IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 25, 2018, accepted March 9, 2018, date of publication March 15, 2018, date of current version May 9, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2815992

An Automatic Document Classifier System Based on Genetic Algorithm and Taxonomy

ALAN DÍAZ-MANRÍQUEZ, ANA BERTHA RÍOS-ALVARADO, JOSÉ HUGO BARRÓN-ZAMBRANO, TANIA YUKARY GUERRERO-MELENDEZ, AND JUAN CARLOS ELIZONDO-LEAL

Facultad de Ingeniera y Ciencias, Universidad Autónoma de Tamaulipas, Ciudad Victoria 87000, México

Corresponding author: Alan Díaz-Manríquez (amanriquez@docentes.uat.edu.mx)

ABSTRACT The use of the Web has increased the creation of digital information in an accelerated way and about multiple subjects. Text classification is widely used to filter emails, classify Web pages, and organize results retrieved by Web browsers. In this paper, we propose to raise the problem of automatic classification of scientific texts as an optimization problem, which will allow obtaining groups from a data set. The use of evolutionary algorithms to solve classification problems has been a recurrent approach. However, there are a few approaches in which classification problems are solved, where the data attributes to be classified are text-type. In this way, it is proposed to use the association for computing machinery taxonomy to obtain the similarity between documents, where each document consists of a set of keywords. According to the results obtained, the algorithm is competitive, which indicates that the proposal of a knowledge-based genetic algorithm is a viable approach to solve the classification problem.

INDEX TERMS Classification algorithms, genetic algorithms, evolutionary computation, optimization.

I. INTRODUCTION

The Web use has strengthened the creation of digital information in an accelerated way and about multiple topics. The classification of text is widely used to filter e-mails, classify web pages and organize the results recovered by the web browsers [1]. In the process of recovering information, which includes work-tasks of representation, organization, storage and access to the information, it is desired to have associations of documents by keywords at all times. In this way, classifying documents in an automatic way would allow us to find information in a more efficient way.

The classification task[1] is a grouping procedure which allows us to group a set of data according to a selected criterion. Generally the objects or data of the same group share similar characteristics with one another, while the objects of different groups will have less similarity among them. For example, in an organization the documents can be classified by a functional criterion, that is to say, grouping the documents by activities inside of the company, or through a criterion of hierarchical order, where the managers have access to different documents that employees have access to.

The goal in the task of classification is to locate the document of an appropriate class. For this, the classification systems mainly include three stages: feature extraction, feature selection and classification [2]. Having a large number of features makes the classification process to be computationally expensive and that the classes not being well defined. The feature selection focuses on reducing the dimensionality, many approaches concentrate on considering only one subset of features extracted from text. Generally, for the selection of characteristics we have techniques based in the collection of documents (information profit, frequency of words, among others) and techniques based in typifying each class [3].

To solve the problem of text classification automatically different approaches have been proposed, among which the supervised [4]–[6] and not supervised algorithms [7]–[9] stand out. The supervised algorithms employ techniques such as Bayesian classifier [4], [5] and k-nearest neighbors [6]. As for the latter, the not supervised approaches use the hierarchical and partitioning grouping algorithms [8], such as K-means [9] and K-medoids [10]. Furthermore, to deal with the ambiguity of text in the sense of the words, it has been incorporated the use of WordNet[2] [11]–[13], as a base for the knowledge to associate phrases and words to the documents.

In both approaches, a sensible piece of information is the number of classes (or groups). For this reason, to solve

[1]In this paper the terms classification, grouping or "clustering" will be used indistinctively.

[2]https://wordnet.princeton.edu/

the classification problem it has been proposed treating it as an optimization problem. There exist a wide variety of classic methods for solving optimization problems. Nonetheless, some classic algorithms tend to stay blocked up in optimal locals or require a great amount of evaluations of the objective function. On the other hand, there exist stochastic optimization techniques, such as the simulated annealing, ant colony optimization, tabu search, evolutionary algorithms, among others, which have been used to efficiently solve optimization problems. In this context, techniques based in *soft computing* have been proposed for the grouping of texts, such as the ant colony optimization [14], particle swarm optimization [7], [15] and genetic algorithms [1], [2].

In this project it is proposed to set the problem of automatic classification of scientific texts as an optimization problem, which allows us to obtain (calculate) the groups from a dataset. The texts correspond to scientific articles associated to the Computing area and the ACM[3] (*Association for Computing Machinery*) taxonomy has been selected, which has been developed by experts in the area to distinguish each document with it keywords set. With the algorithm proposed, it will be found a grouping of the scientific documents with a base on its keywords and using the taxonomy of the ACM the similarity among documents will be measured. This grouping will improve the organization of the documents and will make the process of information recovery easier for the set of documents belonging to an investigation group.

