# Exploring Topic Variants Through an Hybrid Biclustering Approach

Nicolas Médoc, Mohammad Ghoniem, Mohamed Nadif

# Technical Report

## Visual exploration of topic variants through a hybrid biclustering approach

Nicolas Médoc Mohammad Ghoniem Mohamed Nadif

This paper is the full-text English translation of a research paper originally published in French in the proceedings of IHM 2016 conference. Please use the following Bibtex entry to cite it:

@inproceedings{Medoc:2016:VET:3004107.3004116,

**author** = {M{\'e}doc, Nicolas and Ghoniem, Mohammad and Nadif, Mohamed},
**title** = {Visual Exploration of Topic Variants Through a Hybrid Biclustering Approach},
**booktitle** = {Actes De La 28l\`{e}Me Conference Francophone Sur L'InteractionHomme-Machine},
**series** = {IHM '16},
**year** = {2016},
**isbn** = {978-1-4503-4243-8},
**location** = {Fribourg, Switzerland},
**pages** = {103--114},
**numpages** = {12},
**url** = {http://doi.acm.org /10.1145/3004107.3004116},
**doi** = {10.1145/3004107.3004116},
**acmid** = {3004116},
**publisher** = {ACM},
**address** = {New York, NY, USA},
**keywords** = {analytic journalism, biclustering, text visualization},

}

# Exploring Topic Variants Through an Hybrid Biclustering Approach

Nicolas Médoc, Mohammad Ghoniem, and Mohamed Nadif


Fig. 1. *Topic Weighted Map* visualization of 50 topics extracted from online news between Nov. $2^{nd}$ and Nov. $16^{th}$, 2015. Their size and proximity reflect their importance and similarity. Mouse hovering interaction displays 4 mutual links of proximity.

**Abstract**— In large text corpora, analytic journalists need to identify facts, verify them by locating corroborating documents and survey all related viewpoints. This requires them to make sense of document relationships at two levels of granularity: high-level topics and low-level topic variants. We propose a visual analytics software allowing analytic journalists to verify and refine hypotheses without having to read all documents. Our system relies on a hybrid biclustering approach. A new *Topic Weighted Map* visualization conveys all top-level topics reflecting their importance and their relative similarity. Then, coordinated multiple views allow to drill down into topic variants through an interactive term hierarchy visualization. Hence, the analyst can select, compare and filter out the subtle co-occurrences of terms shared by multiple documents to find interesting facts or stories. The usefulness of the tool is shown through a usage scenario and further assessed through a qualitative evaluation by an expert user.

**Index Terms**—Text visualization, biclustering, analytic journalism.

## 1 Introduction

Analytic journalists face a dilemma: while the number of sources of information increases dramatically, the time devoted to investigation is progressively reduced by editorial boards. We propose a visual analytics approach, built in collaboration with an analytic journalist, supporting the exploratory analysis of a corpus of free texts gathered during the journalist's back-grounding work. Even when the corpus includes thousands of documents, analytic journalists seek to be exhaustive without wasting time reading every document. Our solution aims to address this issue by supporting the analyst in carrying out two high-level tasks: 1) find the documents that verify a given hypothesis and, 2) identify new angles or viewpoints prompting him/her to refine or generate a new hypothesis that better fits the facts found in the data.

A challenge is to avoid the analyst to read all documents for identifying useful "assets" for the inquiry and for storytelling, e.g interesting facts, angles or viewpoints. A common approach consists in summarizing text corpora into multiple high-level *topics* [30, 16, 3]. A topic can be characterized by a set of terms associated to the documents where they occur. A collection of a few thousand documents will typically contain about 50 topics (see section 7.1). Yet, the analyst needs to quickly locate topics of interest and understand topic relationships. We hence propose a novel visual map of the corpus, termed a *Weighted Topic Map*. This visualization shows multiple tag-clouds nested in a treemap, a tag cloud for each topic, such that topic area and position reflect their relative size and similarity. Journalists also seek to identify multiple sources sharing common facts, angles or viewpoints. We hence propose an interactive topic visualization supporting the analysis of a selected topic by exploring the term co-occurrence relationships shared by subsets of documents. These relationships, termed as *topic variants*, are lists of terms shared by multiple documents. They represent distinct aspects of the selected *topic* and help journlalists focus their work on a shortlist of documents potentially holding useful assets for their inquiry. We adopt a multi-resolution approach alternating top-down analyses following the Visual Information-Seeking mantra [36], and bottom-up approaches by gradually linking items (terms and documents) through higher level relationships (*topic variants* and *topics*).

Biclustering methods [21, 31] applied on *Terms* × *Documents* matrices are designed with the duality between document vectors and term vectors in mind. The visual analytics tool described in this paper leverages the advantages of two nested structures resulting from two distinct biclustering methods (Figure 2). The first method is based on graph modularity [1] and produces diagonal biclusters which constitute high-level *topics*. The second method, `Bimax` [33], is an overlapping biclustering method. Applied to each matrix block constituting a *topic*, this algorithm reveals all term associations comprised

- *Nicolas Médoc is with University of Paris Descartes and Luxembourg Institute of Science and Technology. E-mail: nicolas.medoc@list.lu*
- *Mohamed Nadif is with University of Paris Descartes. E-mail: mohamed.nadif@mi.parisdescartes.fr.*
- *Mohammad Ghoniem is with Luxembourg Institute of Science and Technology. E-mail: mohammad.ghoniem@list.lu.*

*Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x.*
*For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org.*
*Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx/*

in the related set of documents. These term associations represents *Topic Variants*. This raises the challenge of exploring and interpreting the numerous `Bimax` biclusters which have many terms and documents in common. To address this challenge we designed a hierarchical visualization built on the basis of term overlaps. The analyst can hence explore the *topic variants* and their related documents from the most common terms close to the root, representing the topic in general terms, to the most specific terms closer to the leaf nodes where specific aspects of the topic are found. Coordinated multiple views provide the analyst with filtering, sorting and comparison interactions to identify informative terms and select relevant *topic variants*.

Our contribution is manifold: 1) We combine a diagonal biclustering method based on graph modularity with `Bimax`, a pattern-based overlapping biclustering method. 2) We propose coordinated multiple views to explore and understand numerous overlapping biclusters of documents and terms. 3) We propose the *Topic Weighted Map*, a novel topic visualization based on multiple tag clouds nested in a treemap reflecting topic size and relative similarity.

In the rest of this paper, we describe in section 2 the tasks and data under consideration. In section 3, we survey previous research in text visualization, biclustering methods and bicluster visualization as well as visual analytics solutions for text corpora. In section 4, we provide an overview of the software system relating each visual component to its supported tasks. Our hybrid biclustering approach is presented in section 5. We explain our visualization design in the section 6. We describe the evaluation method of the present work in section 7. Finally, we discuss the results and future work in section 8.

## 2 Tasks and data abstraction

We designed our system by following Munzner's visualization design methodology [32]. Below we characterize the problems, and the related tasks and data. Algorithms and visual encoding are detailed in their respective sections 5 and 6.

### 2.1 User needs and problem characterization

We started our analysis by conducting a semi-supervised interview with an analytic journalist. Our goal was to understand how she worked and to identify the problems she faced when dealing with a large corpus. In general, she expressed a sense of frustration because she lacked the time to be exhaustive in her investigations. She was often forced to reduce the number of documents aggregated during her back-grounding work. A tool that could summarize document content and extract all document relationships could clearly help her identify interesting facts without having to read every document. In complement, we referred to a manual for investigative journalism by Lee Hunter et al. [27], which states that "a hypothesis is a story and a method to test it". The manual proposes a methodology to help journalist work with hypotheses. Based on these two sources, we extracted a general workflow composed of three alternating processes:

- **Map:** the journalist gets an overview of the subject of inquiry;
- **Focus:** she focuses on a specific aspect to identify facts, viewpoints that support, refute or refine her hypothesis;
- **Diversify:** she widens her scope, looking for unexpected information and new analysis angles.

