# Moving from formal towards coherent concept analysis: why, when and how

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.** Formal concept analysis has been largely applied to explore taxonomic relationships and derive ontologies from text collections. Despite its recognized relevance, it generally misses relevant concept associations and suffers from the need to learn from Boolean space models. Biclustering, the discovery of coherent concept associations (subsets of documents correlated on subsets of terms and topics), is here suggested to address the aforementioned problems. This work proposes a structured view on why, when and how to apply biclustering for concept analysis, a subject remaining largely unexplored up to date. Gathered results from a large text collection confirm the relevance of biclustering to find less-trivial, yet actionable and statistically significant concept associations.

**Keywords:** Concept Analysis · Biclustering · Topic Modeling · Unsupervised Knowledge Discovery · Large Digital Libraries

## 1 Introduction

Concept analysis is up to date the most referred unsupervised option for content categorization in large text collections [32]. A concept is an association between attributes (terms or topics) that is coherently verified in a subset of objects (documents). Concept analysis has been largely pursued to explore taxonomic relationships within a corpus, addressing the typical limitations that peer unsupervised approaches face in high-dimensional and sparse spaces [19]. Formal concept analysis (FCA) aims at finding, in Boolean data spaces, concepts as subsets of topics that co-occur in a subset of documents. FCA is the paradigmatic approach to concept analysis [11]. Despite its well-recognized relevance to derive ontologies for content categorization, FCA is hampered by major drawbacks. First, it imposes the selection of binarization thresholds to decide whether a topic is represented in a given document, making it vulnerable to subjective choices and to the item-boundaries problem [13]. As a result, FCA is unable to retain concepts sensitive to the varying predominance of topics in a given document, neglecting the rich nature of vector space models. Also, by focusing on dense regions, FCA neglects potentially relevant concepts, such as where specific topics have a preserved order of importance in a subset of documents [24].

Biclustering aims at finding coherent subspaces (subsets of attributes correlated in a subset of objects), which has been previously suggested for concept analysis in real-valued data spaces derived from text collections [8,5]. The use of biclustering for concept analysis is here termed *coherent concept analysis* (in contrast with formal concept analysis) since concepts are associations that satisfy specific homogeneity criteria of interest, therefore going beyond the strict Boolean formal view. Coherent concepts are sensitive to the predominance of each topic in a given document. In spite of its potentialities, existing research on biclustering text collections pursue specific forms of homogeneity [2,5], not offering a discussion on how different homogeneity and quality criteria affect concept analysis. In addition, existing research leaves aside current breakthroughs in the biclustering domain [12,16]. Finally, a fully structured view on why, when and how to apply biclustering in large text collections remains largely unexplored.

This work offers the first comprehensive view on the use of biclustering to explore large text collections in a fully automated and unsupervised manner, and further discusses its role for content categorization, retrieval, and navigation. The motivation is the need to support search and navigation in the official online publication of a national journal state, a digital library comprising all national laws, regulations and legal acts (NOTE: the real case is anonymized to comply with double blind review criteria; it will be revealed for publication).

This document is organized as follows. Section 2 provides essential background on concept analysis. Section 3 surveys relevant work on the topic. Section 4 discusses why, when and how to apply biclustering. Section 5 gathers results demonstrating the role of 5 biclustering in large text collections. Finally, concluding remarks and future directions are presented.

## 2 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 set using the classic term frequency-inverse document frequency (Tf-idf) metric [29]. Document similarity can be then computed using a loss function such as cosine distance.

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 operations to project data into a new data space along axes (eigenvectors $\alpha_k$) where data mostly vary [20], $w'_{ij}{=}\sum_k^m \alpha_k w_{ik}$. 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 [23];

- *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 [33], is currently a popular method for knowledge representation [19].

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*, $\mathfrak{B}_{(D,T,I)}$, 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).

**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 [24]. 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 [13]. 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 by an arbitrary number of (possibly overlapping) biclusters. Defs.1-2 formalize these concepts, and Fig. 1 illustrates them, contrasting coherent and formal concepts.

