations that govern them or to improve internal selection processes of personnel, associated companies, search for solutions practices, among other objectives. One of the main needs in companies at legal level is the search of criminal records of natural and legal persons with whom they have some employment relationship, in order to minimize the risk of being used in money laundering and terrorist financing operations (LAFT by its acronym in Spanish). With this premise, it is required to have the ability to validate information that is relevant to a particular individual in specialized databases, as well as sources of constant updating and less standardization such as press documents, articles, national and international bulletins, and other online sources.

Usually, systems used for this purpose apply text recognition algorithms, comparing them with their own databases, which have dictionaries of terms similar to those used in news concerning criminal activities [3-4]. The arising premise is the use of these character recognition algorithms, complementing them

with artificial intelligence techniques that allow us not only to identify these syntaxes, but also help assessing the most reliable sources over time and provide results according to what is required. At the same time, visual information of individuals is compared with facial recognition techniques, expanding or delimiting the range of results in cases where the information in written media is not correct.

2. CONCEPTUAL FRAMEWORK

2.1. Definitions

Text-search algorithms: Text-search algorithms are techniques used in order to find the occurrences of a pattern of characters in a given text corresponding to a combination of elements of a defined alphabet [5].

Algorithms of artificial intelligence for searching personalization: Personalized searching are the algorithms that use interests of users to produce fast and relevant search results. Among the input parameters for these algorithms are user profile, analysis of hyperlinks, analysis of pages content, and valuations of collaborative searches [6]. The objective of the mentioned algorithms is to give a weight of relevance to the results of the user's query. It [7] uses some classification and weighting algorithms based on the content of pages selected in user's query, while [8] it does the according work using automatic learning techniques - Referenced in several sources of consultation with the anglicism "machine learning" - which can be classified as artificial intelligence work aimed at the autonomous analysis of data flows.

Facial recognition: Facial recognition is the use of algorithms that take images or models of a face as input parameters in order to process them and generate, as a result, a corresponding identity matching to a database of individuals. As detailed by [9], depending on what is going to be analyzed, images or models, algorithms of different characteristics will take into account various biometric aspects for analysis can be used, each one having different benefits in terms of speed and effectiveness.

2.2. Analysis

Below, a comparative study between algorithms of the categories exposed in definitions section of this article is presented, taking into account indicators concerning the measurement of effectiveness of each one with respect to the particularities of each category they belong to:

a. Text-search algorithms

Brute force search: The aim of brute force search is to make a character-by-character comparison in the text T [s...s + m−1] for all ∈{0,...,n−m + 1} and the P[0...m−1] pattern. The algorithm returns all valid matches. However, as [10] points out, the problem with this approach is effectiveness, since the complexity of the algorithm is the worst possible, being of order O (M x N).

Knuth-Morris-Pratt Algorithm: KMP algorithm is composed of two phases: a text preprocessing in which a branch table based on partial failures of a brute force search is obtained. Using this table, the algorithm will scroll through the text advancing through it, not character by character as in the brute force search, but in the quantities described in the table. The complexity of the algorithm is given by the order O (n + k), where O (n) and O (k) are pre-process and subsequent search complexities.

Boyer-Moore Algorithm: As described by [10], the idea behind the Boyer- Moore algorithm is to perform a process analogous to KMP algorithm, but performing the search from right to left, which allows for larger jumps in the search in the main text, because if the last letter of the pattern to be searched is not found, the following n characters can be discarded, being n the length of the pattern. The complexity of this algorithm is sub-linear, that is, O (N / M).

b. Comparison between algorithms

Based on the order of algorithms analyzed, it is evident that the greatest effectiveness corresponds to Boyer-Moore, due to its complexity order as shown in Table 1. In tests conducted by [10] using an alphanumeric alphabet and several chains generated randomly, the following measurements of execution speed were observed in milliseconds, which corroborate the expected efficiency of each algorithm as shown in Table 2.