### 2.2 Detailed task abstraction

During our preliminary analysis we have identified the need to analyze a large text corpus with an exploratory tool supporting three high-level tasks divided into low-level tasks.

**T1** Summarize the corpus and identify topics of interest and aspects to investigate.
- **T1.1** Understand and locate topics of interest;
- **T1.2** Understand and identify related *topic variants*.

**T2** Find documents supporting an aspect of the working hypothesis.
- **T2.1** Keyword-based search of *topic variants*
- **T2.2** Compare topic variants
- **T2.3** Access textual content related to *topic variants*; show terms in their context.

**T3** Identify unexpected angles prompting the analyst to refine or generate new hypotheses that better fit the facts.
- **T3.1** Discover similar topics and their relationships.
- **T3.2** Suggest interesting terms to refine search queries.
- **T3.3** Suggest *topic variants* having shared documents/terms.

### 2.3 Data abstraction

*Vector Space Models* (*VSM*) use a *Term x Document* matrix to represent corpora. Each document is represented by a vector of distinct terms weighted by the *Term Frequency-Inverse Document Frequency* (*TF-IDF*) [37]. Below we introduce several notations used in the rest of this paper. Given $I$ a set of $n$ documents and $J$ a set of $m$ distinct terms, a *Term* $\times$ *Document* matrix is defined as $X = \{e_{ij}, i \in [1..n], j \in [1..m]\}$. With *TF-IDF* weights in the matrix cells, each entry $e_{ij}$ measures how much a term is representative of the documents yielding low values for empty words which are spread in many documents (*IDF*).We exploit this weighting scheme in both the analytical processing modules and the visualizations. Latent *Topics* and *Topic Variants* are modeled as term-document biclusters, defined formally in section 5.

## 3 Related Work

### 3.1 Visual analytics of text corpora

Several visual analytics tools devoted to corpus exploration are motivated by domain-specific tasks similar to ours. *Feature Lens* [13] supports the exploration of text corpora using text patterns such as frequent words or frequent itemsets of n-grams. Multiple sorting strategies reveal meaningful patterns that can be localized in the documents. Frequent patterns are laid out as large flat lists, without any overview of *topics*.

*Overview* by Brehmer et al. [7] is a visual analytics tool designed to help investigative journalists explore large text corpora. It uses hierarchical clustering to summarize large corpora and is designed to support hypothesis validation (**T2**) and generation (**T3**), like our tool. Expand/collapse interactions offer a good tradeoff between usability and cluster fidelity. But the semantics of document clusters are limited to their $N$ most frequent terms. Our approach focuses the analysis on a selected topic and supports the exploration of multiple angles and viewpoints with the precision and the exhaustiveness offered by the `Bimax` biclustering algorithm [33]. The tool of Alexander et al. [3] was designed to investigate large text corpora. Based on Latent Dirichlet Allocation (LDA) [5], coordinated multiple views and different sorting strategies support both top-down and bottom-up analysis fostering serendipitous discovery, and covering the hypothesis validation and generation tasks of Overview. Both tools address the focus and diversification processes of investigative journalists [27]. Our nested bicluster structure supports a multi-resolution exploration of text corpora, and helps identify *topic variants* within high-level topics.

### 3.2 Text and topic visualization

Text visualization is approached either by considering data features such as terms and documents, or by considering derived data like the topics extracted from the text. In the first category, projection techniques such as PCA or MDS are largely used to visualize clusters of documents with scatterplots, e.g. IN-SPIRE [47]. The study of Brehmer et al. [7] reveals that journalists prefer hierarchical navigation of document clusters to scatterplot representations. We adopt this approach to explore *topic variants*. To understand textual content or extracted topics without reading the documents, *Tag Clouds* [44] visually encodes term importance using their size or color. Many extensions of word clouds have integrated the temporal dimension [28] or a hierarchical structure [18]. *Parallel Clouds* [9] divides the space vertically in a limited number of *topics*, linking common terms horizontally.

In the second category, many visual analytics tools support topic exploration. Based on LDA [5] or its extensions such as hLDA [22] or HDP [43], *Tiara* [46], *ParallelTopics* [14], *TextFlow* [10] and *LeadLine* [15] show the temporal evolution of topics using a variant of

Theme River [25] where each layer represents a topic. *HierarchicalTopics* [16] proposes an adaptable hierarchy of topics to ease the exploration of numerous topics. Topic semantic is conveyed using the list of top N terms. *TopicPanorama* [30] gives an overview of topics, as long as they are common or specific to different sources. It is designed to summarize very large corpora and builds a hierarchy of topics to support multi-resolution exploration. While topic similarity is represented as a graph at each resolution level, only a subset of topics are labeled with two or three terms. *TopicPanorama* unravels a more exhaustive tag cloud on user interaction. Our *Topic Weighted Map* view supports a direct and exhaustive interpretation of all topics using multiple tag clouds, and reflects also their relative size and similarity. By combining two nested biclustering methods, our tool also supports a detailed exploration of *Topic Variants* captured by subtle document relationships based on term co-occurrences.

### 3.3 Biclustering methods and bicluster visualization

Most visual analytics tools for analysis of text corpora rely on topic modeling methods such as LDA [5] and Non-Negative Matrix Factorization (NMF) [29, 8]. Topic semantic is generally represented through the N most frequent terms. Yet, Alexander et al. [2] showed that a topic is more than its top 10 words. They state that the semantic quality of topics relies on the ability of models to reflect subtle term co-occurrence patterns.

Biclustering, also known as co-clustering, is widely used in bioinformatics [31, 21, 33] but may apply to other fields, e.g. text analysis or recommender systems. These methods apply to any matrix whose entries represent a relation between its rows and columns. While biclustering techniques take into account the duality of the relationship between documents and terms, i.e. multiple terms co-occur in a given document and multiple documents share a given term, LDA-based techniques rely on a complete generative model for the documents [35]. Our system uses biclustering to obtain homogeneous biclusters that group similar documents and their most representative terms. Such document-term biclusters can hence be used to model topics in text corpora. Our system combines two clustering methods. A hard biclustering method [1] based on graph modularity has shown better results with textual data than the spectral approach of Dhillon et al. [12]. We use it to extract topics. The *Topic Weighted Map* described in the section 6.1 encodes topic similarity as spatial proximity, and we overlay explicit links betwen topics to alleviate the limitations of hard biclustering as discussed in section 8. Prelić et al. [33] evaluate various biclustering methods for gene expression data. They propose `Bimax`, a pattern-based overlapping biclustering algorithm satisfying a constraint of maximal inclusion. We use it to extract all *Topic Variants* for a given *topic*.

`Bimax` generates a plethora of overlapping biclusters calling for a carefully designed Visual Analytics tool to explore them. In their survey of bicluster visualization techniques, Sun et al. [42] propose a design framework to represent five levels of database-like relationships: entities (1:1), groups (1:n), biclusters (n:m), chains (n:m:...:z) and the schema level. In this work we focus on the bicluster level. BicOverlapper [34] visualizes `Bimax` biclusters as a node-link diagram with a force-directed layout and transparent hulls. But in dense areas with numerous overlaps, node superposition and link crossings preclude the perception of common and distinctive parts of biclusters.