**Definition 1.** *Given a vector space model A, elements in a bicluster* $w_{ij} \in (I, J)$ *have coherence across documents (attributes) if* $w_{ij}$=$c_j$+$\gamma_i$+$\eta_{ij}$ ($w_{ij}$=$c_i$+$\gamma_j$+$\eta_{ij}$)*,* *where* $c_j$ *(or* $c_i$*) is the value of attribute* $t_j$ *(or document* $d_i$*),* $\gamma_i$ *(or* $\gamma_j$*) is the adjustment for document* $d_i$ *(or attribute* $y_j$*), and* $\eta_{ij}$ *is the noise factor of* $w_{ij}$*.*

*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* $\bar{A}$ *be the amplitude of values in* $A$*,* coherence strength *is a value* $\delta \in [0, \bar{A}]$ *such that* $w_{ij} = c_j + \gamma_i + \eta_{ij}$ *where* $\eta_{ij} \in [-\delta/2, \delta/2]$.

**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 ordering* $\pi$ *along the subset of attributes* $J$ *(documents* $I$*).*

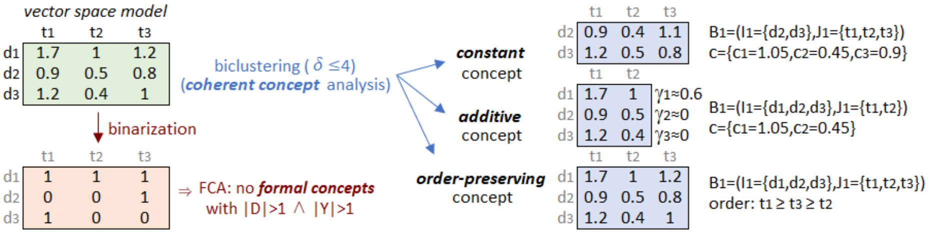

Fig. 1: Formal versus coherent concepts: biclustering with varying homogeneity criteria. Three coherent concepts were found under a constant, additive and order-preserving assumption (Defs.1-2), corresponding to a set of terms with coherent importance (in value, difference and order) on a set documents. Illustrating, $t_1 \geq t_2 \geq t_3$ permutation of terms' relevance is preserved along documents $\{d_1, d_2, d_3\}$. In contrast, no formal concepts were found on the given vector space.

*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 [17].

*Dissimilarity* criteria can be further placed to comprehensively cover the vector space with non-redundant biclusters [14].

## 3 Related work

***FCA in digital collections.*** FCA has been largely applied in Boolean space models given either by terms or (previously extracted) topics. In [4], a method is proposed, guided by both internal clustering quality metrics (Davies-Bouldin Index [7], Dunn Index [9], Silhouette coefficient [31] and The Calinski-Harabasz Index [21]) and external metrics (Reliability, Sensitivity and F-measure [1]). The experimental analysis used a collection of 2200 manually labeled tweets from 61 entities. The binary attributes are given by terms, named entities, references and URLs. A concept lattice is inferred using the Next Neighbours [3] algorithm. Each formal concept is here seen as a topic. Still, a large number of non-relevant topics is generated. The authors thus propose the Stability metric [22] to extract the most promising formal concepts, concluding that, if considering the external evaluation, FCA show a more homogeneous performance than the LDA and Hierarchic Agglomerative Clustering (HAC), with better overall results. Ignatov in [19] and Poelmans et al. in [28] present a survery on different contributions for FCA regarding several applications. Myat and Hla [25] developed a method for web document organization based on FCA. Cimiano et al. [6] presented an approach for the automatic extraction of concept hierarchies from text data. The authors modeled the context of a certain term as a vector representing syntactic dependencies that are automatically acquired from the text corpus with a linguistic parser, producing with the FCA a lattice of partial order that constitutes the concept hierarchy.

***Biclustering digital collections.*** Following the taxonomy of Madeira and Oliveira [24], biclustering algorithms can be categorized according to the pursued homogeneity and type of search. Hundreds of biclustering algorithms were proposed in the last decade, as shown by recent surveys [26, 10]. In recent years, a clearer understanding of the synergies between biclustering and pattern mining paved the rise for a new class of algorithms, referred to as pattern-based biclustering algorithms [13]. Pattern-based biclustering algorithms are inherently prepared to efficiently find exhaustive solutions of biclusters and offer the unprecedented possibility to affect their structure, coherency and quality [14, 12]. This behavior explains why this class of biclustering algorithms are receiving an increasing attention in recent years [13, 18]. BicPAMS [14] consistently combines such state-of-the-art contributions on pattern-based biclustering.

