

Legal Information Retrieval Based on a Concept-Frequency Representation and Thesaurus

Wagner Miranda Costa and Glauco Vitor Pedrosa^a

Graduate Program in Applied Computing (PPCA), University of Brasilia (UnB), Brasilia, Brazil

Keywords: Information Retrieval, Data Mining, Natural Language Processing, Deep Learning, Decision Support System.

Abstract: The retrieval of legal information has become one of the main topics in the legal domain, which is characterized by a huge amount of digital documents with a peculiar language. This paper presents a novel approach, called BoLC-Th (Bag of Legal Concepts Based on Thesaurus), to represent legal texts based on the Bag-of-Concept (BoC) approach. The novel contribution of the BoLC-Th is to generate weighted histograms of concepts defined from the distance of the word to its respective similar term within a thesaurus. This approach allows to emphasize those words that have more significance for the context, thus generating more discriminative vectors. We performed experimental evaluations by comparing the proposed approach with the traditional Bag-of-Words (BoW), TF-IDF and BoC approaches, which are popular techniques for document representation. The proposed method obtained the best result among the evaluated techniques for retrieving judgments and jurisprudence documents. The BoLC-Th increased the mAP (mean Average Precision) compared to the traditional BoC approach, while being faster than the traditional BoW and TF-IDF representations. The proposed approach contributes to enrich a domain area with peculiar characteristics, providing a resource for retrieving textual information more accurately and quickly than other techniques based on natural language processing.


1 INTRODUCTION

The legal domain is characterized by a huge amount of digital documents, having a peculiar language and implicit structures. The large amount of digital has currently pointed out the need of more complex strategies for the management of the embedded unstructured knowledge in terms of analysis and extraction of relevant information. According to (Niebla Zatarain, 2018), computational processes changed the practice of law, specifically by connecting computational models of legal reasoning directly with legal text, generating arguments for and against particular outcomes.

An important task in the legal domain is the retrieval of similar information from a large and diverse dataset. This is a task that can be used to give access to the law to lay people and legal professionals, which increases the transparency and legitimacy of the processes (Sansone and Sperli, 2022). For example, using legal information retrieval it is possible to monitor changes in tax laws and

regulations or use Legal Reasoning in order to check compliance and/or validate the quality of legislative documents.

A fundamental and crucial step for the development of a textual information retrieval system is to extract a compact and discriminative vector representation of the documents. The Bag-of-Words (BoW) model, for example, is one of the most popular techniques for document representation and it is used in many text retrieval applications (Salim and Mustafa, 2022). This approach counts the word frequencies of a document and allows intuitive interpretability of the feature vector, but suffers from the curse of dimensionality and disregards the impact of semantically similar words. To increase the semantic representation of documents, we can encode words by using Word Embeddings techniques, such as the popular word2vec model (Le and Mikolov, 2014). This approach is based on deep neural networks and it encodes the contextual information of each word by capturing information around it. However, the vectors generated from word2vec are difficult to interpret as its values indicate the weight of the neural network used for training.

^a  <https://orcid.org/0000-0001-5573-6830>

To overcome the drawbacks of the BoW and Word Embeddings techniques, the Bag-of-Concepts (BoC) model (Kim et al., 2017) emerged as an alternative approach for document representation by clustering semantically similar words into a common “concept”. Concepts are created by clustering word vectors generated from Word Embedding techniques, such as word2vec (Mikolov et al., 2013) or wang2vec (Ling et al., 2015). Using the idea of “concepts”, the BoC allows to encode the impact of semantically similar words, enriching the representation of a textual document without increasing the dimensionality of the feature vector while offering intuitive interpretability behind the generated document vectors.

In this paper, we introduce the BoLC-Th (Bag of Legal Concepts Based on Thesaurus) to represent legal documents, complying with the vocabulary of legal texts in the BoC representation. The goal of the proposed approach is to emphasize those words/terms that are directly related to the peculiar language of legal documents. For this purpose, the BoLC-Th uses a thesaurus, which is a type of controlled vocabulary whose relationships between its terms are identified through standardized indicators and which are mutually employed. In the proposed approach, each word will have different weight in the final document representation. This weight is based on the distance from the word to the closest term in the thesaurus: the greater the distance the less the weight. In this way, it is possible to represent documents in a more semantically effective way, complying with the domain context.

We evaluated and compared the proposed method with other text representation approaches using a case study with real data. We used a database of jurisprudence texts from the Federal Court of Accounts (TCU) from Brazil, which is the Brazilian federal accountability office responsible for federal public funds as well as the accounts of any person that causes loss to the public treasury. With the help of experts from TCU, we carried out experiments to retrieve similar jurisprudence to a given judgment provided by the users (auditors).