The document is organized in the following way. In Section 2 a brief description of the related papers is presented. In Section 3 some basic concepts are introduced for the proper understanding of this project. In Section 4, the details for the proposed algorithm are presented. Then, in Section 5 the results are shown. Finally, the conclusions and future work are described in Section 6.

II. RELATED WORKS

The topic of text classification has been widely studied with different techniques. In the following section we will describe some work related to the classification task as a problem of optimization and evolutionary techniques for its solution have been used. Furthermore, some work that is supported on the bases for knowledge to improve the representation of the documents is presented.

In [14] a classification of the web pages is proposed where the feature selection process uses the approach of optimization through ant colony to obtain the best features. Once we have the best features the classification process is done using decision trees. Cui *et al.* [15] propose a hybrid algorithm for the grouping of texts. They use the optimization technique through particle swarm and *K-means*. In [7] an algorithm that combines the approach of genetic algorithms with particle swarm optimization is presented with the objective of improving the diversity and convergence of the algorithm of the desired solution. Other approaches like [1], [2] propose genetic algorithms in the process of text grouping. In the case of [2] they combine a genetic algorithm with the technique for the representation of documents, called latent semantics analysis, for improving the selection and transformation of characteristics. On the other hand, in [1] it is proposed to use the genetic algorithm as a classifier, where given a big amount of training documents, rules are generated for classification with a high rate of flexibility that allows to correctly classify a new document.

An important aspect to consider is the representation of the texts, because problems associated to the processing of natural language such as ambiguity and synonymy can be avoided by integrating a set of concepts to the representation. In [11] and [13] it is proposed to extend the representation of words from the document with *synsets*,[4] concepts and sub concepts starting from WordNet. Its experiments show an improvement in the grouping process using *K-means*. In [12] it is proposed to use WordNet to decrease the dimension of the characteristics of the texts and then use a representation called lexical chains that even allow to deal with the problem of ambiguity.

III. BACKGROUND

A. PROBLEM STATEMENT

Given the set $D = \{d_1, d_2, d_3, \ldots, d_n\}$ of scientific documents in which each document d contains m keywords, it is necessary to find from D a set $C = \{C_1, C_2, C_3, \ldots, C_m\}$, in such a way that the similarity between documents in the C_k group to be maximum according to the criterion of grouping.

The grouping problem is one of the most important tasks in data mining. Grouping consists in the act of partitioning a set of not-labelled data inside of groups of similar objects. Each group, called ''cluster'', consists of objects that are similar among them and different to the objects of other groups at the same time.

B. DAVIES-BOULDIN INDEX

A cluster validity index refers to statistic mathematical functions used to evaluate the results of clustering algorithm in a quantitative way. Generally a cluster validity index is useful for two purposes. First, this can be used to determine the number of groups, and second, to determine which is the best partition. In this paper the Davies-Bouldin (DB) index is used. This measure is in function of the dispersion inside the cluster (the more compacted the better) and the distance among clusters (the farther the clusters are among them, the best). To obtain the DB index, it is necessary to define the dispersion inside the cluster (DDC) and the distance among the clusters (DEC). Next, each one them will be defined:

$$DDC_{i,q} = \left[\frac{1}{N_i}\sum_{X \in C_i} D(X, m_i)\right]^{1/q}, \tag{1}$$

[3]https://www.acm.org/

[4]Synset: set of synonyms of a word

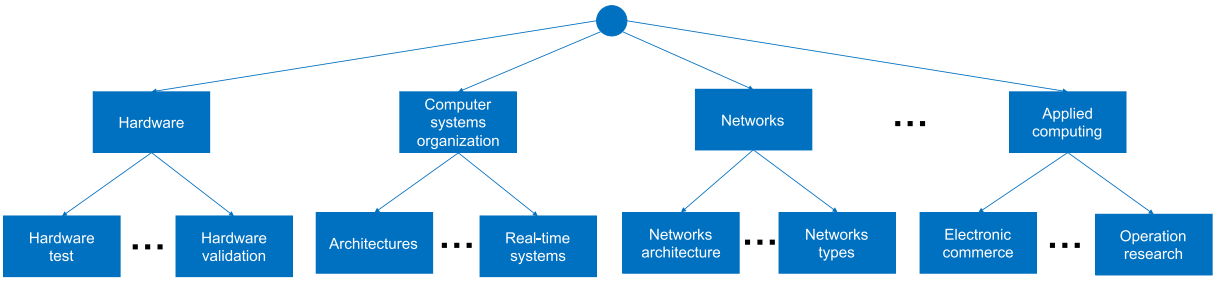

FIGURE 1. Example of the representation of the ACM taxonomy.

