A visual analytics approach for multi-resolution and multi-model analysis of text corpora : application to investigative journalism

Nicolas Médoc

HAL Id: tel-02121464

https://theses.hal.science/tel-02121464

Submitted on 6 May 2019

École Doctorale Informatique, Télécommunications et Électronique
(EDITE de Paris)

Thesis

A Visual Analytics Approach for Multi-resolution and Multi-model Analysis of Text Corpora

Application to Investigative Journalism

By

Nicolas Médoc

This dissertation is submitted in fulfilment of the requirements for the degree of

Doctor of Computer Science

University of Paris Descartes

Supervisor:		
Mohamed Nadif	Professor	University of Paris Descartes
Co-supervisor:		
Mohammad Ghoniem	Senior researcher	Luxembourg Institute of Science and Technology
Examining committee:		
Pascale Kuntz	Professor	University of Nantes
Guy Melançon	Professor	University of Bordeaux
Pierre-François Marteau	Professor	University of Bretagne-Sud
Jean-Daniel Fekete	Research director	INRIA
Arnaud Mercier	Professor	University of Paris Panthéon Assas

École Doctorale Informatique, Télécommunications et Électronique
(EDITE de Paris)

Thèse

Une approche de visualisation analytique pour une analyse multi-résolution de corpus textuels

Application au journalisme d'investigation

Par

Nicolas Médoc

Thèse présentée en vue de l'obtention du grade de

Docteur en Informatique
Université de Paris Descartes

Directeur de thèse :		
Mohamed Nadif	Professeur	Université de Paris Descartes
Co-encadrant :		
Mohammad Ghoniem	Chercheur senior	Luxembourg Institute of Science and Technology
Rapporteurs :		
Pascale Kuntz	Professeur	Université de Nantes
Guy Melançon	Professeur	Université de Bordeaux
Examinateurs :		
Pierre-François Marteau	Professeur	Université de Bretagne-Sud
Jean-Daniel Fekete	Directeur de recherche	INRIA
Membre invité :		
Arnaud Mercier	Professeur	Université de Paris Panthéon Assas

À Christel,

Louis et Nathan,

et ceux sans qui ces mots n’existeraient pas,
Michèle et Jean.

Remerciements

Le travail présenté dans ce manuscrit est le fruit d'une collaboration étroite avec Mohamed Nadif, mon directeur de thèse, Professeur à l'Université de Paris Descartes, et Mohammad Ghoniem, chercheur senior au Luxembourg Institute of Science and Technology, qui a codirigé cette thèse.

Je tiens à remercier en premier lieu Mohamed Nadif pour la confiance et l'intérêt qu'il m'a accordés dès lors que je lui ai présenté mon projet de thèse. Son savoir, sa haute expertise en fouille de données ainsi que ses qualités pédagogiques ont contribué à la bonne orientation de cette thèse. Son soutien, sa générosité et sa disponibilité, malgré la distance, appellent toute ma reconnaissance. Ma reconnaissance va également à Mohammad Ghoniem qui m'a transmis avec patience et bienveillance sa connaissance dans le domaine de la visualisation analytique. Je lui dois notamment sa rigueur, son ambition et ses multiples relectures critiques qui furent essentielles à la réalisation de cette thèse. Ses mots d'encouragement m'ont amené, au cours de ces trois années, à rechercher et développer des compétences insoupçonnées pour mener à bien ce projet.

Je suis trés honnoré par l'intérêt que les membres de mon jury ont porté à mon travail. Je remercie notamment les rapporteurs, Pascale Kuntz, Professeur à l'Université de Nante, et Guy Melançon, Professeur à l'Université de Bordeaux, pour leur lecture et leurs remarques. Je remercie également Jean-Daniel Fekete, Directeur de recherche à L'INRIA, Pierre-François Marteau, Professeur à l'Université de Bretagne-Sud, et Arnaud Mercier, Professeur à l'Université de Paris Panthéon Assas, qui me font l'honneur d'évaluer ce travail.

Ce travail n'aurait pas été possible sans le soutien du Luxembourg Institute of Science and Technology. Je remercie en particulier Benoît Otjacques, responsable de l'unité eScience, pour son appui, ainsi que mes collègues d'eScience. Je remercie le LIPADE pour son accueil. Je porte notamment une pensée particulière à Melissa Ailem et Salah Aghiles qui ont chaleureusement partagé leur bureau lors de mes déplacements au LIPADE. Nos multiples échanges ont grandement enrichi mes recherches. Leurs courses administratives m'ont également fait économiser de nombreuses heures de transport ferroviaire.

J'adresse mes remerciements à Warda Mohamed, journaliste et rédactrice à Orient XXI, pour l'intérêt et le temps qu'elle a consacré à ce projet. Son expertise et son regard critique nous ont permis d'ajuster notre définition des tâches en cohérence avec le métier de journaliste. Par ses retours, nous avons recueilli de précieuses informations pour parfaire notre approche.

Enfin, je ne remercierai jamais assez Christel, Louis et Nathan qui, par leur patience et leur joie, ont su m'épauler et me faire lâcher prise quand il le fallait. Je remercie encore famille et amis pour leurs nombreux messages d'encouragement et de soutien.

Par maints exemples de mes parents, je retiens ceci : tout aussi ardu qu'il se présente, il n'est pas de chemin qui ne saurait être trouvé pour le bonheur de toucher au but.

Abstract

As the production of digital texts grows exponentially, a greater need to analyze text corpora arises in various domains of application, insofar as they constitute inexhaustible sources of shared information and knowledge. We therefore propose in this thesis a novel visual analytics approach for the analysis of text corpora, implemented for the real and concrete needs of investigative journalism. Motivated by the problems and tasks identified with a professional investigative journalist, visualizations and interactions are designed through a user-centered methodology involving the user during the whole development process. Specifically, investigative journalists formulate hypotheses and explore exhaustively the field under investigation in order to multiply sources showing pieces of evidence related to their working hypothesis. Carrying out such tasks in a large corpus is however a daunting endeavor and requires visual analytics software addressing several challenging research issues covered in this thesis.

First, the difficulty to make sense of a large text corpus lies in its unstructured nature. We resort to the Vector Space Model (VSM) and its strong relationship with the distributional hypothesis, leveraged by multiple text mining algorithms, to discover the latent semantic structure of the corpus. Topic models and biclustering methods are recognized to be well suited to the extraction of coarse-grained topics, i.e. groups of documents concerning similar topics, each one represented by a set of terms extracted from textual contents. We provide a new Weighted Topic Map visualization that conveys a broad overview of coarse-grained topics by allowing quick interpretation of contents through multiple tag clouds while depicting the topical structure such as the relative importance of topics and their semantic similarity.

Although the exploration of the coarse-grained topics helps locate topic of interest and its neighborhood, the identification of specific facts, viewpoints or angles related to events or stories requires finer level of structuration to represent *topic variants*. This nested structure, revealed by `Bimax`, a pattern-based overlapping biclustering algorithm, captures in biclusters the co-occurrences of terms shared by multiple documents and can disclose facts, viewpoints or angles related to events or stories. This thesis tackles issues related to the visualization of a large amount of overlapping biclusters by organizing term-document biclusters in a hierarchy that limits term redundancy and conveys their commonality and specificities. We evaluated the utility of our software through a usage scenario and a qualitative evaluation with an investigative journalist.

In addition, the co-occurrence patterns of *topic variants* revealed by `Bimax` are determined by the enclosing topical structure supplied by the coarse-grained topic extraction method which is run beforehand. Nonetheless, little guidance is found regarding the choice of the latter method and its impact on the exploration and comprehension of topics and *topic*

variants. Therefore we conducted both a numerical experiment and a controlled user experiment to compare two topic extraction methods, namely `Coclus`, a disjoint biclustering method, and hierarchical Latent Dirichlet Allocation (`hLDA`), an overlapping probabilistic topic model. The theoretical foundation of both methods is systematically analyzed by relating them to the distributional hypothesis. The numerical experiment provides statistical evidence of the difference between the resulting topical structure of both methods. The controlled experiment shows their impact on the comprehension of topic and topic variants, from analysts' perspective. We found that small topics extracted by `hLDA` contain richer vocabulary, shaping more interpretable topics and topic variants than small topics extracted by `Coclus`. Large `hLDA` topics yield themes with more heterogeneous stories and events in *topic variants* while `Coclus` reveals large topics with more specific and more separable events with *topic variants* eliciting slight differences. Such results allow us to recommend `hLDA` for topic discovery and hypothesis generation, and to recommend `Coclus` for hypothesis verification.

Finally we propose two contributions dealing with text streams. The first, in the visual analytics field, proposes dynamic and coordinated visualizations supporting situation awareness through a sliding time window. The second, in the data mining field, extends the `Bimax` algorithm to support data, typically allowing dynamic addition, update or removal of documents.

Résumé

À mesure que la production de textes numériques croît exponentiellement, un besoin grandissant d'analyser des corpus de textes se manifeste dans beaucoup de domaines d'application, tant ces corpus constituent des sources inépuisables d'information et de connaissance partagées. Ainsi proposons-nous dans cette thèse une nouvelle approche de visualisation analytique pour l'analyse de corpus textuels, mise en œuvre pour les besoins spécifiques du journalisme d'investigation. Motivées par les problèmes et les tâches identifiés avec une journaliste d'investigation professionnelle, les visualisations et les interactions ont été conçues suivant une méthodologie centrée utilisateur, impliquant l'utilisateur durant tout le processus de développement. En l'occurrence, les journalistes d'investigation formulent des hypothèses, explorent leur sujet d'investigation sous tous ses angles, à la recherche de sources multiples étayant leurs hypothèses de travail. La réalisation de ces tâches, très fastidieuse lorsque les corpus sont volumineux, requiert l'usage de logiciels de visualisation analytique se confrontant aux problématiques de recherche abordées dans cette thèse.

D'abord, la difficulté de donner du sens à un corpus textuel vient de sa nature non-structurée. Nous avons donc recours au modèle vectoriel et son lien étroit avec l'hypothèse distributionnelle, ainsi qu'aux algorithmes qui l'exploitent pour révéler la structure sémantique latente du corpus. Les modèles de sujets et les algorithmes de biclustering sont efficaces pour l'extraction de sujets de haut niveau. Ces derniers correspondent à des groupes de documents concernant des sujets similaires, chacun représenté par un ensemble de termes extraits des contenus textuels. Une telle structuration par sujet permet notamment de résumer un corpus et de faciliter son exploration. Nous proposons une nouvelle visualisation, une *carte pondérée des sujets*, qui dresse une vue d'ensemble des sujets de haut niveau. Elle permet d'une part d'interpréter rapidement les contenus grâce à de mutliples nuages de mots, et d'autre part, d'apprécier les propriétés des sujets telles que leur taille relative et leur proximité sémantique.

Bien que l'exploration des sujets de haut niveau aide à localiser des sujets d'intérêt ainsi que leur voisinage, l'identification de faits précis, de points de vue ou d'angles d'analyse, en lien avec un événement ou une histoire, nécessite un niveau de structuration plus fin pour représenter des *variantes de sujet*. Cette structure imbriquée révélée par `Bimax`, une méthode de biclustering basée sur des motifs avec chevauchement, capture au sein des biclusters les co-occurrences de termes partagés par des sous-ensembles de documents pouvant dévoiler des faits, des points de vue ou des angles associés à des événements ou des histoires communes. Cette thèse aborde les problèmes de visualisation de biclusters avec chevauchement en organisant les biclusters terme-document en une hiérarchie qui limite la redondance des termes et met en exergue les parties communes et distinctives des biclusters. Nous avons évalué l'utilité de notre logiciel d'abord par un scénario d'utilisation

doublé d'une évaluation qualitative avec une journaliste d'investigation.

En outre, les motifs de co-occurrence des *variantes de sujet* révélées par Bimax sont déterminés par la structure de sujet englobante fournie par une méthode d'extraction de sujet. Cependant, la communauté a peu de recul quant au choix de la méthode et son impact sur l'exploration et l'interprétation des sujets et de ses variantes. Ainsi nous avons conduit une expérience computationnelle et une expérience utilisateur controlée afin de comparer deux méthodes d'extraction de sujet. D'un coté Coclus est une méthode de biclustering disjointe, et de l'autre, hirarchical Latent Dirichlet Allocation (hLDA) est un modèle de sujet probabiliste dont les distributions de probabilité forment une structure de bicluster avec chevauchement. Les fondements théoriques des deux méthodes sont analysés systématiquement en établissant leur lien avec l'hypothèse distributionnelle. L'expérience computationnelle fournit des preuves statistiques des différences structurelles des sujets obtenus par les deux méthodes. L'expérience controlée fait apparaître des différences concernant l'impact des méthodes sur la compréhension des sujets et des variantes, du point de vue de l'analyste. Nous avons identifié que les petits sujets livrés par hLDA contiennent un vocabulaire plus riche, formant des sujets et des variantes plus compréhensibles que les petits sujets construits par Coclus. Les grands sujets de hLDA forment des thématiques avec des variantes de sujets concernant des histoires et des événements plus hétérogènes, alors que Coclus révèle de grands sujets contenant des événements plus spécifiques et séparés plus nettement, avec des différences plus fines entre les variantes. De tels résultats nous permettent de recommander hLDA pour des tâches de découverte de sujets et de construction d'hypothèses, et de recommander Coclus pour la vérification d'hypothèses.

Enfin, nous présentons deux contributions traitant des flux textuels. La première, dans le domaine de la visualisation analytique, propose des visualisations dynamiques et coordonnées dédiées à la surveillance de la situation courante à travers une fenêtre temporelle glissante. La seconde, dans le domaine de la fouille de données, propose une extension de Bimax qui supporte des données dynamiques, comportant notamment les opérations d'ajout, de suppression ou de mise à jour de documents.

Contents

General Introduction 1

I.1 Context . . . 1

I.2 Challenges and problematics . . . 2

I.3 Contributions . . . 5

I.4 Outline . . . 7

I.5 Notations . . . 9

I State of the Art 11

1 Text Mining and Clustering 13

1.1 Text Models . . . 13

1.2 Topic Models . . . 16

1.3 Clustering methods . . . 18

1.4 Biclustering . . . 21

1.4.1 Generalities . . . 21

1.4.2 Biclustering techniques for topic extraction . . . 22

1.4.3 From coarse-grained topics to fine-grained *topic variants* . . . 25

1.4.4 Pattern-based biclustering for *topic variant* extraction . . . 25

1.5 Topic Models vs. Biclustering . . . 27

1.6 Evaluation of clustering and topic models . . . 29

2 Visual Analytics of text corpora 31

2.1 Feature-based visualizations . . . 32

2.1.1 Document-centric approach . . . 33

2.1.2 Term-centric approach . . . 34

2.2 Topic-based visualization . . . 37

2.2.1 Temporal evolution of topics . . . 37

2.2.2 Exploring semantic relationships . . . 42

2.3 Model-driven design vs. domain-specific design . . . 46

2.3.1 Model-driven design: from coordinated multiple views to interactive clustering . . . 46

2.3.2 Domain-specific visual analysis of text corpora . . . 49

2.4 Bicluster Visualization . . . 52

II Visual Analytics System 57

3 Method and model 59

3.1 Introduction . . . 59

3.2 User centered approach . . . 62

3.2.1 Munzner's Nested Model . . . 62

3.2.2 Data and problem characterization . . . 63

3.2.3 Data and operation abstraction . . . 64

3.3 Multi-model and multi-resolution biclustering approach . . . 66

3.3.1 Text Processing . . . 66

3.3.2 Coarse-grained Topics . . . 66

3.3.3 Fine-grained Topic Variants . . . 71

3.4 Evaluation . . . 74

3.4.1 Protocol . . . 75

3.4.2 Dataset . . . 76

3.4.3 Analysis . . . 77

3.5 Discussion . . . 81

3.6 Conclusion . . . 82

4 Multi-resolution Visual Analysis of Text Corpus 83

4.1 Introduction . . . 83

4.2 System overview . . . 86

4.2.1 Workflow and tasks abstraction . . . 86

4.2.2 Four visual components . . . 87

4.3 Visualization Design . . . 89

4.3.1 Topic Weighted Map . . . 89

4.3.2 Topic Variant Overview . . . 91

4.3.3 Human in the loop . . . 94

4.3.4 Topic Variant Comparator . . . 96

4.4 Evaluation . . . 99

4.4.1 Usage scenario . . . 99

4.4.2 Qualitative evaluation with a domain expert . . . 101

4.4.3 Discussion and future work . . . 109

4.5 User study . . . 110

4.5.1 Properties of topic extraction methods . . . 110

4.5.2 Experimental protocol . . . 117

4.5.3 Results . . . 119

4.5.4 Discussion . . . 123

4.6 Conclusion 124

III Towards Visual Analytics of text streams 127

5 Situation Awareness and Real Time Exploration of Text Streams 129

5.1 Introduction 129

5.2 Related Work 130

5.3 Architecture and Models 131

5.3.1 Event-driven Web-oriented Architecture 131

5.3.2 Text Processing 133

5.3.3 Model for Multiple Dynamic Frequency Matrices 133

5.4 Visualization Design 134

5.4.1 Temporal View 134

5.4.2 Map View 136

5.5 Case Study 136

5.5.1 Data and Tasks 136

5.5.2 The Case-Study 138

5.6 Discussion and future work 142

5.6.1 Discussion 142

5.6.2 Challenges and Future Work 144

5.7 Conclusion 145

6 Dynamic Bimax 147

6.1 Introduction 147

6.2 Dynamic Bimax Algorithm 147

6.3 Proof of condition validity after and before remove 150

6.4 Performance evaluation 151

6.5 Conclusion 153

Conclusion and perspectives 155

Liste des publications 159

Bibliography 161

List of Figures

1.1 The pre-processing pipeline for modeling a text corpus in the Vector Space Model representation. 14

1.2 Example of term co-occurrences of different order. The number at each cell of the matrices corresponds to the number of times the term t_i occurs in the document d_j. 15

1.3 A topic hierarchy built with hLDA. The *nCRP* chooses a node β_k in the tree for each document x_i. If needed, a new path is created from the chosen node to reach depth L. 18

1.4 (a) Bimax biclusters in a term-document matrix. No bicluster is completely covered by another one. (b) Intermediary state of Bimax after the addition of four documents. 26

2.1 The “Galaxy View” of IN-SPIRE [Wise 1995] spatializes documents in a scaterplot where the proximity of the objects represents the similarity of documents. Several clusters of documents can be depicted and some of them are directly interpretable through the three most representative terms. . 33

2.2 A tag cloud created with Wordle [Viegas 2009]. The visual variables such as label size and color can be mapped to metrics such as word frequency or TF-IDF. The orientation of terms can be fixed horizontally to ensure a good readability. 34

2.3 Sparklines of SparkCloud [Lee 2010]. The terms are displayed with a line plot representing the evolution of their importance over time. 34

2.4 The Word Tree visualization [Wattenberg 2008] allows to retrieve all sentences containing a searched keyword and displays in a tree the divergence of the subsequent fragment of sentences. 35

2.5 The TreeCloud visualization based on phylogenetic tree [Gambette 2010] show semantic relationships between terms. The internal nodes representing semantic divergences are not labeled. 36

2.6 The Parallel Tag Cloud visualization [Collins 2009] where each column corresponds to a facet of data. Common terms are explicitly linked to reveal semantic relationships between facets. 36

2.7 The original Theme River visualization [Havre 2002]. Temporal alignment of topics allows to compare evolution of topics relatively to each other. The labels are however limited to one term. 38

2.8 NewsLab, a multi-scale theme river with a burstiness-based filtering and a replay animation [Ghoniem 2007]. The NewsRings view reveals better emergence, disapearence patterns as well as periodic patterns. 38

2.9 Tiara visualization with optimized placement of tag clouds [Wei 2010]. Thin layers have not enough space for labeling. 39

2.10 TextFlow visualization with glyphs representing topic merges and splits in a river metaphor [Cui 2011]. Topic labeling is provided on user interaction in order to prevent visual clutter. 40

2.11 LeadLine for event detection with named entities [Dou 2012]. Each event aspect is depicted in coordinated multiple views. For instance, topics are represented through an EventRiver [Luo 2012], a node-link diagram and a geographical map represent person relationships and location names respectively. 40

2.12 HierarchicalTopics [Dou 2013] allows a hierarchical exploration of topics while depicting temporal patterns. Topics are interpreted through flat lists containing the N most representative terms. 41

2.13 TopicNets[Gretarsson 2012], a node-link diagram representing document-topic relationships obtained with LDA. (a) The topic-similariy layout fixes the topic positions based on their similarity and computes the position of the document with a force-directed layout. (b) The order-preserving layout fixes the documents based on a predefined order, here the timestamps, and computes the topic position with a force-directed layout. 42

2.14 NewsMap: a Tree Map for breaking the news [Weskamp 2004] where rectangle size encodes the number of documents in the topic and each label corresponds to the title of one representative document within the topic. . . 43

2.15 The Word Bridge visualization [Kim 2011] where tag clouds inform topics as well as their relationships. This approach limits the number of displayed topics and relationships. 44

2.16 TwitterCrowd: a treemap of tag clouds [Archambault 2013]. Space filling layout allows to display more terms for important topics and less terms for minor topics. 44

2.17 Topic Panorama for a multi-resolution comparison of multiple sources [Liu 2014]. Each topic is represented as a pie chart depicting its presence in one or several sources. Two or three terms allow to interpret a subset of topics. The radial stacked tree placed at the circumference of the radial layout shows multiple levels of topic clusters. . 45

