An unsupervised method for concept association analysis in text collections

Pavlo Kovalchuk[1,2][0000−0003−1424−6995], Diogo Proença[2][0000−0002−3671−9637], José Borbinha[1,2][0000−0001−5463−8438], and Rui Henriques[1,2][0000−0002−3993−0171]

[1] Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal
[2] INESC-ID, Lisbon, Portugal
{pavlo.kovalchuk,diogo.proenca,jlb,rmch}@tecnico.ulisboa.pt

Abstract. This paper addresses the challenge of content categorization to support document navigation and retrieval. The work is motivated by the need to categorize all legislation of a country, where the existing metadata for each document is not sufficient for effective categorization, as concepts vary considerably among documents, resulting in highly sparse vector-space models. To address this challenge, we survey recent related work and propose a solution that integrates currently dispersed principles in a new unsupervised knowledge discovery process combining principles from topic modeling and formal concept analysis, thus not requiring prior domain knowledge to be applied in large document collections. The results confirm the potential of the proposed approach.

Keywords: Unsupervised Knowledge Discovery · Topic Modeling · Formal Concept Analysis · Concept Associations · Large Digital Libraries

1 Introduction

This work is motivated by the need to support search and navigation in *Diário da República Eletrónico* (DRE)[3], the official on-line publication journal of the Portuguese state. DRE is a digital library updated in continuum, publishing laws, regulations and legal acts. Resource discovery is supported by browsing and search in metadata and full-text, which is effective and efficient for tasks with the objective finding specific documents. However, when the task is knowledge discovery, with the purpose of learning what the collection can hide behind the metadata, the service is not efficient. Given the diversity of topics, DRE.PT results in a very sparse vector-space model, thus motivating the development of a knowledge discovery in text (KDT) process for document categorization. However, generic and fully unsupervised methods for document categorization and KDT are still in demand [37]. Despite the need to support the search and navigation within DRE and digital repositories alike, there are still limitations hampering the user experience. First, existing text mining approaches for the unsupervised categorization of documents are mostly driven by clustering algorithms that take in consideration basic aspects of the documents, such as word

[3] https://dre.pt/

relative frequency and overall content similarity, ignoring more complex relations in documents. In addition, these approaches typically map documents into a vector-space model, representing each document as a high-dimensional vector of weights, hampering clustering performance. Second, most of the existing alternatives for document categorization are based on supervised techniques where the categories for a given document collection are well known [25, 17]. Third, in the domain of (Portuguese) legal documents, there is no related work for the topic of document categorization based on concept associations (although a few contributions have been recently proposed in the context of legal repositories [30, 6], they focus on single-specific aspects of the overall KDT process).

The aim of this work is to propose a unsupervised approach for document categorization and KDT, centered on associations between topics and concepts extracted from document collections. The proposed method combines recent findings from the application of state-of-art techniques on topic modeling and formal concept analysis. As such, it comprehensively tackles all aspects of the KDT process. All the methods proposed in this work will not consider any category for the documents, the entire collection will be treated as equal, and all the document relations and concepts are extracted using automatic methods. The target approach aims at facilitating three major applications of interest: (1) support navigation between documents through the use of hyperlinks; (2) facilitate categorization; and (3) enable summarizing and the comprehensive taxonomic understanding of a set of documents.

This work is organized as follows. Section 2 introduces fundamental topics and theoretical principles behind the methods in the context of the target problem. Section 3 surveys the relevant related work, covering publications that make use of various methods and techniques for the purpose of analyzing textual data. Section 4 describes in detail the solution briefed in Section 3. Section 5 presents some results. Section 6 details the evaluation method for the proposed solution. Finally, Section 7 presents conclusions and future work.

2 Background

We define here KDT as a composition of principles from information retrieval, topic modeling, and concept analysis, aiming at finding relevant relations in a collection of documents $D=\{d_1, .., d_n\}$ in order to provide the necessary knowledge to support document categorization for search and navigation.

To preserve a sound terminology ground, *topic* denotes a semantically related set of *terms*, and *concept* is a (putative) association between terms or topics.

Core concepts. Representing unstructured documents as sets of terms supports their subsequent retrieval by specifying queries on those terms. The *vector space model* represents documents as weighted vectors, $d_i = (w_{i1}, w_{i2}, w_{i3}, ..., w_{im})$ where $w_{ij} \in \mathbb{R}$ and $w_{ij} \geq 0$, so w_{ij} is the frequency of term t_j in document d_i. The weights can be alternatively computed using the classic term frequency-inverse document frequency (Tf-idf) metric [31]. Document similarity can be

easily computed over a vector space model using metrics such as the cosine of the angle between document vectors.

Topic Modeling. The dimensionality of vector space models can be reduced to facilitate subsequent mining tasks while preserving as much useful information as possible.

