Unsupervised Concept Analysis in Legal Documents

Pavlo Kovalchuk
IST, INESC-ID
Lisbon, Portugal
pavlo.kovalchuk@tecnico.ulisboa.pt

Diogo Alexandre Breites de Campos Proença
INESC-ID
Lisbon, Portugal
diogo.proenca@tecnico.ulisboa.pt

Rui Henriques
IST, INESC-ID
Lisbon, Portugal
rmch@tecnico.ulisboa.pt

ABSTRACT
The recent access and consolidated storage of digital legal documents is creating the need for efficiently finding relevant information in large collections of documents. This work addresses the challenge of content categorization to support document navigation and retrieval. The work is motivated by the need to categorize the 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 an associated highly sparse vector-space model. Several authors have proposed methods for content categorization and organization in order to support document navigation and retrieval of relevant information in accordance with user’s needs. However, most of the existing methods are not suitable for documents where the general concepts are unknown. 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, formal concept analysis and biclustering. The proposed process does not require prior domain knowledge to be applied in large document collections. The document collection where the proposed method will be applied is the Portuguese official on-line publication journal repository of legal documents, referred as Diário da República Eletrónico (DRE)\(^1\), the official on-line publication journal of the Portuguese state. DRE is a digital library updated in continuum, publishing laws, regulations and legal acts. The DRE dataset is composed by two different series. The first series concerns all the constitutional laws, international convection’s, supreme court law decisions, and many documents related to the jurisdictional subject. The second series compresses administrative documents and public concurrence. The DRE contains more than 2,000,000 documents. Resource discovery is supported by browsing and search in metadata and full-text, which is effective and efficient for tasks with the objective of 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. In this context, generic and fully unsupervised methods for document categorization and KDT are still in demand [53]. 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 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 [58]. 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 [36, 25]. 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 [46, 6], they focus on single-specific aspects of the overall KDT process). To address the aforementioned challenges, we propose a method that combines recent findings from the application of state-of-art techniques on topic modeling, formal concept analysis and biclustering. 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

\(^{1}\)https://dre.pt/
as equal, and all the document relations and concepts are extracted using automatic methods. The dissertation hypothesis is the following:

(Hypothesis) Combining principles from Topic Modeling, Formal Concept Analysis and Biclustering in a consistent methodology supports the knowledge discovery process in digital libraries with incomplete metadata.

The target approach aims at facilitating three major applications of interest: (1) support navigation between documents through the use of hyperlinks; (2) facilitate categorization; (3) enable summarizations and the comprehensive taxonomic understanding of a given corpus.

The gathered results provide evidence in favor of thesis validation and motivate the applicability of the proposed method for document categorization and navigation.

BACKGROUND

The process of knowledge discovery in text collections (KDT) aims at finding relevant relations in a collection of documents \( D = \{d_1, ..., d_n\} \), a necessary basis for content categorization, search and navigation. To this end, KDT combines principles from information retrieval, topic modeling, and concept analysis.

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

Representing unstructured documents as sets of terms allows subsequent queries on those terms. The vector space model represents documents as weighted vectors, \( d_i = (w_{i1}, w_{i2}, w_{i3}, ..., w_{im}) \) where \( w_{ij} \) is the frequency of term \( t_j \) in document \( d_i \), \( w_{ij} \in \mathbb{R} \) and \( w_{ij} \geq 0 \). Weights can be alternatively computed using the classic term frequency-inverse document frequency (TF-idf) metric [47]. Document similarity can be then computed from the angle between two documents.

Given the common high-dimensionality of vector space models, they can be reduced using principles from topic modeling to facilitate subsequent mining:

- principal component analysis (PCA) uses algebraic operates to project data into a new data space along axes (eigenvectors \( \alpha_k \)) capturing most data variability [28], \( w'_{ij} = \sum_k \alpha_k w_{ik} \). Semantic relations between terms are lost;

- latent semantic analysis (LSA) preserves semantic relations without relying on dictionaries or semantic networks. Terms in a given text document are seen as conceptually independent and linked to each other by underlying, unobserved topics. LSA algorithm identifies those topics considering both their local and global relevance [35];

- latent Dirichlet allocation (LDA) sees documents as probability distributions over latent topics, which in turn are described by probability distributions over terms. To this end, it places multinomial and Dirichlet assumptions to estimate the likelihood of a document to be described by a given topic;

- hierarchical Dirichlet processes (HDP) provides a non-parametric alternative to LDA, enabling the discovery of a non-fixed number of topics from text.