$$DEC_{i,j} = \left\{ \sum_{p=1}^{d} D(m_{i,p}, m_{j,p})^t \right\}^{1/t} = ||m_i - m_j||_t \quad (2)$$

where X is a document belonging to C_i, m_i is the centroid of the i (the centroid m_i contains TC keywords), $D(\mathrm{X}, m_i)$ is the distance between the centroid and X (see Equation 5). On the other hand, $q, t \geq 1$, are whole numbers that can be selected independently. N_i is the size of the i cluster. Additionally, it is necessary to find the biggest proportion ($R_{i,q}$) among the i cluster and the rest:

$$R_{i,q} = \max_{j \in K, j \neq i} \left\{ \frac{DDC_{i,q} + DDC_{j,q}}{DEC_{i,j}} \right\}. \quad (3)$$

Finally, the DB index of a cluster will be defined as the addition of the maximum radius:

$$DB(K) = \frac{1}{K} \sum_{i=1}^{K} R_{i,q} \quad (4)$$

It is necessary to notice that the smaller the $DB(K)$ is, the better the grouping will be.

IV. METHODOLOGY

A. ACM TAXONOMY

The extraction of the taxonomy of the Association for Computing Machinery (ACM) was done automatically. The aforesaid taxonomy can be represented in the form of a tree, that is to say that to obtain the distances between terms, an algorithm to find the minimum distances in a graph can be applied. For such objective the Floyd-Warshall [16] algorithm was used, which requires representing the tree as a graph, and such graph was represented as an adjacency matrix, in order to later apply the algorithm. Once applied the Floyd-Warshall algorithm, in the matrix there will be minimum distances left between each couple of terms, and these distances will be used to obtain the distances between documents. As an example, in Figure 1 it can be observed that the distance between *Hardware* and *Hardware test*, will be 1. While the distance between *Hardware test* and *Hardware validation* will be 2. That is to say, the distance between father and son, will be one. On the other hand, the distance between brothers will be 2.

B. REPRESENTATION OF DOCUMENTS

Each scientific document will be represented as a set of keywords, so that the distance between each pair of documents will be defined as the average of the distances between each couple of keywords:

$$D(i,j) = \frac{\sum_{k=1}^{Ni} \sum_{l=1}^{Nj} dist(K_k, K_l)}{Ni * Nj} \quad (5)$$

where Ni is the number of keywords in the document i, and Nj is the number of keywords in the j document, K_k is the k keyword and K_l is the keyword l. Finally, to obtain the similarity between documents, the following formula is used:

$$S_{i,j} = \frac{1}{D(i,j)} \quad (6)$$

C. EVOLUTIONARY ALGORITHM

The EA used to solve the classification problem as an optimization problem is the Genetic Algorithm (GA). The GA was initially proposed by John H. Holland in his book *Adaptation in Natural and Artificial Systems* [17]. HollandâĂŹs main interest was to study the natural adaptation with the purpose to apply it to machine learning. Holland was convinced that the recombination of a group of genes known as mating (matching) was a critical part in evolution. The GA was developed under the concepts of crossover and mutation. A pseudocode of the simple genetic algorithm is shown in algorithm 1.

Algorithm 1 Simple Genetic Algorithm

1: Initializing population with random prospect solutions (individuals)
2: Evaluate the aptitude of each individual
3: **repeat**
4: Parents selection
5: Apply recombination to parents
6: Apply mutation to each offspring individual
7: Evaluate the new descending individuals
8: Select the most apt individuals to advance to the next generation (replacement)
9: **until** Termination criteria is fulfilled

In the subsequent sections each of the GA components proposed to solve the classification problem are presented.

1) CHROMOSOME REPRESENTATION (INDIVIDUAL)

To represent the clustering problem as an individual from the genetic algorithm an entire representation was used in the following way. In the first phase it was necessary to assign to each word of the ACM taxonomy a whole number (index). In this way, so that each term (keyword) a unique identifier will be held.

Once the identifier for each word is obtained, two parameters will be necessary, the first of them is K, which refers to the number of groups that it is desired to obtain with the genetic algorithm. On the other hand, the second parameter will be TC, which refers to the size of the centroid of each group.

