

Based Systems

Elsevier Editorial System(tm) for Knowledge-

Manuscript Draft

Manuscript Number:

Title: MLP training approaches using similarity-learning for combined terms in Court Text Documents

Article Type: Full Length Article

Keywords: similarity-learning; knowledge management; big data; artificial intelligence; MLP neural networks

Corresponding Author: Mr. Antonio Pires de Castro Júnior, M.D.

Corresponding Author's Institution: Electrical, Mechanical & Computer Engineering School, Federal University of Goias, Goiania, Goias, Brazil

First Author: Antonio Pires de Castro Júnior, M.D.

Order of Authors: Antonio Pires de Castro Júnior, M.D.; Wesley P Calixto; Antonio P Coimbra; Gabriel A Wainer; Viviane M Gomes; Lais F Silva

Abstract: Learning using similarity for combined terms in text document is applied to generate knowledge in real court documents. These combinations of terms found in the corpus are used to train neural networks. Two computational representations are generated from this learning, being used to train the MLP neural networks: binary and frequency. This article compares these two computational forms and presents the one with the best accuracy to predict new lawsuit documents. Although the method can be used in many areas of knowledge, these work is being applied in court documents, judgments, verdicts and pleadings. The method can provide celerity in the judgments, solving the yearning of society as soon as possible, simulating the judicial advisers work on judicial verdicts preparation.

## Cover letter

January 28, 2020

Editorial Department of Knowledge-Based Systems

Dear Editor,

I am submitting a manuscript for consideration of publication in Knowledge-Based Systems Journal. The manuscript is entitled “MLP training approaches using similarity-learning for combined terms in Court Text Documents”.

The automatically similarity-learning technique are used to generate knowledge of a real court corpus. Knowledge is generated using combination of terms, being identified by similarity technique in each document. MLP Neural Networks are applied for court text document categorization. Binary and frequency approaches are compared to train the MLP Neural Networks.

As the method obtained 92% accuracy in the simulations, it is believed that the proposed method meets the objective of the work.

Although the proposed method has been applied in the real court of justice, this work can be used in several branches of knowledge that have large documents volumes and need to automate the process of knowing and establishing intelligent and automatic relations with new data inputs.

It has not been published elsewhere and that it has not been

submitted simultaneously for publication elsewhere.

Thank you very much for your consideration.

Yours Sincerely,

Prof. Antônio Pires de Castro Júnior,

Electrical, Mechanical & Computer Engineering School, Federal University of Goiás,

Goiania, Goiás, Brazil

No. 1488, Universitário Avenue, Universitário

Tel.: +55-62-999776786

E-mail: [apcastrojr@gmail.com](mailto:apcastrojr@gmail.com)

The automatically similarity-learning technique are used to generate knowledge of a real court *corpus*.

Knowledge is generated using combination of terms, being identified by similarity technique in each document.

MLP Neural Networks are applied for Court text document categorization.

Binary and frequency approaches are compared to train the MLP Neural Networks.

The method can be used in many areas of knowledge.

# MLP training approaches using similarity-learning for combined terms in Court Text Documents

Antonio P. de Castro Junior<sup>a,b,e</sup>, Viviane M. Gomes<sup>a,b,d</sup>, Lais F. A. Silva<sup>a</sup>,  
Antonio Paulo Coimbra<sup>c</sup>, Gabriel A. Wainer<sup>d</sup>, Wesley P. Calixto<sup>a,b,c,d,f</sup>

<sup>a</sup>*Experimental & Technological Research and Study Group, Federal Institute of Goias,  
Goiania, Goias, Brazil*

<sup>b</sup>*Electrical, Mechanical & Computer Engineering School, Federal University of Goias,  
Goiania, Goias, Brazil*

<sup>c</sup>*Institute of Systems and Robotics, University of Coimbra, Coimbra, Portugal*

<sup>d</sup>*Visualization, Simulation and Modeling, Carleton University, Ottawa, Canada*

<sup>e</sup>*Corresponding Author e-mail: apcastrojr@gmail.com*

<sup>f</sup>*Corresponding Author e-mail: wpcalixto@pq.cnpq.br*

---

## Abstract

Learning using similarity for combined terms in text document is applied to generate knowledge in real court documents. These combinations of terms found in the corpus are used to train neural networks. Two computational representations are generated from this learning, being used to train the MLP neural networks: binary and frequency. This article compares these two computational forms and presents the one with the best accuracy to predict new lawsuit documents. Although the method can be used in many areas of knowledge, these work is being applied in court documents, judgments, verdicts and pleadings. The method can provide celerity in the judgments, solving the yearning of society as soon as possible, simulating the judicial advisers work on judicial verdicts preparation.

*Keywords:* artificial intelligence, big data, similarity-learning, artificial neural networks, data mining, knowledge management.