Formal Concept Analysis. The theory of FCA, first introduced by Wille [57], is currently a popular method for knowledge representation [23].

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 \( I \subseteq D \times T \) relates \( D \) and \( T \) (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, \( \mathcal{L}_{(D,T,I)} \), is the set of all concepts in a formal context. Concept lattices (also called Galois lattices) related all concepts hierarchically based on the shared elements, from less specific (concepts grouping many objects sharing few attributes) to most specific (fewer objects and more attributes).

Fuzzy Formal Concept Analysis (FFCA) [45] 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.

Biclustering. Given a vector space model \( A \) defined by a set of objects (documents) \( D = \{d_1, ..., d_n\} \), attributes (terms and topics) \( Y = \{t_1, ..., t_m\} \), and elements \( w_{ij} \in \mathbb{R} \) observed in \( d_i \) and \( t_j \):

- a bicluster \( B=(I,J) \) is a \( n \times m \) submatrix of \( A \), where \( I = \{i_1, ..., i_n\} \subseteq D \) is a subset of documents and \( J = \{j_1, ..., j_m\} \subseteq Y \) is a subset of attributes;

- the biclustering task aims at identifying a set of biclusters \( B = (B_1, ..., B_s) \) such that each bicluster \( B_k = (I_k, J_k) \) is a coherent concept that satisfies specific homogeneity, dissimilarity and statistical significance criteria.

Homogeneity criteria are commonly guaranteed through the use of a merit function, such as the variance of the values in a bicluster [38]. Merit functions are typically applied to guide the formation of biclusters in greedy and exhaustive searches. In stochastic approaches, a set of parameters that describe the biclustering solution are learned by optimizing a merit (likelihood) function.

The pursued homogeneity determines the coherence, quality and structure of a biclustering solution [19]. The coherence of a bicluster is determined by the observed form of correlation among its elements (coherence assumption) and by the allowed value deviations from perfect correlation (coherence strength). The quality of a bicluster is defined by the type and amount of accommodated noise. The structure of a biclustering solution is defined by the number, size, shape and positioning of biclusters. A flexible structure is characterized...
A bicluster has constant coherence when \( \gamma_i = 0 \) (or \( \gamma_j = 0 \)) and additive coherence otherwise, \( \gamma_i \neq 0 \) (or \( \gamma_j \neq 0 \)).

Let \( \delta \) be the amplitude of values in \( A \), coherence strength is a value \( \delta \in [0, \bar{A}] \) such that \( w_{ij} = c_{ij} + \gamma_i + \eta_j \) where \( \eta_j \in [-\delta/2, \delta/2] \).

**Figure 1:** Formal versus coherent concepts: biclustering with varying homogeneity criteria.

**Definition 2.** Given a numeric dataset \( A \), a bicluster \( (I,J) \) satisfies the order-preserving coherence assumption iff the values for each object in \( I \) (attribute in \( J \)) induce the same linear ordering \( \pi \) along the subset of attributes \( J \) (objects \( I \)).

**Statistical significance** criteria, in addition to homogeneity criteria, guarantees that the probability of a bicluster’s occurrence (against a null data model) deviates from expectations [22]. **Dissimilarity** criteria can be further placed to comprehensively cover the vector space with non-redundant biclusters [20].

**Related work**

**Structured data representation.** Among diverse work [52, 27], Gonçalves et al. [18] assessed the impact of different representations and preprocessing procedures — including data reduction and term weighting scheme — on the categorization of two collections of legal documents (PAGOD, the Portuguese Attorney General’s Office dataset, and Reuters). Singh et al. [51] 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.

**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. [56] presented a topical n-gram model, an extension of unigram models, to this purpose and extraction of topics and topical phrases.