2.18 In ParallelTopics [Dou 2011], coordinated multiple views help identify what documents contain multiple topics through parallel coordinates on the top-left and a scatterplot on the bottom-right. Multiple interactions order the topics by similarity, highlight terms across topics or reveal topical distribution of documents through pie charts in the scatterplot view. 47

2.19 IVisClustering for interactive topic modeling [Lee 2012]. Coordinated multiple views help explore LDA results. Semantic interactions allow the modification of term weights and changing the topical structure. (A) Node-link diagram show a network of documents colored by cluster/topic. Each cluster is labeled according to its top-6 terms. (B) Cluster Tree View with semantic interactions. (C) Cluster summary view. (D) Parallel coordinates. (E) Term weight view with semantic interactions. (F) Document Tracer View to trace changes of document membership after re-clustering. (G) Document view. 47

2.20 UTOPIAN, a visual analytics software based on semi-supervised NMF. Multiple semantic interactions allow (1) topic merger, (2) document-induced topic creation, (3) topic splitting, (4) keyword-induced topic creation and keyword refinement. 48

2.21 Feature Lens [Don 2007] is devoted to the exploration of books or corpora with many repetitions, through frequent patterns of n-grams. The flat list view does not provide a broad overview of such patterns. 49

2.22 Overview is a visual analytics tool for investigative journalism [Brehmer 2014]. The expandable hierarchical view on the right hand allows progressive disclosure of document clusters. Their labels are comprised of their N most frequent terms. On the right hand, a document viewer provides access to the original content. 50

2.23 Serendip [Alexander 2014] is built in collaboration with literary scholars. Resulting from a user-centered design this software allows top-down and bottom-up analyses and follows principles promoting serendipitous discoveries. 51

2.24 Jigsaw visual analytics approach based on named entities [Stasko 2008]. Parallel coordinates allow to analyze (1:n) relationships of entities through first-order co-occurrences. 52

2.25 BicOverlapper [Santamaría 2008] proposes coordinated multiple views to explore overlapping biclusters in gene-expression data: a) Parallel coordinates reveal similar evolution of genes across coordinates. b) A matrix visualization reveals individual biclusters after reordering both dimensions. d) The node link diagram gives an overview of biclusters delineated by transparent hulls. 53

2.26 In a matrix visualization, Bicluster Viewer duplicates items in column to depict complete biclusters in a compact block [Heinrich 2011]. Common and distinctive parts of biclusters are hardly identifiable. 54

2.27 Furby: hybrid visualization with node-link diagrams and matrices [Streit 2014]. The multiple edge crossings induce visual clutter. 54

2.28 Bixplorer allows exploration of text corpus through the relationships between named entities [Fiaux 2013]. Such a bottom up approach does not convey a broad overview of biclusters. 55

2.29 BiSet: parallel coordinates view with semantic edge bundles [Sun 2015]. BiSet allows top-down exploration of chained relationships between multiple categories of named entities. Bidirectional interactions through edge bundles and entities allow task-oriented rearrangements for sensemaking purposes. 55

2.30 Visual co-cluster analysis of bipartite graphs [Xu 2016]. The space is divided horizontally to display each dimension of biclusters with either an adjacency matrix (A.1) or a 2-level treemap (A.4). The nodes (A.2) of both dimensions of the bipartite graph are connected through aggregated edge bundles (A.3). 56

3.1 The proposed nested biclustering approach builds: (a) in black, diagonal biclusters (coarse-grained topics) and their pairwise confusion blocks; (b) the overlapping biclusters (topic variants) extracted from each topic e.g. $X_{3,3}$. 60

3.2 The four nested levels of the Munzner's model for visualization design (Figure from [Munzner 2009]). 62

3.3 Our data processing pipeline: From left to right, NLP technology is used to select relevant text tokens. A term-document matrix is built. Various topic extraction methods, e.g. `Coclus` and `hLDA`, partition the TF-IDF matrix in topic blocks (in black) with various shapes. The latter are analyzed with Bimax to reveal topic variants. In parallel, topic relationships are extracted from their confusion blocks. Finally, interactive visualizations convey the topical structure and relationships to the analyst. 66

3.4 Two overlapping biclusters (in blue and orange): the columns of the confusion blocks show how many documents share common terms and the rows of the confusion blocks show how many terms occur in the same documents. 70

3.5 (a) Five Bimax biclusters (fine-grained topic variants) and (b) the related term hierarchy built with FPTree. The numbers in orange are document references. Each path from the root to a leaf describes a unique sequence of terms ordered by their overlap degree that represents one *topic variant*. They are grouped in the tree by their common prefix. 73

3.6 Distributions and comparison of different metrics for all topics produced by Coclus and hLDA. Note, for hLDA, the impact of its large topic size on the distributions of DocCoverage, BranchingFactor, TermCoverage, NbDocsByCC and NbCCs. These metrics need also to be compared with topics of similar size in Figure 3.7 and Figure 3.8. 77

3.7 Comparison of Coclus and hLDA through different metrics (Shaded area for p-value>0.05). 78

3.8 Comparison of Coclus and hLDA through metrics concerning the term hierarchy (Shaded area for p-value>0.05). 79

3.9 (a) PCA Projection of topics (blue for hLDA and red for Coclus). (b) PCA projection of metrics with strongly correlated groups (G1 to G7). (c) Interactive histograms to sort the metrics according to their contributions on the X- and/or Y-axis. 80

4.1 The nested structure delivered by our hybrid biclustering approach. (a) The diagonal biclusters (topics) denoted $X_{k,k}, k \in \{1, 2, 3\}$ and the confusion blocks for each pair denoted $X_{k,l}, k \in \{1, 2, 3\}, l \in \{1, 2, 3\}$. (b) Bimax biclusters delivered for one topic. (c) The biclusters organized in a prefix-based term hierarchy and the documents retrieved at each node (the union of the documents sets in the biclusters of the subtree). If the terms of a bicluster do not contains an existing prefix a new branch is created which can involve redundancy, e.g. $\{t5, t2\} \subset \{t3, t4, t5, t2\}$. 84

4.2 The 1^{st} version of our visual analytics tool and its components. The *Topic Weighted Map* visualization (1) shows 50 topics extracted from online news between Nov. 2 nd and Nov. 16 th , 2015. The topic selected in the *Topic Variant Overview* (3) is about the U.S. presidential elections. Five topic variants concerning Hillary Clinton have been sent to the *Comparator* (4) and the term “Obama” is hovered. 88

4.3 (a) The large tag cloud concerns news about U.S. presidential campaign. (b) The small tag cloud contains two distinct stories. The first concerns articles about the danger of betel nuts consumption in Taiwan and the second concerns the bankruptcy of the South Korean firm Handjin Shipping. 90

4.4 The *Topic Variant Overview* resulting from a topic about local elections in Germany in September 2016. Two variants are shown concerning the event. 92

4.5 Label orientation changes automatically according to the angle of the ray and the label alignment. 93

4.6 Interaction Modes. a) The orange biclusters contain any document selected by the clicked node "Israel". b) The biclusters not matching the term "Israel" are filtered. c) The bicluster colored in blue is sent to the topic variant comparator. 95

4.7 Number of biclusters as the parameters of `Bimax` vary. 96

4.8 The 2^{nd} version of our tool. The selected topic concerning an event in France : four women have been arrested after they fail a terrorist attack in Paris near Notre Dame Cathedral. Multiple *topic variants* have been sent to *Topic Variant Comparator*. In the matrix view, the terms are sorted by the number of documents in which they appear and the color of cells are mapped to the TF-IDF metric. Thereby, we hence identify the terms that are representative to few documents. We found "Kassim" (hovered term colored in red) being the name of the instigator of multiple attacks in France. He communicated with the women through "telegram" (the term just before "Kassim"), a messaging application. 97

4.9 The *Weighted Topic Map* shows 50 topics extracted from 3,992 news articles aggregated from multiple online news sources between the 2^{nd} November 2015 and the 16^{th} November 2015. One of the links highlighted from the topic about Ebola targets another topic about health being far away. 100

4.10 The Sunburst and the *Topic Variant Comparator* used for the usage scenario. 102

4.11 The *Topic Variants Overview* shows lists of co-occurring terms that constitutes relationships between 9 documents used by the investigative journalist to write an article. This small set of documents served us as a ground truth to validate with the expert the effectiveness of the tool and to build trust in the system. In the *Topic Variant Comparator*, the terms are sorted by degree of interest. In the Top-10 terms in the list we identify "bentounsi","fuyard", "légitime", "défense". The journalist validated that even if she didn't know the subject matter, she could investigate closely to the top-10 terms for her enquiry. 104

4.12 Two topics concerning Asia are close together. The topic about China includes however news about refugee crisis. 105

4.13 The large orange node on the left of the first ring corresponds to the term "Refugee" and groups many related topic variants. The highlighted one mentions "eritrea", and the complete term sequence is shown on the right hand. 106

4.14 The large orange node on the right of the first ring corresponds to the term "China". The highlighted topic variant concerns a "meeting" between "China" and "Taiwan". 107

4.15 The small orange node highlighted on the left of the first ring corresponds to the term "Island". This term brings documents that also appear in the paths colored in oranges. We can see that the documents containing "Island" cover the branch starting from "China" and the branch starting from "Refugee". Therefore, "Island" contributes to the grouping of a mixture of documents about both themes in the same topic. 108

4.16 Our data processing pipeline: From left to right, NLP technology is used to select relevant text tokens. A term-document matrix is built. Various topic extraction methods can partition the TF-IDF matrix in topic blocks (in black) with various shapes. The blocks are analyzed by Bimax to reveal topic variants and FPTree hierarchizes their term sets. Finally, interactive visualizations convey the topical structure to the analyst. 111

4.17 Examples of treemap tiles representing large and small coarse-grained topics extracted by `hLDA` (a,b) and `Coclus` (c,d) respectively. 112

4.18 The Sunburst obtained from two large topics of `hLDA` and `Coclus` respectively. 114

4.19 Nine unrelated low frequency events separated in distinct subtrees (delineated by dashed lines) in a small `Coclus` topic. 115

4.20 The qualitative scores obtained for `Coclus` and `hLDA` topics. The bar charts show the means and the standard errors of topic scores by size (a,d) or by size and by prior knowledge (b,e). Histograms (c,f) depict the distribution of scores for each size. 119

4.21 (a) Histograms showing user ratings of the quality of the *Topic variant Overview* on a Likert scale (built with R package `HH` [Heiberger 2014]). (b) The distribution of topics by size and by prior knowledge. 120

4.22 Observations of participants collected during the user study. 122

5.1 Event-driven web-oriented architecture 132

5.2 **(a)** Two dynamic Theme Rivers visualize the temporal evolution of different aspects of streaming messages from a microblog. **(b)** A map of Abila City is displayed using the GoogleMap API. Capital letters correspond to the multiple events found in the stream: (A) the POK Rally with (B) the secured street, (C) the "Dancing Dolphin" fire, (D) the black van hit a bike, (E) the black van pursuit with the police, (F) the gun fire and negotiations with the police and (G) the last messages of "trollingsnark" launching the "stage 3" of operations. 135

5.3 Each peak in the Historical Theme River concerns talks delivered by members of the POK organization in a park. The highlighted layer corresponds to Sylvia Marek. In the map, a street is secured by the police. The highlighted alerts are those appearing in the time window selected in the timeline. 137

5.4 The pink circle alert concerns a fire started in the '*Dancing Dolphin*' building. A witness describes the event through a set of messages. 139

5.5 Ironic messages of '*trollingsnark*' among which the message "*False flag operation entering stage 2*". We assume that the "operation" refers to the fire caused by the terrorist as a diversion. 140

5.6 The last messages of '*trollingsnark*' launching the stage 3 of operations, i.e. the deportation of hostages (G), are posted five minutes after the "Black Van has hit a bike" (D) and one minute before the pursuit and the gunfire exchange with the police. 141

6.1 a) The `Bimax` biclusters of the whole static binary matrix. b) The state of B after 5 operations of `DynBimax` through a sliding window (size=3 and shift=1) 148

6.2 computation time of `DynBimax` vs. `Bimax` for different sizes and shifts of sliding widows. 151

6.3 Visualization through a term hierarchy of `DynBimax` applied on news streams. 152

General Introduction

Contents

I.1 Context 1

I.2 Challenges and problematics 2

I.3 Contributions 5

I.4 Outline 7

I.5 Notations 9

I.1 Context

Due to the growing importance of web-enabled devices and applications firmly anchored in our societies, the production of text data has exponentially increased. High availability of data and standardization of data exchanges through Web 2.0 offer unprecedented opportunities to share information and knowledge, be they scientific, cultural or social. In various domains, stakeholders investigate text corpora for different purposes. For instance, one can mention the analysis of scientific articles [Gerrish 2010], the investigation of collections of emails, various document leaks or documents obtained under the Freedom of Information Act at the request of journalists [Brehmer 2014], epidemic surveillance in public health [Broniatowski 2013], sentiment analysis and opinion mining for decision-makers or elected officials [O'Connor 2010], as well as surveillance of criminal networks by the law enforcement authorities [Denef 2013], to name but a few purposes.

Information seeking in a large text corpus is mainly approached by using search engines relying on efficient Information Retrieval technologies. The effectiveness of this approach is widely recognized considering the impact on our societies of Google, Bing or Yahoo!. However keyword-based search requires prior knowledge of the subject matter in order to choose the keywords yielding useful results. Conducting thorough and complete investigations involves a broader process to discover supporting evidence or to generate new hypotheses [Brehmer 2014]. This creates the need to design interactive visualization systems that support the analytical tasks of a given stakeholder having to analyze large text corpora.

The goal of this thesis is to propose a visual analytics approach for the analysis of text corpora and to implement it in a software for a real and concrete application. We specifically focus on the field of investigative journalism. Investigative journalists face a dilemma:

while the number of sources of information increases dramatically, the time devoted to investigation is steadily reduced by editorial boards. During their back-grounding work the journalists gather collections of free texts. They are forced to reduce their size so that they are able to exploit the content within the deadlines. Nonetheless, journalistic work requires exhaustiveness in order to ensure a large coverage of the field under investigation and to multiply sources showing evidence of their working hypotheses. To be more exhaustive when facing large text corpora, they need software supporting their inquiry process. Our approach consists first in giving a clear overview of text contents and their semantic relationships, and second, in allowing the analyst to find specific passages that can validate or disprove her working hypotheses. In the next section we present the challenges associated with the design of such a visual analytics software.

I.2 Challenges and problematics

Text analytics and human knowledge. From the last decades, extensive research has been conducted to deal with text data. Despite the increasing power of computers, their ability to understand human language is very limited [Turney 2010]. This limitation is coined by Cambria and White as "‘cognitive gap of algorithms"’ that only rely on a representation of observed contents, including the *bag-of-words* paradigm falling under the Vector Space Model (*VSM*) [Cambria 2014]. This representation of text corpora is very efficient to extract semantic features and their relationships through word frequencies and co-occurrences [Turney 2010]. Nonetheless, *VSM* representations of observed contents ignore author intention as well as implicit concepts related to physical knowledge, sensory knowledge, psychological knowledge and social knowledge gained from human experience [Cambria 2014]. Research in Natural Language Processing (*NLP*) aims at filling this gap. In their survey, Cambria and White draw three overlapping curves of *NLP* research evolution by focusing on three paradigms, namely the *bag-of-words*, the *bag-of-concepts* and the *bag-of-narratives*. According to them, *NLP* research is currently at the beginning of the semantic curve under the *bag-of-concepts* paradigm. For all of these reasons, carrying out text corpus analysis requires to combine both, powerful computational capabilities and meaningful human knowledge. In this thesis, our work follows precisely the visual analytics approach [Thomas 2005, Keim 2010], placing the expert at the center of the analytic process and enabling him/her to steer algorithms through interactive visualizations until useful insights are found.

Discovering hidden structure. Undoubtedly, the current situation and the ambition of *NLP* research do not preclude the efficiency and the usefulness of the *bag-of-words* paradigm modeled with the *VSM*. It was initially developed for Information Retrieval [Salton 1975] and shows its efficiency every day in search engines. Moreover, the *VSM* has a strong relationship with the distributional hypothesis [Turney 2010] which states that terms are similar if they co-occur in similar contexts [Harris 1954]; it is also the basis of word sense induction [Navigli 2012]. Through this assumption, *VSM* allows to characterize the semantics of text contents, i.e. the meaning of words, sentences or documents. For this reason, *VSM* is leveraged in many machine learning and text mining methods for automatic text summarization, document clustering or topic modeling [Hotho 2005, Aggarwal 2012a]. Besides the unstructured nature of free text, the use of *VSM* to model text data entails several issues, typically the loss of context, the high dimensionality and the sparsity of the resulting term-document matrix [Turney 2010, Aggarwal 2012a]. While the purpose of clustering methods and topic models are to uncover hidden structure to summarize and organize the corpus in categories, they must be carefully chosen by the designer to address the limitations of *VSM* and to support the tasks of the target analyst [Aggarwal 2012b].

Multiple models for a multi-resolution analysis. As part of text corpus investigation work, a first high-level task starts by gaining a broad understanding of the corpus. This tasks requires to draw an overview of its textual contents. On the one hand, one-way clustering approaches deliver a partition either in the document dimension (e.g. document clustering [Steinbach 2000, Aggarwal 2012b]) or in the term dimension (e.g. topic models [Blei 2003, Crain 2012]), making the result difficult to interpret without post-processing. In contrast, biclustering methods consider simultaneously the duality of the term and document dimensions to group similar terms which are representative of similar documents [Madeira 2004, Govaert 2013, Prelić 2006]. These methods extract homogeneous blocks in the term-document matrix, performing dimensionality reduction at the same time and allowing direct interpretation of clusters. Such bicluster structure can provide an overview of coarse-grained topics as a starting point for text corpus investigation. While data and task characterization allows rational choices for determining suitable methods, one unique method does not fully fit user requirements. For instance, in addition to an overview of topics, the analyst needs to drill down into the topics to discover interesting fragments of texts that can support or invalidate her working hypotheses. The flat partition resulting from a single biclustering method does not support this multi-resolution analysis. A nested structure revealing fine-grained *topic variants* is necessary and requires a combination of methods able to handle different bicluster shapes and properties.

Visualization of overlapping biclusters. Term-document biclusters[1] can yield fragments of text revealing useful facts or viewpoints for hypothesis validation or generation

[1] A bicluster in a term-document matrix is a subset of terms shared by multiple documents.

and seem well suited to represent *topic variants*. Moreover, overlapping biclustering is deemed appropriate for handling the multiplicity of topics treated in documents and semantic ambiguity such as word polysemy and synonymy [Shafiei 2006a]. However the visualization of overlapping biclusters raises challenging issues [Sun 2014]. For instance, it is difficult to draw a clear overview of biclusters while conveying a clear picture of common and distinctive elements in both term and document dimensions, especially when a large number of biclusters are obtained. Little attention has been paid to this problem in visual analytics research and no solution has been proposed to represent *topic variants*.

Textual content vs. semantic structure. Text visualization is challenged by the conflicting representation of textual content and the semantic structure of the corpus. Indeed, documents of the corpus can be related to different topics that are more or less linked together and the contents can take multiple angles of reading giving different viewpoints. On the contrary, topics, angles or viewpoints can concern multiple documents. These semantic relationships are of great interest to the analyst and need to be represented visually. At the same time, to understand the content and to interpret these relationships, textual elements, i.e. words, must be displayed in a way that avoids overlaps and reduces visual clutter. Optimal visual encoding and interaction design require a good comprehension of the problems and tasks faced by analysts.

Disjoint biclustering vs. overlapping topic models. In addition, topic models have shown their efficiency in the analysis of text corpora [Dou 2011, Lee 2012]; biclustering methods have also been used successfully to model coarse-grained topics [Shafiei 2006a]. However, the impact of both methods on the topical structure and on the topic comprehension is not well understood. Most evaluation methods assess a document partition against labeled ground truth. They are often accompanied with lists of the top-10 terms from each topic to assess the term partition. However, these approaches give little information about the internal structure and the comprehension of the topics from the analyst perspective. Characterizing these differences through numerical experiments as well as through a user study can greatly help to understand which method better suits which task.

Text Streams. Finally, more and more data is provided in live streams which experts seek to exploit. Confronting huge volumes of streaming data seems however to be a daunting endeavour. The usual tasks consist notably in monitoring the current situation for tracking events while detecting changes and trends [Rohrdantz 2011, Wanner 2014]. Providing the temporal context of the current situation through dynamic historical retrieval can also have an extensive scope of application. Dynamic contextualization of information becomes a crucial need in journalism, in particular for breaking news or fact checking during political debates. In addition, investigative journalists follow targeted user accounts in social media to find sources of information and documents related to their ongoing investigations [Marcus 2011]. Such tasks require dynamic models and visualizations supporting the arrival and obsolescence of documents within a sliding time windows.

I.3 Contributions

The main contribution of this thesis is a novel visual analytics approach for text corpus analysis, applied in the specific field of investigative journalism. This section outlines the multiple contributions of this thesis to the advancement of the visual analytics research.

A multi-resolution visual analytics approach. We propose a visual analytics approach, built in collaboration with a professional 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 are torn between the need to be exhaustive and not affording the time to read every document. In their work, 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: coarse-grained topics and fine-grained *topic variants*. Our approach supports both aspects. A new *Weighted Topic Map* visualization conveys all coarse-grained 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. This visualization organizes the term-document biclusters in a hierarchy that limits term redundancy and conveys their commonalities and specificities, adressing hence the visualization issues of a large amount of overlapping biclusters. Through interactions, the analyst can select, compare and filter the subtle co-occurrences of terms shared by multiple documents in order to find interesting facts or stories. The effectiveness of the tool is shown through a usage scenario and further assessed through a qualitative evaluation by the journalist.