---

## 1. Introduction

There are several areas of expertise working on document classification using machine learning, such as: i) medicine [1], ii) biology [2], iii) engineering [3], iv) law [4], v) education [5] among others. Information retrieval inno-

vations are using techniques to define document type in a corpus, with automatic knowledge generation and automatic information handling [3]. Many studies are moving in this direction, like: i) [6]; ii) [7]; iii) [8]; iv) [9]; v) [4] continue the work in [9]; vi) [3] and vii) [10].

Some studies classify documents using bag-of-words [11] [12] [13] [14], others bag-of-concepts [15] [16] [17] and others applying both solutions [10]. However, it is noticeable that each scenario or area of expertise may have characteristics in its document corpus that do not require much processing exaggeration. Therefore, it is observed that simple solutions, with shorter processing time, can achieve the objective in the specialist demands.

This work presents an artificial intelligence method that performs the recognition of document patterns using similarity-learning to identify documents by their combined terms. The proposed method was applied in the real *corpus* of lawsuits documents.

Although most of the papers report the difficulty in applying techniques to automatically generate knowledge of a database of documents, this study presents a similarity method that allows construct the knowledge of a *corpus* automatically. The knowledge obtained by the similarity-learning are inserted in an artificial neural network, enabling the making of taxonomic forecasts for new documents.

This paper's goal is to present MLP training approaches using similarity-learning for combined terms in court text documents. One of the training approaches allowed neural networks to reach 92% in predictive accuracy.

This paper contains the following structure: Sections 2 and 3 describes the theoretical basis of this work. Sections 4 and 5 details the proposed methodology, Section 6 and Section 7 present the results obtained and the brief discussion of the work, respectively, while the conclusions are provided in Section 8.

## **2. Environment for case study: data from the Judiciary**

According to the Brazilian National Council of Justice, the judiciary in Brazil ends the year 2017 with more than 80 million lawsuits in progress [18]. Figure 1 shows the lawsuit stock increased by 0.3 million (0.4%) in relation to the previous year and it's increasing year by year.

In Figure 1 shows that the judiciary cannot reduce the case currently pending judgment. The judiciary needs to increase the number of judges and their advisers, or construct a software tool capable of speeding up the

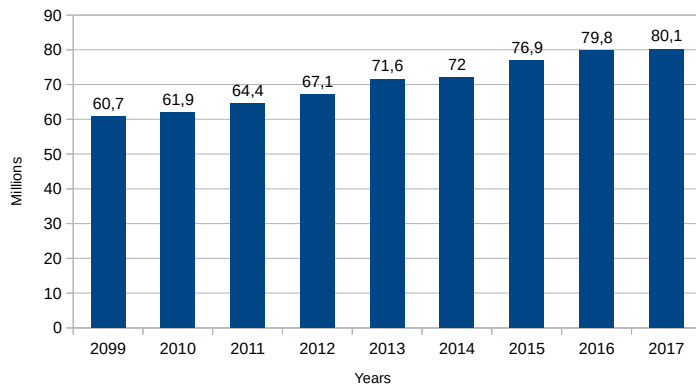


Figure 1: Brazilian judicial branch’s historical lawsuit series.

procedures for judging and filing the lawsuits. This work is trying to use artificial intelligence software to help in this scenario.

### 3. Quantitative model in information retrieve

Information retrieval is responsible for handling and retrieving data objects such as text, images, sounds, and so on. [19], [20], [21] and others describes and evolves information retrieval (IR) making it more sophisticated and interactivity.

The objective of IR is to find and present the correct information, from the contents of the document to the user, satisfying their need in the search expression. The search engine is the most relevant points in the process of retrieving the information, it compares the *query* of the users/systems with *corpus*. Most of the search engines are quantitative in nature, based on disciplines such as: i) logic, ii) statistics, and iii) set theory. Boughanem *et al.* [22] state that quantitative models have boosted the development of information retrieval systems, including: i) Booleans, ii) Vector, iii) Probabilistic, and iv) clustering. The efficiency of the IR system is directly linked to the applied model. These IR models are also known as search engines.

The quantitative models used in information retrieval can associate weights both in terms of indexing and in terms of the search expression. These weights are used to calculate the degree of similarity between search expressions established by the user for each document, or between documents.

Thus, it is possible to obtain documents ordered by degree of similarity based on search expression [23], [24].

The term is the word that represents the concept or meaning present in the document and to identify the relevance of the term in the description of the content of the document is an onerous task [23], [24]. The calculation of the weight is an important aspect and it can be applied in several ways like described by [25], [26], [23], [24] and [26].