Matrix visualizations and parallel coordinates are more effective than node-link diagrams at the bicluster level [42]. Although matrix visualizations are less intuitive than node-link diagrams, they are often more readable for large datasets [20]. Many solutions in bioinformatics propose coordinated multiple views with both a rectangular heatmap and parallel coordinates [4, 26, 34]. They display the items of both dimensions in well separated rows and columns which may be seriated. While these approaches support the interpretation of individual biclusters, the linear arrangement of each dimension fails to convey a clear overview of all overlapping biclusters without splitting individual biclusters (Figure 2b) or duplicating items [26, 40]. Streit et al. [40] address these issues with a hybrid visualization where biclusters are represented by matrices which are linked through their common columns/rows. *Bixplorer* [17] proposes a similar approach for text data to explore chained relationships between different categories of named entities. Following a bottom up approach, this visualization works well on a small number of biclusters of interest. To provide a top-down analysis, BiSet [41] uses a view inspired by Jigsaw [39]. Chained relationships are represented by semantic edge bundles formed by biclusters grouping multiple items from adjacent axes. To support the exploration of term-document bicluster relationships, we organize biclusters in a hierarchy of terms based on term overlaps (see Figure 3). A coordinated rectangular heatmap visualization supports comparative analysis of a reduced selection of biclusters (see Figure 4).

## 4 System overview

The tool presented here supports topic exploration at multiple resolutions, from high-level topics down to raw text. It comprises four visual components that cover all tasks from **T1** to **T3**: the *Topic Weighted Map* in Figure 1, then in Figure 4 the *Topic Variant Overview* (3), the *Topic Variant Comparator* (4) and the *Document Detail View* (5). The journalist starts her work by inspecting the *Topic Weighted Map* to get an overview of all the topics (**T1.1**) extracted from the corpus using a diagonal biclustering algorithm. By selecting a topic of interest, *topic variants* are then captured by an overlapping biclustering algorithm and organized hierarchically with respect to their common terms in Figure 4 (3.1). The analyst can then explore all *topic variants* using a radial tree visualization such as a sunburst visualization (**T1.2**).This visualization is used to start the focus process aiming to verify specific aspects of the working hypothesis. The analyst can then filter the *topic variants* by keyword and hide uninteresting variants (**T2.1**). A subset of *topic variants* can then be chosen for further inspection in the *Topic Variant Comparator* in Figure 4 (4). This view shows a reduced selection of *topic variants* as columns in a matrix representation and depicts the distribution of their terms (4.1) and documents (4.2) (**T2.2**). Various sorting strategies (4.3) provide alternate insights, helping the journalist to find the most informative terms. Finally, the lowest level of detail (5) displays the raw textual content of the retained documents (**T2.3**), to understand the terms in their context. Obviously, documents remain the ultimate asset used by a journalist to prove her hypothesis.

The diversification process is supported using multiple interactions combined to promote serendipity. Starting with the *Topic Weighted Map*, the analyst may browse the vicinity of a topic of interest and follow the suggested links to adjacent topics to deepen the investigation (**T3.1**). Then, the term hierarchy visible in the *Topic Variant Overview* reveals meaningful terms suggesting new keyword combinations for search queries **T3.2**. Finally, *Topic Variants* are highlighted in the *Topic Variant Overview* when they share terms or documents explored in the *Topic Variant Comparator* or in the *Topic Variant Overview* itself (**T3.3**). The system presented here relies on a processing pipe line presented in the next section.

## 5 Processing pipe line

Our system relies on a nested structure built with an hybrid biclustering approach as presented in Figure 2. Our tool implements a Web architecture using Java, Scala and Python for the backend, and in Javascript (*D3.js*) for the client-side visualizations.

### 5.1 Text Processing

Raw text is first processed using Natural Language Processing components (*Stanford CoreNLP*). The traditional pipe line is processed up to the Part of Speech tagging to keep only nouns and verbs. We found them informative enough to support the interpretation and exploration of *topics* and *topic variants*. Then the *Terms*×*Documents* matrix is built with the *TF-IDF* weighting scheme. We keep the 10,000 terms with the highest weights. This number is configurable and the resulting reduced matrix is processed by a diagonal biclustering algorithm.

### 5.2 Topic extraction with diagonal biclustering

Topic extraction. Biclustering methods exploit the duality of terms and documents to reveal consistent patterns. These methods

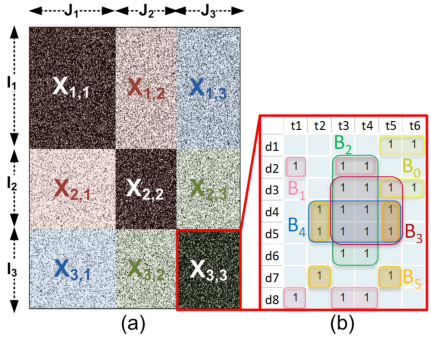


Fig. 2. (a) Diagonal biclusters (*topics*) and confusion blocs used by Equation 2, (b) overlapping biclusters (*topic variants*) built by `Bimax` for each topic, e.g. $X_{3,3}$.

group terms and documents in the same biclusters in such a way that the terms co-occur in the documents. Our *topic extraction* relies on a diagonal biclustering algorithm proposed by Ailem et al. [1]. This algorithm optimizes an objective function based on a graph modularity measure extended to incorporate simultaneously row and column partitions during the optimization process. The resulting biclusters are formally described as follows. For $K$, a given number of biclusters, and $X$, a matrix of size $n \times m$ with $I$ the set of $n$ documents and $J$ the set of $m$ terms, the $k^{th}$ bicluster ($\forall k \in [1..K]$) is described by the block $X_{k,k}$, the submatrix $I_k \times J_k$ where $I_k \subseteq I$ and $J_k \subseteq J$. The following constraint ensures the hard partitioning of rows and columns: $\forall k,l \in [1..K]$ with $k \neq l$, $I_k \cap I_l = \emptyset$ and $J_k \cap J_l = \emptyset$ For a given corpus, the number of topics to extract ($K$) is an important parameter that must be carefully chosen. To discover the best value for $K$, one approach consists in varying $K$ until the objective function of the algorithm reaches its optimum [1]. For each value of $K$, we perform a large number of tests (200 iterations) because the algorithm [1] is randomly initialized and may converge to a local optimum.

Topic relationships. To show topic relationships, we have to measure topic similarity. Hanczar and Nadif [24] use the following Jaccard-based metric to measure similarity between overlapping biclusters:

$$Sim(X_{k,k},X_{l,l}) = \frac{|X_{k,k} \cap X_{l,l}|}{|X_{k,k} \cup X_{l,l}|} = \frac{|X_{k,k} \cap X_{l,l}|_I + |X_{k,k} \cap X_{l,l}|_J}{|X_{k,k} \cup X_{l,l}|_I + |X_{k,k} \cup X_{l,l}|_J} \quad (1)$$

where $|.|_I$ is the cardinality in the row set (i.e. documents) and $|.|_J$ is the cardinality in the column set (i.e. terms). For diagonal biclusters, $|X_{k,k} \cap X_{l,l}|_I = 0$ and $|X_{k,k} \cap X_{l,l}|_J = 0$. This similarity measure must be adapted for diagonal biclusters, e.g. by including information captured in their adjacent blocks. The diagonal partition divides the matrix $X$ in $K \times K$ sub-matrices $X_{k,l}$ (Figure 2), $k \in [1..K]$ referring to the row partition and $l \in [1..K]$ referring to the column partition. Given $(X_{k,k},X_{l,l})$ a pair of diagonal biclusters, $X_{k,l}$ and $X_{l,k}$ constitute the confusion blocks that are likely to share rows or columns with either diagonal bicluster of the pair. So, we want to measure to which extent two diagonal biclusters share rows or columns with their confusion blocks. We adapted Equation 1 to compute the intersection between the biclusters and the confusion blocks row- and column-wise. Given $I_{k,l} = \{i \in I_k : \exists j \in J_l, e_{ij} \in X_{k,l}\}$ and $J_{k,l} = \{j \in J_l : \exists i \in I_k, e_{ij} \in X_{k,l}\}$ the respective non-empty sets of rows and columns for one confusion block, the new similarity metric $Sim'$ is described by Equation 2:

$$Sim'(X_{k,k},X_{l,l}) = \frac{|I_{k,k} \cap I_{k,l}| + |I_{l,l} \cap I_{l,k}| + |J_{k,k} \cap J_{l,k}| + |J_{l,l} \cap J_{k,l}|}{|I_{k,k} \cup I_{l,l}| + |J_{k,k} \cup J_{l,l}|} \quad (2)$$

We use equation (2) to build the similarity matrix of *topics* and visualize topic relationships and to compute the spatial proximity of topics in the *Topic Weighted Map* (section 6.1). The next processing steps build the structure for the second level of detail devoted to *Topic Variants*.