A multi-model system based on a nested bicluster structure. While traditional exploratory text analysis tools are document-centric, term-centric or topic-centric, few solutions consider the duality of terms and documents offered by biclustering methods at different levels of granularity. To extract coarse-grained topics, our system can accommodate multiple topic extraction methods as long as they result in a bicluster structure that may be of different shapes. A thorough understanding of topics relies on the ability to make sense of the related term set and to explain the frontier/relationships between distinct topics. In order to convey such relationships we designed a new similarity metric that can handle both overlapping and disjoint biclusters.

The key structure for performing hypotheses validation and generation is the fine-grained *topic variants* discovered within each topic. Typically, the *topic variants* must capture first-order co-occurrences of terms shared by enough documents, in order to more likely constitute evidence for the hypothesis at hand. These patterns are nonetheless determined by the topical structure delivered by the enclosing partition. The current work characterizes in a systematic fashion the differences between overlapping and disjoint coarse-grained topics. Moreover, objective criteria regarding the quality of topics and the co-occurrence patterns they contain are needed to choose between different topic extraction methods. To this end, we have defined multiple intrinsic metrics, on the basis of which a numerical experiment has been conducted to compare the topics extracted by `Coclus` [Ailem 2016], a disjoint biclustering method, with topics extracted by the probabilistic Hierarchical Latent Dirichlet Allocation (`hLDA`) [Griffiths 2004], a widely used topic extraction method providing overlapping topics. Eventually we aim to judge the suitability of these two families of methods to the tasks carried out by journalists, i.e. the generation and validation of working hypotheses. The goal of this contribution was also to inform the design of a subsequent user study investigating the influence of the choice of both methods on the interpretability of topics by analysts.

Things that matter for topic comprehension: size and topic models. While the document partitions resulting from topic extraction methods can be compared with a labeled ground truth, the semantic of topics is often conveyed by the N most frequent terms. Recent work by Alexander and Gleicher [Alexander 2015] shows that "*a topic is more than the top 10 words*". They argue that, to ensure that a topic model relates document semantics, it must reflect "*the subtle patterns of co-occurrences*". The visual analytics approach we present is precisely designed to explore topics through these subtle co-occurrence patterns called *topic variants*. Our system can hence constitute a common base to investigate the influence of the topical structures elicited by alternate topic extraction methods on the comprehension of topics and their variants by the analyst. Based on the results of the aforementioned numerical experiment, we identified independent variables and elaborated several hypotheses to conduct a user study that compares `Coclus` and `hLDA`, two topic extraction methods representative of biclustering and topic models respectively. Typically, we provide evidence of their differences from an analyst's perspective and identify the characteristics making either method suitable to relevant tasks in investigative journalism.

Towards a Visual Analytics Tool for Situation Awareness and Real-Time Exploration of Text Streams. This work in-progress aims eventually to propose a novel visual analytics approach for text stream analysis. The visual analytics tool we describe in this section was designed to take the IEEE VAST Challenge 2014 [2] (mini-challenge 3) [Médoc 2014]. Our work is driven by two generic tasks: situation awareness and exploratory analysis of text streams. Through an efficient web-oriented architecture, the backend exploits in real time the metadata and named-entities of Twitter-like messages. The frontend offers dynamic and interactive visualizations supporting flexible analyses of text streams with a preliminary solution for giving temporal context of the current situation through dynamic historical retrieval. We discuss several limitations observed during our own analysis of the twitter-like streaming messages provided by the organization comity of the VAST Challenge. We also identified the need to apply dynamic co-clustering to uncover the relationships between the twitter metadata and the named-entities extracted from messages.

Dynamic Bimax for Textual Data. We propose `DynBimax`, a dynamic overlapping biclustering algorithm extending `Bimax`. Our extension handles streaming data with a sliding time window strategy, and supports user-driven modifications of biclusters to better meet user needs. After documents/terms removals or modifications, the algorithm ensures that all remaining biclusters still verify `Bimax`'s maximal inclusion constraint. We prove that certain unnecessary conditions can be pruned from the search space. On a real text stream, we evaluate the gain in computation time of `DynBimax` compared to `Bimax`. The effectiveness of `DynBimax` is shown through data visualization interfaces.

I.4 Outline

This thesis is organized in three parts. The first part proposes a survey of the state of the art in text mining and visual text analytics. The second part describes the visual analytics system we designed for the analysis of text corpora based on the problems and tasks of investigative journalists. The last part is devoted to two dynamic approaches dealing with text streams.

Part I: State of the Art

Chapter 1 starts by describing the *Vector Space Model* and its relationships with the distributional hypothesis. Then, it presents a survey of topic models, clustering and biclustering approaches. We explain which techniques address which issues. We bring out the need of using multiple biclustering approaches and we show the relevance and the novelty of the nested bicluster structure we propose. Through a survey of the evaluation techniques

[2] http://hcil2.cs.umd.edu/newvarepository/benchmarks.php

of clustering and topic models we also justify the need to compare biclustering with topic models from the human comprehension perspective.

Chapter 2 surveys the most relevant visualization approaches involved in the design of a visual analytics software for the investigation of text corpora. Document-centric, term-centric and topic-centric approaches are described and illustrated with figures showing the most representative solutions. We present several visual analytics tools for investigation of text corpora and explain the interest and the novelty of our contributions for drawing up an overview of topics and drilling down into *topic variants*.

Part II: Visual Analytics System

Chapter 3 describes the user-centered approach we adopted for the visualization design. The problems of investigative journalism are characterized and relevant journalist tasks are defined according to the typology of Brehmer and Munzner [Brehmer 2013]. Next, this chapter gives more details of the biclustering methods and the nested structure supporting multi-resolution analysis. We formally describe our similarity measure supporting disjoint as well as overlapping biclusters. Finally a numerical experiment is presented, providing the statistical evidence of structural differences between a topic model approach and a diagonal disjoint biclustering method.

Chapter 4 presents each components of our visual analytics software with a justification of the design rationale for each of the proposed visualizations and interactions. The system is evaluated through a usage scenario and a qualitative evaluation with an expert user. Finally this chapter lays out the controlled experiment we conducted to understand how different topical structures, resulting from topic models and biclustering methods respectively, impact the comprehension of topics by the analyst. The characteristics of both methods and their suitability to the tasks of investigative journalists are discussed.

Part III: Towards Visual Analytics of Text Streams

Chapter 7 and 8 discuss preliminarily work dealing with text streams. Chapter 7 proposes dynamic visualizations of text streams. Chapter 8 proposes a dynamic version of `Bimax`, the pattern-based overlapping biclustering used in Chapter 3 and 4 for extracting *topic variants*.

I.5 Notations

The following notations are used along this thesis:

- X denotes the data as a term-document matrix of size $n \times m$ where $X = \{e_{ij}, i \in [1..n], j \in [1..m]\}$ and $e_{ij} \in \mathbb{R}^+_{\neq 0}$.
- I is the row set of n documents.
- J is the column set of m terms.
- K is the number of biclusters.
- $X_{k,l}$ is the submatrix $I_k \times J_l$ such that $I_k \subseteq I$ and $J_l \subseteq J$, $k \in [1..K]$ being the k^{th} class of row partition, $l \in [1..K]$ being the l^{th} class of column partition.
- $X_k = X_{k,k}, k \in [1..K]$ is the k^{th} diagonal bicluster/submatrix.
- $B_b, b \in [1..\beta]$ is a `Bimax` bicluster $I_b \times J_b$ such that $I_b \subseteq I$ and $J_b \subseteq J$ in any binary matrix X of size $n \times m$, β being the number of biclusters discovered by `Bimax`.
- $\tilde{X}_k, k \in [1..K]$ is the binarized submatrix obtained from X_k with the threshold τ_k.
- $\tau_k, k \in [1..K]$ is the binarization threshold applied to every submatrix X_k to obtain $\tilde{X}_k$.
- $\tilde{B}_{k,b}$, with $k \in [1..K]$ and $b \in [1..\beta_k]$, is a `Bimax` bicluster nested in any binarized submatrix $\tilde{X}_k$. For the sake of readability, the notation $\tilde{B}_{k,b}$ is written B_b in certain figures.

Part I

State of the Art

CHAPTER 1

Text Mining and Clustering

Contents

1.1 Text Models 13

1.2 Topic Models 16

1.3 Clustering methods 18

1.4 Biclustering 21

1.4.1 Generalities 21

1.4.2 Biclustering techniques for topic extraction 22

1.4.3 From coarse-grained topics to fine-grained *topic variants* 25

1.4.4 Pattern-based biclustering for *topic variant* extraction 25

1.5 Topic Models vs. Biclustering 27

1.6 Evaluation of clustering and topic models 29

Textual data holds important properties that must be taken into account, namely the unstructured nature of free texts and their large vocabulary that often carries semantic ambiguity such as word polysemy and synonymy [Aggarwal 2012a]. Investigating a text corpus requires then a processing pipeline capable of extracting a meaningful structure that preserves the content semantics and supports content exploration and comprehension. In this chapter, we propose to briefly survey the techniques involved in such a text processing.

1.1 Text Models

Many text analytics solutions model text corpora using a Vector Space Model representation (*VSM*) [Salton 1975] where each data object is a document whose feature vector corresponds to the words it contains. The resulting term-document matrix shown in Figure 1.1 is also known as a *Bag of Words*. Natural Language Processing (*NLP*) research covers a large range of computational linguistic techniques such as syntactic parsing, semantic analysis as well as sentiment analysis and opinion mining [Cambria 2014]. From this field of computer science, we simply outline the preprocessing steps, listed in Figure 1.1, that are

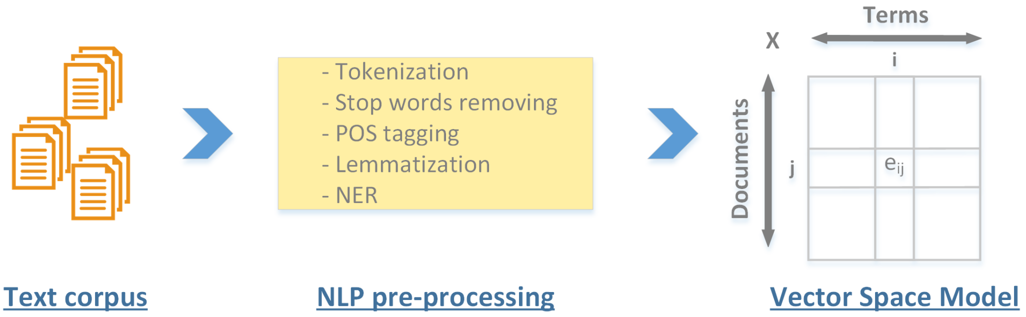


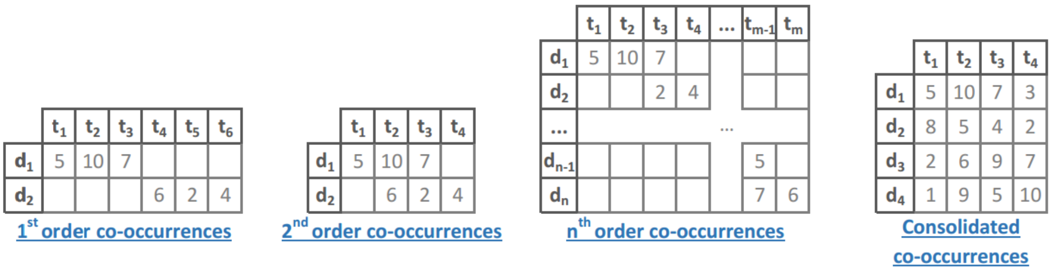
Figure 1.1: The pre-processing pipeline for modeling a text corpus in the Vector Space Model representation.

relevant to this thesis. First, the *tokenization* parses the plain text content of every document to build a list of words. *Stop words* lists are frequently used to remove meaningless words such as articles, prepositions, etc. Then, *part-of-speech tagging* determines the nature of terms (nouns, verbs, adjectives, adverbs, etc.) and prepares the *lemmatization* step that reduces the words to their canonical form, i.e. the dictionary form without inflections. The distinct lemmas so obtained from every document are called "terms" in the rest of this thesis. The set of terms found in the corpus forms a lexicon that constitutes the features of the *VSM*. Finally, *named entity recognition* can be used to assign categories to proper names, i.e. location names, organization names, person names, time or numerical values. The Stanford NLP library [Manning 2014] encompasses a set of algorithms for all of this standard NLP processing pipeline.

VSM has proven its efficiency in Information Retrieval, but it suffers from a drawback: it does not preserve the relative order of terms in document vectors inducing important loss of context for semantic disambiguation. However, *VSM* is well suited to measure term similarity through the distributional hypothesis [Harris 1954]:

Hypthesis 1. *The* ***distributional hypothesis*** *states that the terms tend to be similar if they co-occur in similar contexts.*

The context constitutes one dimension of the *VSM* and can be chosen depending on the needs at different levels of granularity such as the document, the chapter, the paragraph or the sentence. In the present thesis we use the document context. Various weighting schemes make different assumptions to build document vectors [Hotho 2005]. The simplest way to encode document vectors is binary weighting which only considers the presence or absence of the terms in the document. Term Frequency (*TF*) weighting, associates word importance to their frequency in the document but assigns high weights to rather meaningless terms which are frequent in every document in the corpus. Hence, the Term Frequency - Inverse Document Frequency (*TF-IDF*) weighting adapts *TF* by considering



	t_1	t_2	t_3	t_4	t_5	t_6
d_1	5	10	7			
d_2				6	2	4

1st order co-occurrences

	t_1	t_2	t_3	t_4
d_1	5	10	7	
d_2		6	2	4

2nd order co-occurrences

	t_1	t_2	t_3	t_4	...	t_{m-1}	t_m
d_1	5	10	7				
d_2			2	4			
...					...		
d_{n-1}						5	
d_n						7	6

nth order co-occurrences

	t_1	t_2	t_3	t_4
d_1	5	10	7	3
d_2	8	5	4	2
d_3	2	6	9	7
d_4	1	9	5	10

Consolidated co-occurrences

Figure 1.2: Example of term co-occurrences of different order. The number at each cell of the matrices corresponds to the number of times the term t_i occurs in the document d_j.

the rarity of words at the corpus scale to promote surprising events rather than expected events [Turney 2010]. *TF-IDF* weighting scheme is defined as follows:

$$TF\text{-}IDF_{ij} = TF_{ij} \log \frac{1+n}{1+DF_j}, \forall i \in I, \forall j \in J \tag{1.1}$$

, where TF_{ij} is the number of times the term i appears in the document j, n is the number of documents in the corpus and DF_j is the number of documents containing the term j. *TF-IDF* holds a more discriminative power than *TF* and can be viewed as a measure of term representativeness for each document. Finally, the Point-wise Mutual Information (*PMI*) weighting [Church 1990] represents mutual information of the term-document relationship. *PMI* is effective for measuring semantic similarity of words but is biased towards low frequency events [Turney 2010, Role 2011].

To illustrate the assumption underlying the distributional hypothesis (see Hypothesis 1), different patterns of term co-occurrences are presented in Figure 1.2. A first-order co-occurrence is when multiple terms appear in the same context, e.g. when the terms t_1, t_2 and t_3 co-occur in one document d_1. This pattern does not allow finding document relationships. A second-order co-occurrence appears in two contexts that share some of the terms, e.g. when the terms t_1, t_2 and t_3 co-occur in d_1, and t_2, t_3 and t_4 co-occur in d_2. In this case, t_1 and t_4 participate in a 2^{nd} order co-occurrence. The similarity of the documents d_1 and d_2 due to the shared terms t_2 and t_3 induces similarity between t_1 and t_4. Higher-order co-occurrences consider the terms appearing in more similar contexts [Turney 2010] associated in the same way through a chain of shared terms. In addition, we identified that another pattern we call *consolidated co-occurrences* is also important to consider in our analysis. Indeed, the presence of term co-occurrences shared by multiple documents gives a better guarantee of having high commonality between documents so that they constitute similar contexts.

In general, the *VSM* suffers from the high dimensionality and the sparsity of the term-document matrix [Turney 2010, Aggarwal 2012a]. It requires dimensionality reduction techniques and/or summarization such as clustering or topic modeling.

1.2 Topic Models

To extract semantic features from a text corpus and reduce the dimensionality of the term-document matrix, numerous solutions rely on Singular Value Decomposition (SVD) to reveal *latent semantic spaces*. Latent Semantic Indexing (LSI) [Dumais 1992] applies SVD to discover concepts through term similarity. The orthogonal Latent Semantic Allocation [Deerwester 1990] focuses on document similarity to discover document classes. These approaches smooth the term-document matrix in compressed dimensions but the *semantic spaces* are still difficult to interpret [Crain 2012], especially when they yield negative values. Non-negative Matrix Factorization [Paatero 1994] (NMF) relaxes the orthogonal constraint of the semantic spaces derived from SVD and guarantees non-negative values to all of them [Xu 2003] for better interpretation. The probabilistic version of *LSI* (*pLSI*) [Hofmann 1999] also improves over plain *LSI*, but takes a large number of parameters and does not build topic probabilities for the documents [Crain 2012].

Blei et al. [Blei 2003] propose Latent Dirichlet Allocation (LDA) which provides a better representation of documents by generating topics in each of them. The goal of LDA is to find two distributions: β, a distribution of terms for each topic (regulated by hyperparameter η) and, θ, a distribution of topics for each document (regulated by hyperparameter α). To obtain these distributions, LDA fulfills the following generative process:
Let $\beta_k = (\beta_{k1}, \beta_{k2}, ..., \beta_{km})^T$ the distribution of terms in topic k and $\Theta_i = (\theta_{i1}, \theta_{i2}, ..., \theta_{iK})^T$ the distribution of topics in document i.

- for each topic k: choose $\beta_k \sim Dirichlet(\eta)$
- for each document x_i:
  - choose $\Theta_i \sim Dirichlet(\alpha)$
  - for each word j in x_i:
    - choose a topic $k \sim Multinomial(\Theta_i)$
    - choose a word $j \sim Multinomial(\beta_k)$

With the *Gibbs Sampling* mechanism, the process runs through numerous iterations and starts with a random distribution for θ. At each iteration, each document is processed by considering both distributions in order to choose one topic for each term. The β distribution favors the topics where the term has a high probability, i.e. the topics being often assigned to this term in many documents. The Θ distribution can be viewed as an internal topical signature of the document. It favors the most prominent topics with respect to the number of terms assigned to them in the document.

The distributional hypothesis (Hypothesis 1 on page 14) is a central assumption in latent topic models. It is worth noting that preserving high-order co-occurrences in topics enriches the terms with a context that either disambiguates polysemy [Navigli 2012] or groups synonyms.

The internal topical signature of the documents given by Θ tends to assign first-order co-occurrences of terms in the same topics. Then, the β distribution enables higher-order co-occurrences by associating terms in topics shared by different documents at the corpus scale. While LDA clusters terms in topics, it does not provide a mechanism to build document clusters. They can be obtained at the end of the process by grouping the documents by their prominent topics in the final Θ distribution. The number of topics must be fixed arbitrarily.

Hierarchical LDA (hLDA) [Griffiths 2004] adds a mechanism derived from the Chinese Restaurant Process [Aldous 1985] (*CRP*, regulated by γ) to build a topic hierarchy depicted in Figure 1.3 with a fixed depth L. The generative process of hLDA is as follows:

1. For each table $k \in \mathcal{T}$ in the infinite tree:

   (a) $\beta_k \sim Dirichlet(\eta)$

2. For each document x_i

   (a) Draw the path $c_i \sim nCRP(\gamma)$

   (b) Draw an L-dimensional topic proportion vector $\Theta_i \sim Dirichlet(\alpha)$.

   (c) For each word j in x_i:

      i. Choose level $z_{ij} \sim Multinomial(\Theta_i)$.

      ii. Choose word $w_{ij} \sim Multinomial(\beta_{c_i,z_{ij}})$, where $\beta_{c_i,z_{ij}}$ denotes the topic at the level z_{ij} on the path c_i.

Each node is a topic and, at each iteration, the nested *CRP* (*nCRP*) assigns each document x_i to one path c_i, i.e. all the nodes of one branch. As illustrated in figure 1.3, the first document x_1 creates the first branch c_1. Then, the document x_2 creates a new path c_2 from the node β_2, x_3 creates a new path c_3 from β_1 but x_4 is associated to the existing branch c_3. In the *nCRP*, the choice of the node (or a *table* in the *CRP* vocabulary) from which a new path is created or not for a document is influenced by the hyperparameter γ as well as the probabilities of the document terms in every topic of the tree. Once the branch c_i is chosen for a document, the topics of c_i ($\{\beta_{c_i1}, ..., \beta_{c_iL}\}$) are assigned to the terms of the document like LDA does with the joint probability distribution β and Θ.