The *corpus* is commonly represented by matrix with various documents and indexing terms. In the matrix, you can retrieve the information by calculating the similarity, where the intention is to quantify the similarity of content between two documents or between the search expression and each of the documents of the *corpus*. Some of the traditional models for calculating similarity are: i) Jaccard model; ii) the cosine model; iii) coefficient of Dice and iv) other [27].

The Jaccard  $S_{\vartheta}$  similarity is metric used in statistics and returns values in the range [0 1], being indicated to compare significant volumes of documents and terms, as in the case of *Big Data*. Jaccard's expression measures the similarity relation between documents  $D_1$  and  $D_2$  and is given by:

$$S_{\vartheta}(D_1, D_2) = \frac{\sum_{j=1}^n (w_{1,j} \cdot w_{2,j})}{\sum_{j=1}^n (w_{1,j})^2 + \sum_{j=1}^n (w_{2,j})^2 - \sum_{j=1}^n (w_{1,j} \cdot w_{2,j})} \quad (1)$$

where  $w_{1,j}$  is the weight of the  $j$ -th term of document  $D_1$  and  $w_{2,j}$  is the weight of the  $j$ -th term of document  $D_2$ . The result of the expression (1) is presented in percentage of similarity of document  $D_1$  with document  $D_2$ .

#### 4. Similarity-learning for combined terms

A different way of applying the similarity in 1 is presented in this article to generates knowledge and identity to the *corpus*. The similarity technique will not be used to compare documents,  $D_1$  and  $D_2$ , or compare search expressions and documents. The similarity technique will be used to find similarities between terms,  $t_1$  and  $t_2$ , in the documents database. The Jaccard expression 1 is altered to construct the relationship between the terms or combined terms. The altered Jaccard expression  $S_{\alpha}$ , is given by (2).



$$S_\alpha(t_1, t_2) = \frac{\sum_{i=1}^m (w_{i,1} \cdot w_{i,2})}{\sum_{i=1}^m [(w_{i,1})^2 + (w_{i,2})^2] - \sum_{i=1}^m (w_{i,1} \cdot w_{i,2})} \quad (2)$$

In (2), the similarity between the terms  $t_1$  and  $t_2$  is calculated, where  $t_1$  is the first term and  $t_2$  is the second term,  $w_{i,1}$  is the frequency of the term  $t_1$  in the  $i$ th document, and  $w_{i,2}$  is the frequency of the term  $t_2$  in the  $i$ th document. This process is repeated by calculating the similarity between all  $n$  terms  $t_j = t_1, t_2, \dots, t_n$ .

After calculating combinations in all  $n$  terms and in all  $m$  documents, it is possible to infer the similarity relation of the terms in the *corpus* of certain taxonomy, as set out in Table 1, which considers the relationship is automatically constructed between the terms  $t_1$  and  $t_2$ . The expression (2) is constructed to calculate the similarity between two terms ( $2 \times 2$ ), but this expression can be generalized to perform the calculation by combining several terms ( $n \times n$ ).

Table 1: The relationship between terms built automatically, given by  $S_\alpha$  (2) .

	$t_1$	$t_2$	$t_3$	$\dots$	$t_n$
$t_1$	-	$S_\alpha(t_1, t_2)$	$S_\alpha(t_1, t_3)$	$\dots$	$S_\alpha(t_1, t_n)$
$t_2$	$S_\alpha(t_2, t_1)$	-	$S_\alpha(t_2, t_3)$	$\dots$	$S_\alpha(t_2, t_n)$
$t_3$	$S_\alpha(t_3, t_1)$	$S_\alpha(t_3, t_2)$	-	$\dots$	$S_\alpha(t_3, t_n)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$t_n$	$S_\alpha(t_n, t_1)$	$S_\alpha(t_n, t_2)$	$S_\alpha(t_n, t_3)$	$\dots$	-

The Table 1 is the representation of the similarity matrix. It is observed that this matrix has the main diagonal null and that the values above the main diagonal are identical to the values below the main diagonal. Figure 2 illustrates the Algorithm of the proposed model. The result of the application of this methodology creates the relationship between terms, generating the digital fingerprint of the *corpus* for a certain lawsuit taxonomy. This same computational procedure can be implemented for any type of *corpus*, constructing the understanding in the relations between the terms and the documents. This constructed fingerprint is used to train neural networks.