This paper is structured as follows: Section 2 presents related works; Section 3 presents the steps of the proposed method; Section 4 shows the results obtained from an empirical experiment carried out with real data; and Section 5 presents our conclusions.

2 RELATED WORKS

The information retrieval deals with the development of algorithms and models to retrieve information from a large and diverse document repository (Yan, 2009). From a computational point of view, for the implementation of an information retrieval system, first the documents must be represented in a way that the computer can interpret each document within the collection. This representation aims to numerically represent unstructured text documents to make them mathematically computable. In other words, for a given set of text documents $D = \{d_1, d_2, d_3, \dots, d_n\}$, where each d_i represents a document, the textual representation problem is to represent each d_i of D as a point s_i in a feature space S , where the distance between each pair of points in the feature space S is well defined.

To transform textual data into a numerical vector, it is necessary to use a document representation, which can be divided into two categories: word frequency-based and embedding prediction-based (Analytics Vidhya, 2017). In the first group, the document vectors is based on counting words, in the second group, the generation of vectors is supported by neural network mechanisms, which result in non-deterministic numerical representations. Figure 1 shows how these techniques are used in this study. In the following, we discuss the concepts and the related works that motivated the development of this work.

2.1 Bag-of-Words (BoW) Representation

A commonly adopted and effective approach to document representation is the Bag-of-Words model. The BoW model assigns a vector $d = \{x_1, x_2, x_3, \dots, x_l\}$ to a document d , where x_i denotes the normalized number of occurrences of the i -th term (word) in the document, and l is the size of the collection of terms (words) of documents present in the database.

The BoW approach is a simple but effective method for mapping a document to a fixed-length vector. However, the mapping function in the BoW model is hard (or binary), as it only represents the presence or not of a term (word) in the document. The hard mapping function has several limitations: first, the vector generated for each document is extremely sparse, as a document contains only a very small portion of all the basic terms in a database. Second, BoW representations may not effectively capture the semantics of documents, as semantically

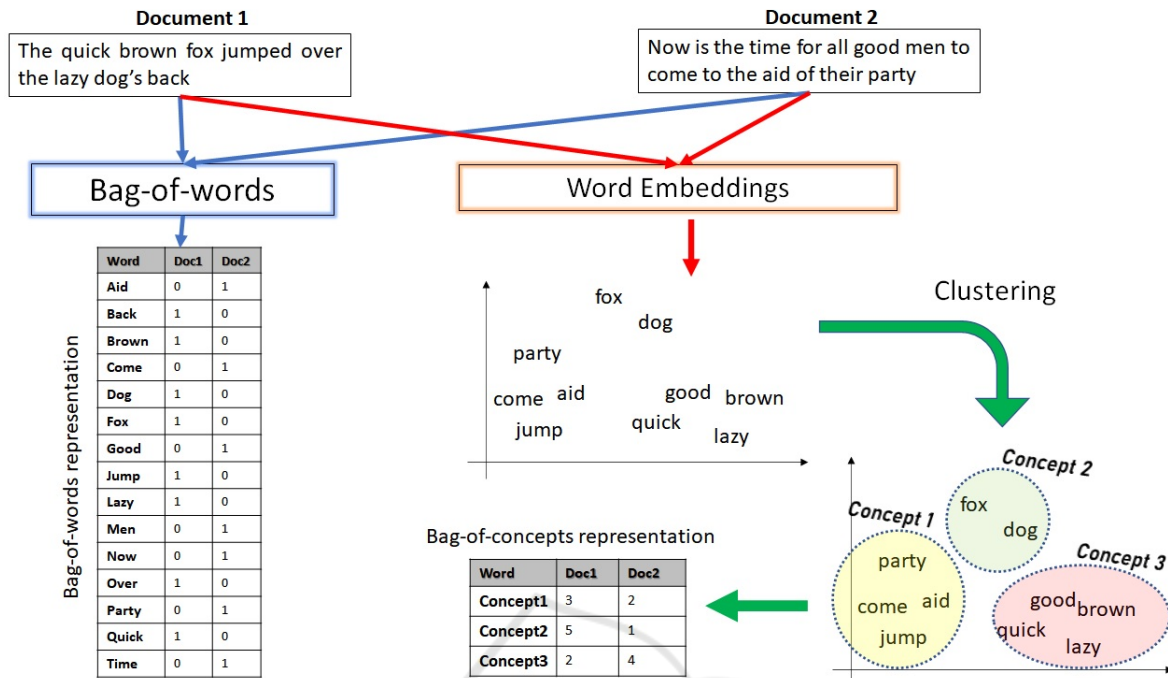


Figure 1: Overview of different modeling-based approaches for textual document representation. The blue lines indicates the Bag-of-Words model, the red line indicates the word embedding techniques and the green lines the Bag-of-Concept representation.

similar documents with different sets of words will be mapped to very different vector spaces. To overcome these drawbacks, we can use word embedding technique, that will be discussed in the following.