### 5.3 Topic Variants

Topic Variant extraction. A journalist verifies her hypotheses by looking for multiple sources that relate the same facts or stories. We assume that the aggregated corpus contains such repetitions. Extracting the co-occurrence relationships between documents using biclustering can, on the one hand, identify informative terms related to facts or stories, and on the other hand, locate multiple documents sharing them. Since terms and documents can be involved in multiple facts or stories, overlapping biclustering methods are fit for this purpose.

Prelić et al. [33] propose such an algorithm, `Bimax`, and a java implementation [4]. It is applied to binary matrices to identify blocks only composed of 1's (see Figure 2b). `Bimax` satisfies a constraint of maximal inclusion [33] to ensure that no bicluster can be covered by another one. Instead, each bicluster is extended row-wise and column-wise to its maximality. In a *Term*×*Document* matrix, `Bimax` extracts all distinct combinations of terms shared by multiple documents. The resulting biclusters, which we term *topic variants*, can be representative of facts, angles or viewpoints shared by multiple documents.

Human in the loop. While the diagonal blocks of the enclosing *topic* structure act as a dimensionality reduction mechanism, they come in various sizes and densities which can dramatically increase the number of `Bimax` biclusters. Since `Bimax` applies to a binary matrix, we defined a configurable binarization threshold on the *TF-IDF* weights. This threshold reduces not only the density of the enclosing *topic* block but also its dimensionality when vectors are entirely set to 0.

`Bimax` has three parameters: the minimum number of rows (*MinRows*) and the minimum number of columns (*MinCols*) belonging to each bicluster. By definition of bi-clusters, $MinRows \geq 2$ and $MinCols \geq 2$. Increasing the *MinCols* ignores the co-occurrence patterns with too few terms, which could be considered irrelevant. Increasing the *MinRows* ignores term co-occurrences found in too few documents. (*MaxBC*) is another parameter of `Bimax` used to set a maximum number of biclusters as a stopping condition. We observed in text data sets that the time performance drops drastically when the number of bi-clusters reaches 10,000 (default value for *MaxBC*). If this limit is reached, the result are hardly usable. The density of the topic matrix block must be reduced by increasing the binarization threshold to keep only the most meaningful terms. This set of parameters lets the analyst drive `Bimax` until interesting insights are found.

## 6 Visualization Design

### 6.1 Topic Weighted Map

To convey a bird's-eye view of the *topics* enclosed in the corpus, we design a novel visualization, termed *Topic Weighted Map* (Figure 1) which combines: 1) an MDS projection of *topics* in the 2D plane generating spatial coordinates for each topic used by 2) a *Weighted Map* visualization (a spatially consistent variant of treemaps) where each node contains 3) a word cloud. For instance, Figure 1 represents 50 *topics* extracted from online news between Nov. $2^{nd}$ and Nov. $16^{th}$, 2015. *Topic*/rectangle size is proportional to the number of terms and the number of documents it contains. *Topic* proximity in the 2D plane reflects *topic* similarity. Below we discuss the building blocks of the *Topic Weighted Map* visualization.

Topic word clouds. The diagonal biclustering step extracts *topics*, consisting of a set of terms that are semantically consistent with respect to a set of documents. Rather than displaying the *Top N* terms for a *topic*, we consider all its terms and show their relative importance within the *topic* using a tag cloud visualization. The size of a term is mapped to an interest criterion (*Interest(j)*) based on the sum of its *TF-IDF* weights across all documents belonging to the *topic*.

$$Interest(j) = \log(1 + \sum_{i \in I_k} e_{ij}), \forall j \in J_k \quad (3)$$

The log transformation is applied to better distinguish the differences in the lower end of the range. Color intensity is mapped to the number of documents containing the term in the *topic*, the darker the more

documents. This visual encoding helps the analyst to quickly identify interesting patterns like: a) contrasted terms which are typical of alternate viewpoints, e.g. those with a strong interest in few documents (large words with a light color); b) terms appearing in many documents, e.g. terms with a small font size and a dark color. We used the word cloud layout implemented by Davies in D3.js [11]

Topic Size and position in the map. We set topic weight to be the size of the corresponding bicluster (the product of the number of terms and the number of documents). An MDS projection of the topic similarity matrix defined in section 5.2 captures topic similarity. The resulting set of 2D coordinates and weights are used to generate a Weighted Map layout, a variant of treemaps proposed by Ghoniem et al. [19] for georeferenced hierarchical data.

In the resulting *Topic Weighted Map* visualization, each rectangle encloses a word cloud, such that rectangle size encodes topic importance and rectangle proximity encodes topic similarity. With this visualization, the analyst can get the broad picture of the topics contained in the corpus, locate interesting topics, and browse the vicinity of a topic to discover similar topics. Topic relationships are also shown on demand by overlaying links to the 5 most similar topics when the mouse hovers over a given topic. Link color encodes the strength of the relationship based on Jaccard similarity given by Equation (2). These links incite the analyst more actively to explore related topics. Because larger topics tend to "attract" smaller ones when using Jaccard similarity, we first filter the *Top 20* similar topics based on link reciprocity (both topics must appear in each other's *Top 20* similar topics). Then, we only show the *Top 5* among the remaining candidate neighbors. At this point, the analyst may click on a topic to drill down and explore all document relationships in the *Topic Variant Overview*. The amount of *topic variants* is shown at the top left corner of each *topic* rectangle.

### 6.2 Topic Variant Overview

*Topic Variants* capture document relationships as extracted by `Bimax` biclustering.Under the *Maximal Inclusion* constraint [33], a `Bimax` bicluster groups together a maximal and unique set of documents sharing a maximal and unique combination of terms. Hence, `Bimax` finds all optimal biclusters satisfying the exhaustiveness required by journalists. Yet, the exhaustiveness of `Bimax` finds an overwhelming number of biclusters that overlap each other with respect to some of the terms and documents they contain, causing major interpretation issues. To make sense of the large number of *topic variants* and their numerous overlaps, we designed an interactive hierarchical visualization (Figure 4). Indeed, the study of Dou et al. [16] shows that the hierarchical exploration of topics is more effective than flat list-based exploration. The hierarchy of clusters proved also to be attractive for journalists in the Overview system [7].

Term hierarchy. The meaning of a *topic variant* can mainly be interpreted by analyzing its terms. So we propose a hierarchy of terms that organizes *topics variants* (the biclusters) based on their overlaps (in Figure 3). We observe that a term having a high overlapping degree (i.e. appearing in a large number of variants) tends to have a generic meaning associated to the high-level topic at hand. Such a term (e.g. "Clinton, Obama, Israël, Netanyahu" in Figure 3) also appears in many documents gathered by the overlapping biclusters. It only makes sense to put these generic terms in the first levels of the hierarchy. Conversely, terms with a low overlap degree are more specific to some *Topic Variants* (e.g. "jewish, secretary") or can be exclusive to one variant. One expects such terms to appear deeper in the branches of the hierarchy.