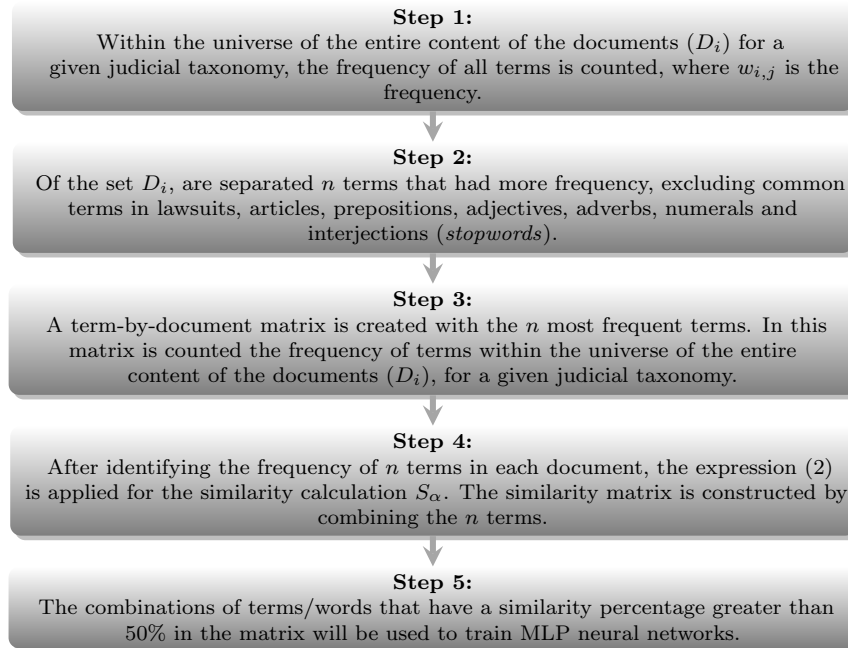


Figure 2: Algorithm used to construct the corpus identity.

## 5. MLP technology

After establishing the knowledge of the *corpus*, vectors of combined terms found by the percentage of similarity given by (2) are used to train a Multi-layer Perceptron (MLP) with backpropagation neural network. With supervised training, MLP can predict the classification of new lawsuits documents.

The algorithm showed in the Figure 2 is used in the Big Data of court decisions of a given taxonomy, generating the vectors of most frequent combination of terms. Since there may be repeated combined terms in different taxonomies or classes of lawsuits documents, it is understood that it is a nonlinearity problem.

The activation function used in the MLP is logistic expression. MLP is made up of 5 hidden layers, each layer with at least 4 neurons.

## 6. Results

Real data from the Judiciary is used, being imported for simulations one hundred thousand documents. Five lawsuits classes/taxonomies were chosen

to carry out the experiments: i) contract; ii) possessory; iii) family/divorce; iv) public pension and v) security mandate.

Programs constructed with Ruby-on-Rails language is used to similarity-learning and programs constructed with R language is used to machine learning.

### 6.1. Applying the similarity-learning combined terms

Documents of contract taxonomy are separated and applied steps 1 and 2, as shows in Algorithm in Figure 2. Ten terms with higher frequency in documents are separated, disregarding common terms in lawsuits and others *stopwords*.

With the terms of highest incidence in *corpus*, a matrix of term is created (step 3 in Figure 2). With this weights created by the expression in (2), the term to term matrix is structured (step 4 in Figure 2), as showed in Table 2 and in Table 3.

Table 2: Court decisions indexations terms - similarity matrix 1/2.

	<i>interest</i>	<i>contract</i>	<i>review</i>	<i>collection</i>	<i>payment</i>
<i>interest</i>	-	<b>0.51</b>	0.04	0.41	0.16
<i>contract</i>	<b>0.51</b>	-	0.13	<b>0.56</b>	0.40
<i>review</i>	0.04	0.13	-	<b>0.56</b>	<b>0.54</b>
<i>collection</i>	0.41	<b>0.56</b>	<b>0.56</b>	-	0.30
<i>payment</i>	0.16	0.40	<b>0.54</b>	0.30	-
<i>capitalization</i>	0.39	<b>0.57</b>	<b>0.51</b>	<b>0.57</b>	0.34
<i>consumer</i>	0.32	0.48	<b>0.58</b>	<b>0.68</b>	<b>0.54</b>
<i>finer</i>	0.17	0.32	0.21	0.37	0.42
<i>installments</i>	0.08	0.22	0.21	0.16	0.37
<i>credit</i>	0.15	0.32	0.12	0.34	0.44

Table 3: Court decisions indexations terms - similarity matrix 2/2.

	<i>capitalization</i>	<i>consumer</i>	<i>finer</i>	<i>installments</i>	<i>credit</i>
<i>interest</i>	0.39	0.32	0.17	0.08	0.15
<i>contract</i>	<b>0.57</b>	0.48	0.32	0.22	0.32
<i>review</i>	<b>0.51</b>	<b>0.58</b>	0.21	0.21	0.12
<i>collection</i>	<b>0.57</b>	<b>0.68</b>	0.37	0.16	0.34
<i>payment</i>	0.33	<b>0.54</b>	0.42	0.37	0.44
<i>capitalization</i>	-	<b>0.61</b>	0.41	0.20	0.35
<i>consumer</i>	<b>0.61</b>	-	0.45	0.20	0.43
<i>finer</i>	0.41	0.45	-	0.30	0.34
<i>installments</i>	0.20	0.20	0.30	-	0.32
<i>credit</i>	0.35	0.43	0.34	0.32	-

With the automatic relation between terms created, the combined terms with similarity percentage  $\geq 50\%$  are used to identify contract taxonomy documents (step 5).