Castro et al. [5] developed BIC-aiNet, an immune-inspired biclustering approach for document categorization that was applied over Brazilian newspapers. Despite its relevance, it is limited to Boolean spaces (presence or absence of topics per document), sharing similar limitations to FCA. Dhillon [8] proposed the use of coclustering (a restrictive variant of the biclustering task that imposes a checkboard structure of biclusters [24]) to explore text collections. Coclustering was applied to vector space models with entries given by $w_{ij} \times log(\frac{n}{n_j})$, where $n$ is the number of documents and $n_j$ the number of statements 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 (exhaustive condition) and to a single concept only (exclusive condition), largely limiting the inherent flexibility of the biclustering task.

## 4 On why, when and how to apply biclustering

As surveyed, pattern-based biclustering approaches provide the unprecedented possibility to comprehensively find concepts in vector space models with parameterizable homogeneity and guarantees of statistical significance [14]. Despite their relevance, their use to explore digital collections remains largely unassessed. This section provides a structured view on why, when and how to bicluster text data.

### 4.1 On *WHY*

As motivated, coherent concept analysis should be considered to:

- avoid the drawbacks of formal concept analysis related with the need to specify thresholds and the item-boundaries problems [11];

- discover concepts in real-valued data spaces sensitive to the representativity of terms and topics per document;
- pursue concepts with desirable properties by parameterizing pattern-based biclustering searches [14] with the aimed coherence, quality, dissimilarity and statistical significance criteria.

Depending on the goal, one or more coherence assumptions (Def.1-2) can be pursued [13, 18]. The classic **constant coherence** can be placed to find groups of documents and topics, where each document has a similar probability to be described by a specific topic. Illustrating, documents $d_1$ and $d_2$ with $p(t_2, t_3, t_7|d_1)$= {0.32,0.90,0.49} and $p(t_2, t_3, t_7|d_2)$={0.29,0.88,0.55} are coherently related under a coherence strength $\delta$=0.1 (allowed deviations from expectations).

The notion of constant association is already a generalization over the traditional Boolean formal concept. Still, it can be further generalized to allow more flexible correlations. One paradigmatic example is the **order-preserving coherence** where a subset of topics have preserved orders of predominance on a subset of documents (Fig.1e). Illustrating, documents $d_1$ and $d_2$ with $p(t_2, t_3, t_7|d_1)$={0.32,0.50,0.47} and $p(t_2, t_3, t_7|d_2)$={0.29,0.97,0.55} are coherently related since they preserve the permutation $w_{i2} \leq w_{i3} \leq w_{i7}$.

Pattern-based biclustering [14] allows the discovery of these less-trivial yet coherent, meaningful and potentially relevant concepts.

### 4.2 On *WHEN*

Coherent concept analysis should be applied when:

- topic representativity matters. Recovering the introduced example, in contrast with coherent concept analysis, FCA under a binarization threshold $\theta$=0.1 is unable to differentiate $p(t_3|d_1)$=$w_{1,3}$=0.12 from $p(t_3|d_5)$=$w_{5,3}$=0.95;
- pursuing less-trivial forms of knowledge (including the introduced constant or order-preserving concepts);
- discretization drawbacks must be avoided;
- pursuing comprehensive solutions of concepts with diverse homogeneity and quality (noise-tolerance) criteria.

In contrast, coherent concept analysis should **not** be applied when:

- text collections are optimally represented as Boolean space models;
- extracting formal ontology structures [11]. Although pattern-based biclustering searches can also explore hierarchical relationships between biclusters, the resulting taxonomies are harder to interpret;
- the desirable binarization thresholds are known in advance and noise-tolerant FCA searches [27] can be applied to handle the noise associated with values near the boundaries of discretization.

### 4.3 On *HOW*

Pattern-based biclustering offers principles to find all potentially relevant concepts as they pursue multiple homogeneity criteria (including multiple coherence assumptions, coherence strength thresholds, and noise tolerance levels), and exhaustively yet efficiently explore different regions of the search space, preventing that regions with large concepts jeopardize the search [14]. As a result, less-trivial (yet coherent) topic associations are not neglected.

The possibility to allow deviations from value expectations (under limits defined by the placed coherence strength) tackles the item-boundaries problem.

Pattern-based biclustering does not require the input of support thresholds as it explores the search space at different supports [12], i.e. we do need to place expectations on the minimum number of documents per concept. Still, the minimum number of (dissimilar) concepts and topics per concept can be optionally inputted to guide the search. Dissimilarity criteria and condensed representations can be placed [14] to prevent redundant concepts.