2.2 Word Embeddings Techniques

It is natural to look for codings that account for semantic relationships between words (Turney, 2006). This leads to the creation of a so-called word embedding (also named as “semantic vector space” or simply “word space”), i.e., a continuous vector space in which the relationships among the vectors is somehow related to the semantic similarity of the words they represent. The goal of word embedding is to capture semantic and syntactic regularities in language from large unsupervised sets of documents, such as Wikipedia. Words that occur in the same context are represented by vectors in close proximity to each other. Fig. 2 shows an example of such embedded space: in this figure, we have embedded words that represent the names of sports car, companies and fruits. While the words with similar meanings are located closer to each other, the words with different meanings are located distant from each other.

The central idea behind word embedding

techniques is to assign a vector representation to each word, so that words that are semantically similar are close to each other in vector space. The merit of the word embedding approaches is that the semantic similarity between two words can be conveniently evaluated based on the measure of cosine similarity between the corresponding vector representations of the two words.

One of the most popular word embedding technique is the word2vec (Mikolov et al., 2013), which is based on a two-layer neural network. The word2vec framework contains two separate models,

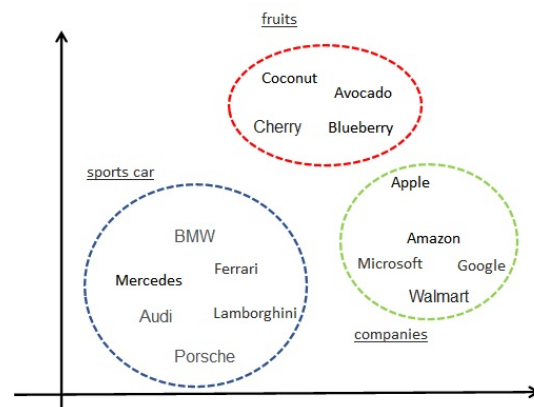


Figure 2: Example of a clustered embedded space.

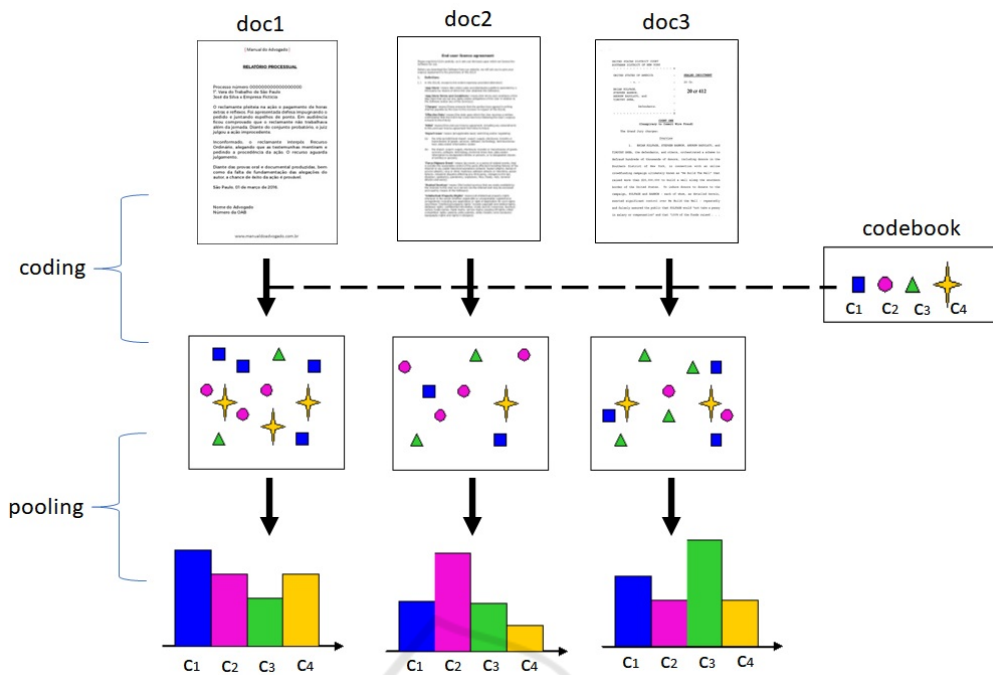


Figure 3: Overall framework of Bag-of-Concepts model.

including Continuous Bag-of-Words (CBoW) and Skip-gram with two reverse training objectives. CBoW tries to predict a word by considering the words around it, while Skip-gram tries to predict a window of words given a single word. Due to its architecture and unsupervised training, word2vec can be efficiently built on a large-scale unannotated corpus. Word2vec is able to encode meaningful linguistic relationships between words in embeddings of learned words.