Table 1. Complexities of text search algorithms

Algorithm	Complexity
Brute force	O (M x N)
Knuth-Morris-Pratt	O (n + k)
Boyer-Moore	O (N / M)

Recovered from [10], algorithms for String matching

Table 2. Test results of text search algorithms

Pattern length	Matches	Brute force	KMP	BM
3	40	225	221	242
10	0	225	221	82
50	0	224	221	25

Recovered from [10], algorithms for String matching, 1-8.

c. Algorithms of artificial intelligence for personalizing searches

LVQ (Learning vector quantization) algorithms. Learning algorithms by vector quantization (LVQ) are widely used in classification of information tasks. According to [11] the strength of this neuronal model is the ability to form characteristic maps in a similar way to what happens in the brain, this algorithm uses reinforced competitive learning, distinguishing a training stage and an exploitation stage.

Naive Bayesian (NB). Based on Bayes conditional probability theorem (1763), it treats different prediction variables independently, while assuming independence between predictor attributes. The algorithm calculates conditional probability for combinations of attributes with the objective. It establishes an independent probability from predictive data. This probability provides the likelihood of each objective, once the instance of each value category is given from each input variable.

Decision trees (C4.5). Using the inductive learning methodology, decision tree algorithm classifies from a set of training data. In each execution of the algorithm, an evaluation of each node is made and it is determined which is the best as a decision parameter. K-Nearest Neighbors (KNN). Known as lazy learning [12]. The parameters of classification by neighborhood are based on the search in a set of prototypes, of k prototypes closest to the prototype to be classified. A metric is specified in order to measure proximity, Euclidean distance is normally used for computational reasons.

Support vector machines (SVM). As [13] says, a Support Vector Machine (SVM) learns the surface of two different classes of entry points. As a one-class classifier, the description given by support vectors data is capable of forming a decision boundary around the learning data domain, with very little or no knowledge of data outside this boundary. Data are mapped by means of a Gaussian kernel, or another type of kernel, to a feature space in a higher dimensional space, where the maximum separation between classes is sought. This border function, when brought back into input space, can separate data by different classes, each forming a grouping.

Comparison between algorithms. For comparing the different algorithms, we take the results obtained in previous works [11] in which a set of articles from news websites from 6 different sources in English is evaluated. They were pre-processed in order to eliminate recurring terms of the language, uppercase and lowercase were normalized, words within the vector or document given a weight or importance, and KEEL simulation tool was used in order to obtain the accuracy percentage of each algorithm in the news collection as shown in Table 3.

Table 3. Predictive accuracy (%) of each algorithm in the news collection varying the number of terms from 20 to 2000

# of Terms	LVQ1	LVQ2.1	LVQ3	SVM	° C	KNN	NB	Average
20	77,5	74,7	78,2	84,5	85,8	75,5	86,9	78,7
40	80,2	77,3	79,2	89,4	91,0	82,5	92,0	81,5
60	78,4	77,5	78,3	88,0	89,1	79,6	91,6	80,6
80	79,0	76,8	77,9	88,8	89,4	79,3	92,3	80,6
100	79,5	76,7	79,1	88,5	89,3	78,4	92,3	80,9
200	85,8	78,9	85,4	91,0	87,4	77,9	92,1	85,3
300	87,1	79,1	84,0	92,6	87,8	69,0	91,9	85,7
400	89,0	82,0	87,7	93,4	86,9	74,6	91,8	88,0
500	89,3	83,1	87,7	94,3	86,8	73,1	91,5	88,6
600	90,0	83,9	87,0	95,5	85,9	72,8	91,4	89,1
700	90,3	82,9	87,3	95,8	85,9	73,3	91,1	89,0
800	91,0	83,4	86,3	96,0	85,9	73,4	90,9	89,1
900	90,3	83,5	85,0	96,6	86,0	78,1	90,5	88,8
1000	90,9	82,1	78,9	96,5	85,9	75,9	90,5	87,1
1100	90,6	81,7	73,3	96,9	85,8	74,5	90,4	85,6
1200	90,5	81,1	69,9	97,1	85,6	73,6	90,1	84,6
1300	91,0	81,3	65,5	97,4	85,6	67,9	89,8	83,8
1400	89,9	80,8	64,4	97,3	85,6	65,9	89,9	83,1
1500	89,8	82,1	61,6	97,4	85,6	64,6	89,9	82,7
1600	90,2	80,8	54,7	97,3	85,6	58,6	89,8	80,7
1700	89,8	80,7	52,8	97,3	85,6	51,5	89,6	80,2
1800	90,3	81,0	49,3	97,6	85,6	48,8	89,6	79,5
1900	89,9	81,6	49,8	97,6	85,6	46,8	89,5	79,7
2000	89,9	81,0	46,3	97,5	85,6	25,0	89,5	78,7