The same process is applied to others taxonomies: i) possessory, ii) family/divorce, iii) public pension and iv) security mandate. The complete result of the process to find the combined terms is stamped in Figure 3.

### *6.2. MLP training approaches using similarity-learning for combined terms*

In order for MLP neural networks to correctly classify court documents, you must first train them with the combined terms found, shown in Figure 3. Two approaches were evaluated for MLP training: binary and frequency. The objective is to verify which approach is best suited for training, so that when the MLP neural network is in production mode it can classify lawsuits documents more accurately.

In this scenario, the court document is represented by a vector, each field in a vector is represented by a combined term. Then, the combined terms shown in the Figure 3 are transformed in vectors.

Thus, we have two vector approaches: i) binary vector and ii) frequency vector. In the binary vector approach, if the court document has the combined terms in your content, the field in a vector receives 1 (one), otherwise 0 (zero). In the frequency vector approach, if the court document has combined terms in your content, the value of the lowest frequency of one of the double terms is inserted in the field in a vector, otherwise 0 (zero). Then, vector is the fingerprint of one court document.

This vector is the input to the MLP neural network. The number of fields in a vector is the same number of combined terms discovery by similarity-learning applied in subsection 6.1, shown in Figure 3.

The MLP with backpropagation has 46 inputs, plus the bias, 5 hidden layers, each layer with at least 4 neurons, and only one output. Since the activation function used is logistic expression, the result of the MLP are between 0 and 1. The neural network has been trained to show the following ratings *tags*: 1 for contract; 0.75 for possessory; 0.5 for family/divorce; 0.25 for public pension and 0 for security mandate, Table 4.

The results should be: between 0.875 and 1.125 is contract; between 0.625 and 0.874 is possessory; between 0.375 and 0.624 is family; between 0.125 and 0.374 is public pension and between -0.125 and 0.124 is security mandate.