Embedding at document level is used in (Renjit and Idicula, 2019) for the similarity assessment of legal texts. Using the paragraph vector (Le and Mikolov, 2014) algorithm, the authors consolidate the text feature vectors and estimate the remaining paragraphs with the stochastic gradient descent method. Then, the angular orientation of the documents is obtained, from the calculation of the cosine similarity of their vectors. Documents that have similar orientations are also considered semantically similar.

2.3 Bag-of-Concepts (BoC) Representation

In the Bag-of-Concepts approach, the words of a text are associated with their semantic representations, obtained by clustering the embedding space of its words. This grouping will generate a codebook of

concepts that will be used to represent the textual documents. Figure 3 shows the overall framework of this model representation, which is divided into two steps: coding and pooling. In the coding step, the words are numerically represented using some word embedding technique and, then, using a codebook of contexts, each word is assigned to a context. The pooling step consist of counting the frequency of each concept within the document.

The Bag-of-Concepts approach overcome some limitations of the BoW approach, such as the lack of semantic representation, high dimensionality and sparseness (Kim et al., 2017). The definition of “concept” as a unit of meaning was presented by (Mourino Garcia et al., 2015), who studied the use of knowledge databases based on the Wikipedia. As a result, a gain of up to 157% was obtained in the modeling of classifiers when compared to BoW.

2.3.1 Extended Models for Bag-of-Concepts

The BoC technique has been used, and sometimes extended, in order to improve the representation of textual document for classification and/or retrieval tasks. The work of (Li et al., 2020), for example, proposes the Bag-of-Concepts-Clusters (BoCCI) model, which groups semantically similar concepts and enhances the BoC representation with entity disambiguation. The problem of synonymy and polysemy is considered in (Wang

et al., 2014), which highlights the disadvantage in vector analysis of short texts, since even texts with similar meanings may not share the same words. (Shalaby and Zadrozny, 2019) uses knowledge bases (KB), such as Wikipedia and Probase, to support the conceptualization. Additionally, neural representations support the densification of BoC vectors to decrease their sparsity without, however, increasing their dimension. The authors indicate a 1.6% improvement in the correlation score and a 5% reduction in categorization errors, and they evaluated the proposed model from three perspectives:

1. Semantic relationship between entities, comparing it to Wikipedia Link-based Measure (WLM), Keyphrase Overlap RElatedness (KORE), Exclusivity-based Relatedness (ExRel) and Combined Information Content (CombIC) models
2. conceptual categorization, comparing it to word embeddings trained on Wikipedia, multiword embeddings trained on Wikipedia, and entity-category embeddings models
3. unsupervised classification of documents against Explicit Semantic Analysis (ESA), Word embeddings best match and Word Embeddings Hungarian algorithm

Instead of comparing documents individually, (Hematialam et al., 2021) seeks to find patterns of similarity between sets of documents of the same nature, such as between full texts and respective summaries, by analyzing the distortion of the graphs resulting from the calculation of the cosine distances of the *embeddings* vectors. This approach allows texts of different sizes and feature vectors of different dimensions to be semantically compared, with satisfactory results. A greater consistency in the semantic interpretability of the *features* of the texts is found in (Li et al., 2020), which in addition to grouping the concepts, uses probabilistic knowledge base (ProBase) and lexicon (WordNet) for disambiguation of entities. Something similar is done by (Rajabi et al., 2020), but in the disambiguation of the grouped concepts themselves.

A characterization method supported by domain-specific ontology (OCTC – *ontology-enabled concept-based text categorization*) is presented by (Lee et al., 2021). Here, a syntactically marked document is converted into a conceptual representation according to the frequency of descriptors of existing concepts in the ontology. This frequency defines the degree of relevance of the concept in a given document. The study states that the conceptual representation requires less computational resources

in its generation and that the accuracy of the OCTC classifier is superior to that of other techniques, such as *Latent Semantic Analysis* (LSA) and *Doc2Vec* combined with *Support Vector Machine* (SVM).

The study in (Lee, 2020) works on the treatment of homonymy by replacing the *word embedding* of the homonyms with their semantic meaning, from the grouping of *Embeddings from Language Models* (ELMo). Locating semantic relationships between terms – hyperonymy and meronymy, for example – was also the approach of (Mehanna and Mahmuddin, 2021), albeit in the context of sentiment analysis. With the use of semantic databases to mark aspects in BoC of texts, this method reached an increase of up to 70% in the accuracy when performing the same task through algorithms based on networks neural and Naive Bayes.