**Topic modeling.** In [46], 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 [55], the authors compared LDA and HDP, concluding HDP shows better results. In addition to the traditional document categorization methods, [9] presents a survey of several probabilistic topic models with soft clustering abilities and their applications for knowledge discovery in text corpora.

**Concept analysis (FCA).** 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 [10], Dunn Index [13], Silhouette coefficient [49] and The Calinski-Harabasz Index [32]) 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 relations among terms and documents. Concept lattices were further discovered to support the analysis of results.
A cluster-based approach is proposed in [8] to browse large document collections. Despite its relevance, it is limited to Boolean spaces (presence or absence of topics per document), sharing similar limitations to FCA. Dhillon [12] proposed the use of coclustering (a restrictive variant of the biclustering task that imposes a checkboard structure) to explore text collections. Coclustering was applied to vector space models with entries given by \( w_{ij} \times \log(\frac{n_j}{n_i}) \), where \( n \) is the number of documents and \( n_j \) is the number of documents containing term \( t_j \) in document \( d_i \). The author was able to identify subsets of words and documents with strong correlation along the Cranfield (1400 aeronautical documents), Medline (1033 medical documents) and Cisi (1460 information retrieval documents) collections. Despite its relevance, coclustering requires all elements to belong to a concept (exclusive condition) and to a single concept only (exclusive condition), largely limiting the inherent flexibility of the biclustering task.

Document categorization. In [29] three approaches for document clustering (HAC [14], 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 [8] 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 [40], 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) [50] is proposed in [24] 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\(^2\), consistently showing lower ADDCC values (mean distance between documents and the clusters centroid).

Krill Herd (KH) algorithms for efficient text document clustering are presented [1]. KH [16] is a nature-inspired clustering method aiming at finding the minimum distance of krill individuals (documents) from foods (centroids) with the highest density. The performance of KH algorithms are compared against standard K-means on four Labic datasets and show superior Purity and Entropy.

Document navigation. The Concept Chain Queries (CCQ) is defined in [26] 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 [37] 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 4, provided in Appendix, synthesizes the contributions and limitations of the surveyed work.

SOLUTION
The proposed unsupervised KDT method, is presented in Fig. 2. Fig. 2 depicts each stage of the pipeline, detailed throughout this section.

Component 1: Structured Data Representation
The proposed method starts by preprocessing each document from the extracted collection. The preprocessing sub component is composed by four different steps. First, all the punctuation, emails, dates, hyper references and numbers are removed. Second, is the legal entity normalization step, many documents in the DRE contain a large amount of legal entities from the Portuguese state, these entities references may be

\(^2\)http://trec.nist.gov/data.html
Component 2: Phrases Document Modeling

The traditional document representation based on BOW, is build around the conceptualization that words are independent units of knowledge found in documents. However this document representation is not able to identify phrases, that carry more information than words. To address this observation, the next component focuses on the extraction of keyphrases from documents that further will provide a lower dimensional space of the real-valued matrix obtained in the previous step. To extract the keyphrases from the documents we will use Phrase\textsuperscript{5} an automatic keyphrase extraction tool from Gensim\textsuperscript{6}. This tool was inspired by previous works developed by Mikolov et al. [39] and Bouma [3]. The process behind the extraction of keyphrases defined in the Phrase stars by computing for each set of two words ($w_i, w_j$) a score defined by:

$$
\text{score}(w_i, w_j) = \frac{\text{count}(w_i w_j) - \delta}{\text{count}(w_i) - \text{count}(w_j)} \tag{1}
$$

This score metric compares the absolute uni-gram and bi-gram count for each set of words. The $\delta$ constant is used as a discounting coefficient in order to avoid too many phrases consisting of very infrequent words to be formed. A set of two words $w_i$ and $w_j$ is considered a phrase if the score is higher then a give threshold, by default this threshold is set to 1. We run this process over our corpus several times in order to capture the maximal sets of phrases with more than two words. After having the keyphrases each document is converted into vector-space model representation based on BOW, where each column corresponds to the extracted phrases.