A) Principal component analysis (PCA). Singular value decomposition (SVD), and the centered PCA variant, are traditional algebraic methods to reduce dimensionality (number of terms) by projecting the original data space in a new data space along the directions where the data varies the most [20]. These directions are determined by the eigenvectors, defining a linear composition of original terms, $w'_{ij} = \sum_k^m \alpha_k w_{ik}$. Although this way of reducing data is effective since only a few directions are commonly needed to capture most data variability, the semantic relations between terms t_j are lost.

B) Latent Semantic Analysis. In natural language processing and distributional semantics, Latent Semantic Analysis (LSA) is widely applied for extracting and analyzing the semantic relations between documents of a given collection. This extraction does not rely on any manually constructed dictionaries or semantic networks. The introduced vector space model assumes that terms in a given text document are conceptually independent of each other, which in real-world problems is not always true. Most terms in a document are linked to each other by underlying, unobserved topics, and the main focus of the LSA algorithm is to identify those topics. The LSA process starts by representing a document as a matrix where each entry represents the frequency of a term (row) in a text passage (column). Depending on the objective of the problem, text passages can be paragraphs, sets of terms or entire documents. For each entry, LSA applies a set of preliminary transformations considering both the terms importance in a particular passage and the degree to which the term carries information in the universe of discourse in general [24]. In order to reduce the number of rows without loosing important information, the LSA applies SVD to obtain a lower dimensional matrix where each feature is a combination of previous values.

C) Latent Dirichlet Allocation. In contrast to LSA, Latent Dirichlet Allocation (LDA) produces topics with a probabilistic frame from a given document by assuming documents with similar topics will use similar groups of terms. Documents are thus defined as a probability distribution over latent topics, and the topics are probability distributions over terms. As defined in [26], the LDA process breaks down a large corpus of documents into three levels: the corpus level, the document level and the term level. At the corpus level, for each topic z_k from a Dirichlet distribution with prior parameter β, LDA generates a topic-terms multinominal distribution ϕ_{z_k}. At the document level, for each document d_i from a Dirichlet distribution with prior parameter α, it generates a document-topics multinominal distribution θ_{d_i}. At the term level, LDA first generates a topic assignment z_k from the document-topics multinominal distribution θ_{d_i}, and for each term t_j in document d_i a term assignment w_{ij} from the topic-terms

distribution ϕ_{z_k} is generated. Accordingly, for a given parameterizable number of K topics,

$$p(w_{ij}) = \int_{\theta} \left(\prod_{i=1}^{n} \sum_{k=1}^{K} p(w_{ij}|z_k;\beta) p(z_k|\theta) \right) p(\theta;\alpha) d\theta \ . \tag{1}$$

D) Hierarchical Dirichlet Processes. Hierarchical Dirichlet Processes (HDP) provide a non-parametric way to discover topics in text data, where each document is interpreted as a set of distributions over topics, and the topics are distributions over terms. Unlike the LDA process, where the number of topics is as a parameter, the HDP infers the number of topics from the data.

Formal Concept Analysis (FCA). The theory of FCA, first introduced by Rudolf Wille [41], became a popular method for knowledge representation [15]. Given a collection of objects (documents), formal concept analysis aims to capture associations between objects (documents) based on the shared attributes (terms or topics). The relationships among objects and attributes can be represented as a concept lattice (also called Galois lattice) where each group of objects can be hierarchically grouped together based on common attributes, from most specific concepts (with fewer objects and more attributes) to less specific ones (concepts grouping many objects but sharing few attributes). A *formal context* is a triplet (D, T, I), where D is the set of documents, T is the set of terms and/or topics, and $I \subseteq D \times T$ is a relation between D and T called an incidence relation. A *formal concept* is a pair (A, O) of a formal context (D, T, I), where A objects (extent) is the set of documents that share O attributes (intent). A *concept lattice*, $\mathfrak{B}_{(D,T,I)}$, is the set of all formal concepts in a formal context.

Fuzzy Formal Concept Analysis (FFCA) [29] incorporates fuzzy logic into FCA to represent vague information. Traditional FCA is suitable for conceptual clustering and the generated lattices disclose relevant information about a given domain. However, information uncertainty may occur in a given domain which results in some attributes being more relevant than others. In order to represent information in these domains, a formal concept's relation between objects and attributes is represented using memberships between 0 and 1, and a confidence threshold placed to eliminate relations that have low membership values.

3 Survey on unsupervised analysis of digital collections

Fig 1. provides a compacted version of our proposed unsupervised method. Relevant related work per step is surveyed in the next subsections, and the underlying design principles and validation described in Section 4 and Section 6 accordingly.

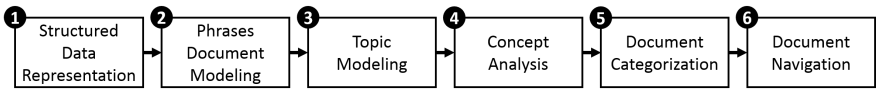

Fig. 1. Compact version of the overall pipeline of the proposed solution.