3 PROPOSED METHOD

Towards an effective representation of documents in the legal domain, in this paper we introduce a novel textual-based representation technique, called BoLC-Th (Bag of Legal Concepts Based on Thesaurus). The goal of the proposed method is to represent textual documents in weighted histograms of concepts. This histogram is the textual document representation that can be used to retrieve similar content-based documents. The BoLC-Th employs a vocabulary of words used in the analysis of legal documents to increase the discriminative power of the document representation. For this purpose, the idea of the proposed technique is to use a thesaurus (which has the terms with the words and their synonyms of the application domain).

Figure 4 shows the steps of the proposed approach, which is divided into two phases: (i) coding and (ii) pooling. The coding phase assigns a concept to each word of the textual document, and the pooling phase summarizes the frequency of each concept in the document. These two phases will be detailed as follows. However, first we need to describe the concept-based dictionary generation, which is used to assign a concept to each word in the document.

3.1 Concept-Based Dictionary Generation

Let $W = \{w_1, w_2, \dots, w_v\}$ be the vocabulary that encodes all the words for a given corpus, and v the size of that vocabulary. Each word $w_i \in W$ is represented by a r -dimensional vector obtained

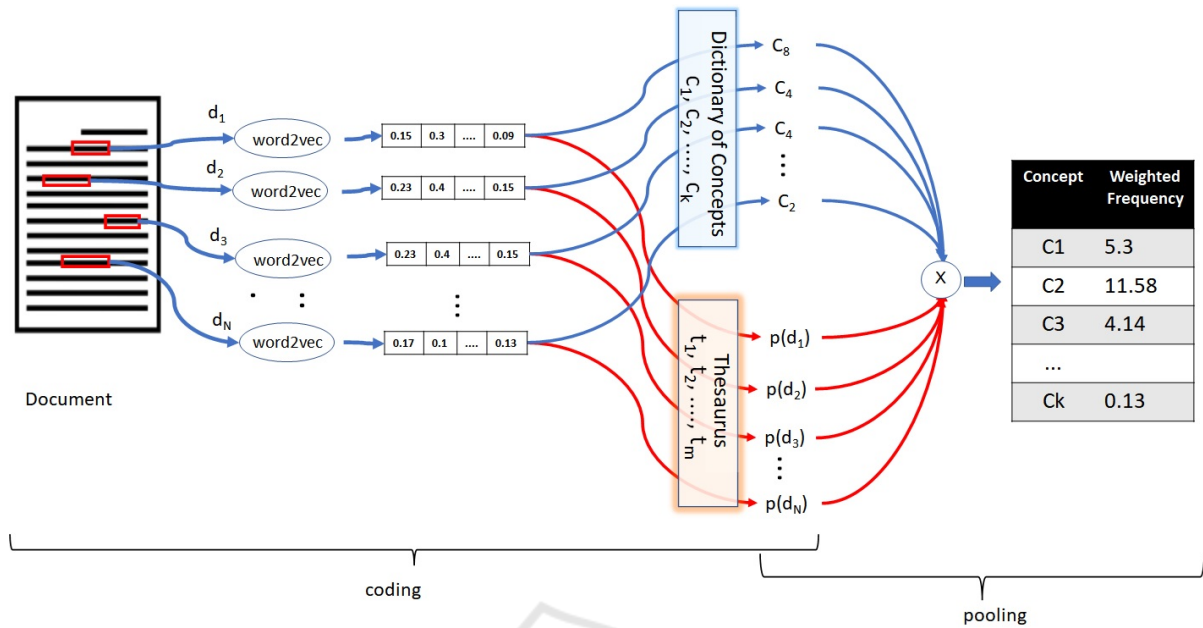


Figure 4: Steps of the proposed method to represent a document in a weighted histogram of concepts based on a thesaurus: the blue lines indicate the generation of concepts for each words in the document and the red lines indicate the weighted of each word according to their nearest similar word in the thesaurus. The final representation of the document is a weighted frequency of concepts.

using any word embedding technique, for example word2vec, generating a matrix $S \in \mathbb{R}^{v \times r}$. Let $T = \{t_1, t_2, \dots, t_m\}$ be a thesaurus with m words, where each t_i corresponds to a term of a specific domain of knowledge that, in the case of this work, involves the words and their synonyms used in the texts of legal documents.

To generated the concept-based dictionary, we need to run a clustering algorithm to cluster semantically similar words. In other words, let $C = \{c_1, c_2, \dots, c_k\}$ be this dictionary: each concept $c_i \in \mathbb{R}^r$ is the centroid of a group $C_i \subset \{S \cap T\}$, such that:

$$c_i = \frac{1}{|C_i|} \sum_{x_j \in \{S \cap T\}} x_j, \quad \forall i \in \{1, 2, \dots, k\} \quad (1)$$

where, $|C_i|$ is the number of words associated with the group C_i and k is the number of groups (concepts). Any clustering algorithm can be used to generate these clusters. However, the use of Spherical k-means is indicated by (Dhillon and Modha, 2001) for clustering sparse and high-dimensional vectors, which is the case of the data analyzed in this work .