Once defined these two parameters, the size of the chromosome can be obtained, which will be represented by a set of whole numbers of the size $K * TC$. Each whole number will make a reference to a keyword from the ACM taxonomy. In this way, the first TC positions from the set, will make reference to the first cluster, the next TC positions will make reference to the second cluster, and so on. Thus, only one individual will contain the K centroids belonging to only one grouping.

2) FITNESS FUNCTION

Once the problem is represented it is necessary to be able to evaluate an individual. For this purpose the Davies-Bouldin index will be used, which allows to verify how good a grouping is. Therefore, the grouping problem will be defined as the minimization of the DB index, and that the aptitude function will consist precisely in the $DB(K)$ value.

3) VARIATION OPERATORS (CROSSOVER AND MUTATION)

Given that the representation used for this problem is the whole representation, there exist in the literature a diversity of methods for crossover and mutation already proposed [18]. For this work the one point crossover proposed by John [17] will be used, crossover of n points and uniform crossover. On the other hand, for the mutation only one type of mutation was tested, in this case it was the random mutation. This type of mutation consists in that for every gene there is a p_m probability of choosing a new random value inside the set of permissible values. In this case it will be a new index corresponding to a new word from the ACM taxonomy.

4) PARENTS SELECTION

The selection of the parents that should be matched will be done through tournament. The selection through tournament is an operator that has de property of not requiring any global knowledge of the population, nor any quantifiable measure of quality. This type of selection is fast and easy to implement and apply. The application of the tournament selection chooses λ members of a population of mu individuals. The pseudocode to apply the selection through tournament is shown in the Algorithm 2.

Algorithm 2 Pseudocode for the Selection Through Tournament

```
/*Assuming that it is required to select λ individuals from a population of μ*/
actual_member = 1
while (actual_member ≤ λ) do
    Selecting k individuals from the population randomly
    Comparing the individuals and selecting the best from them (we will name this individuals as i)
    selecte[actual_member] = i
    actual_member = actual_member + 1
end while
```

For this work, binary tournament was used, and the number of individuals to select are the size of the population, that is to say $\lambda = \mu$.

5) SURVIVOR SELECTION

The selection mechanism of survivors, also known as replacement, is the responsible for reducing the memory used during the evolutionary process, that is to say, reducing the size of μ parents and λ children to only μ individuals that will form the parents of the next generation.

For this work the selection $\mu + \lambda$ was used. This type of selection refers to the case where the set of parents and children are mixed and hierarchized according to their aptitude, in this way, the best μ individuals are kept as parents for the next generation.

V. RESULTS

In this section the experiments and results obtained are shown to evaluate the efficiency of the proposed algorithm.

A. EXPERIMENT IN TEST DATA

1) TEST CASES GENERATOR

In order to evaluate the performance of a proposed algorithm, a test cases generator was created. Given that the ACM taxonomy has 11 main categories, an algorithm able to create a maximum of 11 different groups was created (one for each category). Thus, for each category a set of representative words was chosen, taking mainly the words that are found as leaves in a tree from the taxonomy. For this reason, if it is desired to create a set of tests with m clusters, m categories were chosen randomly, and for each one of them n documents were created, for which for each k keywords belonging to the chosen category were selected. Such procedure can be seen in algorithm 3.

2) EXPERIMENTAL DESIGN

With the purpose of evaluating the results of the clustering algorithm in an internal way, the Precision measure (A) is used. To measure the Precision of the resulting clusters Equation 7 will be used, which determines the total amount of

Algorithm 3 Test Cases Generator

```
Generating representative keywords for each category.
Let m the number of groups
Let n the number of documents per group
Let k the number of keywords per document
Categories ←Selecting m groups randomly
for i = 1 to m do
    for j = 1 to n do
        /* documents[i] will be a vector with the k keywords */
        documents[j + (i − 1) ∗ m] ←Selecting k keywords from category[i]
    end for
end for
```

TABLE 1. Set of parameters used in the algorithms proposed.