3.1 Related work

Step 1: Structured data representation. Among diverse work [35, 19], Gonalves et al. [13] assessed the impact of different representations and preprocessing procedures – including data reduction and term weighting schema – on the categorization of two collections of legal documents (PAGOD, the Portuguese Attorney Generals Office dataset, and Reuters). Singh et al. [34] explored the impact of placing different data representations (with/without stop words, with/without stemming), schemes (term frequency, Tf-idf and Boolean), clustering algorithms (K-means, Heuristic K-means and Fuzzy C-means) and algorithmic variants (different heuristics for initial seed selection) to categorize documents from Reuters-21578, Classic-7095 and 20 Newsgroups collections. Fuzzy C-means, unlike K-means, provides a degree of membership of each document for each cluster. Using Residual Sum of Squares (RSS) and Purity metrics, it was concluded that the Tf-idf scheme with stemming is the best setup to represent documents, that heuristic K-means produced better results than the standard K-means, and Fuzzy C-means proves to be the most robust clustering algorithm.

Step 2: Phrases document modeling. Modeling word order and phrases can be used to enrich term representations based on the classic bag-of-words assumption. Wang et al. [40] presented a topical n-gram model, an extension of unigram models, to this purpose and extraction of topics and topical phrases.

Step 3: Topic modeling. In [30], we find an approach to organize legal judgments from topics obtained using LDA aiming at minimizing distances between topics and documents (so each cluster of documents relates to a given topic). Using legal judgments manually categorized, this work aimed improving retrieval by finding topics using LDA and then computing the cosine similarity between each document and the extracted topics to find the closest topic for each document. In [39], the authors compared LDA and HDP , concluding HDP shows better results. In addition to the traditional document categorization methods, [8] presents a survey of several probabilistic topic models with soft clustering abilities and their applications for knowledge discovery in text corpora.

Step 4: Concept analysis. FCA has been both applied on terms and topics. In [5], a method for topic detection based on FCA is proposed, guided by both internal clustering quality metrics (Davies-Bouldin Index [9], Dunn Index [10], Silhouette coefficient [32] and The Calinski-Harabasz Index [22]) and external metrics (Reliability, Sensitivity and F-measure [2]). The experimental analysis used a collection of 2200 manually labeled tweets from 61 entities, where the binary attributes associated with terms, named entities, references and URLs. To produce a smaller and denser formal context maintaining the relevant relations, attributes with high frequency were removed. Afterwards, the concept lattice are computed with the Next Neighbours [4] algorithm. Each formal concept is here considered a topic. Still, a large number of topics is generated, meaning that not all concepts are meaningful topics. The authors thus propose the Stability metric [23] to extract the most promising formal concepts. From the obtained

results the authors concluded that, if considering the external evaluation, FCA demonstrates a more homogeneous performance than the LDA and Hierarchic Agglomerative Clustering (HAC), obtaining better overall results independent of the parameter setting. In [28], we find an application of FCA for web document categorization. Tf-idf was used to extract the relevant terms to build the formal context, and a threshold considered to capture the absence or presence of a given attribute (term) in a specific object (document). Several text mining techniques were applied on the frequent term-sets of the given domain to analyse the relations among terms and documents. Concept lattices and clusters of formal concepts were further discovered to support the analysis of results.

Step 5: Document categorization. In [21] three approaches for document clustering (HAC [11], K-means and the bisecting K-means) are compared over the datasets Reuters-21578, WebAce and TREC. Each document is represented using a vector-space model based on term frequency. Results collected using entropy, F-measure and overall similarity based on a weighted cosine formula (to measure cluster cohesiveness) indicate that bisecting K-means is better than the standard K-means and as good or better than the hierarchical approach.

A cluster-based approach is proposed in [7] to browse large document collections (Scatter/Gather). It starts to scatter the collection into a small number of documents clusters, presenting short summaries of the obtained results to the user. Clusters can be selected, gathered together to form a sub-collection and clustered to generate smaller clusters. This is repeated until the groups of documents become small enough. Results produced on 5000 articles posted on the New York Times News Service during August of 1990 show that the Scatter/Gather method can be an effective.

In [27], an inter-passage approach for text document clustering is proposed based on the discovery of multiple topic segments per document. The method removes stop-words, applies stemming, and computes a score based on the Tf-idf and SentiNetWord for each word in each topic segment per document. The word with the highest score in a segment will be treated as representative keyword for that segment. Once having the representative keyword for each segment, the overall segment score is computed by averaging the score of all words in a segment. Finally, segments are clustered together by applying the K-means algorithm. The result are clusters of segments that relate to a given topic, and the original document associated to each segment.