With hLDA, the branch-wise restriction of topics for each document increases the sparsity of Θ. Therefore, the terms occurring together in a document are distributed in fewer topics, which can increase the likelihood of obtaining consolidated co-occurrences in the

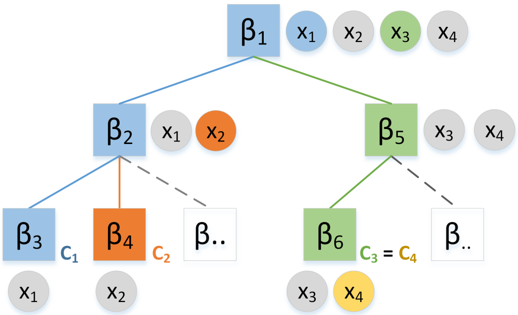


Figure 1.3: A topic hierarchy built with hLDA. The *nCRP* chooses a node β_k in the tree for each document x_i. If needed, a new path is created from the chosen node to reach depth L.

topics. Such a mechanism can be viewed as a form of document clustering which, compared to LDA, can improve the similarity of the contexts and consequently the similarity of the terms, according to the distributional hypothesis. However, the *nCRP* mechanism considers only groups of term-topic relationships [Wang 2009] which actually guarantee only high-order co-occurrences. In addition, generic terms tend to appear at the first levels of the hierarchy because they share many documents. The specific terms shared by less documents are placed more in depth. hLDA discovers automatically the number of topics from the data, by only setting the hierarchy depth. Both LDA and hLDA need their hyperparameters to be set; the interpretation and impact of the latter on topic shapes are not easy to anticipate by a lay user audience such as journalists.

1.3 Clustering methods

Text corpus summarization is mainly approached by using clustering methods [Aggarwal 2013]. Belonging to the family of unsupervised learning methods, clustering aims to group similar objects to discover latent structure from the data itself by optimizing criteria such as intra-cluster compactness and inter-cluster separability [Liu 2010]. Clustering is largely studied in various domains [Everitt 2011, Aggarwal 2013], but for the specific case of text clustering problems the techniques must be adapted to the characteristics of textual data, especially the high dimensionality and the sparsity of the term-document matrix [Aggarwal 2012b]. Similarity/dissimilarity measures are well suited criteria to represent cluster compactness/separability. A comparative study of similarity measures has been conducted with text data where *Pearson Correlation Coefficient*, *Jaccard Coefficient* and *Cosine similarity* proved to be most effective options [Huang 2008]. We summarize in this section the main categories of existing clustering methods.

Hierarchical methods [Murtagh 2012, Steinbach 2000] build a hierarchical structure where at each level all data objects are partitioned at a different level of resolution. *Agglomerative approaches* start with unitary clusters (i.e. containing only one object) and apply multiple pair-wise merges with a linkage criterion until one all-inclusive cluster is obtained. In contrast, *divisive approaches* such as MONA and DIANA [Kaufman 2009] start with one all-inclusive cluster and split the cluster recursively until obtaining a singleton for each object. One advantage of these methods is that the desired number of clusters can be chosen by cutting the hierarchy horizontally at one depth level. Hierarchical structures are preferable for text corpus exploration [Dou 2013], but the high complexity ($O(n^3)$ and $O(2^n)$ of agglomerative and divisive approaches respectively) limits their applicability to small corpora [Steinbach 2000].

Partitioning methods include the *k-means* algorithm [MacQueen 1967], one of the most popular clustering methods due to its simplicity and performance. This approach starts by initializing k centroids (mean vectors of clusters' objects). *k-means* iterates the two following steps until convergence:

1) objects are first associated to their closest centroid with a distance metric;

2) new centroids are computed based on the objects in the new clusters.

Identifying the optimal number of clusters is not straightforward and the choice of the distance/similarity metric is of great importance since it shapes the result [Huang 2008]. While the *Euclidian distance* is commonly used with k-means, *spherical k-means* using cosine similarity is more effective for textual data [Dhillon 2001b]. Instead of using centroids, alternative cluster representations have been proposed. For instance, k-medoids method [Kaufman 1987] represents the cluster with one of its objects, and k-medians [Bradley 1997] uses the cluster's median instead of the mean.

Density-based methods such as DBSCAN [Ester 1996] suppose that the density of objects within a cluster is higher than the density of objects in the remaining data. DBSCAN requires a density threshold characterized by two parameters: a radius ε and the minimum number of objects *minPts* to be found in the circle area delimited by ε. Every data object is scanned to determine its cluster membership. *Core objects* that verify the density threshold are member of a cluster and transmit their membership to all the *density-reachable objects* appearing in their ε area. A first cluster starts with one *core object* and its cluster membership is hence extended by transmission over the neighborhood of the *core objects*. Other clusters are formed among unreached objects in the same way until all objects are scanned. This approach has the advantage of revealing clusters of various shapes and discovering the number of clusters automatically from the data.

Graph-based methods have recently received a great attention in the field of social sciences and social networks analysis. A graph is represented by a set of nodes linked by edges, often weighted by a similarity measure (e.g. from the corresponding adjacency matrix). Graph clustering aims to identify a community structure by maximizing the intra-cluster connectivity and minimizing inter-cluster connectivity. The most straightforward formulation of graph clustering consists in solving the minimum cut problem by minimizing the adjacent nodes belonging to different clusters (inter-cluster connectivity) [Leighton 1988, Ding 2001, Schaeffer 2007]. The more recent modularity criterion [Newman 2004] optimizes intra-cluster connectivity. The principle is to maximize the difference between the number of observed intra-cluster edges and this same number expected in a random distribution of edges in the graph. The *Modularity* measure is formulated as follows:

$$\frac{1}{2|E|}\sum_{i,i'=1}^{n}(e_{ii'}-\frac{e_{i.}e_{i'.}}{2|E|})\sum_{k=1}^{K}z_{ik}z_{i'k} \tag{1.2}$$

where $e_{ii'}=1$ if there is an edge between node i and i' and $e_{ii'}=0$ otherwise, $e_{i.}$ is the degree of the node i (number of adjacent nodes), $e_{i'.}$ is the degree of the node i', and $\frac{e_{i.}e_{i'.}}{2|E|}$ is the expected number of edges between i and i', $|E|$ is the number of edges and thus $2|E|$ is the sum of node degrees. Since this formulation of the problem is NP-Hard, an efficient alternative is the spectral approach [Ng 2001] that computes eigenvectors of the Laplacian matrix of the graph.

Model-based methods consider that data can be generated by a mixture model, which gives a great flexibility with respect to the data types and structures. The model relies on the assumption that data objects are independent and identically distributed conditionally to their cluster membership. Given X a $n\times m$ matrix, the number of clusters K and a probabilistic model family, the density function of a mixture model is formulated as follows:

$$f(X,\Theta)=\prod_{i=1}^{n}\sum_{k=1}^{K}\pi_k\varphi(x_i|\alpha_k) \tag{1.3}$$

where π_k is the probability that any object belongs to the k^{th} cluster with $\sum_{k=1}^{K}\pi_k=1$, $\varphi(x_i|\alpha_k)$ is the probability density function of an observation x_i from the k^{th} cluster that takes the parameter α_k (e.g. $\alpha_k=\{\mu_k,\Sigma_k\}$ for a Gaussian distribution). The clustering problem consists in estimating $\Theta=\{\pi_1,\pi_2,...,\pi_k,\alpha_1,\alpha_2,...,\alpha_k\}$ that better shapes the structure of the observed data X. To this end, *Expectation Maximization* (*EM*) [Dempster 1977] and its classification variant *CEM* [Celeux 1992] are both two usual algorithms that maximize the log-likelihood of a mixture model from observed data X. For more details the interested readers may refer to the following surveys [McLachlan 2004, Govaert 2009].

Conclusion. Clustering methods applied on a text corpus group similar documents in homogeneous clusters. Clustering is useful in several applications such as document organization and browsing as well as corpus summarization [Aggarwal 2012b]. However, such a one-way clustering remains difficult to interpret directly because it does not reveal the features (the terms) that better characterize the clusters. A straightforward solution consists in labeling the clusters with the "top-N" frequent terms [Brehmer 2014] or with representative sentences [Krstajić 2013]. Beyond cluster labeling, using semantic graphs, taking into account term relationships within clusters [Role 2014], can make them easier to interpret. However, such traversal graphs are only built in post-processing once the clusters are supplied.

1.4 Biclustering

In this thesis we propose to exploit biclustering approaches that simultaneously reveal row and column partitions. The purpose is to organize the document collection while selecting at the same time the relevant terms for a thorough comprehension of cluster contents.

1.4.1 Generalities

Biclustering, also known as co-clustering, is widely used in bio-informatics [Madeira 2004, Prelić 2006, Govaert 2013] but may apply to other fields, e.g. text analysis or recommender systems. It takes as input any 2D matrix whose entries represent a relation between its rows and columns, and delivers homogeneous sub-matrices or blocks that reveal consistent row-column relationships. In a term-document matrix, these methods define an optimization criterion that considers simultaneously the term and document dimensions to take into account the duality of their relationships, i.e. multiple terms co-occur in a given document and multiple documents share a given term. Biclustering brings certain advantages for text analysis: it deals with high dimensionality and sparsity of term-document matrices through an adaptive dimensionality reduction at each iteration; it delivers a consistent latent structure that preserves the document/term duality; it allows a thorough comprehension of document clusters by simultaneously clustering the terms.

Bicluster structures fall in two main categories, having their pros and cons. On the one hand, hard biclusters assume that rows and columns are allocated exclusively to one bicluster. This approach does not account for term polysemy and the multiplicity of topics treated in documents [Shafiei 2006a]. But the terms remain semantically consistent within a cluster of documents and can entail more specificity. The hard partitioning can serve the purpose of dimensionality reduction and is expected to produce more specific and separable topics. On the other hand, overlapping biclustering avoids strict partitioning by assigning rows and columns to multiple biclusters. However, overlapping biclusters lead to visualization issues as it complicates the creation of comprehensive overviews and the identification of the common and distinctive items of the biclusters [Sun 2014]. Another distinction to mention concerning bicluster shapes is the *fuzzy* (or *soft*) assignment of objects to biclusters, i.e. weighted assignment to biclusters, versus *crisp* (or binary) assignment to biclusters. The fuzzy approach supplies more nuanced and accurate results, albeit more difficult to analyze by a lay audience. In contrast, the crisp assignment is easier to interpret but confers to the algorithm the responsibility of deciding cluster membership.

1.4.2 Biclustering techniques for topic extraction

Different approaches have been proposed for biclustering.

Metric-based methods optimize a criterion that measures the coherence of the bicluster structure with the original data. For instance, a criterion suitable for continuous data is the least-squares [Govaert 1995] formulated as follows:

$$Q(C, Z, W) = \sum_{i,j,k,l} z_{ik} w_{jl} (x_{ij} - c_{kl})^2 \tag{1.4}$$

where Z and W are respectively the row and column partitions, and $C = (c_{kl})$ is the matrix summarized by representative values of biclusters. The biclustering problem consists in minimizing criterion 1.4 to find the optimal partitions Z and W as proposed in *double k-means* [Govaert 1995]. Other methods optimize the mean square residue [Cho 2004] for biological data. However for contingency tables which are well suited for textual data, the chi-squared statistic and mutual information are more adapted as proposed in `CROKI2` [Govaert 1995] or `ITCC` [Dhillon 2003].

Graph-based biclustering methods consider the data matrix as a bipartite graph and look for diagonal biclusters. For instance, spectral relaxation optimizes the *minimum cut* criterion [Dhillon 2001a] or the *modularity* measure [Labiod 2011] in a term-document bipartite graph, but the recent `Coclus` algorithm [Ailem 2016] avoids the eigenvector computation through an alternating optimization procedure that outperforms the previous ones. In this thesis we used this algorithm to extract topics grouping similar documents and their most representative terms. This is why we describe `Coclus` in more details below.

Given the rectangular matrix X defined on $I \times J$ (a bipartite graph), the diagonal biclustering problem consists in finding simultaneously the row and column partition with a block seriation relation $C = ZW^T$ defined on $I \times J$ where $c_{ij} = 1$ if object i appears in the same block as feature j and $c_{ij} = 0$ otherwise. Through this formulation, $c_{ij} = \sum_k z_{ik}w_{jk}$ and the modularity criterion becomes:

$$Q(X, C) = \frac{1}{e_{..}} \sum_{i,j,k} (e_{ij} - \frac{e_{i.}e_{.j}}{e_{..}}) z_{ik}w_{jk}, \tag{1.5}$$

where $e_{..} = \sum_{ij} e_{ij} = |E|$ is the total edge weight (number of edges for a binary matrix X), $e_{i.}$ is the sum of the edge weights of i (or degree of i in the binary case), $e_{.j}$ is the sum of the edge weights of j (or degree of j in the binary case).
Equation 1.5 can be rewritten as follows:

$$Q(X, C) = \frac{1}{e_{..}} Trace[(X - \delta)^t ZW^t] = Q(X, ZW^t), \tag{1.6}$$

where $\delta = (\delta_{i,j})$ is the $n \times m$ matrix defined by $\forall i, j \delta_{ij} = \frac{e_{i.}e_{.j}}{e_{..}}$.
The alternating maximization of *Modularity* proposed in `Coclus` starts with the following proposition proven in [Ailem 2016]:

Proposition 1. *Let X be a $n \times m$ positive data matrix and C be a $n \times m$ matrix defining a block seriation, the modularity measure $Q(X, C)$ can be rewritten as*

1. $Q(X^W, Z) = \frac{1}{e_{..}} Trace[(X^W - \delta^W)^t Z] = Q(X, C)$
 where $X^W := \{e^W_{ik} = \sum_{j=1}^m w_{jk}e_{ij}; i = 1, .., n; k = 1, .., K\}$
 and $\delta^W := \{\delta^W_{ik} = \frac{e_{i.}e^W_{.k}}{e_{..}}; i = 1, ..., n; k = 1, ..., K\}$ *with* $e^W_{.k} = \sum_{j=1}^m w_{jk}e_{.j}$
2. $Q(X^Z, W) = \frac{1}{e_{..}} Trace[(X^Z - \delta^Z)^t W] = Q(X, C)$
 where $X^Z := \{e^Z_{kj} = \sum_{i=1}^n z_{ik}e_{ij}; j = 1, .., m; k = 1, .., K\}$
 and $\delta^Z := \{\delta^Z_{kj} = \frac{e_{.j}e^Z_{k.}}{e_{..}}; j = 1, ..., m; k = 1, ..., K\}$ *with* $e^Z_{k.} = \sum_{i=1}^n z_{ik}e_{i.}$

In Algorithm 1, `Coclus` maximizes alternatively $Q(X^W, Z)$ and $Q(X^Z, W)$ until convergence. This method does not have hyperparameters but requires a predefined number of biclusters. `Coclus` is computationally efficient for sparse matrices and requires a maximum of 20 iterations [Ailem 2016]. It surpasses other commonly used diagonal biclustering algorithms with respect to the quality of the document partition in terms of Accuracy, Normalized Mutual Information (NMI) [Strehl 2002], and Adjusted Rand Index (ARI) [Rand 1971], for most data sets [Ailem 2015, Ailem 2016].

Even though hard biclustering produces disjoint topics, the analyst still needs to understand topic frontiers and topic relationships. Existing similarity metrics use pairwise overlaps to compare a bicluster partition with a labeled ground truth [Horta 2014], and fail to build confusion matrices for disjoint partitions. In the case of disjoint biclusters, we lack similarity metrics needed to capture topic relationships. This is why we propose in this work a new similarity metric that works for both disjoint and overlapping biclusters in section 3.3.2.2.

Model-based biclustering methods consider that data can be generated by a mixture model also called Latent Block Model (LBM) [Govaert 2013] for the specific case of the biclustering problem. Given X a $n \times m$ matrix with the rows set $I = \{1,..,n\}$ and the columns set $J = \{1,..,m\}$, a latent block (or bicluster) is a set of matrix entries e_{ij} generated by a probability density function. *LBM* assumes that

1. data objects e_{ij} are independent and identically distributed conditionally to their bicluster membership z_i and w_j;

2. z and w are independent latent variables, i.e. $p(z,w) = p(z)p(w)$ generated by a $Multinomial$ distribution.

Given the matrix X, the number K of row clusters, the number L of column clusters and a probabilistic model family, the density function of the *LBM* framework is as follows:

$$f(X|\Theta) = \sum_{z,w} \prod_{i}^{n} \alpha_{z_i} \prod_{j}^{m} \rho_{w_j} \prod_{ij} \varphi(e_{ij}|\Theta_{z_i w_j}) \tag{1.7}$$

where the parameters $\Theta = \{\alpha, \rho, \Theta_{11}, ..., \Theta_{KL}\}$ are composed of $\alpha = \{\alpha_1, ..., \alpha_K\}$ and $\rho = \{\rho_1, ..., \rho_L\}$ the probabilities that any entry e_{ij} appears in row clusters, respectively column clusters, and Θ_{kl} is the parameter of the probability function φ related to the bicluster labeled by kl. The generative process is described as follows:

Algorithm 1 Coclus [Ailem 2016]

```
Input: binary or contingency matrix $X$, number of biclusters $K$
Output: partition matrices $Z$ and $W$
Initialization of W
repeat
    Compute $X^W = XW$
    Compute Z maximizing $Q(X^W, Z)$ by
        $z_{ik} = \arg\max_{1 \leqslant l \leqslant K} (e^W_{il} - \frac{e_{i.} e^W_{.l}}{e_{..}}) \forall i = 1, ..., n; k = 1, ..., K$
    Compute $X^Z = Z^t X$
    Compute W maximizing $Q(X^Z, W)$ by
        $w_{jk} = \arg\max_{1 \leqslant l \leqslant K} (e^Z_{lj} - \frac{e^Z_{l.} e_{.j}}{e_{..}}) \forall j = 1, ..., m; k = 1, ..., K$
    Compute $Q(X, ZW^t)$
until no change of $Q(X, ZW^t)$
```

1. For each row i: choose a row cluster $z_i = k \sim Multinomial(\alpha_1, ..., \alpha_K)$
2. For each column j: choose a column cluster $w_j = l \sim Multinomial(\rho_1, ..., \rho_L)$
3. For each entry (i, j): choose a value $e_{ij} \sim \varphi(e_{ij}|\Theta_{z_i w_j})$

Resolving the biclustering problem consists in estimating Θ that better shapes the structure of the observed data X. Similarly to one-way clustering, *Expectation Maximization* (*EM*) [Dempster 1977] and its classification variant *CEM* [Celeux 1992] are also used under the *LBM* to maximize the log-likelihood of observed data X. *EM* delivers fuzzy bicluster assignment in w and z while *CEM* delivers crisp bicluster assignment. The interested readers may refer to [Govaert 2013] for more details and a complete survey.

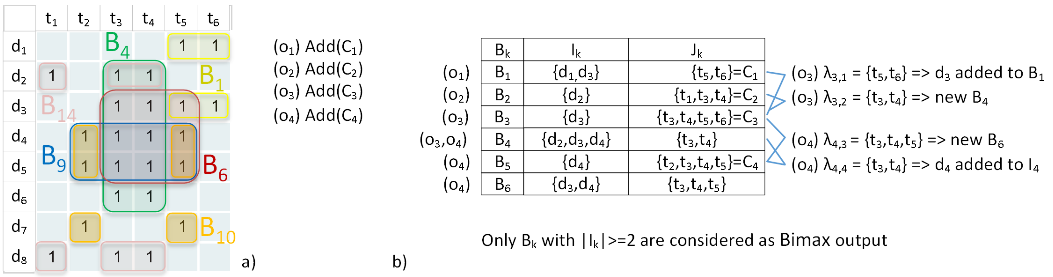
1.4.3 From coarse-grained topics to fine-grained *topic variants*

For textual data, the biclustering techniques presented above consider both the document and term dimensions to deliver homogeneous biclusters, that is a set of terms consistently grouped to describe a set of similar documents. Such biclusters can be considered as topics but they are not granular enough to identify specific viewpoints or facts shared by multiple sources; they are rather well-suited to model coarse-grained topics that capture high-order co-occurrences.

During the biclustering process, the optimization of the document partition improves their similarity and thus the likelihood to have co-occuring terms in similar contexts. In contrast, the optimization of the term partition improves their similarity and thus the likelihood to have multiple documents sharing similar terms in biclusters. Thereby the duality of the two alternated optimizations increases the intra-cluster commonality between both terms and documents, and increases the likelihood to have consolidated co-occurences in the biclusters, i.e. co-occurrences of terms found in multiple documents. Such fine-grained patterns reveal meaningful *topic variants* that can represent angles, viewpoints or facts shared by multiple sources and constitute a suitable structure for thorough text corpus investigation, typically for hypothesis validation, refining or generation. For this reason, we propose in this thesis a nested bicluster structure combining both fine-grained *topic variants* nested in coarse-grained topics. The pattern-based biclustering method presented below delivers the fine-grained *topic variants* of this structure.