The proposed approach for the generation of the dictionary of concepts differs from the traditional approach by considering only those words from the corpus that are present in the thesaurus. Our goal is

to restrict the generation of concepts to words strictly related to the application context.

3.2 Weighted Pooling of Concepts

Let $D = \{d_1, d_2, \dots, d_N\}$ be a textual document, where N is the number of words in that document and each word $d_i \in D$ is represented by a r -dimensional vector obtained by the same word embedding technique used to generate the concept-based dictionary. Each word $d_i \in D$ is associated with each of the existing k concepts in C using a function $\phi : \mathbb{R}^d \rightarrow \mathbb{N}^k$, defined as:

$$d_i \rightarrow \phi(d_i) = \{\alpha_1^i, \alpha_2^i, \dots, \alpha_k^i\} \quad (2)$$

where α_j^i can be seen as an activation function, associating the i -th word of the document to the j -th concept of the dictionary of concepts. This activation can be *hard* or *soft*. Classic coding is based on *hard* coding where each word d_i is assigned to only one concept, that is:

$$\alpha_j^i = \begin{cases} 1, & \text{if } j = \arg \min_{j \in \{1, \dots, k\}} \|d_i - c_j\|_2^2 \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

The innovative approach of this work consists of using a weighted concept-based pooling. The idea is to calculate the distance of each word $d_i \in D$ to its nearest similar word $t_r \in T$. In other words, the final

vector that will represent the document D will be a weighted histogram of concepts $H = [h_1, h_2, \dots, h_k]$, where k is the number of concepts in the concept dictionary and each $h_j \in H$ is defined as:

$$h_j = \sum_{i=1}^N p(d_i) \alpha_j^i, \quad \forall j \in \{1, 2, \dots, k\} \quad (4)$$

where,

$$p(d_i) = 1 - \arg \min_{t_r \in T} \|d_i - t_r\|_2^2 \quad (5)$$

The weighting function $p : \mathbb{R}^d \rightarrow \mathbb{R}$ returns the distance from the word d_i to its nearest similar word in T . If $d_i \in T$, then $p(d_i) = 1$ and, therefore, the concept associated with the word d_i will be fully accounted in the histogram. Otherwise, the greater the distance of word d_i is from the its nearest word in the thesaurus, the smaller the contribution of the concept of the word d_i in the final histogram.

4 EXPERIMENTAL RESULTS

The proposed method BoLC-Th was compared with three other traditional approaches for textual document representation: BoW, BoC and TF-IDF. The performance of each technique was evaluated on a case study, involving the retrieval of similar judgments and jurisprudence of the Federal Court of Accounts (TCU), which is the Brazilian federal accountability office. It is tasked with assisting Congress in its Constitutional incumbency to exercise external audit over the Executive Branch.

4.1 Database and Ground-Truth

The database for carrying out the comparative experiments is composed of a set of 15,000 (fifteen thousand) statements of jurisprudence from the Federal Court of Auditors (TCU), available on its Portal¹. In the process of preparing case law, TCU experts group judgments according to the areas of activity of the External Control and consolidate the Court's understanding of a given subject. These texts are called *statements*. In turn, the judgments, which represent the decisions of the Court, have their texts summarized into *summaries*.

In order to establish a benchmark for our comparative tests, a ground truth was built to compare the results obtained. This base is composed by the selection of 10 (ten) summaries of judgments handed down by the TCU. For each of these

summaries, TCU experts manually indicated the utterances with the greatest semantic similarity. This set (summaries of judgments and their respective statements) constitutes the ground truth of this work.

4.2 Thesaurus

Thesaurus can be defined as a list of terms with their synonyms – or some other equivalent semantic relationship. Commonly, it deals with a domain of knowledge, which makes its use very specific. For the scope of this work, the TCU developed a Thesaurus with the aim of standardizing the terminology used in the Court's institutional activities, in addition to supporting the handling of information in organization. This Thesaurus is structured by government functions and has Subject, Entity and Locality descriptors. It is composed of 10,525 synonym-term tuples and Table 1 shows some sample of the Thesaurus used in this work.

Table 1: Examples (in Portuguese language) of the term-synonym list of the Thesaurus developed by TCU and used in this work.