An improved K-means algorithm combined with Particle Swarm Optimization (POS) [33] is proposed in [16] for efficient web document clustering. POS is considered to obtain the best initial cluster centroids for the K-means algorithm. This method was tested against other clustering methods on various text document collections[4], consistently showing lower ADDCC values (mean distance between documents and the clusters centroid).

Krill Herd (KH) algorithms for efficient text document clustering is presented in [1]. KH [12] is a nature-inspired clustering method aiming at finding the minimum distance of krill individuals (documents) from foods (centroids) with

[4] http://trec.nist.gov/data.html

the highest density. The performance of KH algorithms are compared against standard K-means on four Labic datasets and show superior Purity and Entropy.

Step 6: Document navigation. The Concept Chain Queries (CCQ) is defined in [18] as a text mining technique focused on detecting links between topics across text documents. It generates a Concept Association Graph (CAG) where the nodes correspond to concepts and the links to associations. Queries are interpreted as finding the most meaningful evidence trails across documents. A cross-document knowledge discovery solution is proposed in [26] using Semantic Concept Latent Dirichlet Allocation (SCLDA) and Semantic Concept Hierarchical Dirichlet Process (SCHDP) methods, where documents are represented as meaningful Bag-of-Concepts, rather than words. The methods were applied to the CCQ problem, where the objective is to discover new relations between concepts across documents. Tests on 9/11 counter-terrorism data show superior performance over other LDA and HDP-based approaches.

Table 1. Overview on the different contributions and limitations on surveyed works.

Ref.	Contributions	Limitations
(1) Structured Data Representation		
[13]	Analysis on representation and preprocessing procedure for Portuguese legal documents	Overlook of syntactical and semantic information on the representation of documents
[34]	Performance comparison of different representations and hard versus soft clustering algorithms	Document categorization based on the overall content similarity
(2) Phrases Document Modeling		
[40]	Topic modeling integrated with n-grams models	Topic quality not compared with standard metrics (Coherence or Perplexity)
(3) Topic Modeling		
[30, 39]	Document categorization based on topic modeling techniques	Tests on small document collection where overall domains are known; hard clustering type results
(4) Concept Analysis		
[5]	FCA as a topic modeling method; formal context reduction; external and internal topic quality metrics	No comparison with other well known topic modeling methods namely, HDP or LSA
[28]	FCA applied in the context of IR; Concept-based document categorization	Non-exhaustive exploration of different attributes for the documents in the formal context
(5) Document Categorization		
[21]	Overview on several clustering algorithms and their performance on document categorization	Fairly basic clustering algorithms
[7]	Cluster-based method to support document search in large collections of documents	No cluster quality measures; no comparison with other clustering methods
[27]	Document categorization based on text segments related to a specific topic	Non-exhaustive exploration of different methods for text segmentation
[16]	Improved K-means algorithm with POS approach	Hard clustering type results; non-exhaustive quality comparison with other clustering algorithms
[1]	KH algorithm applied for document categorization	The KH algorithm require a large number of iterations not suitable for large collections of documents
(6) Document Navigation		
[18]	Combining text retrieval and link analysis for KDT; exploring the concept of CCQ for topic associations	No clear evidence for the scalability of the proposed method; no explicit description on the tools used
[26]	New topic modeling methods based on the principle of BOC; exploring different preprocessing methods	No direct comparison of topic quality based on standard metrics

3.2 Concluding remarks

Table 1 synthesizes the contributions and limitations of the surveyed work over each step of our proposed pipeline. Most methods only consider standard vector based representations, overlooking conceptual aspects of the documents. The methods that take into consideration concept associations within documents

either rely on prior knowledge, omit relevant specifics to guarantee a generic and implementable KDT process, or do not guarantee the scalability of the proposed approach for large collections.

4 The proposed unsupervised KDT process

The proposed unsupervised method, first depicted in Fig. 1, is here detailed. The extended pipeline of our proposed solution is presented in Fig. 2. Fig. 2 depicts each stage of the pipeline, detailed throughout this section.