Algorithm	N	$\%_c$	$\%_m$	gens	TC
GA-1	100	0.8	$1/tam\_crom$	100	5
GA-2	100	0.8	$1/tam\_crom$	100	5
GA-3	100	0.8	$1/tam\_crom$	100	5

correctly assigned elements in each cluster [19].

$$A = \frac{\sum_{i=1}^{k} c_i}{n}. \tag{7}$$

Where k is the final number of clusters, n is the number of instances of the dataset, c_i is the number of correctly classified instances in the i cluster and in its corresponding class. Furthermore, the error (E) is calculated which is defined by Equation 8.

$$E = 1 - A \tag{8}$$

In order to measure the efficiency of the proposed algorithm three sets of different data were created with 3, 4 and 5 groups, for each class in each dataset 50 documents were created, that is to say, the dataset with 3 classes will contain 150 documents, the dataset with 4 classes 200 documents and the dataset with 5 classes 250 documents. In the proposed algorithm three different versions of the GA were evaluated, GA-1 (one point crossover), GA-2 (two points crossover) and GA-3 (Uniform crossover). The parameters used for the algorithms are observed in Table 1. The percentage of crossover ($\%_c$) used was 0.8 and the percentage of mutation ($\%_m$) $1/n$, where n is the size of the chromosome. Each algorithm was executed 31 times, and the E error was measured. The results obtained for the dataset with 3 clusters can be seen in Figure 2. As it can be observed, the best result for the three clusters was achieved by the GA with uniform crossover. The second place was for the algorithm with two points crossover and the worst results were achieved by the algorithm that uses the one point crossover. Nonetheless, in the three versions of the algorithm it can be observed that the error is very small, the maximum error obtained is approximately 0.23.

On the other hand, the evaluation with four clusters can be observed in Figure 3. In such Figure a similar behavior can be

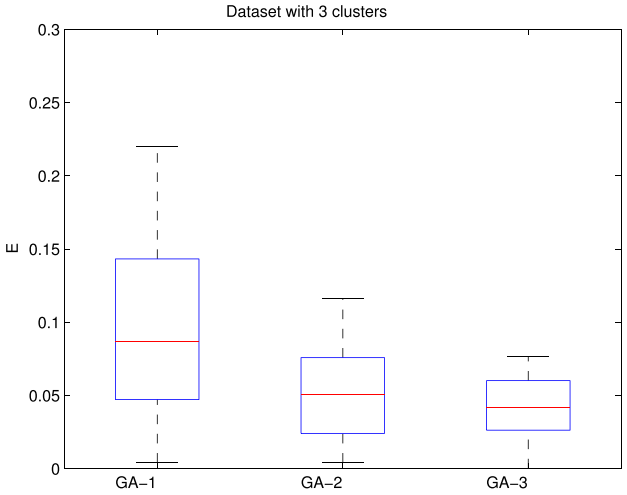

FIGURE 2. Results of the execution of the three versions of the proposed algorithm: set of data with 3 groups.

TABLE 2. Results of precision, exhaustiveness and measure *F*1 for the set of data with 3 groups.

Algorithm	P	R	F1
GA-1	0.9133	0.9144	0.9138
GA-2	0.9400	0.9401	0.9400
GA-3	0.9600	0.9603	0.9601

TABLE 3. Results of precision, exhaustiveness and measure F1 for the set of data with 4 groups.

Algorithm	P	R	F1
GA-1	0.8850	0.8885	0.8867
GA-2	0.9000	0.9015	0.9007
GA-3	0.9400	0.9414	0.9407

TABLE 4. Results of precision, exhaustiveness and measure F1 for the set of data with 5 groups.

Algorithm	P	R	F1
AG-Cruza1	0.8600	0.8635	0.8617
AG-Cruza2	0.9120	0.9119	0.9119
AG-Cruza3	0.9400	0.9409	0.9404

observed, the algorithm with the best performance was again for the genetic algorithm with uniform crossover. On the contrary, it can also be observed that the errors were less than 0.3. Generally, it can be mentioned that for four groups the results obtained are acceptables.

Finally, in Figure 4 the results obtained with five clusters are shown. It can be observed that once more the algorithm with better behavior was the GA with uniform crossover. Additionally, it can be validated that even though the growth in number of clusters the behavior of the algorithm keeps itself stable, with errors less to 0.3.

In order to evaluate the results of the grouping in an external way, seen as a series of decisions, the precision, exhaustiveness and measure-F were calculated, such measures have been widely used for the evaluation of the task

TABLE 5. Matrix of confusion for the classifier AG-3.