<i>Term</i>	<i>Synonym</i>
Tributo	Obrigaç�o fiscal
Sentena	Decis�o judicial
Ren�ncia de receita	Ren�ncia tribut�ria
Governo eletr�nico	e-GOV
Patrim�nio mobili�rio	Bens patrimoniais m�veis
Incentivo fiscal	Gastos tribut�rios
Whitelist	Remetentes confi�veis
Preju�zo	Dano contratual
Capacitao	Qualificao profissional
Racismo	Discriminao racial

4.3 Parameter Settings

The generation of feature vectors and the execution of similarity queries were performed in the Python 3.8 programming language, with support from the scikit-learn, Gensim, Pandas, Matplotlib and NumPy libraries. The execution environment for this work was Jupyter Notebook 6.4.12, processed on a computer with the following configuration: Intel(R) Core(TM) i7-10610U CPU @1.80GHz 2.30GHz, 20GB RAM, Windows 10 Pro 64 Operating System -bit.

For the implementation of the BoC and BoLC-Th approaches, two parameters need to be defined:

1. the size of the concept-based dictionary
2. the size of the vector-of-words generated by the word2vec technique.

¹<https://www.tcu.gov.br/>

Table 2: Vector dimensionality of each technique for representing a textual document.

Technique	Vector Dimensionality
BoW	10.007
TF-IDF	10.005
BoC	100
BoLC-Th	100

Empirically, we performed exhaustive tests with different values for these two parameters in order to investigate which combination had the more accurate values. The best parameters for the two approaches were obtained with the vector-of-words (word2vec) with a dimension equal to 100 and a dictionary with 100 concepts. Therefore, these two parameters were used, both by the BoC and BoLC-Th approaches.

Table 2 shows the vector dimensionality used for each technique to represent a textual document: the BoC and the BoLC-Th techniques have the smallest vector size. The vector size used by the BoC and BoLC-Th approach is 99% smaller than the BoW and TF-IDF approaches. This is reflected in storage space savings and, mainly, in the low computational cost to perform the calculation of the similarity between the vectors in the process of retrieving similar documents.

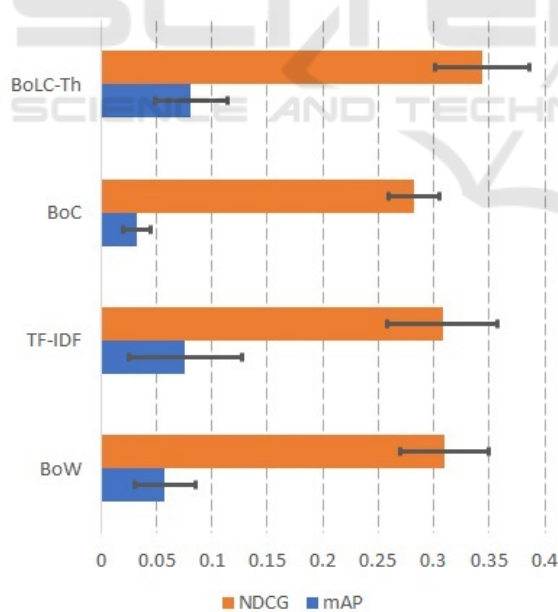


Figure 5: Best mAP and NDCG values obtained for each evaluated techniques.

4.4 Obtained Results

The metrics used to evaluate the performance of the techniques were: the *mean Average Precision* (mAP) and *Normalized Discounted Cumulative Gain* (NDCG). Both measures are often used to measure effectiveness of web search engine algorithms or related applications.

Figure 5 shows the best values obtained for each evaluated technique. We can see that the BoLC-Th and TF-IDF achieved superior values compared to the BoC and BoW techniques. Comparing the BoLC-Th and TF-IDF, the first one achieved superior performance in both measures (mAP and NDCG).

One of the main drawback of information retrieval systems is the execution time used to compare the query document with all other documents in the database. For this reason, it is also important to evaluate the efficiency of each techniques. Table 3 also shows the processing time of each technique to retrieve similar documents. The BoW and TF-IDF techniques have the highest processing time to retrieve similar documents as they have the vector with the highest dimensionality among those evaluated. Both approaches, BoC and BoLC-Th, have similar execution times.

Table 3: Time query processing used for each technique.

Technique	Time Query Processing (ms)
BoW	0.205 ± 0.15
TF-IDF	0.285 ± 0.08
BoC	0.154 ± 0.03
BoLC-Th	0.1411 ± 0.01

An advantage of using the BoLC-Th technique is the fast execution in calculating the similarity between the documents in the database: as it has a compact vector, fewer calculations are needed. This is reflected in the scalability of the proposed technique: the search for similar documents becomes more efficient in an area whose database has grown considerably over time, due to the increasing digitization of legal processes.

5 CONCLUSIONS

This work presented a new approach, called BoLC-Th, to represent legal textual documents. The proposed method is an extension of the traditional BoC approach, considering a weighted frequency of the most important concepts in the final representation of documents according to the application domain. The BoLC-Th technique makes use of a thesaurus

(specialized vocabulary), with the terms/words of the application context, which allows to semantically enrich the BoC method.