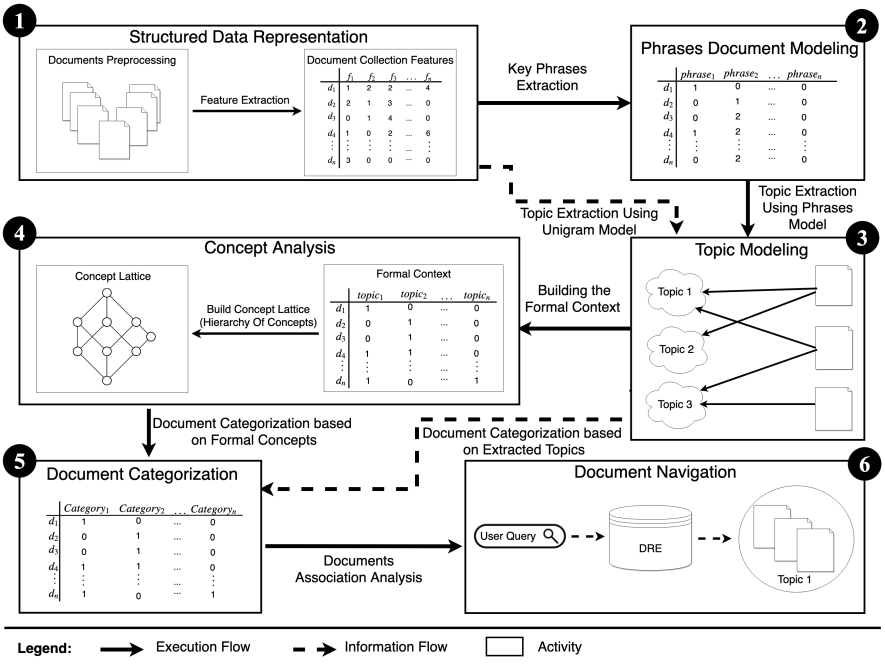

Fig. 2. Extended pipeline for the proposed solution.

(1) Structured Data Representation. Each document in the collection is preprocessed, and words not carrying relevant information removed, including stop words, punctuation, alphanumeric words, numbers, and highly frequent words. Then, each document is converted into a vector representation based on BOW-word for the subsequent extraction of features. Each document is represented as an entry in a real-valued matrix, offering a high-dimensional and structured representation of the corpus.

(2) Phrases Document Modeling. In the next step, a representation combining BOW and keyphrases (BOW-phrase) is produced for each document using PKE, a keyphrase extraction method presented by Boudin and Florian [3].

(3) Topic Modeling. Using representations based on BOW-word (step 1) and the BOW-phrase (step 2), different sets of topics can be extracted using LSA,

LDA and HDP methods for topic modeling. The topics produced from the different approaches can be combined and subsequently mapped into a formal context. If topics are to be readily used for the aimed categorization end, topics with highly overlapping terms should be merged or removed.

(4) Concept Analysis. This step starts by creating a formal context based on a fuzzy FCA frame, where objects correspond to documents and attributes correspond to the set of topics obtained in step 3. The relations between documents and topics are represented by a membership value between [0,1] that corresponds to the likelihood of a given document being characterized by a specific topic. In order to avoid meaningless formal concepts in the next step, all entries from the formal context that do not satisfy a membership threshold can be binarized and the AddIntent algorithm [38] (or equivalent) applied over the reduced formal context. To facilitate interpretation, the obtained formal concepts can be organized in a concept lattice that corresponds to a hierarchical structure based on a generalization-specialization relationship, where those at the top (bottom) of the concept lattice will represent general (specific) topics. To guarantee the relevance of the obtained formal concepts, an additional pruning procedure (step 5) is required.

(5) Document Categorization. The quality of each topic or formal concept can be measured using Stability, a metric of cohesiveness [5]. Stability is used to remove from the concept lattice the formal concepts that do not represent cohesive groups of documents. The pruned lattice provides an organized and informative view over the document collection, where each document is categorized by the hierarchy of topics where it appears. In other words, the entire collection is factorized into groups of documents that are correlated in accordance with a given set of topics that can be interpreted as a general concept. Unlike hard clustering algorithms described in the literature, the proposed method adequately tackles the difficulty of learning from high-dimensional and sparse data structures, and further enables the identification of overlapping groups of documents by allowing a given document to appear in multiple groups. The possibility to not only rely on concepts, but to augment these with the original topics extracted in stage **3** (Fig. 2) guarantees that all potentially relevant content is considered, enabling a comprehensive characterization and full traceability of document categories from concepts, topics, and phrases.

(6) Document Navigation. Once the document collection is organized by topics and concepts, searching and browsing can be narrowed to the groups of documents that share identical contents (step 5), instead of the entire collection.

5 Results

This section presents initial results gathered from 5000 legal documents mentioning the Ministry of Agriculture ("Ministrio da Agricultura", in Portuguese). Following the pipeline presented in *section 4*, in step 1 each document was preprocessed to remove uninformative words (including words that are not noun

phrases or proper nouns) and frequent words on most documents. Next, in step 2 all phrases per document were extracted and used to convert each document into a BOW-phrase vector representation. In step 3, topic modeling was applied using the six setups described in *section 4* (recalling: LSA, LDA and HDP, each with BOW-word and BOW-phrase representations). Both LSA and LDA take as input the number of topics k to be extracted from the document collection. Fig.3 measures the impact of k on the topic's coherence [36]. The best k corresponds to 82 topics (coherence of 0.43) for setup (A), 82 topics (coherence of 0.32) for setup (B), 113 topics (coherence of 0.51) for setup (C), and 3 topics (coherence of 0.56) for setup (D). HDP extracts a total of 150 topics for both setups (E) and (F) (coherence of respectively 0.474 and 0.627), a constant regardless of k.