**Statistical significance**. A sound statistical testing of concepts is key to guarantee the absence of spurious relations, and ensure concept relevance when categorizing contents and making other decisions. To this end, the statistical tests proposed in BSig [17] are suggested to minimize false positives (outputted concepts yet not statistically significant) without incurring on false negatives. This is done by approximating a null model of the target vector space and appropriately testing each bicluster in accordance with its underlying coherence.

**On robustness to noise and missing values.** Similarly to some FCA extensions, pattern-based biclustering can pursue biclusters with a parameterizable tolerance to noise [12]. This possibility ensures robustness to the algorithm-specific fluctuations on topic likelihood per document. Also, and similarly to general FCA approaches, pattern-based biclustering is robust to missing data as it allows the discovery of biclusters with an upper bound on the allowed amount of missing values [16]. This is particularly relevant to handle topic uncertainties.

**Other opportunities**. Additional benefits of pattern-based biclustering that can be carried towards concepts analysis include: 1) the possibility to remove uninformative elements in data to guarantee a focus, for instance, on coherent concepts with non-residual topic probabilities [16]; 2) incorporation of domain knowledge to guide the task in the presence of background metadata [15]; and 3) support classification and regression task in the presence of document annotations by guaranteeing the discriminative power of biclusters [13].

## 5 Results

To illustrate the enumerated potentialities of coherent concept analysis, results are gathered in four major steps. First, we introduce the pursued methodology and 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 concepts, providing a trustworthy means for concept analysis.

**Methodology**. The target forms of concept analysis should be preceded by the preprocessing of text collections to find a proper structured data representation of relevant topics, and succeeded by the statistical and domain-driven assessment of the found concepts, which then serve as basis to support categorization and navigation by linking documents with shared concepts.

*Dataset.* Over 35000 legal documents issued by state bodies in the domain of agriculture were extracted from the *Dirio da Repblica Eletrnico* (DRE), the official on-line publication journal of the Portuguese state. This collection has a total of 24018518 tokens (213868 unique tokens).

*Preprocessing.* Each document was pre-processed 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. Fig.2 depicts the word distribution of the documents before (green histogram) and after (blue histogram) preprocessing.

*Topic modeling.* We further used *Phrase*[3] to extract the combined words (phrasing) per document. From the obtained feature matrix, topics were extracted using LSA, LDA and HDP methods. Fig.3 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. 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.

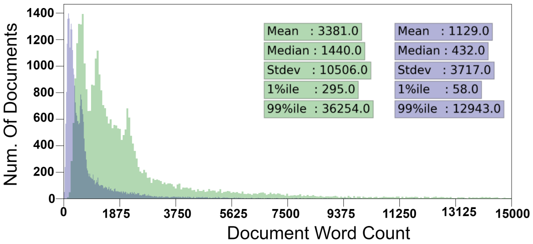

Fig. 2: Word count distribution over documents before and after preprocessing.

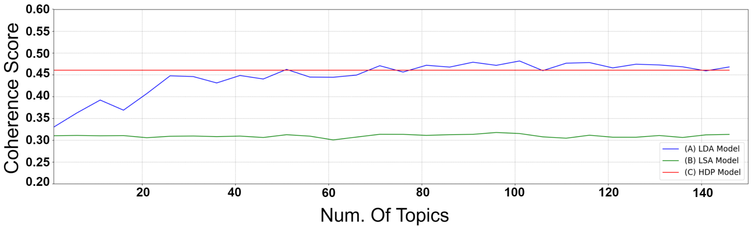

Fig. 3: Comparing topic modeling methods (LSA, LDA, HDP) w.r.t. coherence score.

[3] Automatic keyphrase extraction tool from Gensim: https://radimrehurek.com/gensim/

**Formal concept analysis**. Fig.4 applies FCA [11] to the preprocessed dataset – a vector space model with 35000 documents and 120 topics – under a variable binarization 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 [30]. 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 [17] reveal that only a small fraction of the returned concepts are statistically significant.