1.4.4 Pattern-based biclustering for *topic variant* extraction

Prelić et al. [Prelić 2006] evaluate various biclustering methods for gene expression data. They propose *Bimax* a pattern-based overlapping biclustering algorithm satisfying a constraint of maximal inclusion (*MIC*). On a term-document matrix, *Bimax* identifies all



	t_1	t_2	t_3	t_4	t_5	t_6
d_1					1	1
d_2	1		1	1		
d_3			1	1	1	1
d_4		1	1	1	1	
d_5		1	1	1	1	
d_6			1	1		
d_7		1			1	
d_8	1		1	1		

(o_1) Add(C_1)
(o_2) Add(C_2)
(o_3) Add(C_3)
(o_4) Add(C_4)

	B_k	I_k	J_k
(o_1)	B_1	$\{d_1,d_3\}$	$\{t_5,t_6\}=C_1$
(o_2)	B_2	$\{d_2\}$	$\{t_1,t_3,t_4\}=C_2$
(o_3)	B_3	$\{d_3\}$	$\{t_3,t_4,t_5,t_6\}=C_3$
(o_3,o_4)	B_4	$\{d_2,d_3,d_4\}$	$\{t_3,t_4\}$
(o_4)	B_5	$\{d_4\}$	$\{t_2,t_3,t_4,t_5\}=C_4$
(o_4)	B_6	$\{d_3,d_4\}$	$\{t_3,t_4,t_5\}$

(o_3) $\lambda_{3,1} = \{t_5,t_6\}$ => d_3 added to B_1
(o_3) $\lambda_{3,2} = \{t_3,t_4\}$ => new B_4
(o_4) $\lambda_{4,3} = \{t_3,t_4,t_5\}$ => new B_6
(o_4) $\lambda_{4,4} = \{t_3,t_4\}$ => d_4 added to I_4

Only B_k with $|I_k|>=2$ are considered as Bimax output

Figure 1.4: (a) Bimax biclusters in a term-document matrix. No bicluster is completely covered by another one. (b) Intermediary state of Bimax after the addition of four documents.

distinct combinations of terms shared by multiple documents, i.e. consolidated co-occurrences. *Bimax* takes as input a binary matrix that can be obtained from the original matrix by applying a threshold. It delivers a *crisp* assignment of objects to multiple biclusters. The formal definition of Bimax biclusters is described hereafter.

Given I a set of n documents in rows, J a set of m terms in columns and X a binary $n \times m$ matrix defined as $X = \{e_{ij} \in \{0,1\}, \forall i \in [1..n], \forall j \in [1..m]\}$, a *Bimax* bicluster $B_k, k \in [1..\beta]$ is a submatrix $I_k \times J_k$ where $I_k \subset I$ and $J_k \subset J$ and for which all entries $e_{ij} = 1, \forall i \in I_k, \forall j \in J_k$. The *MIC* ensures that no bicluster is completely covered by another one (Figure 1.4a) and is defined as follows [Prelić 2006]:

Proposition 2. *For any $B_k, k \in [1..\beta]$ described by the pair (I_k, J_k) the* MIC *ensures that $\nexists\, B_{k'}$, $k' \in [1..\beta]$ with $I_{k'} \subseteq I$ and $J_{k'} \subseteq J$ such that (1) $\forall i' \in I_{k'}, \forall j' \in J_{k'} : e_{i'j'} = 1$ and (2) $I_k \subseteq I_{k'} \wedge J_k \subseteq J_{k'} \wedge (I_{k'}, J_{k'}) \neq (I_k, J_k)$.*

Prelić et al. propose an incremental approach defined in Algorithm 2. For any new document $i \in I$ described by the set of terms $C_i = \{j \in J | e_{ij} = 1\}$, all biclusters

Algorithm 2 Incremental Bimax [Prelić 2006]

```
var:$B$
for $i := 1$ to $n$ do
    $C_i := \{j | e_{ij} = 1 \wedge 1 \leqslant j \leqslant m\}$
    for all $k \in [1..|B|]$ do
        $\lambda_{i,k} := J_k \cap C_i$
        if $\exists k' \in [1..|B|]$ with $J_{k'} = \lambda_{i,k}$ then
            $I_{k'} := I_{k'} \cup \{i\}$
        else
            $B := B \cup \{(I_k \cup \{i\}, \lambda_{i,k})\}$
    if $\nexists k' \in [1..|B|]$ with $J_{k'} = C_i$ then
        $B := B \cup \{(\{i\}, C_i)\}$
return B
```

$B_k \in B$ are scanned to look for any intersection $\lambda_{i,k}$ between C_i and the terms J_k. $\lambda_{i,k}$ is the maximal set of terms shared between C_i and I_k. The condition in line 6 extends $B_{k'}$ to its maximality by adding i in $I_{k'}$ (in Figure 1.4b, o_3 adds d_3 to B_1 due to $\lambda_{3,1} = J_1 = \{t_5, t_6\}$). Otherwise, a new bicluster is added, with the documents $I_k \cup \{i\}$ (B_6 during o_4 with $J_6 = \{t_3, t_4, t_5\}$). Finally, after updating B, the condition in line 10 creates a new unitary bicluster (UB) with $I_k = \{i\}$ and $J_k = C_i$ (B_5 during o_4). Note that $\forall i \in I, \exists B_k \in B | J_k = C_i$. This ensures that any new row will also be compared with all previously added rows. In the result, only the biclusters with a minimum number of rows (`MinR`) and columns (`MinC`) are kept.

This algorithm works indifferently by either adding exclusively row vectors or exclusively column vectors and the number of biclusters β is discovered by the algorithm. Since the time complexity of `Bimax` ($O(nm\beta \min\{n, m\})$) is sensitive to the size and density of the matrix at hand, this method cannot be directly applied to the high-dimensional term-document matrix resulting from the entire corpus.

In this work, we rely on a top-level partitioning method to identify matrix blocks where `Bimax` is applied. We describe this nested biclustering method in the chapter 3. To our knowledge, the present work is the first attempt to use `Bimax` with textual data.

1.5 Topic Models vs. Biclustering

The purpose of this thesis is to propose a visual analytics system supporting investigation of text corpora. Due to the unstructured nature of free texts, such a system must rely on multiple techniques to reveal meaningful structures and to preserve the semantic of textual content.

This chapter has presented an overview of text mining techniques from which two main approaches stand out for coarse-grained topic extraction, namely *topic models* and biclustering methods. Indeed, they are able to group similar documents in topics and to select the most representative terms for direct interpretation. However exploring the topics through a set of terms and a set of documents allows only a superficial analysis of the dominant themes. A thorough investigation requires searching meaningful angles, viewpoints and facts treated in documents and at the same time identifying their commonalities and differences. Moreover identifying low-frequency events in a corpus can be of great interest for investigative journalists searching for alternate viewpoints. The fine-grained structure revealed by `Bimax` is well suited to this purpose. Through the proposed nested bicluster structure this exact pattern-based biclustering method reveals the internal structure of the topics, shaped actually by the upper-level topic extraction method. Therefore, we summarize below the theoretical differences between the topical structures obtained with both families of methods.

The topical structure delivered by any topic extraction method must preserve the semantic cohesion of the terms as well as their consistency with document contents. To this end, *topic models* as well as biclustering methods rely on the distributional hypothesis (Hypothesis 1 on page 14). To give a clear comprehension of what the topical structure fitting the distributional hypothesis is, we state the guarantees given by the different kinds of co-occurence patterns depicted in Figure 1.2. A topic with simple first-order co-occurrences does not give any guarantee that documents share terms and constitute similar contexts in topics. When topics contain high-order co-occurrences, the terms shared between documents increase the likelihood to have similar contexts but does not guarantee a large commonality between many documents. Therefore, we also consider the consolidated co-occurrences, i.e. term co-occurrences shared by multiple documents that ensure stronger commonality (or similarity) between multiple contexts.

As described in section 1.2, `LDA` provides overlapping clusters of terms but relies on a complete generative model for the documents [Shafiei 2006b]. Therefore `LDA` does not exploit the document-term duality and considers only high-order co-occurrences. `hLDA` extends the original `LDA` process by grouping documents in topics. While this could increase the similarity of the contexts, `hLDA` does not consider the term-document duality either; the document-topic assignments do not result from document relationships but actually from multiple groups of word-topic relationships [Wang 2009] (one group for each document) that deliver only high-order co-occurrences. In contrast, biclustering methods exploit the term-document duality during the process by alternating optimizations of criteria in both dimensions as proposed through the *Latent Block Model* [Govaert 2013]. Thereby biclustering captures high-order co-occurrences while promoting consolidated co-occurrences. To this end, Wang et al. propose a biclustering extension of `LDA` named the Latent Dirichlet Bayesian Co-clustering method [Wang 2009].

In this thesis we designed a numerical experiment to provide statistical evidence of the differences between the two families of methods we retained for topic extraction, i.e. *topic models* and *biclustering*. Moreover we conducted a user study to observe how these differences impact topic comprehension through our visual analytics system. In addition we were also concerned with the differences between disjoint and overlapping shapes of topics in terms of their interpretability. To this end, we compared `Coclus` [Ailem 2016] a graph-based diagonal biclustering approach with the hierarchical topic model `hLDA` [Griffiths 2004].

The next section surveys different existing evaluation approaches for clustering, biclustering and topic models.

1.6 Evaluation of clustering and topic models

Evaluating mathematical models and algorithms can be approached by both *validation* and *verification* processes [Isenberg 2013]. While a *validation* process ensures that a model is correct with respect to some aspects of reality, a *verification* process assesses the accuracy of algorithms with respect to the mathematical model. Both *validation* and *verification* are of empirical nature and rely on a ground truth as well as on recognized baseline models or algorithms. In this way, reproducibility of research is of great importance to improve confidence in the proposed models.

Hence, clustering methods are mainly evaluated by comparing new algorithms with state-of-the-art algorithms through extrinsic metrics measuring the agreement of a partition against labeled ground truth. Such metrics can be Accuracy, Normalized Mutual Information (NMI) [Strehl 2002], and Adjusted Rand Index (ARI) [Rand 1971].

Besides extrinsic quality measures, intrinsic measures can be used to assess the quality of a partition when ground trough is unavailable [Liu 2010]. These metrics are often used to optimize the parameters of clustering methods applied on unknown data, e.g. to find the appropriate number of clusters. The chosen metric can be the objective criterion optimized by the clustering method itself [Ailem 2016]. More general metrics are proposed in the literature for one-way clustering [Deborah 2010, Liu 2010, Vendramin 2010, Saber 2015] and for biclustering [Horta 2014]. They generally characterize intra-cluster compactness and inter-cluster separation [Liu 2010]. For a bicluster partition, existing approaches examine overlapping elements in both dimensions [Horta 2014] and are not fit to measure inter-cluster similarity/dissimilarity in a hard-partition such as diagonal biclustering. In this thesis we propose a similarity measure designed for both overlapping and disjoint biclusters. Neither extrinsic or intrinsic metrics take the semantic quality of the term partition into consideration.

Topic models are often evaluated through predictive metrics [Wallach 2009] measuring to which extent a learned model applies to new documents. These measures are based on *perplexity* or *log likelihood* which do not evaluate the usefulness or the meaningfulness of topics. Chang et al. [Chang 2009] evaluate topic semantic through analyst efficiency in accomplishing two tasks: searching for intruding terms in topics and searching for intruding topics in documents. In the user study of Newman et al. [Newman 2010], analysts were asked to assign a score to judge the usefulness of topics in the light of their top 10 terms. These scores are compared with those obtained automatically with *Pointwise Mutual Information* (*PMI*), by measuring the cohesion of the top 10 terms in light of external data such as Wikipedia or Google n-grams. Newman et al. made the assumption that the usefulness of topics is assessed by judging the semantic association of terms. *PMI* is specifically well adapted to such semantic cohesion of terms through their co-occurrences [Turney 2010].

The results of the experiment show that the correlation between the human scores and *PMI* scores exceeds the inter-rater correlation, telling that *PMI* is a good candidate to measure the usefulness of topics based on external data. Several other similar metrics have been used to measure the semantic cohesion of topic terms [Mimno 2011, Aletras 2013], but for all these experiments the analysis was limited to the top 10 terms.

However, according to Alexander and Gleicher [Alexander 2015], the analysis of document relationships (a prerequisite of investigation tasks) cannot be limited to so few terms per topic. They compare the achievement of such analysis without limitation of the number of terms and with limitation to the top 50 terms. Their study shows that the whole term distribution must be taken into account. According to them, the quality of topic models relies on their ability to reveal subtle co-occurrence patterns representing document relationships. The distributional hypothesis (Hypothesis 1 on page 14) states the same idea in the following terms : the terms tends to be similar if they co-occur in similar contexts. Hence, in our controlled experiment in the section 4.5 we studied the human comprehension of topics through the term co-occurrence patterns revealed by `Bimax` biclusters (a.k.a. *topic variants*) and their hierarchical structure for the tasks undertaken in journalistic work.

The next chapter presents related work in the area of text visualization and visual analytics systems for the exploratory analysis of text corpora.

CHAPTER 2

Visual Analytics of text corpora

Contents

2.1 Feature-based visualizations 32

2.1.1 Document-centric approach 33

2.1.2 Term-centric approach 34

2.2 Topic-based visualization 37

2.2.1 Temporal evolution of topics 37

2.2.2 Exploring semantic relationships 42

2.3 Model-driven design vs. domain-specific design 46

2.3.1 Model-driven design: from coordinated multiple views to interactive clustering 46

2.3.2 Domain-specific visual analysis of text corpora 49

2.4 Bicluster Visualization 52

Information Visualization employs computer graphics and interactions to support human data analytic tasks. As described by Jacques Bertin [Bertin 1977], information is visually encoded by using basic graphical units called *marks* (points, lines, areas, surfaces, volumes) whose attributes, referred to as *visual variables* (position, length, shape, value, color, orientation, texture), can be modified according to data and tasks. In practice, the design of a visualization software requires a good comprehension of user tasks to properly map data types and operations to visual encoding and interactions. In chapter **??**, we present the methodology we adopted for the visual design proposed in chapter 4.

The purpose of text corpus visualization is first to speed up the analysis of textual content by avoiding reading all the documents. During investigative work, the analyst seeks passages of texts that allow to verify, refine or generate her working hypothesis. The widely recognized Visual Information-Seeking mantra “Overview first, zoom and filter, then details on demand” [Shneiderman 1996] is well suited to this purpose. However, the visual exploration of textual data relies, to a great extent, on data processing algorithms presented in chapter 1 that must be painstakingly configured to address user needs. The more recent Visual Analytics paradigm [Thomas 2005] places the user at center-stage in a process supporting analytic reasoning through the combination of automated and visual analysis [Keim 2010]. Ideally, such an approach relies on semantic interactions [Endert 2012], i.e. visual manipulations that control the underlying model and enhance the sensemaking process.

In the present chapter we describe the most relevant research work related to this thesis, pertaining to the visual analysis of text corpora, including text, topic and bicluster visualization, and visual analytics solutions supporting the investigation of text corpora. Text visualization has intensively been studied in the last decade and several surveys give an overview of the large scope of this research field [Šilić 2010, Alencar 2012, Gan 2014, Wanner 2014]. Recently an online interactive survey[1] proposes a taxonomy of text visualizations involving multiple aspects. They concern high-level analytic tasks, low-level tasks for interactions and visual representations, the source (*Single document*, *Corpora*, *Streams*), the properties (*Geospatial*, *Timeseries*, *Networks*) and the domain of data (*Social Media*, *Scientific papers*, *Editorial Media*, *Literature*, etc.), as well as visualization techniques [Kucher 2015]. The present survey emphasizes the advantages and drawbacks of state-of-the-art approaches in light of three other aspects of differentiation. First, we differentiate the structuration level of text corpora leveraged by the visualizations, from the low-level feature-based visualizations up to topic-based visualizations. Secondly, we differentiate two kinds of patterns emphasized by visualizations of text corpora, i.e. temporal patterns or semantic relationship. Finally, visual analytics tools with coordinated multiple views are differentiated through their design orientation, i.e. model-oriented design or domain-oriented design. We start this survey with the lower level of structuration of text corpora, qualified as feature-based visualizations.

2.1 Feature-based visualizations

Many text visualizations are based on data features such as terms, documents or sentences. Document- or term-centric visualizations are mainly approached by the spatialization of items in the form of scatterplots, node-link diagrams or tag clouds. Based on abstract shapes and multiple visual variables, these visualizations aim to convey structural patterns revealed by text mining algorithms. However, the interpretation of patterns is only possible through labels displaying text features, i.e. terms. The main difficulty in text visualization is to combine, in the same visual space, the representation of structural patterns such as semantic relationships or temporal context, and the uncluttered representation of meaningful text features.

[1] http://textvis.lnu.se

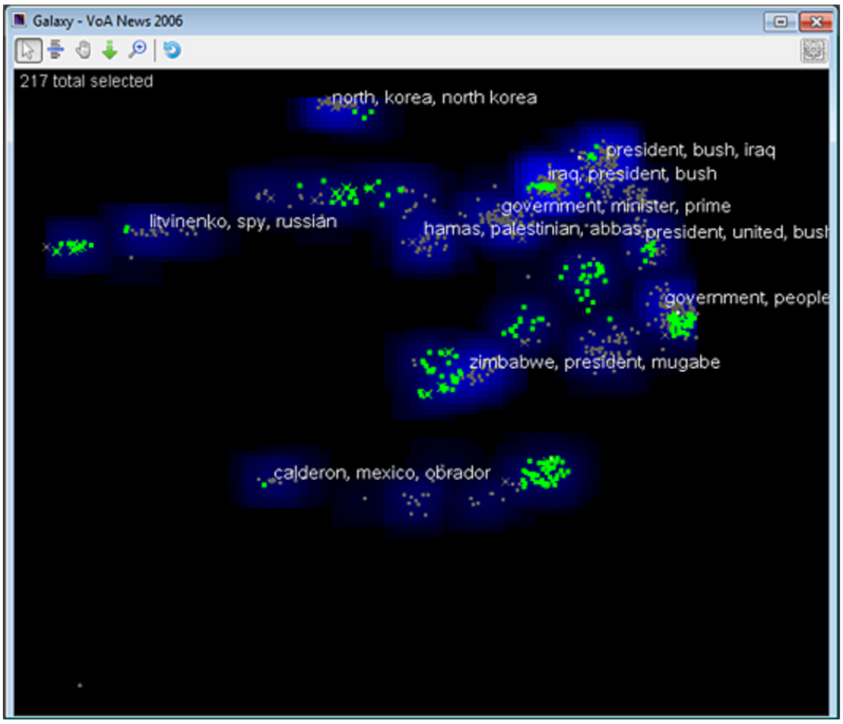


Figure 2.1: The "Galaxy View" of IN-SPIRE [Wise 1995] spatializes documents in a scaterplot where the proximity of the objects represents the similarity of documents. Several clusters of documents can be depicted and some of them are directly interpretable through the three most representative terms.

2.1.1 Document-centric approach

Projection techniques such as Principal Component Analysis (*PCA*), Correspondance Analysis (*CA*) or Multi-Dimensional Scaling (*MDS*) are widely used to visualize clusters of documents with scaterplots. IN-SPIRE [Wise 1995] proposes a "Galaxy View" where similarity between documents is represented by spatial proximity. The document clusters are labeled at their position in the plot with the most representative terms. We observe in Figure 2.1 that the number of clusters having a label and the number of terms displayed for each are limited to a few in order to prevent label overlaps. More recently, ForceSPIRE [Endert 2012] improves sensemaking through an interactive node-link representation of documents with a force-directed layout. Semantic interactions enable direct manipulation of spatialization and modification of a keyword weighting scheme with respect to the user's analytical reasoning. Such document-centric approach is too granular to give a broad overview of content. Moreover, the study of Brehmer et al. [Brehmer 2014] reveals that journalists prefer hierarchical navigation of document clusters to scatterplot representations.

Figure 2.2: A tag cloud created with Wordle [Viegas 2009]. The visual variables such as label size and color can be mapped to metrics such as word frequency or TF-IDF. The orientation of terms can be fixed horizontally to ensure a good readability.

2.1.2 Term-centric approach

To understand textual content or extracted topics without reading the documents, Tag Clouds [Viegas 2009] emphasize the most important terms through their size or color as shown in Figure 2.2. Tag clouds are commonly adopted in many Web 2.0 sites and some of them propose to generate tag clouds from texts. It was shown that flat lists are more efficient than tag clouds for tasks involving a ranking strategy such as alphabetical search or identifying the "top-N" frequent terms [Halvey 2007, Rivadeneira 2007, Archambault 2013]. However, when the ranking strategy is unknown or is not constitutive of the task, having to scroll through the lists limits user efficiency and no significant difference is found between lists and tag clouds representations. In this case, the latter becomes preferable due to the more engaging interface [Archambault 2013]. Recently, many extensions of tag clouds have been proposed to lay out temporal patterns or semantic structure. For instance, Spark

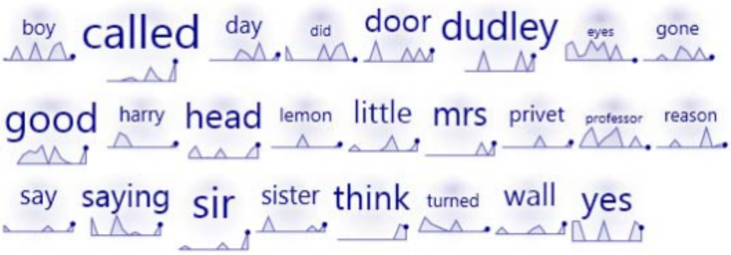


Figure 2.3: Sparklines of SparkCloud [Lee 2010]. The terms are displayed with a line plot representing the evolution of their importance over time.

Clouds [Lee 2010] incorporates sparklines [Tufte 2006] in tag clouds, i.e. line charts depicting the temporal evolution of terms in Figure 2.3. Without temporal alignment, it is difficult to identify common trends between terms. Giving the temporal context of text contents is crucial for investigative journalism [Fulda 2015]. We discuss some preliminary work in the field of visual analytics of text stream in chapter 5, but it does not constitute the core contribution of this thesis.