The nodes of the hierarchy correspond to the terms of *topic variants* (biclusters). Each path in the hierarchy is a unique sequence of terms describing one *topic variant*, starting with the most common terms close to the root level and ending with the most distinctive terms in the leaves (**T2.2**). We build such a term hierarchy using the *FPTree* algorithm [23]. The terms of the enclosing *topic* are first sorted in the decreasing order of the overlapping degree of terms. In case of equality an alphabetical sort is applied to ensure a unique global order for each topic. Every bicluster is inserted in the tree, starting from the root, by maximizing the prefix commonality of every bicluster with regard to the global order of terms. The insertion of a new term in an existing path causes the creation of a new branch starting from the last matching term. Biclusters are placed on the leaf of their respective path (see Figure 3).

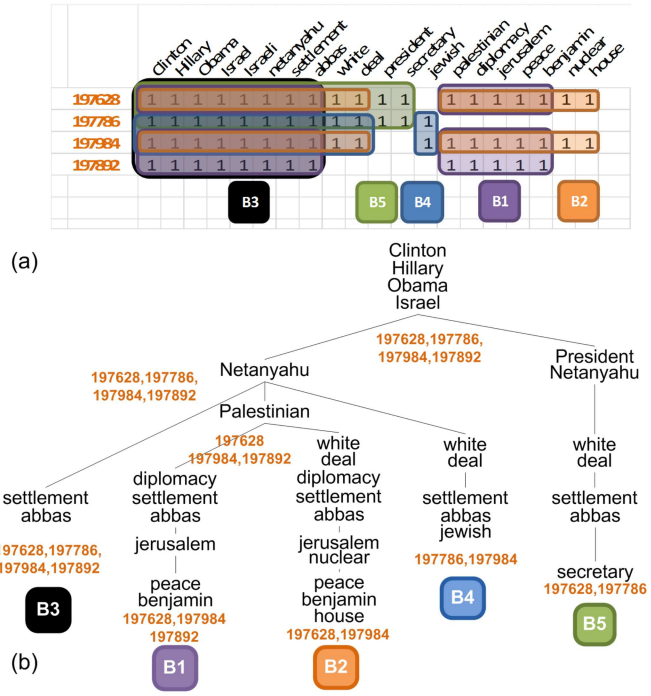


Fig. 3. a) Matrix model showing biclusters ($B1$ to $B5$) selected in the comparator (4) in Figure 4. b) The term hierarchy: each bicluster is inserted in the tree by matching common terms sorted by their overlapping degree (prefix numbers of term labels). Document ids appear in orange. The node *"palestinian"* belongs to two overlapping biclusters: $B_1$,$B2$ and the union of document sets given by these biclusters is *"197628,197984,197892"*. These documents share the whole sequence of terms along the path *"Clinton, ..., Israel, Netanyahu, Palestinian"*.

Nodes at each level of the hierarchy correspond to terms having different overlapping degrees. Terms from which multiple branches fan out can be viewed as articulation points that progressively guide the analyst to find meaningful *topic variants* (**T1.2**). Moreover, each node in the hierarchy selects all the documents of overlapping biclusters (**T2.3**). As the analyst moves deeper along the branches towards the leaves, the document set under consideration is gradually reduced and matches an increasing but more specific set of terms. This top-down exploration focuses the analysis progressively around a precise angle.

Interactive visualization. We represent the resulting term hierarchy using a *Sunburst* visualization [38] implemented with D3.js [6]. In Figure 4 (3.1), each branch represents the complete term sequence of one *topic variant*. As the analyst mouses over a node, the associated term appears as a tooltip and the complete term sequence corresponding to the node path is shown on the right hand side (3.2). The term is also highlighted across all branches (3.3) where it occurs to ease the comparison across *topic variants*. By clicking on a node, the matching documents are listed in the *Document Detail View* (5). A first goal is to find informative terms that can confirm or disprove the working hypothesis. A second goal is to suggest combinations of terms to help the analyst find unexpected viewpoints or to express queries.

To further support tasks **T1** to **T3**, we propose three interaction modes (see component (2) in Figure 4). 1) The *Filtering* mode shows in shades of blue the paths that contain all search query terms (**T1.2, T2.1**); the other nodes remain in gray shades. The *topic variants* that are filtered/grayed out can be hidden/unhidden on simple click. 2) The *Document distribution* mode allows to highlight in shades of orange

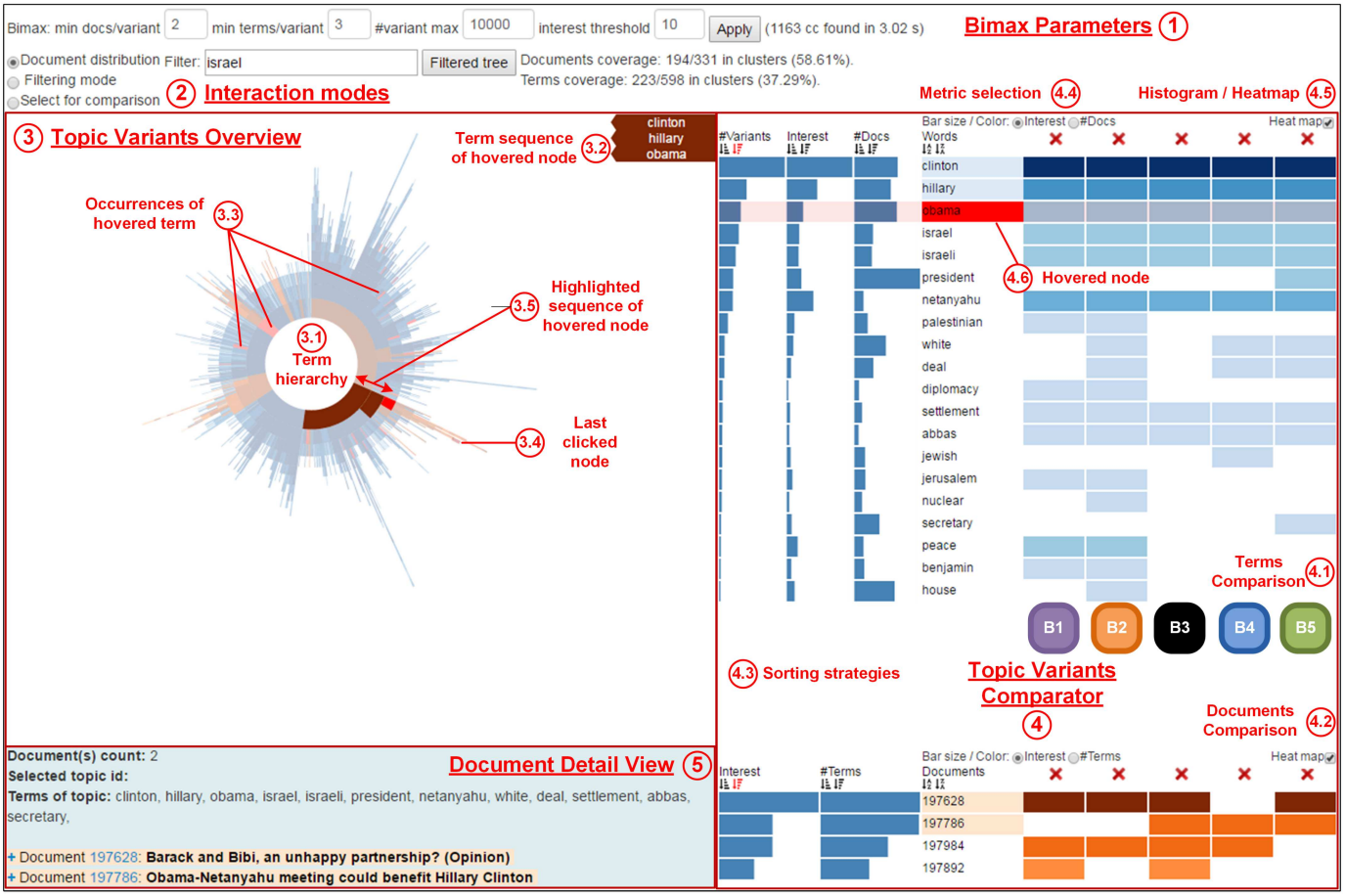


Fig. 4. Three components to explore the Topic Variants: *Topic Variant Overview*, *Topic Variants Comparator* and *Document Detail View*. Five *topic variants* have been selected to be compared.

all paths containing at least one of the documents associated with the clicked node. This points the analyst to new *topic variants* that share the selected documents (**T3.3**). These new *topics variants* can bring new documents revealing new angles, new facts or viewpoints shared with already known documents. 3) The *Select for comparison* mode selects the paths sent to the *Topic Variant Comparator* (**T2.2**).