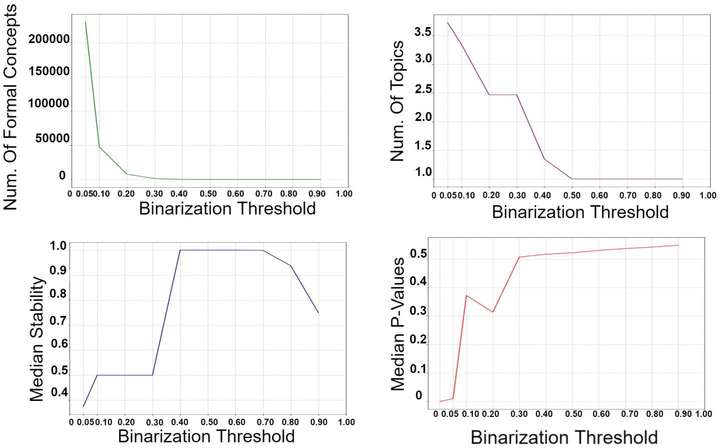

Fig. 4: FCA: binarization threshold impact on the: a) number of concepts, b) average number of topics per concept, c) solution stability, and d) median stat. significance.

**Coherent concept analysis**. BicPAMS [14] 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 [17] applied to compute the statistical significance of each concept.

Table 1 synthesizes the results produced by BicPAMS [14] on the preprocessed dataset. BicPAMS is able to efficiently find 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 ($p$-value<1) with constant coherence ($|\mathcal{L}|$=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 2 provides the details of four constant biclusters (their respective pattern, topics, coherence strength and statistical significance)

Table 1: Biclustering solutions found in DRE dataset using BicPAMS with varying homogeneity criteria.

| *Homogeneity* | $\|\mathcal{L}\|$ | *#biclusters* | *average #rows* | *median p*-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%) |

Table 2: Coherence concepts: zoom-in on 4 constant and 4 order-preserving concepts. For simplicity sake, the values of the concepts are presented in a discrete manner: $|\mathcal{L}|$ for constant coherence and 0 to 20 for order-preserving coherence. Illustrating, consider the constant concept $B_1$ with elements {2,0.5,1} for document $x_{3662117}$ in topics $\{t_{14}, t_{43}, t_{47}\}$: 0.5, 1 and 2 values correspond to topics with respectively residual, low and high probability to occur in $x_{3662117}$ document.