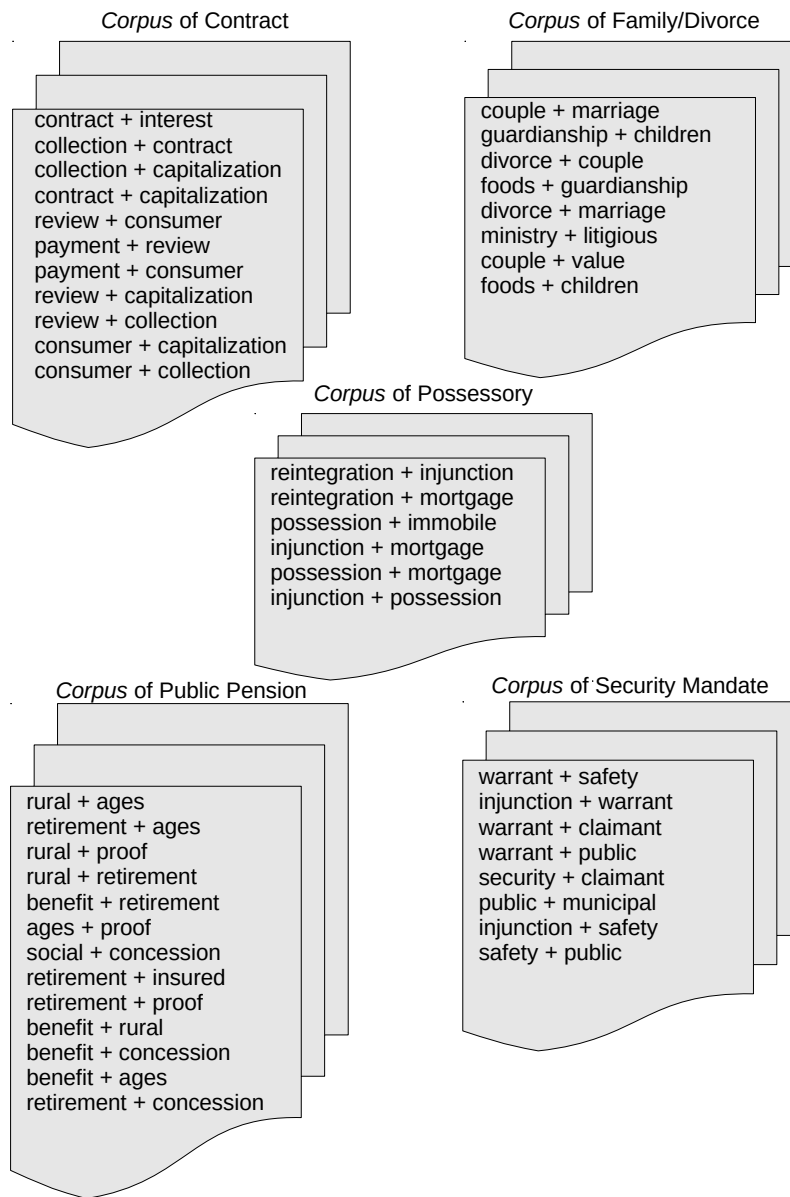


Figure 3: Knowledge generated by the automatic method presented in Subsection ??.

### 6.3. Simulations with real court documents

The neuralnet package was used in the R language to implement back-propagation MLP. The purpose of the simulation is to use the MLP neural

Table 4: Definition of the *tags* in the data set to represent the taxonomies of documents.

Taxonomies	Tags
Contracts	1
Possessory	0.75
Family/Divorce	0.5
Public Pension	0.25
Security Mandate	0

network, trained by similarity-learning of the combined terms, to automatically classify the new lawsuits, after reading the entire content of the complaint. Thus, 25 real court documents were used, five for each taxonomy.

Figure 4 shows the results of the simulations of the MLP neural network trained using binary vector and Figure 5 shows the results of the simulations of the MLP neural network trained using frequency vector. The dots on each line, in both Figures, show the MLP outputs for each taxonomy of the court document.

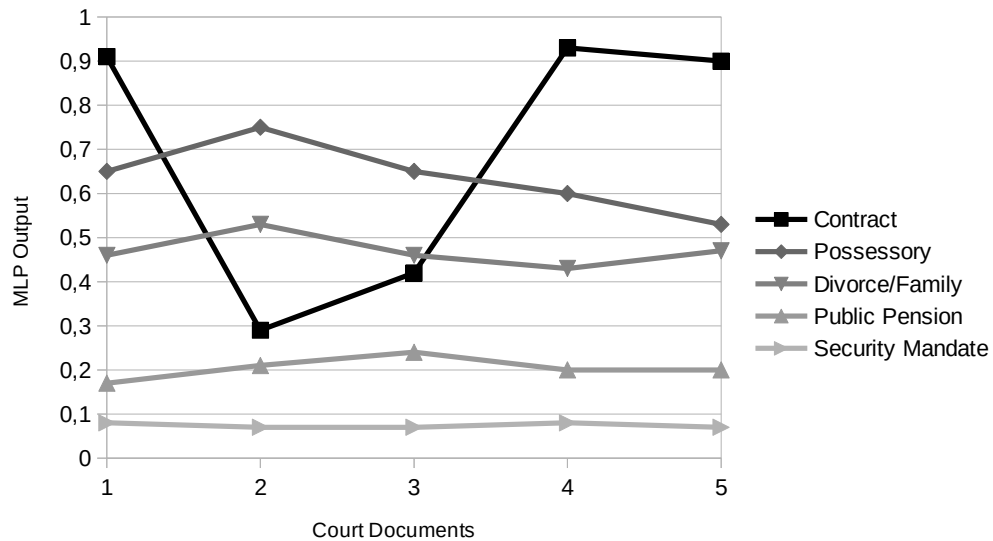


Figure 4: MLP outputs for each court document after training using binary vector. Each line presents the results for five document taxonomies.

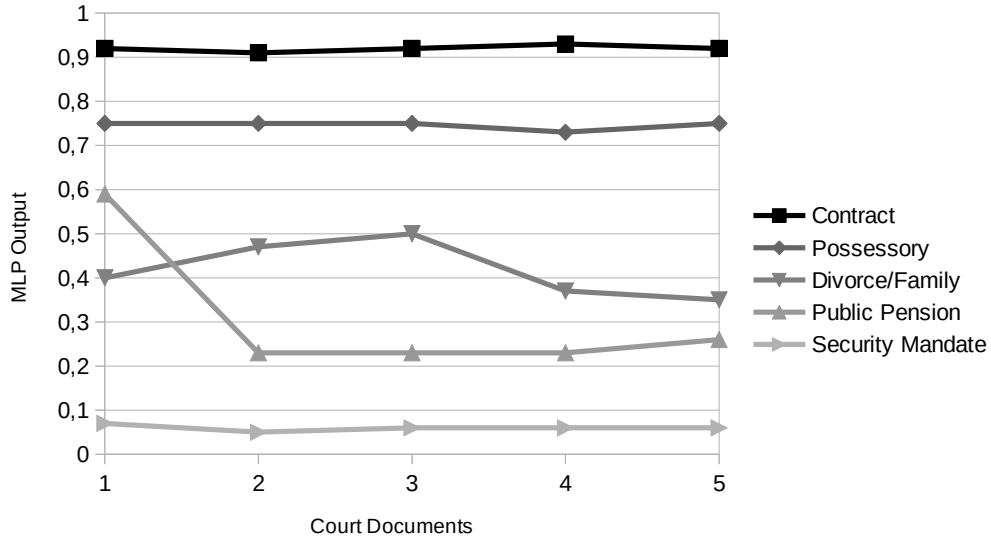


Figure 5: MLP outputs for each court document after training using frequency vector. Each line presents the results for five document taxonomies.