Component 3: Topic Modeling

From the previously obtained feature matrix, different sets of topics can be extracted using LSA, LDA and HDP methods for topic modeling, as presented in Background. In this component we will use the implementations found in Gensim for each topic modeling algorithms (LSA\textsuperscript{7}, LDA\textsuperscript{8}, HDP\textsuperscript{9}). Each method will applied over the feature matrix produced on the previous component. The topics produced from the different approaches can be combined and subsequently mapped into attributes for the original documents. If topics are to be readily used for the aimed categorization end, topics with highly overlapping terms should be merged or removed.

Component 4: Concept Analysis

In the following section, we present two parallel methods for concept analysis.

Formal Concept Analysis

This component 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 component 3. The relations between documents and topics are represented by a value between $[0,1]$ that corresponds to the

\textsuperscript{5}https://radimrehurek.com/gensim/models/phrases.html
\textsuperscript{6}https://radimrehurek.com/gensim/
\textsuperscript{7}https://radimrehurek.com/gensim/models/lsimodel.html
\textsuperscript{8}https://radimrehurek.com/gensim/models/ldamodel.html
\textsuperscript{9}https://radimrehurek.com/gensim/models/hdpmodel.html

Figure 2: Extended pipeline for the proposed solution.

represented as abbreviations or in full. In order to minimize the error on the next stages of the proposed method and to guarantee that different representations of the same entity are considered as the same, these different representations must be normalized. This normalization is carried out with the support of an external file provided by the DRE that compiles different entities used throughout the document collection, each entry in this file provides a entity id, the abbreviated form and the full form. Each entity is identified in the corpus using a tool from the Spacy\textsuperscript{3} library based on regular expressions Phrase-Matcher and converted to a normalized form, based on the following notation "TOKEN_xxx", where “xxx” is the id of the entity. Third, in order to avoid topics and concepts that do not improve the understating of the corpus words that are not, proper noun or verb are removed. To remove these words, first the POS for each word is computed using the POS tagger from Spacy, next all the words that are not noun, proper noun or verb are removed based on the POS. Forth, after the POS step, words not carrying relevant information are removed for each document, including stop words, punctuation, alphanumeric words, numbers, and highly frequent words. Additionally, all the proper names that correspond to deputies are also removed this is achieved using an exhaustive list of Portuguese first and last names\textsuperscript{4}. Then each document is converted into a vector-space model representation where the weights are computed using the traditional BOW. All vector-space model representations are combined into a feature matrix, where the rows correspond to the documents and the columns to each word, this high-dimensional features matrix offer a structured representation of highly dimensional heterogeneous data.
To facilitate interpretation, the obtained formal concepts are then 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 is required. The quality of each formal concept can be measured using Stability [48, 5, 33], a metric of cohesiveness. Stability measures how much the intent (attributes) in a concept depends on particular set of objects in the extent. The algorithm used is defined by:

$$\sigma(A, B) = \frac{|C \subseteq A| C' = B|}{2^{|A|}}$$  \hspace{1cm} (2)

where $|A|$ is the intent’s set cardinality ($A$) and $C$ corresponds to each subset of $A$ whose concept’s intent ($C'$) is equal to the concept intent of $A$, that is $C' = B$. Please refer to [48] for the original definition of equation 2.

Stability is used to remove from the concept lattice the formal concepts that do not represent cohesive groups of documents and topics. 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.

**Biclustering**

Despite the well-recognized relevance of FCA to explore taxonomical relations and deriving concept hierarchies from text collections [30], FCA is hampered by major drawbacks. First, FCA is only able to derive formal hierarchies of concepts from symbolic data representation, when the challenge of concept extraction is applied to a real-valued data spaces the selection of a discretization threshold is required to discretize the real-valued data spaces associated to the text collection. This need to select a discretization threshold imposed by the FCA brings discretization drawbacks and generally missing relevant associations observed in the original real-valued data spaces [30]. Biclustering, the discovery of concept associations given by subsets of documents correlated on subsets of topics, is suggested to address the aforementioned problems and further allow the unprecedented discovery of order-preserving associations. To this end, we selected the BiC PAMS algorithm [21]. The BiC PAMS algorithm was selected to drive coherent concept analysis since it consistently integrates state-of-art algorithms on pattern-based biclustering and guarantees the efficiency of the underlying searches. Additionally with BiC PAMS we will explore different homogeneity criteria, evaluate the degree of quality and control the amount of noise aloud in the obtained biclusters.