Using an explicit hierarchy, Word Tree [Wattenberg 2008] represents an interactive form of the "keyword-in-context" method for information retrieval. In Figure 2.4, the searched keyword is placed at the root of a tree where the children are the subsequent terms of the sentences retrieved as the result of a search query. Quite informative, this approach shows the multiple contexts in which one term or multiple collocated terms appear. Nonetheless this bottom-up approach does not give a broad overview of the corpus contents as recommended by the Visual Information Seeking mantra [Shneiderman 1996].

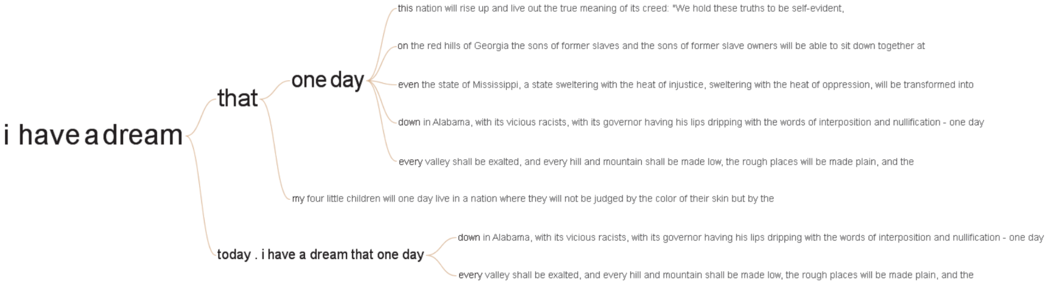

i have a dream
that
one day
this nation will rise up and live out the true meaning of its creed: "We hold these truths to be self-evident,
on the red hills of Georgia the sons of former slaves and the sons of former slave owners will be able to sit down together at
even the state of Mississippi, a state sweltering with the heat of injustice, sweltering with the heat of oppression, will be transformed into
down in Alabama, with its vicious racists, with its governor having his lips dripping with the words of interposition and nullification - one day
every valley shall be exalted, and every hill and mountain shall be made low, the rough places will be made plain, and the
my four little children will one day live in a nation where they will not be judged by the color of their skin but by the
today . i have a dream that one day
down in Alabama, with its vicious racists, with its governor having his lips dripping with the words of interposition and nullification - one day
every valley shall be exalted, and every hill and mountain shall be made low, the rough places will be made plain, and the

Figure 2.4: The Word Tree visualization [Wattenberg 2008] allows to retrieve all sentences containing a searched keyword and displays in a tree the divergence of the subsequent fragment of sentences.

Tree Cloud [Gambette 2010] uses a node-link diagram to organize the terms like a phylogenetic tree to reflect their semantic proximity (Figure 2.5). This solution places the terms at the leaves of the tree and internal node ramifications represent semantic divergences between terms. Hence, terms are spatially grouped by semantic proximity to depict common lexical fields. However this approach requires a drastic feature selection to keep the most representative terms only and tree ramifications are difficult to interpret because only leaves are labeled. Spark Clouds, Word Tree and Tree Clouds are term-centric representations where topics are not clearly delineated.

Parallel Tag Clouds [Collins 2009] is a mixture of tag clouds and parallel coordinates as depicted in Figure 2.6. Each column represents a list of terms representing a subset of the data (the facets) and term relationships are displayed explicitly on user interaction. While multiple topics can be shown simultaneously, it does not scale to many topics due to the spatial organization of terms in columns.

We observe in this section that, on the one hand, marks with abstract shapes in scatter-

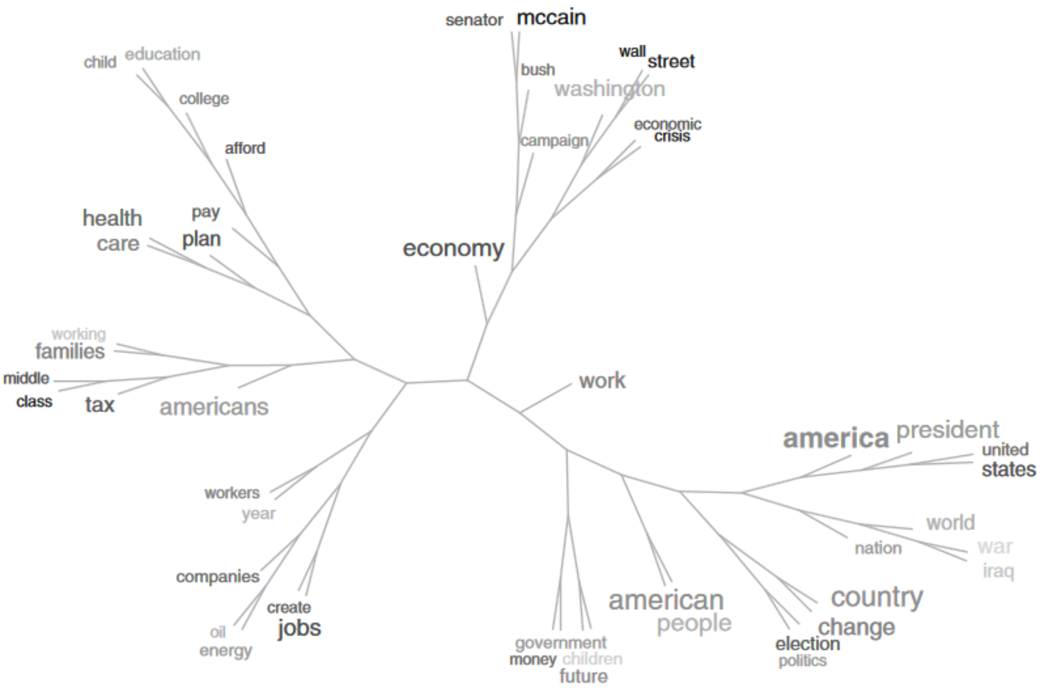


Figure 2.5: The TreeCloud visualization based on phylogenetic tree [Gambette 2010] show semantic relationships between terms. The internal nodes representing semantic divergences are not labeled.

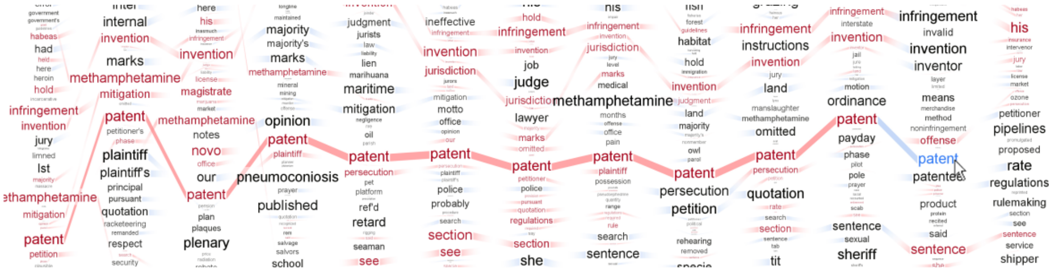


Figure 2.6: The Parallel Tag Cloud visualization [Collins 2009] where each column corresponds to a facet of data. Common terms are explicitly linked to reveal semantic relationships between facets.

plots or node-link diagrams convey the structure revealed by analytic processing, typically the semantic proximity of terms or documents. A remaining problem is the interpretation of such a structure that can be done by overlaying the terms of the underlying content. Indeed, the difficulty is to give enough context allowing correct interpretation while avoiding the visual clutter induced by label overlaps. One solution relies on feature selection to keep only the most representative terms. Another solution consists in displaying labels on demand based on user interaction. While a straightforward dynamic labeling approach reveals single labels of hovered items by showing a tooltip, excentric labeling [Fekete 1999] can simultaneously lay out multiple labels included in a larger range giving more context for a more precise and correct interpretation. Such an on-demand labeling remains suboptimal to provide a broad overview of contents. One the other hand, tag clouds optimize the spatialization of labels by avoiding overlaps. Although the number of terms with a readable size is limited, fisheye interaction can be added to enhance the readability of small labels. Tag clouds favor hence the interpretation of text contents but does not support a thorough understanding of the underlying semantic structure. To conclude, a challenging issue with text corpus visualization is to deal with this trade-off: giving an overview of the textual content allowing a precise semantic interpretation while depicting the topical structure as a meaningful starting point for deeper analysis.

2.2 Topic-based visualization

In complement to term/document-centric visualization discussed above, many visual analytics tools are devoted to topic exploration. Indeed topic extraction methods surveyed in chapter 1 provide a useful structure for text summarization or categorization, helping corpus exploration. Depending on user needs, topic visualizations emphasize either the analysis of temporal context or semantic relationships. Since in many situations experts need both, visual analytics solutions often resort to coordinated multiple views due to the difficulty of simultaneously representing data structure and time in the same 2D visual space [Bach 2016]. We distinguish two general forms of visual text analytics approaches. The first relies on text mining algorithms and display the result with interactive visual manipulation that only change the visual encoding. The second follows a human-centered machine learning approach [Sacha 2016] that relies on semantic interactions, i.e. visual interactions incorporating human knowledge in machine learning algorithms.

2.2.1 Temporal evolution of topics

NewsLab [Ghoniem 2007], Tiara [Wei 2010], TextFlow [Cui 2011] and Parallel-Topics [Dou 2011] show the temporal evolution of topics using variants of Theme

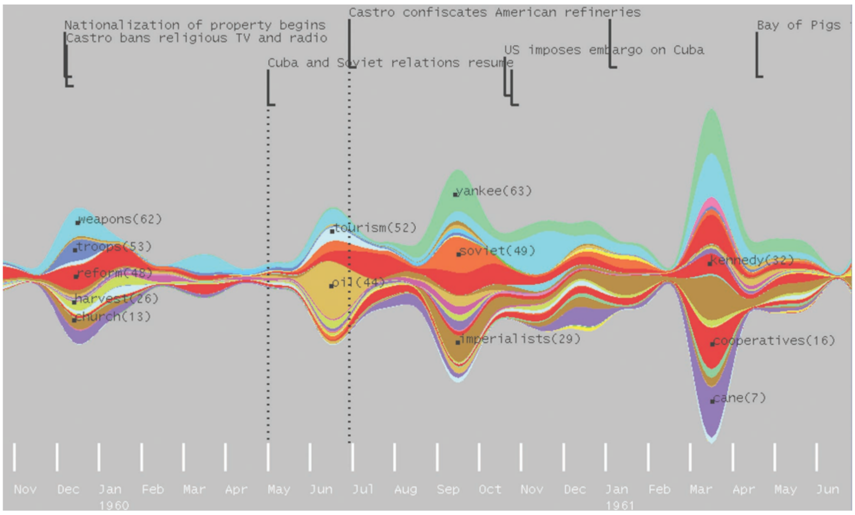


Figure 2.7: The original Theme River visualization [Havre 2002]. Temporal alignment of topics allows to compare evolution of topics relatively to each other. The labels are however limited to one term.

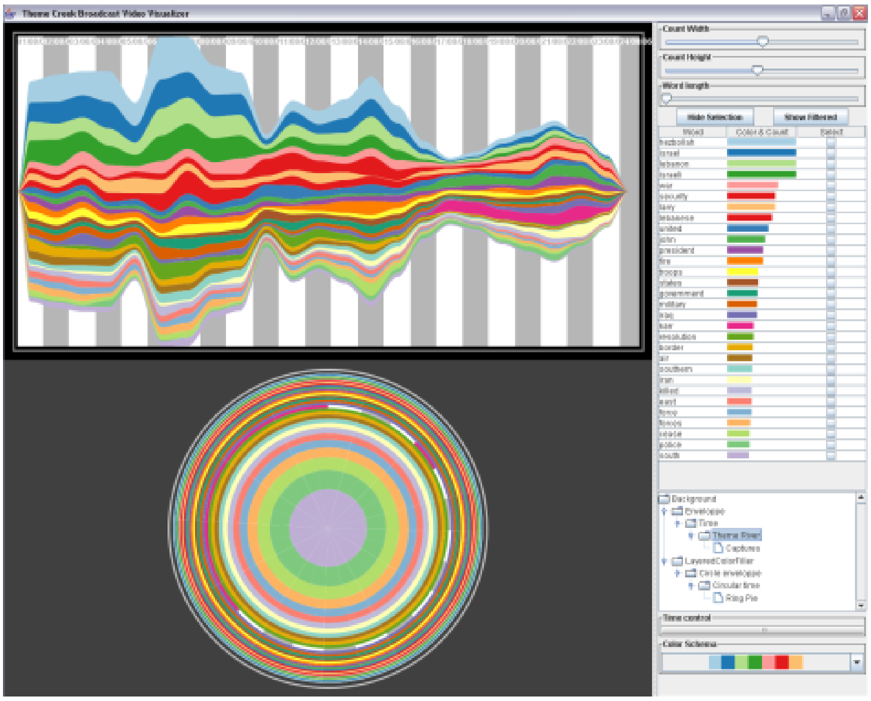


Figure 2.8: NewsLab, a multi-scale theme river with a burstiness-based filtering and a replay animation [Ghoniem 2007]. The NewsRings view reveals better emergence, disapearence patterns as well as periodic patterns.

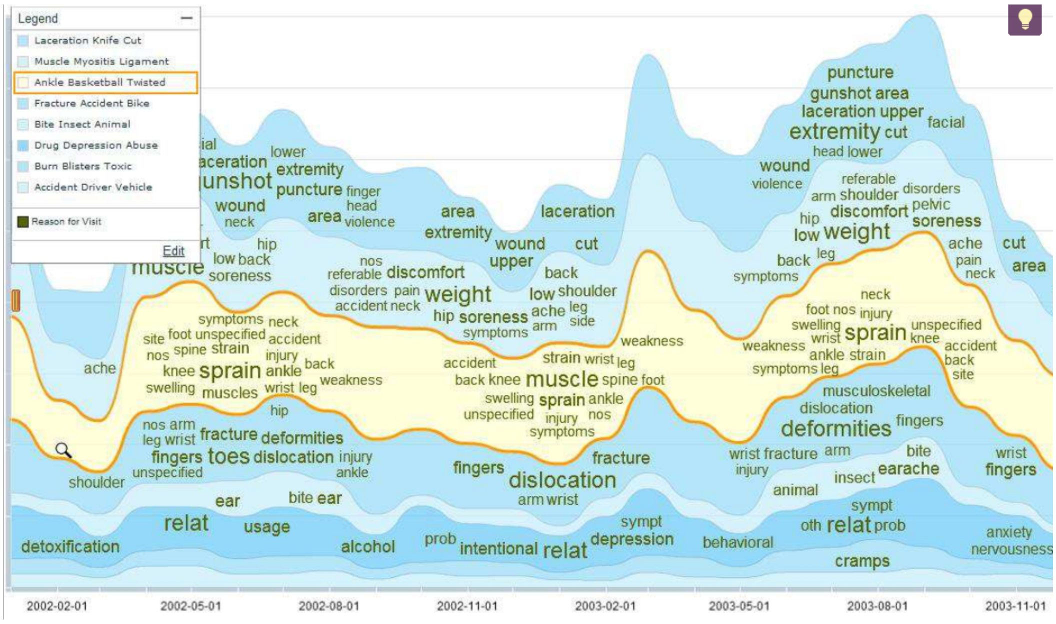


Figure 2.9: Tiara visualization with optimized placement of tag clouds [Wei 2010]. Thin layers have not enough space for labeling.

River [Havre 2002]. Dedicated to temporal data, Theme River represents topics as stacked areas as shown in Figure 2.7. Time is represented on the x-axis and the thickness of the layers on the y-axis encodes the relative importance of the topics in the corpus. The *Theme River* visualization reveals global temporal patterns such as the evolution of the importance of topics, but also local patterns such as the emergence or disappearance of topics. In addition the temporal alignment with respect to the x-axis as well as the relative thickness of topics allow comparative analysis and offer a great insight for contextualization. However the difficulty remains in the interpretation of topics due to the lack of place for labeling.

NewsLab [Ghoniem 2007], as shown in Figure 2.8, supports the exploration of text streams through a multi-scale theme river. While global patterns can be explored through a coarse time scale, a replay animation allows to explore patterns at a finer time scale. While this approach can deals with hundreds of topics in a theme river the interpretation relies on a colored legend and interactive labeling. Multiple sorting and filtering are proposed based on keywords, time or event burstiness. A NewsRings complementary view allows to better identify emergence/disapearence of topics as well as periodic patterns.

Tiara [Wei 2010] incorporates tag clouds in the layers of a *Theme River* by considering three criteria [Liu 2009]: temporal proximity, content legibility and content amount. Their algorithm optimizes the placement of the tag clouds (Figure 2.9), but the main limitation stems from the limited number of layers whose thickness is large enough to fit readable labels. Like NewsLab, Tiara proposes multiple topic ordering criteria including topic cov-

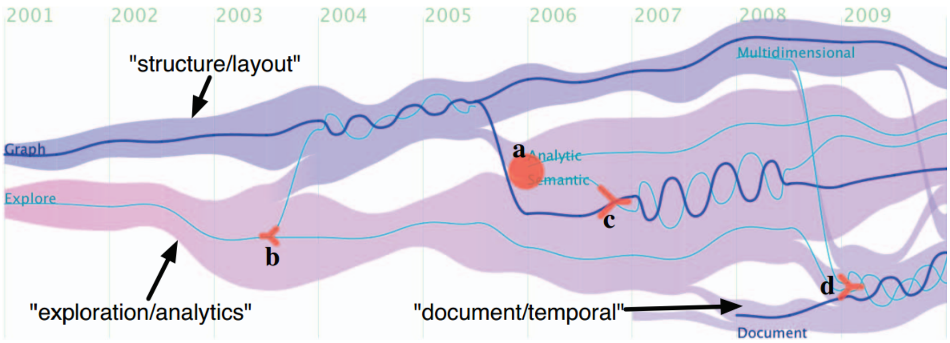


Figure 2.10: TextFlow visualization with glyphs representing topic merges and splits in a river metaphor [Cui 2011]. Topic labeling is provided on user interaction in order to prevent visual clutter.

erage and variance, distinctiveness or information gain.

In contrast, TextFlow [Cui 2011] displays glyphs representing meaningful patterns such as topic merges and splits in a river flow metaphor (Figure 2.10). In this case, direct interpretation of topics in the main view is not possible. Topic semantic is shown in a coordinated tag cloud view.

Focusing on event detection, LeadLine [Dou 2012] relies on the extraction of topics and named entities. Events are characterized from the content by the *4W* questions: *What*=topics, *Where*=location names, *Who*=person/organization names and *When*=documents' timestamps. In Figure 2.11, topics are arranged in an EventRiver [Luo 2012], i.e. fixed-size areas where event burstiness is detected through bubbles

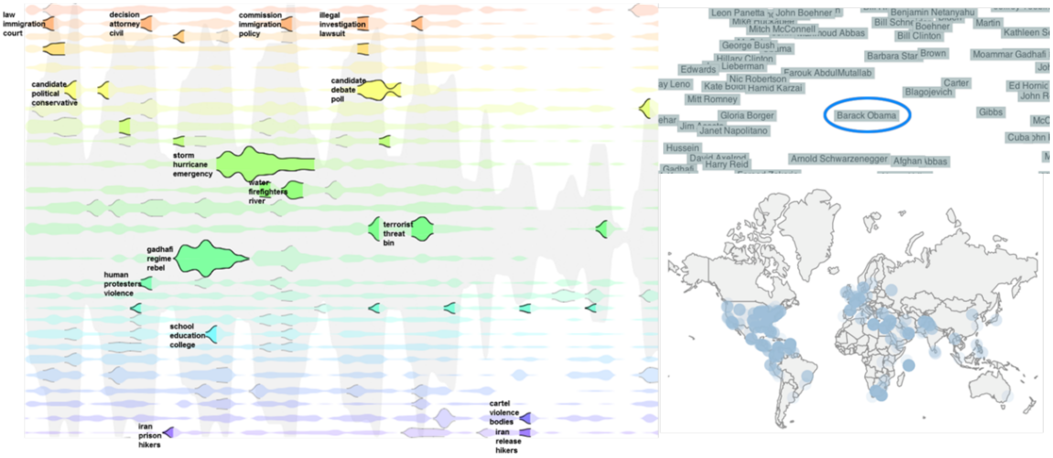


Figure 2.11: LeadLine for event detection with named entities [Dou 2012]. Each event aspect is depicted in coordinated multiple views. For instance, topics are represented through an EventRiver [Luo 2012], a node-link diagram and a geographical map represent person relationships and location names respectively.

labeled with two or three keywords following the time axis. Other aspects of the events are depicted through coordinated multiple views, namely a graph of entity relationships, a geographical map placing location names, and a topic cloud view showing more topic terms line by line.

As described above, time-oriented visualization has raised a great deal of interest in visual text analytics in view of the importance of time for the contextualization of text streams. We also present in chapter 5 an approach based on dynamic *Theme Rivers* and a time-synchronized geographical map to support the investigation of twitter-like text streams. A drawback of explicit time representations is that one dimension (the x-axis) of the 2D space is used for time, making more challenging the representation of complex patterns such as topic similarity in the remaining dimension (the y-axis).

HierarchicalTopics [Dou 2013] address this issue with an adaptable hierarchy grouping the topics by similarity (Figure 2.12). Their user study shows that a hierarchical structure improves the exploration of large numbers of topics compared to flat ordered lists or layers. However, the semantic of topics is depicted through horizontal flat lists with only a few terms which limit their interpretation. In addition, the internal nodes are not informative. They have to be annotated manually after the analysis of the term lists placed on the leaves of the tree.