## 7. Discussion

Simulations have shown that the application of altered Jaccard expression (2), given by  $S_\alpha$ , to find similarity in combined terms is able to recognition court documents and can be applied together with MLP neural networks in order to predict the correct classification of new documents.

In scenario using binary vector training, the method was able to correctly classify 21 documents of the 25 loaded in the simulator, accuracy of 84%. In Figure 4 it is possible to see that MLP misclassified two court documents for contract taxonomy and two court documents for possessory taxonomy.

In scenario using frequency vector training, the method was able to correctly classify 23 documents of the 25 loaded in the simulator, accuracy of 92%. In Figure 5 it is possible to see that MLP misclassified one court document for divorce/family taxonomy and one court document for public pension taxonomy.

Both scenarios used were satisfactory for the proposed objective. However, in the tests performed, frequency vector-trained MLP showed better prediction accuracy than binary vector-trained MLP. The results presented

are indicative of the potential of the propose method.

The combination of the terms founded by calculation of similarity, given by  $S_\alpha$  in (2), was performed two by two ( $2 \times 2$ ), this combination can be modified in order to perform the calculation with larger combinations of terms and producing more accurate results.

## 8. Conclusion

The obtained knowledge by altered Jaccard, in the expression (2), was inserted in the artificial neural network, aiming to make correct predictions to establish the classification of the lawsuits. With the accuracy of 92%, by training using frequency vector, it is concluded that the proposed method meets the objective of this work, which is to establish methodology to generate automatic knowledge and to indicate the correct taxonomy of new court documents.

Among the several contributions of this work, the main ones are: i) knowledge obtained by altered Jaccard, in the expression (2) and in the proposed scenario, present the best way to train an MLP, using binary vector or frequency vector, to predict new court documents.

The method is still under construction, undergoing improvements and refinements. Studies are under way to use Wordnet database to match the combined terms founded with their respective synonyms and their meanings. This is expected to increase the ability to find documents related to the already known taxonomy.

Another point of interest is the possibility of finding and creating new classifications, in cases where the documents do not fit in any of the known classes or that have scores of ties between two or more taxonomies. In these situations the solution can automatically create other taxonomies of document and start the knowledge process. However, the problem is in defining the value that allows identifying when the document does not fit into any of the established *tags* taxonomies.

The method proposed in this work can be applied in several branches of knowledge that have large data volumes and need to automate the process of knowing and establishing intelligent and automatic relations with new data inputs.



## References

- [1] O. Arsene, I. Dumitrache, I. Miha, Medicine expert system dynamic bayesian network and ontology based, *Expert Systems with Applications* 38 (2011) 15253–15261.
- [2] J.-B. Lamy, Owlready: Ontology-oriented programming in python with automatic classification and high level constructs for biomedical ontologies, *Artificial intelligence in medicine* 80 (2017) 11–28.
- [3] M. Rani, A. K. Dhar, O. Vyas, Semi-automatic terminology ontology learning based on topic modeling, *Engineering Applications of Artificial Intelligence* 63 (2017) 108–125.
- [4] N. Zhang, Y.-F. Pu, S.-Q. Yang, J.-L. Zhou, J.-K. Gao, An ontological chinese legal consultation system, *IEEE Access* 5 (2017) 18250–18261.
- [5] A. Grubišić, S. Stankov, I. Peraić, Ontology based approach to bayesian student model design, *Expert systems with applications* 40 (2013) 5363–5371.
- [6] M. Ceci, A. Gangemi, An owl ontology library representing judicial interpretations, *Semantic Web* 7 (2016) 229–253.
- [7] B. Fawei, A. Wyner, J. Z. Pan, M. Kollingbaum, Using legal ontologies with rules for legal textual entailment, in: *AI Approaches to the Complexity of Legal Systems*, Springer, 2015, pp. 317–324.
- [8] M. A. Calambás, A. Ordóñez, A. Chacón, H. Ordoñez, Judicial precedents search supported by natural language processing and clustering, in: *2015 10th Computing Colombian Conference (10CCC)*, IEEE, pp. 372–377.
- [9] N. Zhang, P. Wang, Y. Pu, Challenges and related issues for building chinese legal ontology, in: *2015 International Conference on Mechatronics, Electronic, Industrial and Control Engineering (MEIC-15)*, Atlantis Press, pp. 1260–1265.
- [10] L. Huang, D. Milne, E. Frank, I. H. Witten, Learning a concept-based document similarity measure, *Journal of the American Society for Information Science and Technology*, ISSN 1532–2882 (2012).