Experimental results were carried out with the objective of analyzing the performance of the proposed approach in the retrieval of legal documents by calculating the semantic similarity of their vector representations. The proposed BoLC-Th method was compared with the traditional BoW, TF-IDF and BoC approaches. The proposed method achieved better performance when compared to the BoW model and, on average, 40% more efficient than that obtained with the BoC. Thus, a significant advantage was demonstrated with the use of the BoLC-Th technique for the analyzed case study. As future work, other clustering techniques for the generation of concepts should be considered.

The main contribution of the BoLC-Th is that it incorporates the advantages of the BoC approach, such as the compact representation, while enriching legal documents representation as it considers specific words/terms from the context area. This is a valuable contribution to a domain area with peculiar characteristics, providing a valuable tool for retrieving textual information accurately and quickly.

REFERENCES

- Analytics Vidhya, N. (2017). An intuitive understanding of word embeddings: From count vectors to word2vec.
- Dhillon, I. S. and Modha, D. S. (2001). Concept decompositions for large sparse text data using clustering. *Machine learning*, 42(1):143–175.
- Hematialam, H., Garbayo, L., Gopalakrishnan, S., and Zadrozny, W. W. (2021). A method for computing conceptual distances between medical recommendations: Experiments in modeling medical disagreement. *Applied Sciences*, 11(5).
- Kim, H. K., Kim, H., and Cho, S. (2017). Bag-of-concepts: Comprehending document representation through clustering words in distributed representation. *Neurocomputing*, 266:336–352.
- Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In Xing, E. P. and Jebara, T., editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1188–1196, Beijing, China. PMLR.
- Lee, Y. (2020). Systematic homonym detection and replacement based on contextual word embedding. 53(1):17–36.
- Lee, Y.-H., Hu, P. J.-H., Tsao, W.-J., and Li, L. (2021). Use of a domain-specific ontology to support automated document categorization at the concept level: Method development and evaluation. 174:114681.
- Li, P., Mao, K., Xu, Y., Li, Q., and Zhang, J. (2020). Bag-of-concepts representation for document classification based on automatic knowledge acquisition from probabilistic knowledge base. *Knowledge-based systems*, 193:105436.
- Ling, W., Dyer, C., Black, A. W., and Trancoso, I. (2015). Two/too simple adaptations of Word2Vec for syntax problems. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1299–1304, Denver, Colorado. Association for Computational Linguistics.
- Mehanna, Y. S. and Mahmuddin, M. B. (2021). A semantic conceptualization using tagged bag-of-concepts for sentiment analysis. *IEEE access*, 9:118736–118756.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Mourino Garcia, M. A., Perez Rodriguez, R., and Anido Rifon, L. E. (2015). Biomedical literature classification using encyclopedic knowledge: a wikipedia-based bag-of-concepts approach. *PeerJ (San Francisco, CA)*, 3:e1279–e1279.
- Niebla Zatarain, J. M. (2018). Artificial intelligence and legal analytics: New tools for law practice in the digital age. *SCRIPT-ed*, 15:156–161. doi: <https://doi.org/10.2966/scrip.150118.156>.
- Rajabi, Z., Valavi, M. R., and Hourali, M. (2020). A context-based disambiguation model for sentiment concepts using a bag-of-concepts approach. *Cognitive computation*, 12(6):1299–1312.
- Renjit, S. and Idicula, S. M. (2019). Cusat nlp@ aila-fire2019: Similarity in legal texts using document level embeddings. In *FIRE (Working Notes)*, pages 25–30.
- Salim, M. N. and Mustafa, B. S. (2022). A survey on word representation in natural language processing. In *1ST Samara International Conference for Pure and Applied Science (SICPS2021): SICPS2021*. AIP Publishing. doi: <https://doi.org/10.1063/5.0121147>.
- Sansone, C. and Sperlí, G. (2022). Legal information retrieval systems: State-of-the-art and open issues. *Information Systems*, 106:101967.
- Shalaby, W. and Zadrozny, W. (2019). Learning concept embeddings for dataless classification via efficient bag-of-concepts densification. *Knowledge and Information Systems*, 61(2):1047–1070.
- Turney, P. D. (2006). Similarity of semantic relations. *Computational Linguistics*, 32(3):379–416.
- Wang, F., Wang, Z., Li, Z., and Wen, J.-R. (2014). Concept-based short text classification and ranking. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM '14*, page 1069–1078, New York, NY, USA. Association for Computing Machinery.
- Yan, J. (2009). *Text Representation*, pages 3069–3072. Springer US, Boston, MA.