	Category	Obtained: Hardware	Networks	Software and its engineering	Applied Computing	Total
Real	Hardware	20	0	0	0	20
	Networks	0	15	3	2	20
	Software and its engineering	1	1	16	2	20
	Applied Computing	0	2	1	17	20
	Total	21	18	20	21	80

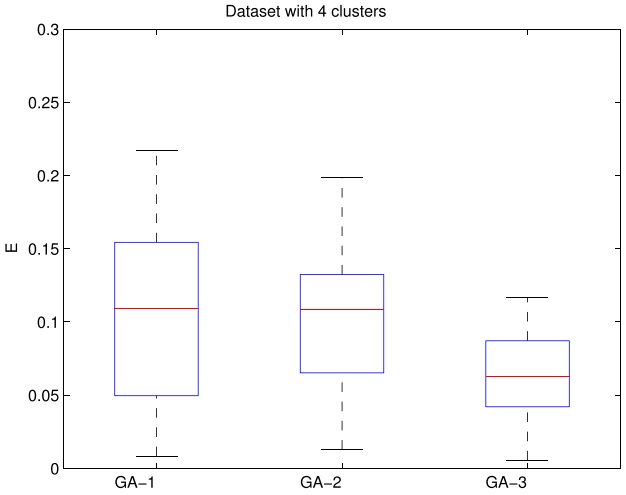

FIGURE 3. Results of the execution of the three versions of the proposed algorithm: set of data with 4 groups.

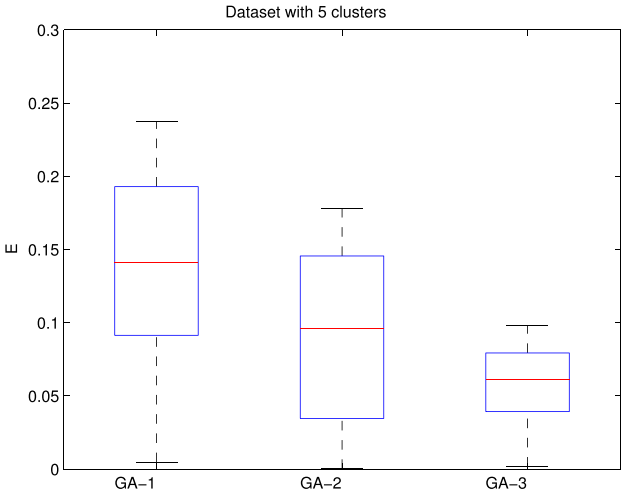

FIGURE 4. Results of the execution of the three versions of the proposed algorithm: set of data with 5 groups.

of text grouping [1], [20], [21]. The precision (P) (see Equation 9) is calculated as the fraction of documents correctly placed in the same group; the Exhaustiveness (R) (see Equation 10) is the fraction of real documents that are identified; and finally the F measurement ($F1$), calculated by Equation 11, is a harmonic median between P and R. To calculate P, R and $F1$ the following data is defined:

- *True Positives (TP)* are those documents that were located by the algorithm in the same cluster that indicates the class that was known beforehand.
- *False Positives (FP)* make reference to those documents that were located by the algorithm in the cluster i and that actually belonged to another cluster.
- *False Negatives (FN)* are those elements of the cluster i that were located in a different cluster that was indicated in its label.
- *True Negatives (TN)* is the amount of documents that were located correctly outside the cluster i, that is to say, those unaffiliated documents to the cluster in matter and that indeed did not correspond to this one.

$$P = \frac{TP}{TP + FP} \tag{9}$$

$$R = \frac{TP}{TP + FN} \tag{10}$$

$$F1 = \frac{2 * P * R}{P + R} \tag{11}$$

Tables 2, 3, 4 show the results of the measures P, R and $F1$ calculated for the testing documents with 3, 4 and 5 clusters, respectively. It can be observed that the genetic algorithm with uniform crossover (GA-3) reaches the highest values of $F1$ in all cases, this value being higher to 94%.

B. REAL DATA TESTS

In order to evaluate the results of the clustering in real data, 4 groups of scientific articles of different areas were created. The categories selected were *Hardware*, *Networks*, *Software and its engineering* and *Applied Computing*. For each category 20 different articles were selected. The keywords associated to each article do not shovel among them. The algorithm that was evaluated was the GA-3, which is the one that got the best results in the tests with fictitious data.

Table 5 corresponds to the matrix of confusion that shows the results of the classifier GA-3. The distribution of the elements related to the real classes and the forecasted classes can be observed. For the *Hardware* class, its 20 elements were correctly classified. In the *Networks* class 15 elements were correctly classified and from the remaining 5, 3 were classified in the *Software and its engineering* class and 2 in the *Applied computing* class. On the other hand, for the *Software and its engineering* class, 16 elements were correctly classified and from the remaining 4, 1 was assigned to the *Hardware* category, 1 to the *Networks* category and 2 to the *Applied computing* category. Finally, for the *Applied computing* class, 17 were correctly classified, 2 were located in the *Networks* category and 1 was assigned to the *Software*

TABLE 6. Results of the evaluation from the GA-3 classifier with real data.