**Component 5: Document Categorization**

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 of concepts. These concepts obtained in component 4 (Figure 2) will enable a comprehensive multi-layered categorization of the documents based on the different subjects that each document refers to. The possibility to not only rely on these concepts, but to augment these with the original topics extracted in component 3 (Figure 2) guarantees that all potentially relevant content is considered, once again enabling a comprehensive characterization and full traceability of document categories from concepts, topics, and phrases.

**Component 6: Document Navigation**

Once the document collection is categorized by concepts, it is possible to draw hyperlinks from document to document that share the same concepts. This multi-layered categorization based on concepts will narrow down the task of searching and browsing from the entire collection to well defined conceptual groups of documents that share the same concepts.

**RESULTS**

To illustrate the potential applicability of the proposed method, results are gathered in four major steps. First, we analyze the target corpus. Second, we empirically delineate general differences of FCA and biclustering. Third, we provide evidence for the relevance of finding non-trivial (yet meaningful) concepts with constant and order-preserving forms of coherence. Finally, we show that biclustering guarantees the statistical significance of relations, providing a trustworthy means for concept analysis.

**Dataset.** Over 35000 legal documents issued by state bodies in the domain of agriculture were extracted from the ministry of agriculture for experimental analysis.

**Preprocessing.** Each document was preprocessed to remove stop words, punctuation, numbers, links, emails and dates. Next, the Part-Of-Speech (POS) for each word is extracted, and all words that are not nouns or proper nouns are removed. Finally, words with high frequency and low TF-IDF scores are also removed.

**Topic modeling.** We further used Phrase\(^{10}\) to extract the combined words (phrasing) per document. From the obtained feature matrix, topics were extracted using LSA, LDA and HDP methods. Fig.4 shows for LDA and LSA how the quality of the approaches vary with the number of topics (HDP is non-parametric). The coherence score establishes the quality of the obtained topics by computing the probability of pairs of words in a given topic appearing together on the documents associated with a given topic. Figures 12 13 and 14 visually depict the separation between the topics generated by the

\(^{10}\) Automatic keyphrase extraction tool from Gensim: https://radimrehurek.com/gensim/
different methods. The orange nodes correspond to the topics and the blue nodes correspond to the terms associated to each topic. In accordance, LDA was selected. A document is then seen as a vector of probabilistic values that corresponds to the likelihood (predominance) of a given topic appear in the document.

**Formal concept analysis.** In Table 1 and in Fig. 5 we present the impact of different discretization thresholds when FCA [17] is applied to the preprocessed dataset – a vector space model with 35000 documents and 120 topics – under a variable discretization threshold $\theta$. $\theta$ parameterization is a highly sensible choice as evidenced by its impact on the number of formal concepts (from 230k concepts when $\theta=0.05$ to 48k when $\theta=0.1$ and 122 when $\theta = 0.5$), average number of topics per concept, and the stability criterion [48]. Elements in the vector space model close to $\theta$ are excluded from the concepts. By seeing topics as Bernoulli variables in a Boolean data space, binomial tail statistics [22] reveal that only a small fraction of the returned concepts are statistically significant.

Table 1: Discretization threshold impact on number and size of formal concepts.

<table>
<thead>
<tr>
<th>Discretization threshold</th>
<th>#Formal Concepts</th>
<th>AVG #Rows</th>
<th>AVG #Columns</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>230065</td>
<td>6.07</td>
<td>1.73</td>
</tr>
<tr>
<td>0.10</td>
<td>47765</td>
<td>8.37</td>
<td>3.35</td>
</tr>
<tr>
<td>0.20</td>
<td>7729</td>
<td>14.29</td>
<td>2.47</td>
</tr>
<tr>
<td>0.30</td>
<td>1395</td>
<td>41.55</td>
<td>2.47</td>
</tr>
<tr>
<td>0.40</td>
<td>188</td>
<td>245.59</td>
<td>1.35</td>
</tr>
<tr>
<td>0.50</td>
<td>122</td>
<td>341.01</td>
<td>1.00</td>
</tr>
<tr>
<td>0.60</td>
<td>122</td>
<td>314.12</td>
<td>1.00</td>
</tr>
<tr>
<td>0.70</td>
<td>117</td>
<td>313.68</td>
<td>1.00</td>
</tr>
<tr>
<td>0.80</td>
<td>109</td>
<td>329.97</td>
<td>1.00</td>
</tr>
<tr>
<td>0.90</td>
<td>62</td>
<td>575.89</td>
<td>1.00</td>
</tr>
</tbody>
</table>