### 6.3 Topic Variant Comparator

The *Topic Variant Comparator* offers a workspace in which the analyst can store and remove *topic variants* of his choice. Multiple interactions and sorting strategies help the analyst to find meaningful terms and documents (**T3.2**). In Figure 4, the *Topic Variant Comparator* displays *Topic Variants* as columns in a matrix visualization split in two parts. In the top (4.1), rows correspond to the terms occurring in the *Topic Variants*. In the bottom (4.2), they correspond to documents containing the *Topic Variants*. We keep the same color palettes as in the *Topic Variant Overview* for consistency (term-related information in blue and document information in orange). For each term three metrics are computed: (1) the bicluster overlapping degree (#Variants) reproduces term order in the branches of the hierarchy, (2) the degree of interest of terms is based on the *TF-IDF* weights and (3) the last metric is the number of documents where the term occurs within the topic (see section 6.1 for more details on the two last metrics). These metrics, shown on the left hand side as barcharts, support multiple term sorting strategies (4.3). We also provide an alphabetical sorting for fast term lookup. The user can also choose either of these metrics to display inside the matrix cells (4.4), either in a barchart mode or in heatmap mode (4.5). The *Topic Variant Comparator* and the *Topic Variant Overview* are coordinated: the hovered terms are colored in red in both (**T3.2, T3.3**) and their entire path highlighted in the *Topic Variant Overview* and colored in blue in the column *Words* in the *Topic Variant Comparator*. By clicking on a matrix cell, the selected documents are listed in the *Document View* (5) (**T2.3**).

The bottom part of the matrix (4.2) shows the documents distribution. Only two metrics are proposed: 1) an interest score based on the *TF-IDF* aggregated row-wise and 2) the number of terms in the topic variant for each document. In this part of the matrix, the interactions take place column-wise (i.e. *topic variant* wise). The titles of the selected documents are colored in orange in the list if they appear in the comparator. These interactions are designed to help the user compare *Topic Variants* (**T2.2**), identify terms and documents that are shared or specific (**T3.2, T3.3**).We also support the diversification process. The analyst can hover over an interesting term for further investigation, and identify new variants to add in the workspace (**T3.3**).

## 7 Evaluation

### 7.1 Usage scenario

This usage scenario shows how our tool is used to explore a large data set, and demonstrates its ability to generate, refine and verify hypotheses. The data set consists of 3,992 news articles aggregated from multiple online news sources (BBC, CNN, Reuters, France24, Egypt Independant and Der Spiegel) between November, $2^{nd}$ and November $16^{th}$ 2015. We extracted 50 *topics* ($K = 50$), depicted in the *Topic Weighted Map* in Figure 1. From 200 tests performed for each $k \in [10, 20, ..., 100, 150, ..., 500]$, the optimal graph modularity criterion of Ailem et al. [1] was obtained with $K = 50$.

On the left side in Figure 1, the large topics depict the hot news in the period covered by the news corpus (the crash of a Russian airplane in Egypt on the $31^{th}$ October 2015, the American presidential elec-

tions, the war in the Middle East, the immigrant crisis, and terrorist attacks in Paris). The topic related to the American elections draws our attention. The tag cloud contains large terms in dark blue such as *"president", "debate", "candidate", "Clinton", "Trump"*. It contains also large terms in lighter shades of blue such as *"Netanyahu", "Israel", "Palestinian"*. In Figure 4, this *topic* has 1,163 *topics variants*. Mousing over the nodes at the center of the hierarchy reveals the most shared terms among the *topic variants*: *"republican", "Clinton", "Rubio", "Isral", "Trump", "debate", "candidate"*. The terms are all clearly related to the US presidential campaign except "israel". We hypothesize that candidates discuss about Israel.

The *Document distribution* mode reveals in orange *topic variants* sharing documents with the clicked node. With the central node *"israel"*, this mechanism brings up *topic variants* sharing a common sequence of terms (*"Clinton-Hillary-Obama-Isral-israeli-Netanyahu"*) linking the US election to Israel. Next, we use the *Filtering* mode to focus the analysis on *topic variants* containing *"Isral"*. Among the remaining *topic variants* we send those containing *"Clinton"* to the *Topic Variant Comparator*. We chose to color encode the degree of interest in the heatmap matrix view ((4.4) in Figure 4) and to sort the terms by the number of enclosing variants (*#Variants* on top of barcharts). The terms shared by all variants helps focus on the topic: "clinton", "hillary", "obama", "israel", "israeli", "netanyahu", "settlement" and "abbas". In the middle, the *topic variant B3* groups all these shared terms and all documents selected in the comparator (4.2). By quickly reading the titles of the documents, we identify an event : "Netanyahu meets Obama at the White House". Two *topic variants* on the left (*B1* and *B2*) have specific terms: *"palestinian"*, *"diplomacy"* and *"peace"*. These variants group three documents relating difficult diplomacy between Obama and Netanyahu. The terms *"nuclear"* and *"deal"* of *B*2 refers to their divergence about the nuclear deal with Iran. On the right, two *topic variants* bring other specific terms: *"president,secretary"* (*B*5)) and *"jewish"* (*B*4). These two variants bring in a new document (4.2) from CNN (id=197786) titled "Obama-Netanyahu could benefit Hillary Clinton". The author anticipated that a successful meeting between Obama and Netanyahu could influence Jewish voters to vote for Hilary Clinton, the US Secretary of State, and candidate for president. This new document leads us to refine our hypothesis: "successful diplomacy between Obama and Netanyahu benefits the democratic candidates". This usage scenario shows the ability of our tool to drill down into a topic, generate and refine hypothesis, identify document relationships that reveal facts and stories while distinguishing multiple angles or viewpoints.

### 7.2 Qualitative evaluation with a domain expert

Our Visual Analytics tool was designed in collaboration with Warda Mohamed, a professional analytic journalist and editor at *Orient XXI*. She also writes for a number of French media, including *Le Monde diplomatique* and *Mediapart*. We refer to her as "the expert" in the sequel. We met the expert three times, for two to three hours each time. First, we conducted a semi-structured interview to understand the needs of analytic journalists and identify high-level tasks. During the second meeting we presented a first version of the *Topic Weighted Map* and *Topic Variant Overview* visualizations. We aimed to validate and refine our task definition and collect her feedback about our system. In the third meeting we carried out a two-fold qualitative evaluation. We made a voice recording of the entire meeting to thoroughly analyze expert feedback.

In the first part, we gave a 30-minutes live demo of the system using a small set of 9 documents corresponding to material she handpicked to prepare a previously published paper. The paper was investigating the blunders made by the French police. The aim was to confront our findings to hers and validate the tool with a form of ground truth that was known to the expert. We incited her to ask questions and comment the results shown by the tool. For a corpus of 9 documents the *Topic Weighted Map* was of little use. We moved quickly to the *Topic Variant Overview*. After a quick glance, the expert validated the relevance of the terms of the hierarchy saying: *"I know the subject and all interesting and important points do stand out"*. For instance, the term *"arm"* (the limb) refers to how Ali Ziri has been bent during his arrest. The term *"fugitive"* has widely been debated in the media. The term *"April"* corresponds to the month where Amine Bentounsi has been killed in 2012. This event led to debates between the two rounds of the French presidential election that followed shortly.