	Hardware	Networks	Software and its engineering	Applied Computing
True Positives (TP)	20	15	16	17
False Positives (FP)	0	5	4	3
False Negatives (FN)	1	3	4	4
P	1.0	0.75	0.8	0.85
R	0.95	0.83	0.8	0.8
F1	0.97	0.79	0.8	0.88

and it engineering category. In Table 6 the results of P, R and $F1$ for each class are shown. An average of 85% in precision was achieved.

VI. CONCLUSIONS

The use of evolutionary algorithms to solve classification problems has been a recurrent approach. Nonetheless, there are scarce approaches in which problems of classification are solved where the attributes of the data to classify are of text kind. In this sense, the use of the ACM taxonomy was proposed in order to obtain thy similarity among documents, where each document is formed by a set of keywords. Additionally, a genetic algorithm that improves the index of validity of a cluster was proposed, which also makes use of the measure of similarity already proposed between documents. Furthermore, a methodology to obtain the distance between the words in the ACM taxonomy was designed. Such methodology makes use of the Floyd-Warshall algorithm, which is typically used to obtain the minimum distance between two nodes in a graph. On the contrary, the grouping problem of scientific documents was proposed as a problem of optimization of one objective, for which it was designed a genetic algorithm in order to solve the before mentioned problem of optimization. Moreover, a test cases generator was designed, with the objective of having instances to carry out an experimental study.

On the other hand, for the proposed algorithm three functions of crossover were implemented, obtaining three versions of the same algorithm. They were compared in three sets of different data. According to the results obtained it can be concluded that the algorithm with the best performance was the one that used the uniform crossover. Nevertheless, it was also able to be observed that in general all the versions achieve to solve the clustering problem in a very close way to the best.

Furthermore, the results obtained from the $F1$ measure given by the clustering algorithm with fictitious data and with real data are competitive, which indicates that the proposal of a genetic algorithm for the clustering of documents based in knowledge (taxonomy) is viable for a later process of information recovery.

Finally, as future work it is expected to evaluate the efficiency of the algorithm in a set with a higher amount of real scientific articles. Additionally, it is expected to propose another measure of similarity between scientific documents that does not make use of the ACM taxonomy so in this way we do the text grouping of document sets of any domain and not being restricted to the taxonomy.

REFERENCES

[1] M. I. Khaleel, I. I. Hmeidi, and H. M. Najadat, "An automatic text classification system based on genetic algorithm," in *Proc. 3rd Multidiscipl. Int. Social Netw. Conf. SocialInform., Data Sci.*, 2016, Art. no. 31.

[2] A. K. Uysal and S. Gunal, "Text classification using genetic algorithm oriented latent semantic features," *Expert Syst. Appl.*, vol. 41, no. 13, pp. 5938–5947, 2014.

[3] L. Özgür and T. Güngör, "Two-stage feature selection for text classification," in *Information Sciences and Systems 2015*. Kidlington, U.K.: Springer, 2016, pp. 329–337.

[4] Y.-H. Chang and H.-Y. Huang, "An automatic document classifier system based on Naíve Bayes classifier and ontology," in *Int. Conf. Mach. Learn. Cybern.*, Jul. 2008, pp. 3144–3149.

[5] K. A. Vidhya and G. Aghila, "Hybrid text mining model for document classification," in *Proc. 2nd Int. Conf. Comput. Autom. Eng. (ICCAE)*, Feb. 2010, pp. 210–214.

[6] V. Bijalwan, P. Kumari, J. Pascual, and V. B. Semwal, "Machine learning approach for text and document mining," *CoRR*, vol. 6, pp. 115–123, Jun. 2014.

[7] K. Premalatha and A. M. Natarajan, "Hybrid PSO and GA models for document clustering," *Int. J. Adv. Soft Comput. Appl*, vol. 2, no. 3, pp. 1–19, 2010.

[8] L. F. D. C. Nassif and E. R. Hruschka, "Document clustering for forensic analysis: An approach for improving computer inspection," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 1, pp. 46–54, Jan. 2013.

[9] N. Kamel, I. Ouchen, and K. Baali, "A sampling-PSO-K-means algorithm for document clustering," in *Genetic and Evolutionary Computing*. Prague, Czech Republic: Springer, 2014, pp. 45–54.

[10] R. Barbosa, D. Januario, A. E. Silva, R. Moraes, and P. Martins, "An approach to clustering and sequencing of textual requirements," in *Proc. IEEE Int. Conf. Dependable Syst. Netw. Workshops (DSN-W)*, Jun. 2015, pp. 39–44.