- [11] V. Garg, S. Vempati, C. V. Jawahar, Bag of visual words: A soft clustering based exposition, Third National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (2011).
- [12] X. M. A. Bosch, A. Zisserman, Scene classification using a hybrid generative/discriminative approach, IEEE Trans. Pattern Analysis and Machine Intelligence, 30 (2008).
- [13] J. P. S. Lazebnik, C. Schmid, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories (2006).
- [14] A. Z. R. Fergus, P. Perona, Object class recognition by unsupervised scale-invariant learning (2003).
- [15] D. Milne, I. Witten, D. Nichols, A knowledge-based search engine powered by wikipedia, In Proceedings of the 16th Association for Computing Machinery (ACM) Conference on Information and Knowledge Management (2007).
- [16] R. Mihalcea, C. Corley, C. Strapparava, Corpus-based and knowledge-based measures of text semantic similarity, In Proceedings of the 21st National conference on Artificial Intelligence (2006).
- [17] E. Gabrilovich, S. Markovitch, Feature generation for text categorization using world knowledge, In Proceedings of the 19th International Joint Conference on Artificial Intelligence (2005).
- [18] C. L. A. Rocha, Justice in numbers: document produced by the brazilian judiciary (in portuguese: Justiça em números), Digital Magazine of the National Council of Justice (in Portuguese: Justiça em Números 2018 – CNJ) 1 (2018).
- [19] C. N. Mooers, Zatocoding applied to mechanical organization of knowledge, American documentation 2 (1951) 20–32.
- [20] M. B. Almeida, Revisiting ontologies: A necessary clarification, Journal of the American Society for Information Science and Technology 64 (2013) 1682–1693.
- [21] F. Coimbra Delicato, L. Pirmez, L. Fernando Rust da Costa Carmo, Fenix–personalized information filtering system for www pages, Internet Research 11 (2001) 42–48.

- [22] M. Boughanem, A. Brini, D. Dubois, Possibilistic networks for information retrieval, *International Journal of Approximate Reasoning* 50 (2009) 957–968.
- [23] J. M. Ponte, W. B. Croft, A language modeling approach to information retrieval, Ph.D. thesis, University of Massachusetts at Amherst, 1998.
- [24] J. M. Ponte, W. B. Croft, A language modeling approach to information retrieval, in: *ACM SIGIR Forum*, volume 51, ACM, pp. 202–208.
- [25] G. Salton, M. J. McGill, *Introduction to modern information retrieval*, McGraw-Hill, Inc., 1986.
- [26] G. Salton, *Automatic text processing: The transformation, analysis, and retrieval of*, Reading: Addison-Wesley (1989).
- [27] V. Thada, V. Jaglan, Comparison of jaccard, dice, cosine similarity coefficient to find best fitness value for web retrieved documents using genetic algorithm, *International Journal of Innovations in Engineering and Technology* 2 (2013) 202–205.

## **Acknowledgements**

The authors would like to thank National Council for Scientific and Technological Development (CNPq), Foundation for Research Support of the State of Goiás (FAPEG), Brazilian Federal Agency for Support and Evaluation of Graduate Education (CAPES) and Federation of the Industry of the State of Goiás (FIEG/SENAI/FATESG).

## **Additional information**

### **Competing interests**

The author(s) declare no competing interests.

## AUTHOR DECLARATION TEMPLATE

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.


We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us.

We confirm that we have given due consideration to the protection of intellectual property associated with this work and that there are no impediments to publication, including the timing of publication, with respect to intellectual property. In so doing we confirm that we have followed the regulations of our institutions concerning intellectual property.

We understand that the Corresponding Author is the sole contact for the Editorial process (including Editorial Manager and direct communications with the office). He is responsible for communicating with the other authors about progress, submissions of revisions and final approval of proofs. We confirm that we have provided a current, correct email address which is accessible by the Corresponding Author and which has been configured to accept email from [apcastrojr@gmail.com](mailto:apcastrojr@gmail.com).

Signed by all authors as follows:

Antonio P. de Castro Junior



---

Wesley Pacheco Calixto



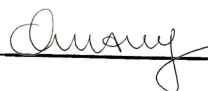
---

Viviane M. Gomes




---

Lais F. A. Silva




---

Antonio Paulo Coimbra



---

Gabriel A. Wainer



---