In the second part of the evaluation which lasted one hour and a half, she manipulated the visualizations with a larger corpus, the one used for our usage scenario in section 7.1. She manipulated the tool driven by a semi-directed interview to cover tasks **T1** to **T3**. At any time she could ask questions about the meaning of the visualizations or the way to carry out a particular task. We invited her to comment what she understood, what she found interesting, what the difficulties were or ways to improve the tool. The first assignment we gave the expert was to find two topics respectively about: football, astronomy, health/medicine and Asia'. For three of these themes she found the two topics in a few seconds ans was able to explain their differences. Concerning Asia, she found only one topic. We explain this by the fact that one of the two topics mixed news about China with news about refugees (the topic outlined in red border in Figure 1). Globally she enjoyed the *Topic Weighted Map* view: *"I like this tool, because 3,000 documents is a lot for me"*. *"even though there is some noise in the topics, there is always a link with the terms"*. She added: *"The color and position of terms within the topics make sense and the overlayed links are relevant."* She found it strange that the largest topics were stacked on the left hand side. We explain this phenomena by the large number of shared terms which increases their similarity and brings them close together in the *Topic Weighted Map*.

The expert decided to explore the topic about China and refugees more in depth. Some *topic variants* concern various European countries, but also Eritrea. She commented that: "*Eritrea is rather uncommon in the European refugee crisis. This shows that the* Topic Variants Overview *covers a broad range of detailed aspects. For instance, we spot the countries and the debated questions such as the Shengen Agreement, and asylum claims*". She complained about feeling lost with this visualization, even though she understood its benefits. The labeling interaction of the sunburst view needed to be improved, as discussed in section 8. To explain the relationship between China and the European refugee crisis, we invited the expert to use *Document distribution* mode. Clicking alternatively on the roots nodes "China" then "refugee", distinct parts of the hierarchy turn orange. But, clicking on the term "island", a majority of *topic variants* turns orange. Further scrutiny of the related documents revealed that the association of China and island refers to Taiwan, while the association refugee and island refers to the Greek islands at the forefront of the European refugee crisis. She then commented: "*We see that things are grouped from a certain angle that could be very narrow and could make sense or not. But it's good that the system shows these links, this stirs up curiosity*".

Next we asked her to identify meaningful *topic variants* through the *Topic Variant Comparator*. We noticed that she followed a repetitive analytic scheme. After manipulating the sorting strategies of the *Topic Variant Comparator*, she checked the meaning of terms in the context of the enclosing documents. She commented: "*It can save me a lot of time. Even with an unfamiliar topic, if I know that the first terms are the most relevant, I will look closely at the first ten terms only*." We also noticed that she did not spontaneously use the document comparison available in the bottom part of the *Topic Variant Comparator*. She finally explained that among numerous documents, there is lot of redundancies and she has to keep only the ones matching the core of the subject from her angle. She would then build a master file gathering all the material she will use to write her article. We believe that the document comparator view could be the precursor of this master file used by many journalists.

Finally she suggested the following improvements of our tool. First, the ability to save the workspace may support the analysis of the corpus from different angles and reopen the previous ones. Next, the exclusive assignment of terms to topics (due to hard partitioning) will oissbly lead someone who is unfamiliar with the subject to miss important aspects.

## 8 Discussion and future work

The usage scenario presented above shows that our visual analytics tool supports the exploration of multiple angles and view points shared by documents, and proves the feasibility of tasks **T1** to **T3** described in section 2.2. The qualitative evaluation is still preliminary. It relies on one expert user and a semi-structured interview. Yet, the participation of the expert at the begining of the design guaranteed a good characterization of the problem and tasks [32]. While the qualitative evaluation gave us an appreciation of the usefulness of our tool, it needs to be completed to reach more general conclusions.

The expert is indeed the sole judge of the importance and the novelty of the *topic variants* found with our tool. To evaluate the effectiveness of our tool to support hypothesis verification or generation, we face the issue that most journalists limit their corpora to a manageable number of documents. Without large corpora holding a known ground truth, we need to observe the long-term adoption of the tool by several journalists in their new enquiries, while avoiding the bias of the semi-directed approach. We plan to conduct quantitative studies with a significant number of participants to explore the feasibility of tasks **T1** to **T3** using our visualizations by comparing biclustering algorithms with `hLDA` [22]. We also aim to compare the effectiveness of the sunburst view to other tree visualizations such as Word Tree [45].

To draw an overview of *topic variants*, we need to choose between 1) avoiding duplication of redundant terms, by linking items with explicit links or contours (node-link diagrams and parallel coordinates [34], bipartite graphs [41]) which complicates the identification of biclusters due to node superposition and edge crossings; 2) avoiding node superposition and edge crossings (matrix visualization [26, 40]) which impedes the identification of common and distinctive terms. Our hierarchical approach is a trade-off that avoids occlusions and reduces term duplication. *Topic variants* can be outlined individually with their common and distinctive terms. The various interactions give multiple perspectives to appreciate the distribution of duplicated items, and the matrix view of the *Topic Variant Comparator* avoids term duplication by placing column-wise a small selection of *topic variants* of interest.

We have seen earlier in this paper that hard partitioning of data is problematic since terms and documents can belong to multiple topics. Probabilistic approaches providing overlapping partitions are difficult to apprehend by a lay audience [2]. We have alleviated this problem in the *Topic Weighted Map* by computing and overlaying topic relationships to support the diversification process in analytic journalism. In future work, we would like to further exploit topic overlap as captured by the confusion blocks (see section 5.2) to let the user shape new topics by merging partially/completely several topics.

Finally, the qualitative evaluation showed that the sunburst view may be difficult to use by journalists. One reason may be the on-demand labeling strategy we adopted. Even if the complete sequence of terms is displayed on the right hand, the user's focus is often near the mouse pointer. This complicates the whole interpretation of term sequences. We are currently investigating alternate labeling strategies better suited for multiple term sequences.

## 9 Conclusion

In this paper, we described a visual analytics tool supporting analytic journalists in dealing with large corpora. Our hybrid biclustering method supports multi-resolution analyses. To provide high-level topic overview, we designed a new visualization, the *Topic Weighted Map*. We proposed a new approach to explore overlapping *Term*×*Document* biclusters based on a term hierarchy.A qualitative evaluation showed that this term hierarchy, coordinated with the *Topic Variant Comparator*, supports the exploratory analysis of a large number of *topic variants* and finding useful facts or viewpoints corroborating facts and stories enclosed in the documents.


## Acknowledgments

The authors wish to thank Warda Mohamed for her contribution to the design and the evaluation of the system, and the anonymous reviewers who helped improve this paper.


## References


[1] M. Ailem, F. Role, and M. Nadif. Co-clustering Document-term Matrices by Direct Maximization of Graph Modularity. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 1807–1810, New York, NY, USA, 2015. ACM.

[2] E. Alexander and M. Gleicher. Task-Driven Comparison of Topic Models. *IEEE Transactions on Visualization and Computer Graphics*, PP(99):1–1, 2015.

[3] E. Alexander, J. Kohlmann, R. Valenza, M. Witmore, and M. Gleicher. Serendip: Topic model-driven visual exploration of text corpora. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 173–182, Oct. 2014.

[4] S. Barkow, S. Bleuler, A. Preli, P. Zimmermann, and E. Zitzler. BicAT: a biclustering analysis toolbox. *Bioinformatics*, 22(10):1282–1283, May 2006.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[6] M. Bostock. D3-Sunburst. http://bl.ocks.org/mbostock/4063423.

[7] M. Brehmer, S. Ingram, J. Stray, and T. Munzner. Overview: The Design, Adoption, and Analysis of a Visual Document Mining Tool for Investigative Journalists. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2271–2280, Dec. 2014.