Recovered from [11]. An evaluation of the LVQ algorithm in a text collection. *Cuban Journal of Computer Science, 10* (4).

d. Facial recognition

Due to the fact that individualization will be made taking into account only images contained in sanction lists databases and images of web press articles, only methods of image analysis will be taken into account and those based on 3D models will have to be discarded. The following two facial recognition techniques are proposed:

Principal component analysis (PCA). Known as PCA technique. This facial recognition technique employs an initial processing of a face image to convert the matrix of pixels into a set of vectors, then, they will be projected in a space of smaller values. These values are compared with those stored in a database of facial information taking into account a tolerance value.

Locality preserving projections (LPP). This algorithm is known as LPP. LPP performs the same reduction of initial data that PCA performs, but in addition, it performs another process which results in almost identical values in the small projected space of values when dealing with the face of the same person in consecutive images taken from the same video source. Such additional processing may result in lower computation speed, but the accuracy in results will be greater.

e. Comparison between algorithms

To verify the success rate and speed with which both methods positively identify individuals by feeding algorithms with face images, results obtained by [9] will be taken, who developed the software tests taking into account return values of instantaneous results and accumulated results; that is, those that required more processing time before generating a positive result as shown in Table 4.

Table 4. Test results with facial recognition algorithms

Method	Instant results	Accumulated results
DCT	78.744%	81.7%
LPP	77.456%	85.4%

Recovered from [9]. Estudio de técnicas de reconocimiento facial, 86.

f. Background

As can be observed in the theoretical framework and in the articles referenced, algorithms intended to be used for a more efficient individualization process have already been widely developed by various experts in computer science. So it can be said that the development of this work is based on the compilation of previous works results rather than in development. On the other hand, despite the antiquity of some of the algorithms discussed in this article, they are widely used nowadays, because they have proven their effectiveness over time, such as KMP algorithm for searching text patterns, which is still used in current browsers when the user wants to search for a text in a web document. The optimization of processes described in this article would

3. RESULTS

From the analysis of text search algorithm complexities, as well as tests proposed and developed by [10], the result is that Boyer-Moore algorithm is the most recommended to carry out searches, being superior to the others in terms of execution speed. In the analysis of individualized algorithms results, we appreciate that SVM performance presents a consistency superior to other algorithms, always exhibiting a behavior above the average, regardless of the number of terms of the sample, and showing a behavior without negative fluctuations in cases of greater number of terms.

When comparing facial recognition algorithms proposed and evaluated by [9], it is concluded that the method generating the most positive results was locality preserving projections, reaching more than 85% identifications, so it would be the most recommended one when implementing a system. With analysis of results and obtaining the best algorithmic methods, the future incorporation of different methodologies is considered as a future work in order to optimize results in individualization process, leading to ideal results and with a lower error coefficient and false positives. Artificial intelligence algorithms allow not only to improve results searching databases and in documents of news websites, but also to classify different sources in order to give priority to those that have greater relevance in the individualization of subjects.

4. CONCLUSION

In this paper, three categories of algorithms to be used in the process of implementing an individualization system for criminal records searches were examined. Th algorithmic categories examined here were search in texts, artificial intelligence for personalization of searches, and facial recognition.

They were compared using the metrics proposed in previous works, such as Hernández, Gou, and Betancour in order to obtain the best techniques from each category. Finally, it was found that the most recommendable algorithms for use in an individualization system are Boyer-Moore for text search, vector support machines for artificial intelligence, and locality preserving projections for facial recognition.