In Figure 6, we present the concept lattice after pruning using a stability threshold of 0.9, we were able to reduce our original concept lattice from 44765 formal concepts to 7492 formal concepts. In Figure 7, we provide a concept lattice that displays the formal concepts that contain the document with the id 157436. Document 157436 contains six topics (topic_40, topic_75, topic_44, topic_20, topic_31 and topic_69) with probabilities of at least 10% of appearing in this document.

**Biclustering.** BicPAMS [20] is applied as it combines state-of-the-art principles on pattern-based biclustering. BicPAMS is below used with default parameters: varying coherence strength ($\delta = \bar{A}/|\mathcal{L}|$ where $|\mathcal{L}| \in \{2, 3, 4, 5\}$), decreasing support until 100 dissimilar biclusters are found, up to 30% noisy elements, 0.05 significance level, and constant and order-preserving coherence assumptions. Two search iterations were considered by masking the biclusters discovered after the first iteration to ensure a more comprehensive exploration of the data space and a focus on less-trivial concepts. Topic-based frequency distributions were approximated, and the statistical tests proposed in [22] applied to compute the statistical significance of each concept.

Table 2 synthesizes the results produced by BicPAMS [20] on the preprocessed dataset. BicPAMS is able to efficiently find

12http://web.tecnico.ulisboa.pt/rmch/bicpams/
homogeneous, dissimilar and statistically significant concepts (subsets of topics with coherent predominance on a subset of documents). Illustrating, a total of 327 statistically significant concepts \((p\text{-value}<1)\) with constant coherence \((|\mathcal{Z}|=3)\) and an average of 112 supporting documents were found. These initial results show the impact of placing coherence assumptions and coherence strength criteria on concept analysis.

**Constant concepts.** Table 3 provides the details of four constant biclusters (their respective pattern, topics, coherence strength and statistical significance) using BicPAMS. Each bicluster shows a unique pattern of topic predominance. These results motivate the relevance of finding constant biclusters (their respective pattern, topics, coherence and strength criteria on concept analysis.

A closer analysis of the found biclusters further shows their robustness to the item-boundaries problem: topics with slightly deviating likelihoods from pattern expectations are not excluded. This allows the analysis of vector space models without the drawbacks of discrete views placed by FCA approaches.

**Order-preserving concepts.** Non-constant patterns are suggested if the focus is not on determining levels of performance but to assess the relative representativity among topics. BicPAMS [20] was applied to find such less-trivial yet relevant concepts. Table 3 details 4 order-preserving biclusters. Fig.9 visually depicts 2 of these concepts. Understandable, FCA is unable to recover such concepts given their flexible (yet meaningful) homogeneity criteria.

| Homogeneity | \(|\mathcal{Z}|\) | #clusters | average #rows | median #documents | \(p\text{-value}\) | \%most freq. pattern |
|-------------|-----------------|-----------|---------------|-------------------|----------------|---------------------|
| constant    | 2               | 121       | 647.62        | 0.00              | \(I = [0,0,0\{100\%\}]\) |
| constant    | 3               | 327       | 112.07        | 2.34e-152         | \(I = [0,0,0\{23\%\}]\) |
| constant    | 4               | 165       | 77.72         | 6.18e-122         | \(I = [1,0,0\{24\%\}]\) |
| constant    | 5               | 161       | 44.78         | 1.97e-74          | \(I = [0,0,0\{30\%\}]\) |
| order preserving | NA       | 163       | 201.66        | 0.99              | \(I = [7,13,5\{4\%\}]\) |

**Robustness.** Tolerance to noise can be customized to find concepts with desirable bounds on quality. In addition to noise tolerance, \(\eta_{ij}\), coherence strength, \(\delta=\overline{A}/|\mathcal{Z}|\), can be further explored to comprehensively model associations with slight-to-moderate deviations from expectations. Fig. 10 shows the impact of quality on the number of biclusters, average number of documents per bicluster and median \(p\)-values when BicPAMS is applied with constant coherence.