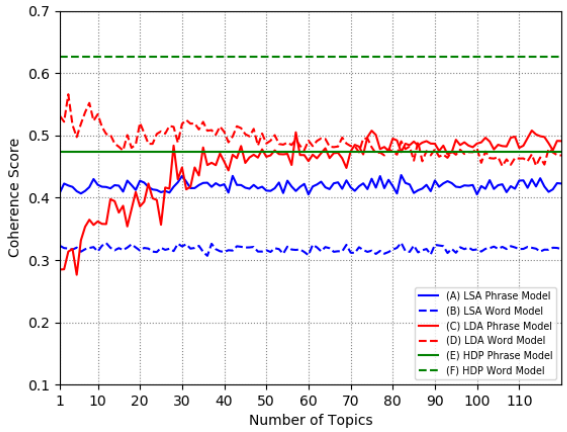

Fig. 3. Comparison of the Coherence Score for the Different Models.

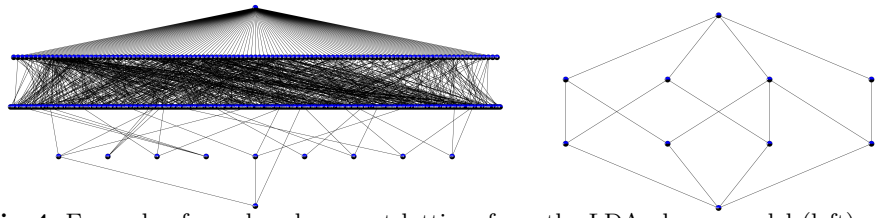

Fig. 4. Example of a reduced concept lattices from the LDA-phrase model (left) and for the document **3535010**, contained in that concept lattices (right).

In Table 2 we compare the results for the six setups, considering for each the 10 most frequent topics. In the upper side we can see, on the left, the total number of documents assigned to each topic, and on the right, the number of unique documents assigned. To support the analysis, in the lower side of the table we see the same information but normalized in relation to the size of the collection (5000 documents). We can realize LDA is the methods that creates topics that

	HDP		LDA		LSA		HDP		LDA		LSA	
#	word	phrase	word	phrase	word	phrase	word	phrase	word	phrase	word	phrase
1	4489	3768	1696	690	1192	1369	2191	2387	466	288	1192	1369
2	1385	732	1508	628	1096	1163	93	170	347	71	1096	1163
3	751	382	1014	588	716	313	152	46	134	308	716	313
4	513	302	968	577	413	233	47	149	109	104	413	233
5	358	258	955	570	324	142	3	47	93	90	324	142
6	234	213	886	568	177	106	8	73	139	87	177	106
7	221	166	838	553	90	80	14	14	45	91	90	80
8	139	150	770	519	84	80	3	23	90	85	84	80
9	119	134	747	507	81	77	0	23	50	154	81	77
10	85	132	733	481	80	76	0	21	71	251	80	76
1	90%	75%	34%	14%	24%	27%	49%	63%	27%	42%		
2	28%	15%	30%	13%	22%	23%	7%	23%	23%	11%		
3	15%	8%	20%	12%	14%	6%	20%	12%	13%	52%		
4	10%	6%	19%	12%	8%	5%	9%	49%	11%	18%		
5	7%	5%	19%	11%	6%	3%	1%	18%	10%	16%	100%	
6	5%	4%	18%	11%	4%	2%	3%	34%	16%	15%		
7	4%	3%	17%	11%	2%	2%	6%	8%	5%	16%		
8	3%	3%	15%	10%	2%	2%	2%	15%	12%	16%		
9	2%	3%	15%	10%	2%	2%	0%	17%	7%	30%		
10	2%	3%	15%	10%	2%	2%	0%	16%	10%	52%		

Table 2. Documents assigned to each topic for the 10 most frequent topics, counting the total of documents (left) and unique documents (right).

are assigned to groups of documents of more uniform sizes (ranging from 15% to 30% for BOW-word and, very impressing, from an 10% to 14% for BOW-phrase). On the other side, LSA assures, by its nature, that each topic has only unique documents. Finally, HDP seems not so interesting, at least considering the specific collection we used for this test. Finally, the AddIntent algorithm is used to extract the formal concepts from a formal context, maintaining only the formal concepts with high stability. In Fig. 4 we can see the example of the resulted concept lattices over the topics resulted from the LDA-phrase model, and the concepts for the document 3535010[5] (The *Lattice Visualization*[6] tool was used for this purpose). This way, we are able to categorize each document based on its related topics and formal concepts, as well as link it to documents with similar concepts (to support document navigation, for example).