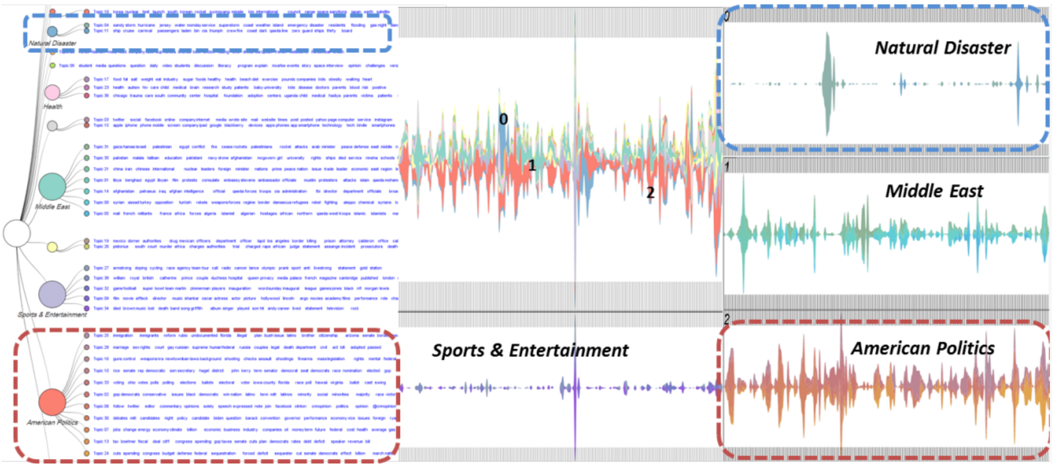


Figure 2.12: HierarchicalTopics [Dou 2013] allows a hierarchical exploration of topics while depicting temporal patterns. Topics are interpreted through flat lists containing the N most representative terms.

In conclusion, the depiction of topical temporal patterns can be achieved to the detriment of the semantic interpretation of textual content. Indeed explicit representations of time in a 2D space leave only one remaining dimension for the spatialization of semantic structure and labels. Various solutions handling the visual representation of time and data structure are surveyed in [Bach 2016] with the help of the *Space-Time Cube*, including small multiple and animations [Rufiange 2013, Bach 2014, Archambault 2016]. In the present thesis, we propose two dynamic approaches for the investigation of text streams in the fields of visualization and data mining. Nonetheless we mainly focused our work on the visual exploration of *topics* and *topic variants* for the investigation of static text corpora. The following section surveys topic visualizations that favor semantic analysis of topical structure to the detriment of temporal contextualization.

2.2.2 Exploring semantic relationships

TopicNet [Gretarsson 2012] allows the visual analysis of large corpora through a document-topic graph and a force-directed layout built on the output of LDA. Document-topic edges are created with a threshold applied on the topic probability distribution of each document. This threshold can be modified interactively by the user through a slider to control the number of edges, and consequently the amount of visual clutter. A first layout preserves topic similarity by fixing their position based on their similarity computed with Kulback-Leiber divergence and MDS. Then, document positions are computed with the force-directed algorithm as shown in Figure 2.13a. The second layout preserves the timestamp order of documents by fixing their position on the circumference of a circle. Then, topic position is given by a force-directed algorithm, as shown in Figure 2.13b. Multiple interactions allows the expert to filter and explore data from different perspectives. In

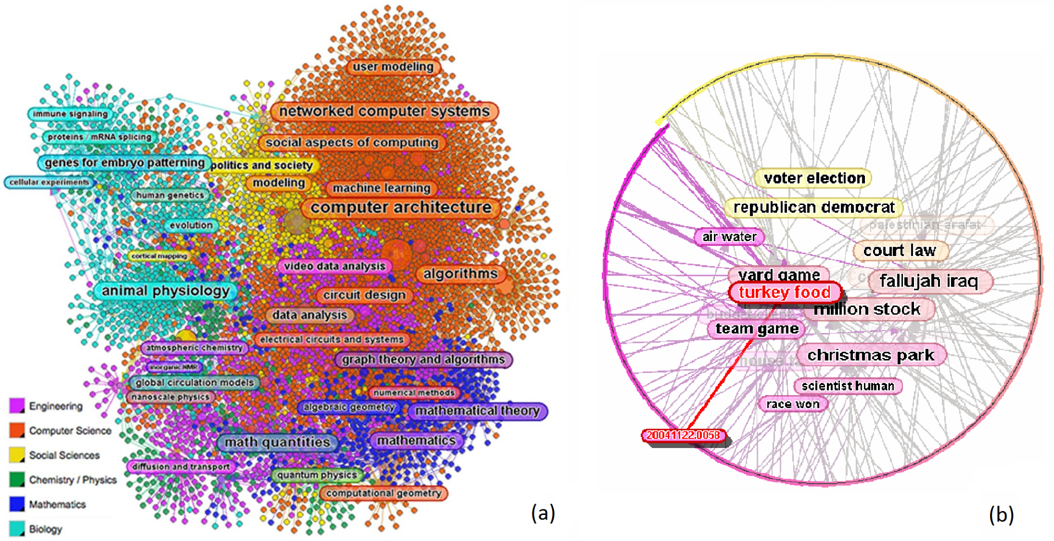
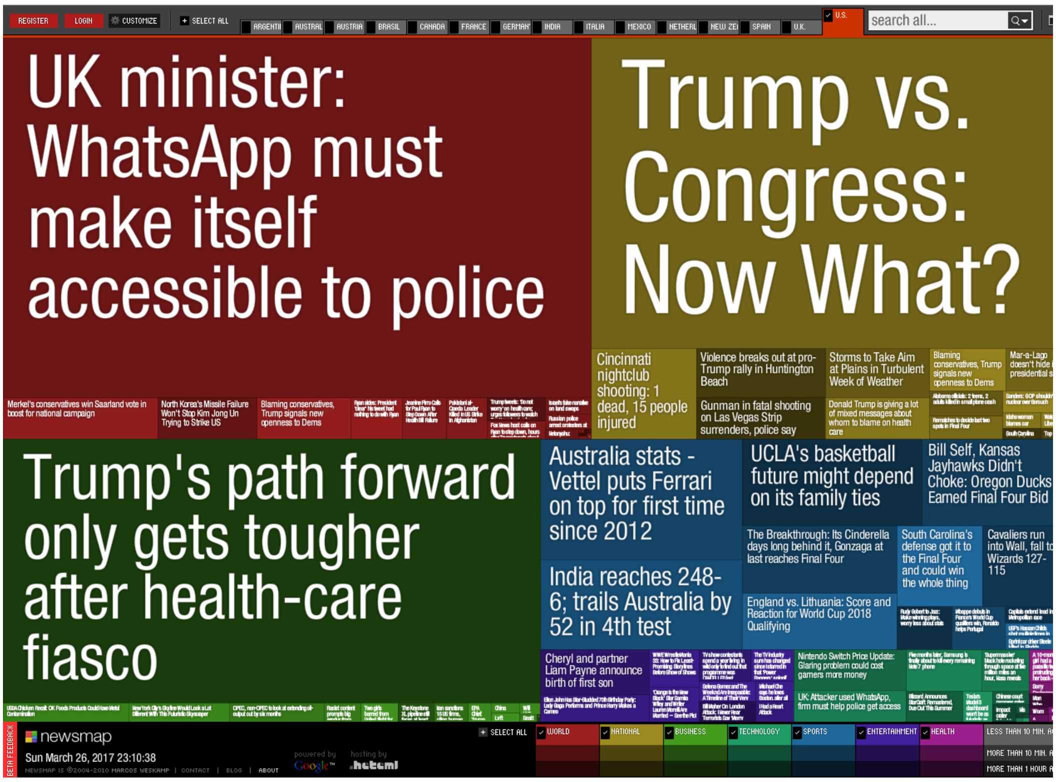


Figure 2.13: TopicNets[Gretarsson 2012], a node-link diagram representing document-topic relationships obtained with LDA. (a) The topic-similariy layout fixes the topic positions based on their similarity and computes the position of the document with a force-directed layout. (b) The order-preserving layout fixes the documents based on a predefined order, here the timestamps, and computes the topic position with a force-directed layout.

addition LDA can iteratively be applied on filtered data, revealing sub-topics of a selection of interest. Node-link diagrams provide an intuitive representation of node connections, clusters and outliers but often result in visual clutter in dense areas inducing node overlaps and edge crossing. Moreover, the number of terms describing each topic must be limited to reduce node overlaps.

News Map [Weskamp 2004] in Figure 2.14 organizes a larger number of topics in a two-level treemap where the size of each tile is mapped to the number of enclosed documents. In the first level, the topics are grouped in predefined color-encoded categories (e.g. Sports, Business, etc.). In the second level, the topics are labeled using the title of a representative document. While single titles are quite informative, they reflect a limited and incomplete view of the actual topic content.



UK minister: WhatsApp must make itself accessible to police

Trump vs. Congress: Now What?

Merkel's conservatives win Saarland vote in boost for national campaign

North Korea's Missile Failure Won't Stop Kim Jong Un Trying to Strike US

Blaming conservatives, Trump signals new openness to Dems

Cincinnati nightclub shooting: 1 dead, 15 people injured

Violence breaks out at pro-Trump rally in Huntington Beach

Storms to Take Aim at Plains in Turbulent Week of Weather

Blaming conservatives, Trump signals new openness to Dems

Gunman in fatal shooting on Las Vegas Strip surrenders, police say

Donald Trump is giving a lot of mixed messages about whom to blame on health care

Trump's path forward only gets tougher after health-care fiasco

Australia stats - Vettel puts Ferrari on top for first time since 2012

UCLA's basketball future might depend on its family ties

Bill Self, Kansas Jayhawks Didn't Choke: Oregon Ducks Earned Final Four Bid

The Breakthrough: Its Cinderella days long behind it, Gonzaga at last reaches Final Four

South Carolina's defense got it to the Final Four and could win the whole thing

Cavaliers run into Wall, fall to Wizards 127-115

India reaches 248-6; trails Australia by 52 in 4th test

England vs. Lithuania: Score and Reaction for World Cup 2018 Qualifying

Cheryl and partner Liam Payne announce birth of first son

Nintendo Switch Price Update: Glaring problem could cost gamers more money

UK: Attacker used WhatsApp, firm must help police get access

Sun March 26, 2017 23:10:38

NEWSMAP IS ©2004-2010 MARCOS WESKAMP | CONTACT | BLOG | ABOUT

powered by Google™ hosting by .hotcml

Figure 2.14: NewsMap: a Tree Map for breaking the news [Weskamp 2004] where rectangle size encodes the number of documents in the topic and each label corresponds to the title of one representative document within the topic.

Based on tag clouds, WordBridge [Kim 2011] is a composite view of tag clouds incorporated in a node-link diagram (Figure 2.15). The topic and their links are represented by tag clouds, eliciting the nature of the relationships. The drawback of this approach is that the whole graph cannot be visualized and expand/collapse interactions are needed. This approach is then limited to local analyses and does not convey a broad overview.

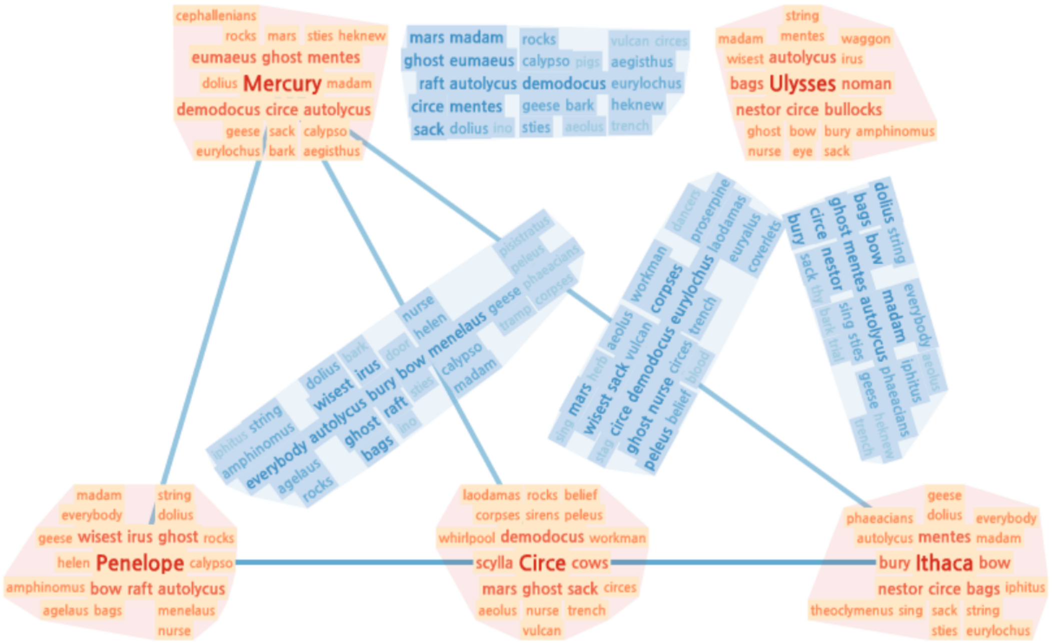


Figure 2.15: The Word Bridge visualization [Kim 2011] where tag clouds inform topics as well as their relationships. This approach limits the number of displayed topics and relationships.

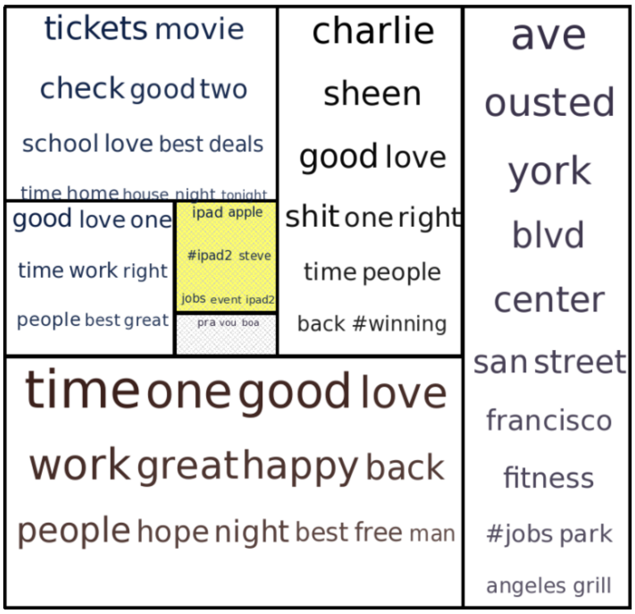


Figure 2.16: TwitterCrowd: a treemap of tag clouds [Archambault 2013]. Space filling layout allows to display more terms for important topics and less terms for minor topics.

TwitterCrowd [Archambault 2013] arranges tag clouds in a one-level treemap as shown in Figure 2.16. The advantage of the treemap space filling approach is that it optimizes the available space to extend the tag clouds to their maximality. The relative size of tiles allows to display more terms for important topics and less terms for minor topics. However topic similarity is not visually encoded.

TopicPanorama [Liu 2014] 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. An incremental graph matching algorithm merges topic graphs from multiple sources by integrating user feedback into the algorithm. While topic similarity is represented as a graph at each resolution level, only a subset of topics are labeled with two or three terms (Figure 2.17). TopicPanorama unravels a more exhaustive tag cloud through user interaction.

The *Weighted Topic Map* view we present in this thesis supports exhaustive interpretation of all topics through multiple tag clouds while reflecting their relative importance and similarity.

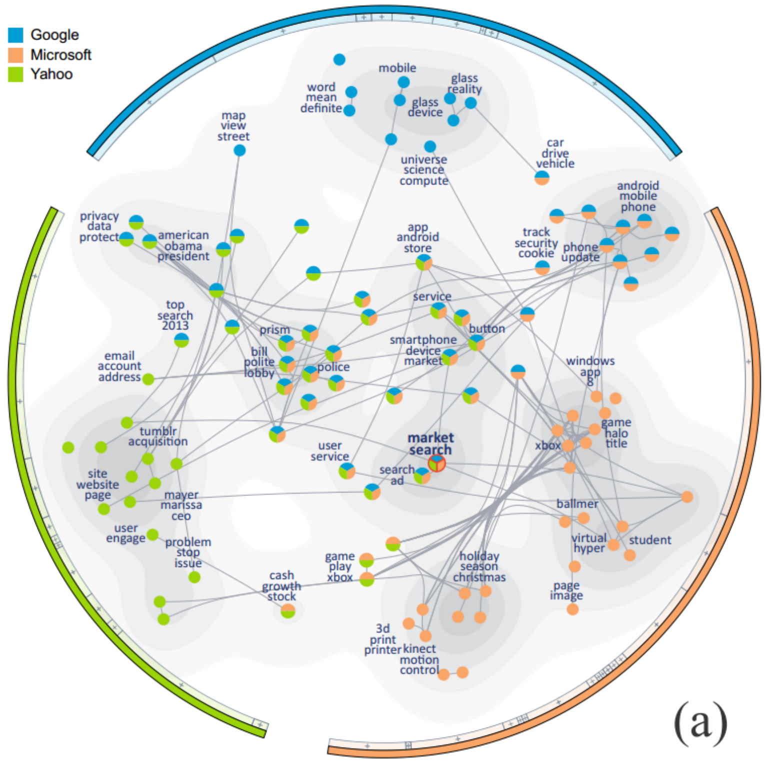


Figure 2.17: Topic Panorama for a multi-resolution comparison of multiple sources [Liu 2014]. Each topic is represented as a pie chart depicting its presence in one or several sources. Two or three terms allow to interpret a subset of topics. The radial stacked tree placed at the circumference of the radial layout shows multiple levels of topic clusters.

2.3 Model-driven design vs. domain-specific design

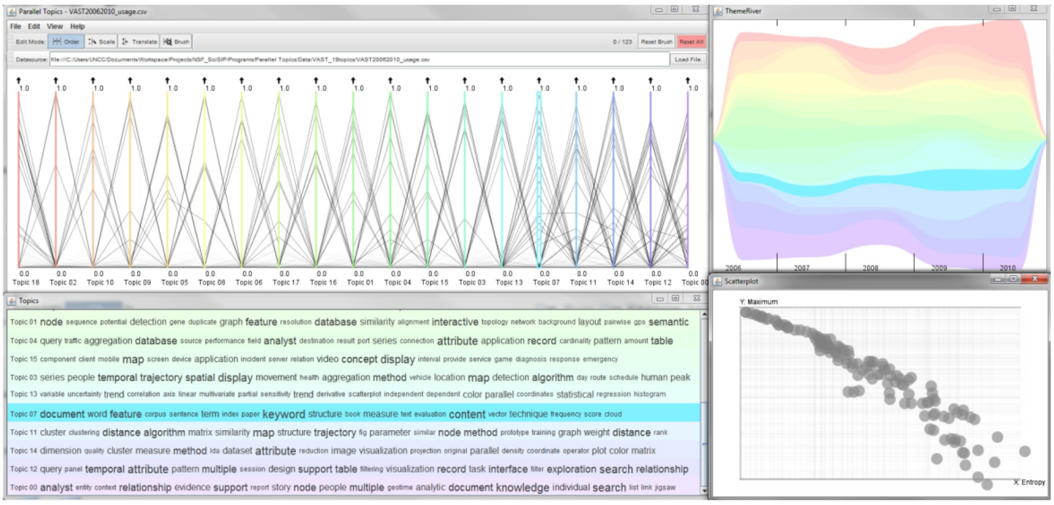
Soft clustering, i.e. weighted assignment of terms/documents to topics, as generated by probabilistic topic models or NMF involves challenging analysis of the results which requires sophisticated visual analytics tools with multiple coordinated views. Ensuring analyst's interpretation and trust against model-driven visualization of text corpora also depends on model alignment [Chuang 2012], i.e. how the model, the visual encoding and the tasks of the analyst are well-aligned together. Chuang et al. propose other design recommendations such as model verification, model modification and progressive disclosure that we took into account along this thesis. The next subsection exemplifies such model-driven designs where visual encoding is aligned to the model. Subsection 2.3.2 presents several approaches where the design is motivated by domain-specific tasks.

2.3.1 Model-driven design: from coordinated multiple views to interactive clustering

Based on the probability distribution of LDA [Blei 2003], ParallelTopics [Dou 2011] help identify what documents contain multiple topics through coordinated multiple views (Figure 2.18). In the parallel coordinates view each line shows the distribution of topics in one document. In the scatterplot view, the position of documents encodes the number of topics modeled using Shanon entropy, from single-topic documents on the top-left to multi-topic documents on the bottom-right. Topic evolution is depicted through a theme river view and horizontal flat lists expose the most representative terms of each topic.

In Figure 2.19, IVisClustering [Lee 2012] is a sophisticated tool providing semantic interactions taking into account user intention in an iterative process. A hard clustering of documents is derived from the LDA probability distribution by choosing the most prominent topic for each document. Each color-encoded cluster/topic is depicted within a node-link diagram and is labeled with its top-6 terms. The parallel coordinates view supports the analysis of soft-clustering through the topic distribution of documents, as in ParallelTopics. On the one hand, semantic interactions allow the modification of term weights, influencing the computation of the probability distributions of LDA. On the other hand, modifications of the topical structure, i.e. deleting, merging, re-clustering and sub-clustering, are supported by drag-and-drop interactions in the "Cluster Tree View". Such an iterative process is supported by multiple views that trace changes applied by LDA processing after user modifications.