**Statistical Significance.** Table 1 shows the biclustering ability to find statistically significant concepts. A bicluster is statisti-
11 provides a scatter plot of the statistical significance and area...

This analysis suggests the presence of a soft correlation be-

Figure 9: Visuals of order-preserving concepts B1 and B4 (Table 3): chart-heatmap views.

Figure 10: Impact of the allowed noise tolerance in coherent concept analysis (BicPAMS under constant coherence and $\mathcal{G} \in \{2, 3, 4, 5\}$): number of concepts, average number documents per concept, and median $p$-value.

Figure 11: Statistical significance versus size of constant (up) ($\mathcal{G}=\{0.0,10.0,0.045,1.0\}$) and order-preserving biclusters (using statistical tests proposed in [22]).

**CONCLUSION**

This work addressed the problem of discovering knowledge from a collection of documents without background knowledge. Accordingly an unsupervised KDT process is proposed to support document categorization, 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 and a comprehensive pipeline was developed and tested. In this work we propose FCA and Biclustering as two parallel methods to identify and extract meaningful concepts to categorize the document collections and establish links between the documents. This work also proposes for the first time a comprehensive principles on how to apply biclustering for content categorization in large and heterogeneous text collections. Biclustering, a form of coherent concept analysis, is suggested to tackle the limitations of FCA since it explores all potentially relevant information available in vector spaces by focusing the searches on less-trivial, yet meaningful and statistically significant concepts. Pattern-based biclustering searches are suggested since they hold unique properties of interest: efficient exploration; optimality guarantees; discovery of concepts with parameterizable coherence; tolerance to noise and missing data; incorporation of domain knowledge; complete biclustering structures without positioning restrictions; and sound statistical testing. Results provide preliminary evidence of their relevance for document categorization and navigation within the legal document collection. Additionally the results also confirm the unique role of biclustering in finding relevant associations between topics and documents, providing enough feedback for content categorization in large text collections.

**Future work**

In future work, we aim to extend the proposed method to 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. In particular, we will attempt to validate the proposed pipeline on a corpus categorized by EUR-Lex 13. EUR-Lex is the official online journal of European Union (EU), which contains legal and other public documents of the EU member states. Second, we aim at developing a ranking metric to establish the most relevant documents in each concept. This creates the possibility of establishing weighted links between the documents in order to identify document priorities for the document navigation. In addition, we want to comprehensively evaluate the importance of using concept extraction methods to support the categorization of DRE near field experts in tasks of IR. Furthermore, we aim at exploring alternative merit functions and other homogeneity criteria in order to identify biclusters with other properties. Finally, we suggest the integration of word embeddings to represent the documents (ELMO embeddings [44]) with topic modeling [41].

**Scientific Contributions**

With the development of this work, we achieved the following scientific contributions:

- TPDL 2019 (Accepted): "An Unsupervised Method for Concept Association Analysis in Text Collections" [31];
- ECIR 2020 (Submitted): "Moving from formal towards coherent concept analysis: why, when and how. [30]."