[8] J. Choo, C. Lee, C. K. Reddy, and H. Park. UTOPIAN: User-Driven Topic Modeling Based on Interactive Nonnegative Matrix Factorization. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):1992–2001, Dec. 2013.

[9] C. Collins, F. Viegas, and M. Wattenberg. Parallel Tag Clouds to explore and analyze faceted text corpora. In *IEEE Symposium on Visual Analytics Science and Technology, 2009. VAST 2009*, pages 91–98, 2009.

[10] W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu, and X. Tong. TextFlow: Towards Better Understanding of Evolving Topics in Text. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2412–2421, Dec. 2011.

[11] J. Davies. D3-cloud. https://github.com/jasondavies/d3-cloud.

[12] I. S. Dhillon. 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*, KDD '01, pages 269–274, New York, NY, USA, 2001. ACM.

[13] A. Don, E. Zheleva, M. Gregory, S. Tarkan, L. Auvil, T. Clement, B. Shneiderman, and C. Plaisant. Discovering Interesting Usage Patterns in Text Collections: Integrating Text Mining with Visualization. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, pages 213–222, New York, NY, USA, 2007. ACM.

[14] W. Dou, X. Wang, R. Chang, and W. Ribarsky. ParallelTopics: A probabilistic approach to exploring document collections. In *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 231–240, 2011.

[15] W. Dou, X. Wang, D. Skau, W. Ribarsky, and M. Zhou. LeadLine: Interactive visual analysis of text data through event identification and exploration. In *2012 IEEE VAST*, pages 93–102, 2012.

[16] W. Dou, L. Yu, X. Wang, Z. Ma, and W. Ribarsky. HierarchicalTopics: Visually Exploring Large Text Collections Using Topic Hierarchies. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2002–2011, 2013.

[17] P. Fiaux, M. Sun, L. Bradel, C. North, N. Ramakrishnan, and A. Endert. Bixplorer: Visual Analytics with Biclusters. *Computer*, 46(8):90–94, Aug. 2013.

[18] P. Gambette and J. Vronis. Visualising a Text with a Tree Cloud. In H. Locarek-Junge and C. Weihs, editors, *Classification as a Tool for Research*, pages 561–569. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[19] M. Ghoniem, M. Cornil, B. Broeksema, M. Stefas, and B. Otjacques. Weighted maps: treemap visualization of geolocated quantitative data. In *IS&T/SPIE Electronic Imaging*, pages 93970G–93970G. International Society for Optics and Photonics, 2015.

[20] M. Ghoniem, J.-D. Fekete, and P. Castagliola. On the Readability of Graphs Using Node-Link and Matrix-Based Representations: A Controlled Experiment and Statistical Analysis. *Information Visualization*, 4(2):114–135, June 2005.

[21] G. Govaert and M. Nadif. *Co-Clustering*. ISTE & Wiley, series editor

edition, 2013.

[22] T. L. Griffiths, M. I. Jordan, J. B. Tenenbaum, and D. M. Blei. Hierarchical Topic Models and the Nested Chinese Restaurant Process. In S. Thrun, L. K. Saul, and B. Schlkopf, editors, *Advances in Neural Information Processing Systems 16*, pages 17–24. MIT Press, 2004.

[23] J. Han, J. Pei, Y. Yin, and R. Mao. Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach. *Data Mining and Knowledge Discovery*, 8(1):53–87, Jan. 2004.

[24] B. Hanczar and M. Nadif. Using the bagging approach for biclustering of gene expression data. *Neurocomputing*, 74(10):1595–1605, May 2011.

[25] S. Havre. ThemeRiver: visualizing thematic changes in large document collections. *Visualization and Computer Graphics, IEEE Transactions on*, 8(1):9–20, 2002.

[26] J. Heinrich, R. Seifert, M. Burch, and D. Weiskopf. BiCluster Viewer: A Visualization Tool for Analyzing Gene Expression Data. In G. Bebis, R. Boyle, B. Parvin, D. Koracin, S. Wang, K. Kyungnam, B. Benes, K. Moreland, C. Borst, S. DiVerdi, C. Yi-Jen, and J. Ming, editors, *Advances in Visual Computing*, number 6938 in Lecture Notes in Computer Science, pages 641–652. Springer Berlin Heidelberg, Sept. 2011. DOI: 10.1007/978-3-642-24028-7_59.

[27] M. L. Hunter, N. Hanson, S. Rana, L. Sengers, D. Sullivan, and P. Thordsen. *Story-Based Inquiry: A manual for investigative journalists. http://markleehunter.free.fr/documents/SBI_english.pdf*.

[28] B. Lee, N. Riche, A. Karlson, and S. Carpendale. SparkClouds: Visualizing Trends in Tag Clouds. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1182–1189, 2010.

[29] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[30] S. Liu, X. Wang, J. Chen, J. Zhu, and B. Guo. TopicPanorama: A full picture of relevant topics. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 183–192, Oct. 2014.

[31] S. C. Madeira and A. L. Oliveira. Biclustering Algorithms for Biological Data Analysis: A Survey. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 1(1):24–45, Jan. 2004.

[32] T. Munzner. A Nested Model for Visualization Design and Validation. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):921–928, Nov. 2009.

[33] A. Prelić, S. Bleuler, P. Zimmermann, A. Wille, P. Bhlmann, W. Gruissem, L. Hennig, L. Thiele, and E. Zitzler. A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, 22(9):1122–1129, May 2006.

[34] R. Santamaría, R. Thern, and L. Quintales. A visual analytics approach for understanding biclustering results from microarray data. *BMC Bioinformatics*, 9(1):247, May 2008.

[35] M. Shafiei and E. Milios. Latent Dirichlet Co-Clustering. In *Sixth International Conference on Data Mining, 2006. ICDM '06*, pages 542–551, Dec. 2006.

[36] B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *IEEE Symposium on Visual Languages, 1996. Proceedings*, pages 336–343, 1996.

[37] K. Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972.

[38] J. Stasko, R. Catrambone, M. Guzdial, and K. McDonald. An evaluation of space-filling information visualizations for depicting hierarchical structures. *International Journal of Human-Computer Studies*, 53(5):663–694, Nov. 2000.

[39] J. Stasko, C. Görg, and Z. Liu. Jigsaw: Supporting Investigative Analysis through Interactive Visualization. *Information Visualization*, 7(2):118–132, June 2008.

[40] M. Streit, S. Gratzl, M. Gillhofer, A. Mayr, A. Mitterecker, and S. Hochreiter. Furby: fuzzy force-directed bicluster visualization. *BMC Bioinformatics*, 15(Suppl 6):S4, May 2014.

[41] M. Sun, P. Mi, C. North, and N. Ramakrishnan. BiSet: Semantic Edge Bundling with Biclusters for Sensemaking. *IEEE Transactions on Visualization and Computer Graphics*, PP(99):1–1, 2015.

[42] M. Sun, C. North, and N. Ramakrishnan. A Five-Level Design Framework for Bicluster Visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1713–1722, Dec. 2014.

[43] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476):1566–1581, Dec. 2006.

[44] F. B. Viegas, M. Wattenberg, and J. Feinberg. Participatory Visualization with Wordle. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1137–1144, Nov. 2009.

[45] M. Wattenberg and F. B. Viégas. The word tree, an interactive visual concordance. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1221–1228, Nov. 2008.

[46] F. Wei, S. Liu, Y. Song, S. Pan, M. X. Zhou, W. Qian, L. Shi, L. Tan, and Q. Zhang. TIARA: a visual exploratory text analytic system. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pages 153–162, New York, NY, USA, 2010. ACM.

[47] J. Wise, J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow. Visualizing the non-visual: spatial analysis and interaction with information from text documents. In *Info. Vis.*, pages 51–58, Oct. 1995.