[11] K. Sarkar and R. Law, "A novel approach to document classification using WordNet," *CoRR*, vol. 1, pp. 259–267, Oct. 2015. [Online]. Available: https://arxiv.org/abs/1510.02755

[12] T. Wei, Y. Lu, H. Chang, Q. Zhou, and X. Bao, "A semantic approach for text clustering using WordNet and lexical chains," *Expert Syst. Appl.*, vol. 42, no. 4, pp. 2264–2275, 2015.

[13] A. Hotho, S. Staab, and G. Stumme, "Ontologies improve text document clustering," in *Proc. 3rd IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2003, pp. 541–544.

[14] E. Saraç and S. A. Özel, "An ant colony optimization based feature selection for Web page classification," *Sci. World J.*, vol. 2014, Jul. 2014, Art. no. 649260.

[15] X. Cui, T. E. Potok, and P. Palathingal, "Document clustering using particle swarm optimization," in *Proc. IEEE Swarm Intell. Symp. (SIS)*, Jun. 2005, pp. 185–191.

[16] E. W. Weisstein. "Floyd-Warshall algorithm." From MathWorld—A Wolfram Web Resource. [Online]. Available: http://mathworld.wolfram.com/Floyd-WarshallAlgorithm.html

[17] H. J. Holland, *Adaptation in Natural and Artificial Systems*. Cambridge, MA, USA: MIT Press, 1992.

[18] A. E. Eiben and J. E. Smith, *Introduction to Evolutionary Computing*. New York, NY, USA: Springer, 2003, vol. 53.

[19] A. Almuhareb and M. Poesio, "Attribute-based and value-based clustering: An evaluation," in *Proc. EMNLP*, 2004, pp. 1–8.

[20] R. Jensi and D. G. W. Jiji. (2014). "A Survey on optimization approaches to text document clustering." [Online]. Available: https://arxiv.rg/abs/1401.2229
[21] S.-S. Hong, W. Lee, and M.-M. Han, "The feature selection method based on genetic algorithm for efficient of text clustering and text classification," *Int. J. Adv. Soft Comput. Appl.*, vol. 7, no. 1, pp. 1–19, 2015.

ALAN DÍAZ-MANRÍQUEZ received the B.Sc. degree in electronics engineering from the Tecnológico Nacional de México Campus, Ciudad Victoria, Mexico, in 2007, and the M.Sc. and Ph.D. degrees from the Information Technology Laboratory, Cinvestav Tamaulipas, Ciudad Victoria, in 2009 and 2014, respectively. His research interests include evolutionary computation and multiobjective optimization.

ANA BERTHA RÍOS-ALVARADO received the Ph.D. degree in computer science from Cinvestav Tamaulipas in 2013. She is currently a full-time Research Professor with the Faculty of Engineering and Sciences, Universidad Autónoma de Tamaulipas, México. Her research interests include text mining, semantic Web, and knowledge representation and management.

JOSÉ HUGO BARRÓN-ZAMBRANO received the master's degree in computer science information from the National Institute for Astrophysics Optics and Electronics in 2008 and the Ph.D. degree from the Computer Science Information Technology Laboratory, Center for Research and Advanced Studies, CINVESTAV, in 2014. He has more than several years of experience in designing dedicated parallel hardware architectures using the FPGAs. He is currently a full-time Professor with the Universidad Autónoma de Tamaulipas, Ciudad Victoria, México. He has published several papers in international conferences, journals, and book chapters. His main research interests include reconfigurable computing and the computational applications of FPGA devices in different domains, such as computer vision, digital signal processing, and neural computing. He has served as a regular reviewer for international conferences and peer review journals.

TANIA YUKARY GUERRERO-MELENDEZ received the B.S. degree in telematics engineering from the Autonomous University of Tamaulipas, Mexico, in 2002, and the master's degree from the Polytechnic University of Catalonia, Barcelona, Spain, in 2007, where she is currently pursuing the Ph.D. degree in telematics engineering. She is a full-time Professor with the Autonomous University of Tamaulipas. Her current research interest includes business process management, semantic Web, and telematics services.

JUAN CARLOS ELIZONDO-LEAL received the B.Sc. degree in electronics engineering from the Tecnológico Nacional de México Campus, Ciudad Victoria, México, in 2006, and the M.Sc. and Ph.D. degrees from the Information Technology Laboratory, Cinvestav Tamaulipas, Ciudad Victoria, México, in 2008 and 2013, respectively. His research interests include the robot path planning and multi-robot coordination.