<table>
<thead>
<tr>
<th>Ref.</th>
<th>Contributions</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>Structured Data Representation</td>
<td></td>
</tr>
<tr>
<td>[18]</td>
<td>Analysis on representation and preprocessing procedure for Portuguese legal documents</td>
<td>Overlook of syntactical and semantic information on the representation of documents</td>
</tr>
<tr>
<td>[51]</td>
<td>Performance comparison of different representations and hard versus soft clustering algorithms</td>
<td>Document categorization based on the overall content similarity</td>
</tr>
<tr>
<td>(2)</td>
<td>Phrases Document Modeling</td>
<td></td>
</tr>
<tr>
<td>[56]</td>
<td>Topic modeling integrated with $n$-grams models</td>
<td>Topic quality not compared with standard metrics (Coherence or Perplexity)</td>
</tr>
<tr>
<td>(3)</td>
<td>Topic Modeling</td>
<td></td>
</tr>
<tr>
<td>[46, 55]</td>
<td>Document categorization based on topic modeling techniques</td>
<td>Tests on small document collection where overall domains are known; hard clustering type results</td>
</tr>
<tr>
<td>[7]</td>
<td>Text segmentation method based on LSA</td>
<td>Text segmentation method tested over a well known dataset with 500 documents, where the major topics and concepts are well known, which facilitates the estimation of the LSA parameters.</td>
</tr>
<tr>
<td>(4)</td>
<td>Concept Analysis - a) Formal Concept Analysis</td>
<td></td>
</tr>
<tr>
<td>[5]</td>
<td>FCA as a topic modeling method; formal context reduction; external and internal topic quality metrics</td>
<td>No comparison with other well known topic modeling methods namely, HDP or LSA</td>
</tr>
<tr>
<td>[42]</td>
<td>FCA applied in the context of IR; Concept-based document categorization</td>
<td>Non-exhaustive exploration of different attributes for the documents in the formal context</td>
</tr>
<tr>
<td>(4)</td>
<td>Concept Analysis - b) Biclustering Concept Analysis</td>
<td></td>
</tr>
<tr>
<td>[11]</td>
<td>Comprehensive example of applying Biclustering in text mining</td>
<td>The proposed algorithm was tested on small set of documents, using binary relations between documents and terms</td>
</tr>
<tr>
<td>[12]</td>
<td>Exhaustive demonstration over different datasets of the potential applicability of Biclustering in text mining, using continuous relations between objects and attributes</td>
<td>Non-exhaustive analysis on the assessment of the quality on the obtained Biclusters</td>
</tr>
<tr>
<td>[21]</td>
<td>State-of-the-art biclustering algorithm (BicPAMS) able to extract biclusters with parameterizable coherence (including non-constant types such as additive and order-preserving types) and quality.</td>
<td>No clear evidence on the potential applicability of BicPAMS in textual data</td>
</tr>
<tr>
<td>(5)</td>
<td>Document Categorization</td>
<td></td>
</tr>
<tr>
<td>[29]</td>
<td>Overview on several clustering algorithms and their performance on document categorization</td>
<td>Fairly basic clustering algorithms</td>
</tr>
<tr>
<td>[8]</td>
<td>Cluster-based method to support document search in large collections of documents</td>
<td>No cluster quality measures; no comparison with other clustering methods</td>
</tr>
<tr>
<td>[40]</td>
<td>Document categorization based on text segments related to a specific topic</td>
<td>Non-exhaustive exploration of different methods for text segmentation</td>
</tr>
<tr>
<td>[24]</td>
<td>Improved K-means algorithm with POS approach</td>
<td>Hard clustering type results; non-exhaustive quality comparison with other clustering algorithms</td>
</tr>
<tr>
<td>[1]</td>
<td>KH algorithm applied for document categorization</td>
<td>The KH algorithm require a large number of iterations not suitable for large collections of documents</td>
</tr>
<tr>
<td>(6)</td>
<td>Document Navigation</td>
<td></td>
</tr>
<tr>
<td>[26]</td>
<td>Combining text retrieval and link analysis for KDT; exploring the concept of CCQ for topic associations</td>
<td>No clear evidence for the scalability of the proposed method; no explicit description on the tools used</td>
</tr>
<tr>
<td>[37]</td>
<td>New topic modeling methods based on the principle of BOC; exploring different preprocessing methods</td>
<td>No direct comparison of topic quality based on standard metrics</td>
</tr>
</tbody>
</table>
Figure 13: 120 topics extracted from the agriculture dataset using the LDA method, orange nodes correspond to the topics and the blue nodes to the corresponding words.

Figure 12: 120 topics extracted from the agriculture dataset using the LSA method, orange nodes correspond to the topics and the blue nodes to the corresponding words.

Figure 14: 120 topics extracted from the agriculture dataset using the HDP method, orange nodes correspond to the topics and the blue nodes to the corresponding words.