ACKNOWLEDGEMENTS

This work was supported by the Universidad Nacional de Colombia, Sede Medellín under the projects HERMES-34671 and HERMES-36911. The authors thank the School of Physics for their valuable support to conduct this research.

REFERENCES

[1] A F. Z. Salmam, A. Madani, and M. Kissi, "Emotion recognition from facial expression based on fiducial points detection and using Neural Network," *Int. J. Electr. Comput. Eng*., vol. 8, no. 1, p. 52, Feb. 2018.
[2] B L. Deshpande and M. N. Rao, "Concept Drift Identification using Classifier Ensemble Approach," *Int. J. Electr. Comput. Eng.*, vol. 8, no. 1, p. 19, Feb. 2018.
[3] C A. L. H. P.S and U. Eranna, "An Efficient Activity Detection System based on Skeleton Joints Identification," *Int. J. Electr. Comput. Eng*., vol. 8, no. 6, p. 4995, Dec. 2018.
[4] V. Balajichandrasekhar M., T. S. Rao, and G. Srinivas, "An Improvised Methodology to Unbar Android Mobile Phone for Forensic Examination," *Int. J. Electr. Comput. Eng*., vol. 8, no. 4, p. 2239, Aug. 2018.
[5] Pandey, S. K., Dubey, N. K., & Sharma, S, "A Study on String Matching Methodologies," 5(3), 4732-4735, 2014.
[6] Joshi, C., Jaiswal, T., & Gaur, H, "An Overview Study of Personalized Web Search," 3(1), 1-3, 2013.
[7] Salmela, L., & Tarhio, J., "Algorithms for Weighted Matching," Work, 276-286, 2007.
[8] Hapfelmeier, A., Mertes, C., Schmidt, J., & Kramer, S, "Towards Real-Time Machine Learning," *Department of Computer Science, Technische Universität München*, 85748 Garching, Germany, 2012.
[9] Hernández, R., "Estudio de técnicas de reconocimiento facial," 86. Recuperado de http://teocom.googlecode.com/svn-history/r39/trunk/docs/papers/PFC_RogerGimeno.pdf, 2010.
[10] Gou, M, "Algorithms for String matching," 1-8, 2014.
[11] Guerrero Enamorado, A., & Ceballos Gastell, D, "Una evaluación del algoritmo LVQ en una colección de texto," *Revista Cubana de Ciencias Informáticas*, 10(4), 154-170, 2016.
[12] García, C., & Gómez, I, "Algoritmos de aprendizaje: knn & kmeans," *Universidad Carlos III de Madrid*, 1-8. Retrieved from http://www.it.uc3m.es/jvillena/irc/practicas/08-09/06.pdf, 2006.
[13] Betancour, G, "Las máquinas de soporte vectorial (SVMs)," *Scientia Et Technica*, (27), 67-72. https://doi.org/10.22517/23447214.6895, 2005.

BIOGRAPHIES OF AUTHORS

Gonzalo M. Arias: Informatics Engineer graduated from Politécnico Colombiano Jaime Isaza Cadavid, Medellín, Colombia. Worked as professor of C# in the Tecnológico de Antioquia, Medellín, Colombia. He is a software programmer with experience in C#, Visual Basic and C++. His research interests cover mostly security topics. E-Mail: Mauricio.arias@outlook.com

Pablo A. Peláez: Systems Engineer graduated from Universidad de Antioquia, Medellín, Colombia. He is a documentation manager and database administrator.
E-Mail: andresgris82@gmail.com

Fredy Edimer Hoyos: received his BS and MS degree from the National University of Colombia, at Manizales, Colombia, in Electrical Engineering and Industrial Automation, in 2006 and 2009, respectively, and Industrial Automation Ph.D. in 2012. Dr. Hoyos is currently an Associate Professor of the Science Faculty, School of Physics, at National University of Colombia, at Medellin, Colombia. His research interests include nonlinear control, system modelling, nonlinear dynamics analysis, control of nonsmooth systems, and power electronics, with application within a broad area of technological process. Dr. Hoyos is an Associate Researcher in Colciencias and member of the Applied Technologies Research Group - GITA at the Universidad Nacional de Colombia. https://orcid.org/0000-0001-8766-5192