| Bicluster properties | Pattern | | | | Bicluster properties | Pattern | | | |
|---|---|---|---|---|---|---|---|---|---|
| $B_1$ (with $\|\mathcal{L}\|$=4) | 3662117 | 2.0 | 0.5 | 1.0 | $B_1$ | 2177091 | 4.5 | 8.5 | 1.5 |
| topics = [14,43,47] | 3384398 | 2.0 | 0.5 | 0.5 | topics = [37,76,93] | 293285 | 2.5 | 15.0 | 1.5 |
| #documents = 147 | | | … | | #documents = 299 | | | … | |
| $p$-value = 6.79e-170 | 979773 | 2.0 | 0.0 | 1.5 | $p$-value = 0.08 | 1181178 | 2.5 | 15.5 | 1.5 |
| | 1438820 | 2.0 | 0.0 | 1.0 | | 74661197 | 15.0 | 17.0 | 5.5 |
| | 3459557 | 1.5 | 0.0 | 0.5 | Order of difficulty: | 3189813 | 8.5 | 13.5 | 1.5 |
| | 75740163 | 2.0 | 0.0 | 1.5 | $t_{76} \geq t_{37} \geq t_{93}$ | 385434 | 12.5 | 15.0 | 1.5 |
| $B_2$ (with $\|\mathcal{L}\|$=3) | 235558 | 1.5 | 1.0 | 1.0 | $B_2$ | 453556 | 9.5 | 2.5 | 8.5 |
| topics = [76,103,118] | 1762073 | 2.0 | 1.5 | 1.0 | topics = [14,43,47] | 2806956 | 7.5 | 4.5 | 5.5 |
| #documents = 337 | | | … | | *#documents* = 290 | | | … | |
| $p$-value = 0 | 553876 | 2.0 | 1.0 | 1.0 | $p$-value = 0.21 | 494218 | 13.5 | 1.5 | 2.5 |
| | 632429 | 1.5 | 1.0 | 1.5 | | 75740163 | 15.5 | 1.5 | 12 |
| | 196216 | 2.0 | 1.5 | 1.5 | Order of difficulty: | 279258 | 8.5 | 1.5 | 1.5 |
| | 250617 | 2.0 | 1.0 | 1.5 | $t_{14} \geq t_{47} \geq t_{43}$ | 3551103 | 16.5 | 1.5 | 1.5 |
| $B_3$ (with $\|\mathcal{L}\|$=3) | 221325 | 1.5 | 2.0 | 1.0 | $B_3$ | 421452 | 9.5 | 2.5 | 9.5 |
| topics = [37,76,118] | 547844 | 1.5 | 1.5 | 1.0 | topics = [76,93,118] | 547844 | 8.5 | 1.5 | 5.5 |
| #documents = 363 | | | … | | *#documents* = 283 | | | … | |
| $p$-value = 0 | 572890 | 2.0 | 1.5 | 1.0 | $p$-value = 0.36 | 3189813 | 13.5 | 1.5 | 8.5 |
| | 3189813 | 1.5 | 2.0 | 1.5 | | 553876 | 13.5 | 1.5 | 1.5 |
| | 156660 | 2.0 | 1.5 | 1.0 | Order of difficulty: | 385434 | 15.0 | 1.5 | 2.5 |
| | 553876 | 2.0 | 2.0 | 1.0 | $t_{76} \geq t_{118} \geq t_{93}$ | 196216 | 17.5 | 2.5 | 7.5 |
| $B_4$ (with $\|\mathcal{L}\|$=4) | 361504 | 1.5 | 1.0 | 2.5 | $B_4$ | 3595682 | 15.0 | 4.5 | 8.5 |
| *topics* = [37, 46, 76] | 221325 | 1.5 | 1.0 | 2.5 | topics = [14,19,109] | 2806956 | 7.5 | 1.5 | 5.5 |
| #documents = 183 | | | … | | #documents = 186 | | | … | |
| $p$-value = 1.81e-252 | 168871 | 2.0 | 1.5 | 1.5 | $p$-value = 0.99 | 2645902 | 14.5 | 2.5 | 10.5 |
| | 512991 | 1.5 | 1.0 | 2.5 | | 67412614 | 17.5 | 2.5 | 10.5 |
| | 324968 | 1.0 | 1.0 | 2.5 | Order of difficulty: | 341633 | 17.0 | 1.5 | 1.5 |
| | 148432 | 1.5 | 1.5 | 3.0 | $t_{14} \geq t_{109} \geq t_{19}$ | 3551103 | 16.5 | 2.5 | 14.5 |

using BicPAMS. Each bicluster shows a unique pattern of topic predominance. Fig.5 visually depicts these concepts using line charts and heatmaps. Each line in the chart (and row in the heatmap) represents a document and the values (colors) show the representivity of its topics. These results motivate the relevance of finding constant concepts to group topics in accordance with their representivity in a document, a possibility neglected by FCA.

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.

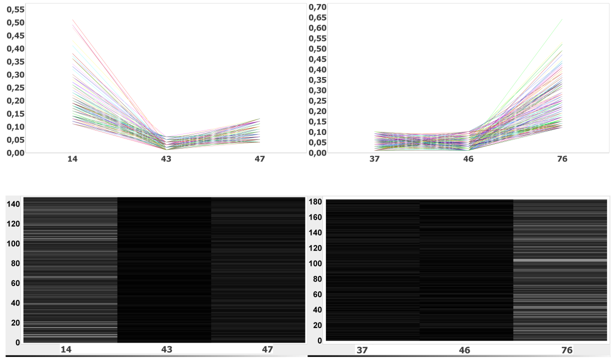

Fig. 5: Visuals of constant concepts **B1** and **B3** (Table 2): chart and heatmap views.

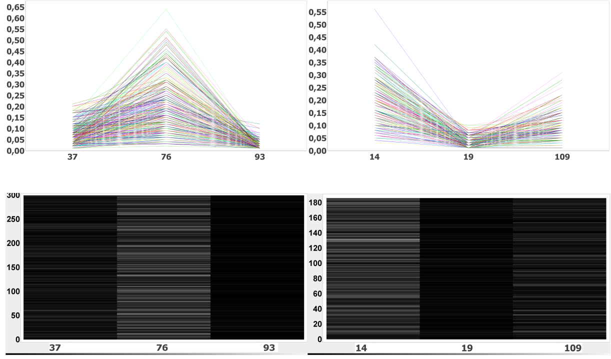

Fig. 6: Visuals of order-preserving concepts **B1** and **B4** (Table 2): chart-heatmap views.