6 Principles for quality assessment of the pipeline

Our future work must now research how the best way to make use of each method, eventually resulting in a solution combining the best of each (or at least from LSA and LDA). In addition to well-known clustering metrics (such as accuracy-based views in the presence of ground truth on document categories or cohesion-separation views) that can be considered at the end of pipeline to assess

[5] https://dre.pt/web/guest/home/-/dre/3535010/details/maximized
[6] http://latviz.loria.fr/

document categorization, we suggested additional metrics to test the adequacy of the proposed unsupervised KDT process over a given document collection. Motivated by the results in [5] that show internal clustering measures are adequate to estimate the quality of topic modeling methods, the performance of each setup in step 3 should be promptly assessed using: coherence [36] to measure the degree of similarity between high scoring words in a topic; Dunn index [10] to measure how dense and well-separated are the obtained topics; and Calinski-Harabasz index [22], a ratio between topic variance and the overall within-topic variance. These measures should be applied to estimate the optimal number of topics in LSA and LDA procedures. Measures of homogeneity and quality should be applied to evaluate formal concepts in step 4. For example, the amount of tolerated noise (fraction of 0), and statistical tests based on the Bernoulli distribution can be placed to access if a formal concept is statistically significant (i.e. it deviates from expectations) [14]. To evaluate how closely related are the documents in a formal concept, the cohesion metric is suggested.

7 Conclusion and Future Work

This work addressed the problem of creating a structured and categorized view of a collection of documents without background knowledge. Accordingly an unsupervised KDT process is proposed to support document indexing for retrieval and general navigation in the entire collection of documents. Relevant related work was surveyed to this end, and their limitations and contributions identified. Building upon these findings, the principles underlying the target KDT process were introduced. The results demonstrate the applicably and the potential of our proposed method. In future work, we aim to use alternative collections of legal documents to study the adequacy of the proposed approach and the impact of selecting alternative document representations and topic modeling approaches; explore different ways of preprocessing Portuguese legal documents in order to overcome some linguistic barriers present in the Portuguese language that may reduce the quality of the extracted topics; and also to further analyze the obtained hierarchy of term associations to dynamically infer an ontology, which can then be considered to support content categorization.

Acknowledgement. This work was supported by Imprensa Nacional Casa da Moeda (INCM) and national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2019.

References

1. Abualigah, L.M., Khader, A.T., Al-Betar, M.A., Awadallah, M.A.: A krill herd algorithm for efficient text documents clustering. In: 2016 IEEE symposium on computer applications & industrial electronics (ISCAIE). pp. 67–72. IEEE (2016)
2. Amigó, E., Gonzalo, J., Verdejo, F.: A general evaluation measure for document organization tasks. In: Proceedings of the 36th international ACM SIGIR confer-

ence on Research and development in information retrieval. pp. 643–652. ACM (2013)
3. Boudin, F.: Pke: an open source python-based keyphrase extraction toolkit. In: COLING. pp. 69–73. Osaka, Japan (2016)
4. Carpineto, C., Romano, G.: Concept data analysis: Theory and applications. John Wiley & Sons (2004)
5. Castellanos, A., Cigarrán, J., García-Serrano, A.: Formal concept analysis for topic detection: a clustering quality experimental analysis. Information Systems **66**, 24–42 (2017)
6. Chen, Y.L., Liu, Y.H., Ho, W.L.: A text mining approach to assist the general public in the retrieval of legal documents. IJ American Soc. for Info. Science and Tech. **64**(2), 280–290 (2013)
7. Cutting, D.R., Karger, D.R., Pedersen, J.O., Tukey, J.W.: Scatter/gather: A cluster-based approach to browsing large document collections. In: ACM SIGIR. pp. 318–329. ACM (1992)
8. Daud, A., Li, J., Zhou, L., Muhammad, F.: Knowledge discovery through directed probabilistic topic models: a survey. Frontiers of computer science in China **4**(2), 280–301 (2010)
9. Davies, D.L., Bouldin, D.W.: A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence **PAMI-1**(2), 224–227 (1979)
10. Dunn, J.C.: Well-separated clusters and optimal fuzzy partitions. Journal of cybernetics **4**(1), 95–104 (1974)
11. El-Hamdouchi, A., Willett, P.: Comparison of hierarchic agglomerative clustering methods for document retrieval. The Computer Journal **32**(3), 220–227 (1989)
12. Gandomi, A.H., Alavi, A.H.: Krill herd: a new bio-inspired optimization algorithm. Communications in nonlinear science and numerical simulation **17**(12), 4831–4845 (2012)
13. Gonçalves, T., Quaresma, P.: Evaluating preprocessing techniques in a text classification problem. São Leopoldo, RS, Brasil: SBC-Sociedade Brasileira de Computação (2005)
14. Henriques, R., Madeira, S.C.: Bsig: evaluating the statistical significance of biclustering solutions. Data Mining and Knowledge Discovery (2017)
15. Ignatov, D.I.: Introduction to formal concept analysis and its applications in information retrieval and related fields. In: Russian Summer School in IR. pp. 42–141. Springer (2014)
16. Jaganathan, P., Jaiganesh, S.: An improved k-means algorithm combined with particle swarm optimization approach for efficient web document clustering. In: ICGCE. pp. 772–776. IEEE (2013)
17. Jiang, S., Pang, G., Wu, M., Kuang, L.: An improved k-nearest-neighbor algorithm for text categorization. Expert Systems with Applications **39**(1), 1503–1509 (2012)
18. Jin, W., Srihari, R.K., Ho, H.H., Wu, X.: Improving knowledge discovery in document collections through combining text retrieval and link analysis techniques. In: ICDM. pp. 193–202 (2007)
19. Kadhim, A.I., Cheah, Y.N., Ahamed, N.H.: Text document preprocessing and dimension reduction techniques for text document clustering. In: 2014 4th International Conference on Artificial Intelligence with Applications in Engineering and Technology. pp. 69–73. IEEE (2014)
20. Kalman, D.: A singularly valuable decomposition: the svd of a matrix. The college mathematics journal **27**(1), 2–23 (1996)
21. Karypis, M.S.G., Kumar, V., Steinbach, M.: A comparison of document clustering techniques. In: IW on Text Mining at SIGKDD (2000)