**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 [14] was applied to find such less-trivial yet relevant concepts. Table 2 details 4 order-preserving biclusters. Fig.6 visually depicts 2 of these concepts. Understandable, FCA is unable to recover such concepts given their flexible (yet meaningful) homogeneity criteria.

**Robustness**. Tolerance to noise can be customized to find concepts with desirable bounds on quality. In addition to noise tolerance, $\eta_{ij}$, coherence strength, $\delta = \bar{A}/|\mathcal{L}|$, can be further explored to comprehensively model associations with slight-to-moderate deviations from expectations. Fig. 7 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 statistically significant if the number of documents with a given pattern or permutation of topics is unexpectedly low [17]. Fig. 8 provides a scatter plot of the statistical significance and area ($|I| \times |J|$) of constant ($|\mathcal{L}|$=3) and order-preserving biclusters. This analysis suggests the presence of a soft correlation between size and statistical significance. A few order-preserving concepts have low statistical significance (upper dots) and should therefore be discarded for not incorrectly bias decisions.

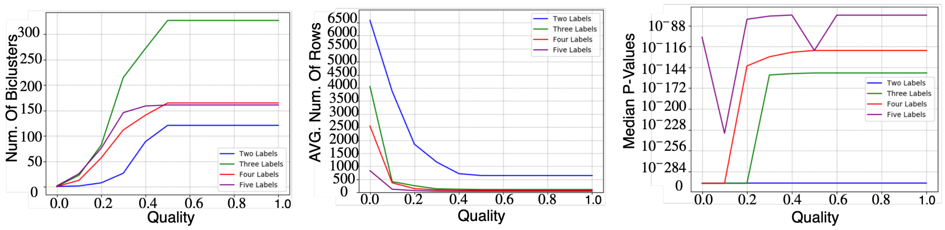

Fig. 7: Impact of the allowed noise tolerance in coherent concept analysis (BicPAMS under constant coherence and $\mathcal{L} \in \{2, 3, 4, 5\}$): number of concepts, average number documents per concept, and median $p$-value.

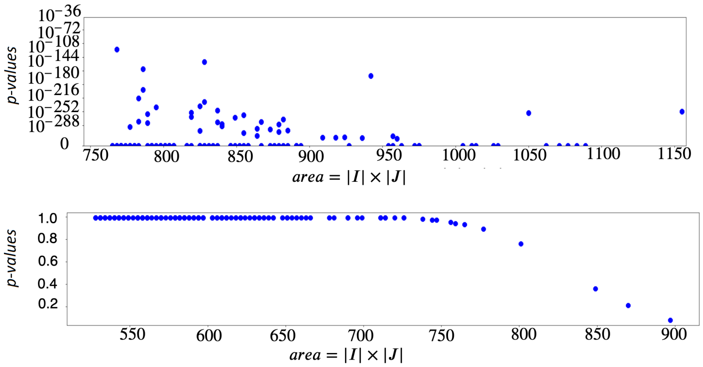

Fig. 8: Statistical significance *versus* size of constant (up) ($\mathcal{L}$={{0,0.10},{0.045,1}}) and order preserving (down) biclusters (using statistical tests proposed in [17]).

## 6 Concluding remarks

This work proposes 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 from a real corpus confirm the unique role of biclustering in finding relevant associations between topics and documents. Results further evidence the ability to unveil interpretable concepts with guarantees of statistical significance and robustness, thus providing a trustworthy context with enough feedback for content categorization in large text collections.

**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 references DSAIPA/DS/0111/2018 and UID/CEC/50021/2019.

## References

1. Amigó, E., Gonzalo, J., Verdejo, F.: A general evaluation measure for document organization tasks. In: Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. pp. 643–652. ACM (2013)
2. Banerjee, A., Dhillon, I., Ghosh, J., Merugu, S., Modha, D.S.: A generalized maximum entropy approach to bregman co-clustering and matrix approximation. In: ACM SIGKDD international conference on knowledge discovery and data mining. pp. 509–514. ACM (2004)
3. Carpineto, C., Romano, G.: Concept data analysis: Theory and applications. John Wiley & Sons (2004)
4. Castellanos, A., Cigarrán, J., García-Serrano, A.: Formal concept analysis for topic detection: a clustering quality experimental analysis. Information Systems **66**, 24–42 (2017)
5. de Castro, P.A., de França, F.O., Ferreira, H.M., Von Zuben, F.J.: Applying biclustering to text mining: an immune-inspired approach. In: Artificial immune systems, pp. 83–94. Springer (2007)
6. Cimiano, P., Hotho, A., Staab, S.: Learning concept hierarchies from text corpora using formal concept analysis. Journal of artificial intelligence research **24**, 305–339 (2005)
7. Davies, D.L., Bouldin, D.W.: A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence **PAMI-1**(2), 224–227 (1979)
8. Dhillon, I.S.: Co-clustering documents and words using bipartite spectral graph partitioning. In: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 269–274. ACM (2001)
9. Dunn, J.C.: Well-separated clusters and optimal fuzzy partitions. Journal of cybernetics **4**(1), 95–104 (1974)
10. Eren, K., Deveci, M., Küçüktunç, O., Çatalyürek, Ü.V.: A comparative analysis of biclustering algorithms for gene expression data. Briefings in bioinformatics **14**(3), 279–292 (2013)
11. Ganter, B., Wille, R.: Formal concept analysis: mathematical foundations. Springer Science & Business Media (2012)
12. Henriques, R., Madeira, S.: Bicpam: Pattern-based biclustering for biomedical data analysis. Alg. for Molecular Biology **9**(1), 27 (2014)
13. Henriques, R., Antunes, C., Madeira, S.C.: A structured view on pattern mining-based biclustering. Pattern Recognition **4**(12), 3941–3958 (2015)
14. Henriques, R., Ferreira, F.L., Madeira, S.C.: Bicpams: software for biological data analysis with pattern-based biclustering. BMC bioinformatics **18**(1), 82 (2017)
15. Henriques, R., Madeira, S.C.: Bic2pam: constraint-guided biclustering for biological data analysis with domain knowledge. Algorithms for Molecular Biology **11**(1), 23 (2016)
16. Henriques, R., Madeira, S.C.: Bicnet: Flexible module discovery in large-scale biological networks using biclustering. Algorithms for Molecular Biology **11**(1), 1–30 (2016)
17. Henriques, R., Madeira, S.C.: Bsig: evaluating the statistical significance of biclustering solutions. Data Mining and Knowledge Discovery **32**(1), 124–161 (2018)
18. Henriques, R., Madeira, S.C.: Triclustering algorithms for three-dimensional data analysis: A comprehensive survey. ACM Comput. Surv. **51**(5), 95:1–95:43 (Sep 2018)

19. 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)
20. Kalman, D.: A singularly valuable decomposition: the svd of a matrix. The college mathematics journal **27**(1), 2–23 (1996)
21. 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)
22. Kuzuetsov, S.: Stability as an estimate of thie degree of substantiation of hypotheses derived on the basis of operational, similarity (1990)
23. Landauer, T.K., Foltz, P.W., Laham, D.: An introduction to latent semantic analysis. Discourse processes **25**(2-3), 259–284 (1998)
24. Madeira, S.C., Oliveira, A.L.: Biclustering algorithms for biological data analysis: A survey. IEEE/ACM Transactions on Computational Biology and Bioinformatics **1**(1), 24–45 (2004)
25. 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)
26. Oghabian, A., Kilpinen, S., Hautaniemi, S., Czeizler, E.: Biclustering methods: biological relevance and application in gene expression analysis. PloS one **9**(3), e90801 (2014)
27. Pensa, R.G., Boulicaut, J.F.: Towards fault-tolerant formal concept analysis. In: Congress of the Italian Association for Artificial Intelligence. pp. 212–223. Springer (2005)
28. Poelmans, J., Kuznetsov, S.O., Ignatov, D.I., Dedene, G.: Formal concept analysis in knowledge processing: A survey on models and techniques. Expert systems with applications **40**(16), 6601–6623 (2013)
29. Rajaraman, A., Ullman, J.D.: Data Mining, p. 117. Cambridge University Press (2011)
30. Roth, C., Obiedkov, S., Kourie, D.: Towards concise representation for taxonomies of epistemic communities. In: International Conference on Concept Lattices and Their Applications. pp. 240–255. Springer (2006)
31. 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)
32. Tan, P.N.: Introduction to data mining. Pearson Education India (2018)
33. Wille, R.: Restructuring lattice theory: an approach based on hierarchies of concepts. In: Ordered sets, pp. 445–470. Springer (1982)