22. Kozak, M.: a dendrite method for cluster analysis by caliński and harabasz: A classical work that is far too often incorrectly cited. Communications in Statistics-Theory and Methods **41**(12), 2279–2280 (2012)
23. Kuzuetsov, S.: Stability as an estimate of thie degree of substantiation of hypotheses derived on the basis of operational, similarity (1990)
24. Landauer, T.K., Foltz, P.W., Laham, D.: An introduction to latent semantic analysis. Discourse processes **25**(2-3), 259–284 (1998)
25. Li, C.H., Yang, J.C., Park, S.C.: Text categorization algorithms using semantic approaches, corpus-based thesaurus and wordnet. Expert Systems with Applications **39**(1), 765–772 (2012)
26. Li, X., Jin, W.: Cross-document knowledge discovery using semantic concept topic model. In: ICMLA. pp. 108–114. IEEE (2016)
27. Mishra, R.K., Saini, K., Bagri, S.: Text document clustering on the basis of inter passage approach by using k-means. In: IC on Comp., Communication & Automation. pp. 110–113. IEEE (2015)
28. Myat, N.N., Hla, K.H.S.: Organizing web documents resulting from an information retrieval system using formal concept analysis. In: Asia-Pacific Symposium on Info. and Telec. Technologies. pp. 198–203. IEEE (2005)
29. Quan, T.T., Hui, S.C., Cao, T.H.: A fuzzy fca-based approach to conceptual clustering for automatic generation of concept hierarchy on uncertainty data. In: CLA. pp. 1–12 (2004)
30. Raghuveer, K.: Legal documents clustering using latent dirichlet allocation. IAES Int. J. Artif. Intell **2**(1), 34–37 (2012)
31. Rajaraman, A., Ullman, J.D.: Data Mining, p. 117. Cambridge University Press (2011)
32. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics **20**, 53–65 (1987)
33. Shi, Y., Eberhart, R.C.: Parameter selection in particle swarm optimization. In: IC on evolutionary programming. pp. 591–600. Springer (1998)
34. Singh, V.K., Tiwari, N., Garg, S.: Document clustering using k-means, heuristic k-means and fuzzy c-means. In: IC on Computational Intelligence and Communication Networks. pp. 297–301. IEEE (2011)
35. Srividhya, V., Anitha, R.: Evaluating preprocessing techniques in text categorization. International journal of computer science and application **47**(11), 49–51 (2010)
36. Stevens, K., Kegelmeyer, P., Andrzejewski, D., Buttler, D.: Exploring topic coherence over many models and many topics. In: Joint Conference on Empirical Methods in NLP and Computational Natural Language Learning. pp. 952–961. Association for Comp. Linguistics (2012)
37. Tan, P.N.: Introduction to data mining. Pearson Education India (2018)
38. Van Der Merwe, D., Obiedkov, S., Kourie, D.: Addintent: A new incremental algorithm for constructing concept lattices. In: IC on Formal Concept Analysis. pp. 372–385. Springer (2004)
39. Venkatesh, R.K.: Legal documents clustering and summarization using hierarchical latent dirichlet allocation. IAES International Journal of Artificial Intelligence **2**(1) (2013)
40. Wang, X., McCallum, A., Wei, X.: Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In: ICDM. pp. 697–702. IEEE (2007)
41. Wille, R.: Restructuring lattice theory: an approach based on hierarchies of concepts. In: Ordered sets, pp. 445–470. Springer (1982)