UTOPIAN [Choo 2013] in Figure 2.20 models topics using a semi-supervised NMF algorithm. Document clusters are represented as a node-link diagram that supports multiple user interactions influencing the algorithm. Semantic interactions include keyword weight refinement in existing topics, splitting/merging of topics for interactive adjustment of their


Parallel Topics - VAST20062010_usage.csv
File Edit View Help
Edit Mode Order Scale Translate Brush
0 / 123 Reset Brush Reset All
Load File
Topic 18 Topic 02 Topic 10 Topic 09 Topic 05 Topic 08 Topic 06 Topic 17 Topic 16 Topic 01 Topic 04 Topic 15 Topic 03 Topic 13 Topic 07 Topic 11 Topic 14 Topic 12 Topic 00
1.0
0.0
ThemeRiver
2006 2007 2008 2009 2010
Scatterplot
Y: Maximum
X: Entropy
Topics

Topic 01 node sequence potential detection gene duplicate graph feature resolution database similarity alignment interactive topology network background layout pairwise gps semantic

Topic 04 query traffic aggregation database source performance field analyst destination result port series connection attribute application record cardinality pattern amount table

Topic 15 component client mobile map screen device application incident server relation video concept display interval provide service game diagnosis response emergency

Topic 03 series people temporal trajectory spatial display movement health aggregation method vehicle location map detection algorithm day route schedule human peak

Topic 13 variable uncertainty trend correlation axis linear multivariate partial sensitivity trend derivative scatterplot independent dependent color parallel coordinates statistical regression histogram

Topic 07 document word feature corpus sentence term index paper keyword structure book measure text evaluation content vector technique frequency score cloud

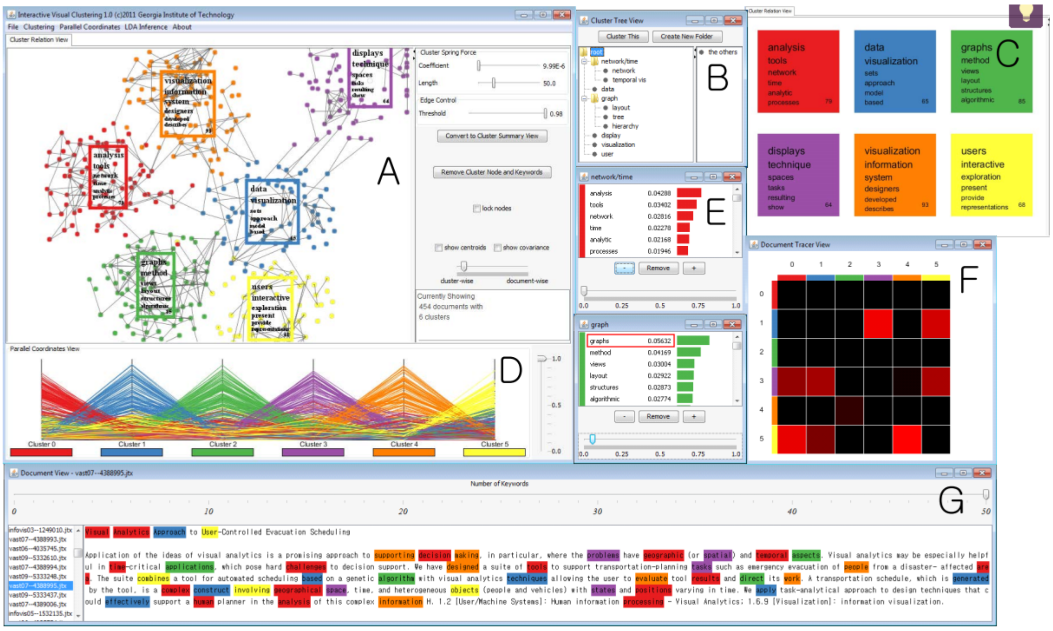
Topic 11 cluster clustering distance algorithm matrix similarity map structure trajectory fig parameter similar node method prototype training graph weight distance rank

Topic 14 dimension quality cluster measure method lda dataset attribute reduction image visualization projection original parallel density coordinate operator plot color matrix

Topic 12 query panel temporal attribute pattern multiple session design support table filtering visualization record task interface filter exploration search relationship

Topic 00 analyst entity context relationship evidence support report story node people multiple geotime analytic document knowledge individual search list link jigsaw

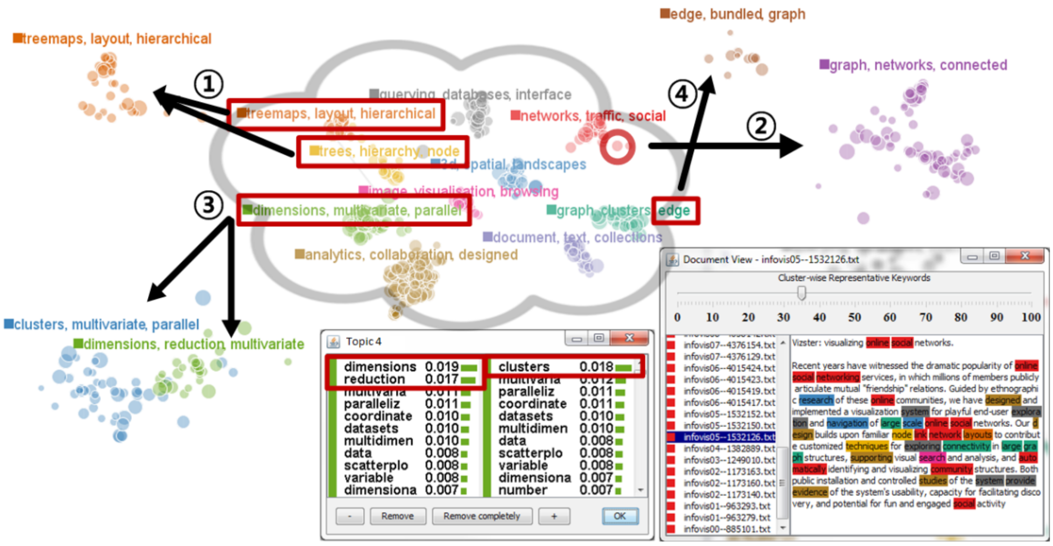
Figure 2.18: In ParallelTopics [Dou 2011], coordinated multiple views help identify what documents contain multiple topics through parallel coordinates on the top-left and a scatterplot on the bottom-right. Multiple interactions order the topics by similarity, highlight terms across topics or reveal topical distribution of documents through pie charts in the scatterplot view.



Visual Analytics Approach to User-Controlled Evacuation Scheduling

Application of the ideas of visual analytics is a promising approach to supporting decision making, in particular, where the problems have geographic (or spatial) and temporal aspects. Visual analytics may be especially helpful in time-critical applications, which pose hard challenges to decision support. We have designed a suite of tools to support transportation-planning tasks such as emergency evacuation of people from a disaster- affected area. The suite combines a tool for automated scheduling based on a genetic algorithm with visual analytics techniques allowing the user to evaluate tool results and direct its work. A transportation schedule, which is generated by the tool, is a complex construct involving geographical space, time, and heterogeneous objects (people and vehicles) with states and positions varying in time. We apply task-analytical approach to design techniques that could effectively support a human planner in the analysis of this complex information H. 1.2 [User/Machine Systems]: Human information processing - Visual Analytics; 1.6.9 [Visualization]: information visualization.

Figure 2.19: IVisClustering for interactive topic modeling [Lee 2012]. Coordinated multiple views help explore LDA results. Semantic interactions allow the modification of term weights and changing the topical structure. (A) Node-link diagram show a network of documents colored by cluster/topic. Each cluster is labeled according to its top-6 terms. (B) Cluster Tree View with semantic interactions. (C) Cluster summary view. (D) Parallel coordinates. (E) Term weight view with semantic interactions. (F) Document Tracer View to trace changes of document membership after re-clustering. (G) Document view.


treemaps, layout, hierarchical
edge, bundled, graph
graph, networks, connected
①
querying, databases, interface
treemaps, layout, hierarchical
networks, traffic, social
④
②
trees, hierarchy, node
2d, spatial, landscapes
image, visualisation, browsing
③
dimensions, multivariate, parallel
graph, clusters
edge
document, text, collections
analytics, collaboration, designed
clusters, multivariate, parallel
dimensions, reduction, multivariate
Topic 4
-
Remove
Remove completely
+
OK
Document View - infovis05--1532126.txt
Cluster-wise Representative Keywords
0
10
20
30
40
50
60
70
80
90
100

dimensions	0.019	clusters	0.018
reduction	0.017	multivaria	0.012
multivaria	0.011	paralleliz	0.011
paralleliz	0.011	coordinate	0.011
coordinate	0.010	datasets	0.010
datasets	0.010	multidimen	0.010
multidimen	0.010	data	0.008
data	0.008	scatterplo	0.008
scatterplo	0.008	variable	0.008
variable	0.008	dimensiona	0.007
dimensiona	0.007	number	0.007

infovis07--4376154.txt
infovis07--4376129.txt
infovis06--4015424.txt
infovis06--4015423.txt
infovis06--4015419.txt
infovis06--4015417.txt
infovis05--1532152.txt
infovis05--1532150.txt
infovis05--1532126.txt
infovis04--1382889.txt
infovis03--1249010.txt
infovis02--1173163.txt
infovis02--1173160.txt
infovis02--1173140.txt
infovis01--963293.txt
infovis01--963279.txt
infovis00--885101.txt

Vizster: visualizing online social networks.

Recent years have witnessed the dramatic popularity of online
social networking services, in which millions of members publicly
articulate mutual "friendship" relations. Guided by ethnographi
c research of these online communities, we have designed and
implemented a visualization system for playful end-user explora
tion and navigation of large scale online social networks. Our d
esign builds upon familiar node link network layouts to contribut
e customized techniques for exploring connectivity in large gra
ph structures, supporting visual search and analysis, and auto
matically identifying and visualizing community structures. Both
public installation and controlled studies of the system provide
evidence of the system's usability, capacity for facilitating disco
very, and potential for fun and engaged social activity

Figure 2.20: UTOPIAN, a visual analytics software based on semi-supervised NMF. Multiple semantic interactions allow (1) topic merger, (2) document-induced topic creation, (3) topic splitting, (4) keyword-induced topic creation and keyword refinement.

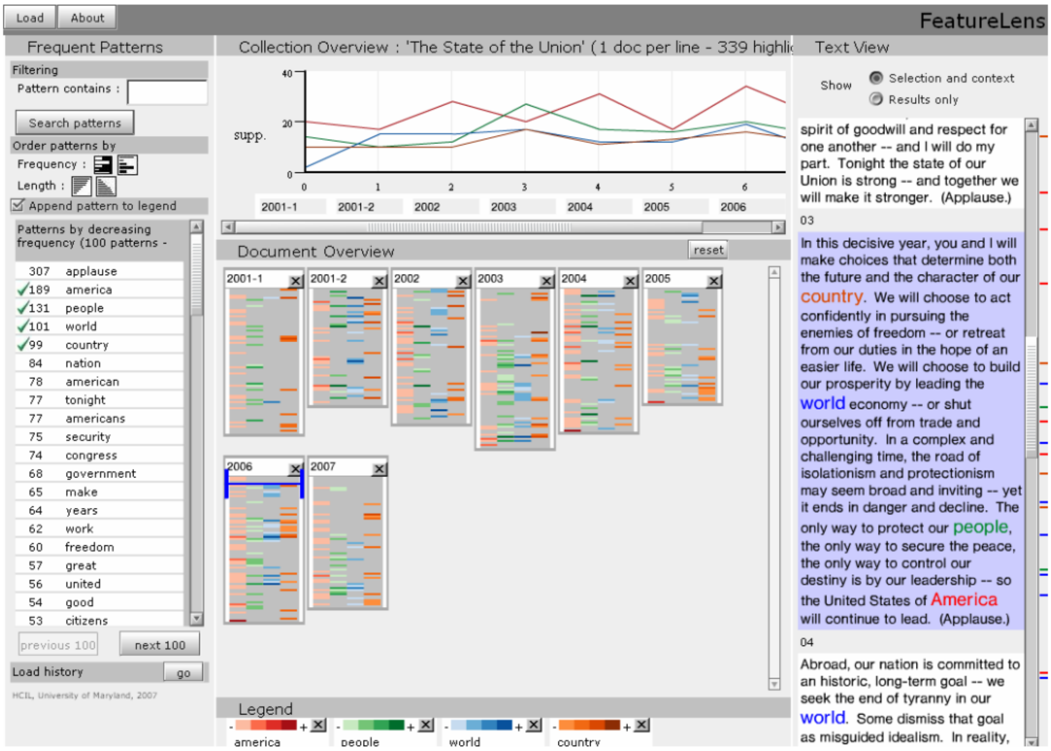
number, or even keyword/document-induced topic creation by giving a small set of exemplar terms/documents.

Visual text analytics solutions presented in this section, e.g. [Dou 2011, Lee 2012, Alexander 2014] use topic models derived from LDA [Blei 2003] or Non-Negative Matrix Factorization [Lee 1999, Choo 2013]. While the semantic of topics is often represented by their N most representative terms, recent work [Alexander 2015] shows that "a topic is more than the top 10 words". The authors found that, for a topic model to relate the semantic of the documents, it must reflect "the subtle patterns of term co-occurrences". This observation corroborates the distributional hypothesis (Hypothesis 1 on page 14). This is why we do our best to preserve all topic terms in our visualizations. BY doing so, we are also able to extract low-frequency co-occurrences that can constitute alternative and interesting facts, angles or viewpoints for text corpora investigation.

In addition, these solutions propose coordinated and multiple views with semantic interactions that integrate the expert knowledge and intention into the model. But the task definition arises from the structure of the model itself. Moreover, these approaches require a good comprehension and anticipation of the behavior underlying the models and algorithms. Model alignment, as recommended in [Chuang 2012], must also consider user tasks. The following section surveys visual analytics approaches for text corpora designed for domain-specific tasks.

2.3.2 Domain-specific visual analysis of text corpora

Feature Lens [Don 2007] has been motivated by students in English literature working on *The Making of Americans*, a book written by Gertude Stein, that extensively uses repetitions. Hence, Feature Lens is devoted to investigate text collections through text patterns such as frequent words or frequent itemsets of n-grams. The *topic variants* extracted in this thesis by the `Bimax` algorithm are very close to these patterns. In Figure 2.21, multiple sorting strategies reveal meaningful patterns that can be located in the context of documents. However the myriad of patterns are displayed as a flat list. Instead, we propose an overview of co-occurrence patterns, i.e. *topic variants*, presented through a hierarchy based on term overlaps.



307	applause
✓189	america
✓131	people
✓101	world
✓99	country
84	nation
78	american
77	tonight
77	americans
75	security
74	congress
68	government
65	make
64	years
62	work
60	freedom
57	great
56	united
54	good
53	citizens

spirit of goodwill and respect for one another -- and I will do my part. Tonight the state of our Union is strong -- and together we will make it stronger. (Applause.)

03

In this decisive year, you and I will make choices that determine both the future and the character of our country. We will choose to act confidently in pursuing the enemies of freedom -- or retreat from our duties in the hope of an easier life. We will choose to build our prosperity by leading the world economy -- or shut ourselves off from trade and opportunity. In a complex and challenging time, the road of isolationism and protectionism may seem broad and inviting -- yet it ends in danger and decline. The only way to protect our people, the only way to secure the peace, the only way to control our destiny is by our leadership -- so the United States of America will continue to lead. (Applause.)

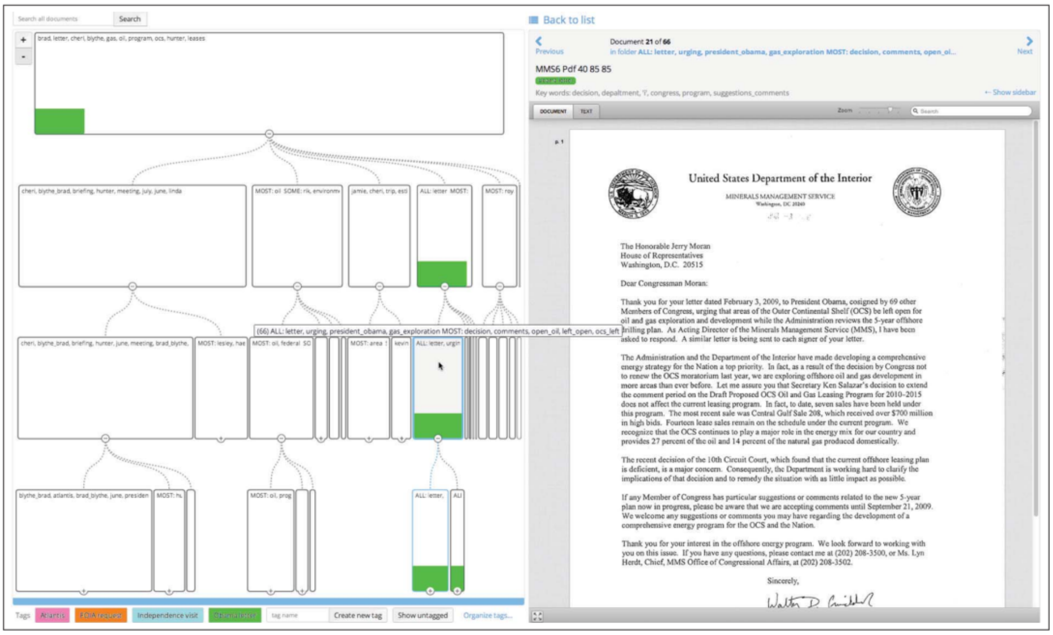
04

Abroad, our nation is committed to an historic, long-term goal -- we seek the end of tyranny in our world. Some dismiss that goal as misguided idealism. In reality,

Figure 2.21: Feature Lens [Don 2007] is devoted to the exploration of books or corpora with many repetitions, through frequent patterns of n-grams. The flat list view does not provide a broad overview of such patterns.

Designed for journalists, the Overview system [Brehmer 2014] supports hypothesis verification and generation (Figure 2.22). It relies on an agglomerative hierarchical clustering of the documents. Interactions such as expand/collapse offer a well-balanced tradeoff between usability and cluster fidelity. Hierarchical clustering does not intrinsically extract clusters of representative terms; the N most frequent terms are used to interpret each document cluster, which was found to be semantically suboptimal [Alexander 2015].

However, this tool summarizes large collections of documents and is designed to support hypothesis verification and generation, two tasks which are also supported by our solution. Both tasks and visual encoding are defined following Munzner's user-centered methodology [Munzner 2009] and the Brehmer's typology of tasks [Brehmer 2013]. The design of the Overview system results from multiple iterations and adjustments integrating expert feedbacks collected through a long-term adoption study. Our approach supports similar tasks through progressive disclosure of *topic variants* to search facts and viewpoints relevant to the working hypothesis.


Search all documents
Search
+
-
brad, letter, cheri, blythe, gas, oil, program, ocs, hunter, leases
cheri, blythe_brad, briefing, hunter, meeting, july, june, linda
MOST: oil SOME: rik, environm
jamie, cheri, trip, esti
ALL: letter MOST:
MOST: roy
(66) ALL: letter, urging, president_obama, gas_exploration MOST: decision, comments, open_oil_left_open, ocs_left
cheri, blythe_brad, briefing, hunter, june, meeting, brad_blythe,
MOST: lesley, hae
MOST: oil, federal SO
MOST: area
kevin
ALL: letter, urgin
blythe_brad, atlantis, brad_blythe, june, presiden
MOST: hu
MOST: oil, prog
ALL: letter,
ALL
Tags
Atlantis
FOIA request
Independence visit
Obama letter
tag name
Create new tag
Show untagged
Organize tags...
Back to list
Previous
Next
DOCUMENT
TEXT
Zoom
Search
← Show sidebar
p. 1

Document 21 of 66
in folder ALL: letter, urging, president_obama, gas_exploration MOST: decision, comments, open_ol...

MMS6 Pdf 40 85 85

Key words: decision, depaltment, 'I', congress, program, suggestions_comments

United States Department of the Interior

MINERALS MANAGEMENT SERVICE
Washington, DC 20240

The Honorable Jerry Moran
House of Representatives
Washington, D.C. 20515

Dear Congressman Moran:

Thank you for your letter dated February 3, 2009, to President Obama, cosigned by 69 other Members of Congress, urging that areas of the Outer Continental Shelf (OCS) be left open for oil and gas exploration and development while the Administration reviews the 5-year offshore drilling plan. As Acting Director of the Minerals Management Service (MMS), I have been asked to respond. A similar letter is being sent to each signer of your letter.

The Administration and the Department of the Interior have made developing a comprehensive energy strategy for the Nation a top priority. In fact, as a result of the decision by Congress not to renew the OCS moratorium last year, we are exploring offshore oil and gas development in more areas than ever before. Let me assure you that Secretary Ken Salazar's decision to extend the comment period on the Draft Proposed OCS Oil and Gas Leasing Program for 2010-2015 does not affect the current leasing program. In fact, to date, seven sales have been held under this program. The most recent sale was Central Gulf Sale 208, which received over $700 million in high bids. Fourteen lease sales remain on the schedule under the current program. We recognize that the OCS continues to play a major role in the energy mix for our country and provides 27 percent of the oil and 14 percent of the natural gas produced domestically.

The recent decision of the 10th Circuit Court, which found that the current offshore leasing plan is deficient, is a major concern. Consequently, the Department is working hard to clarify the implications of that decision and to remedy the situation with as little impact as possible.

If any Member of Congress has particular suggestions or comments related to the new 5-year plan now in progress, please be aware that we are accepting comments until September 21, 2009. We welcome any suggestions or comments you may have regarding the development of a comprehensive energy program for the OCS and the Nation.

Thank you for your interest in the offshore energy program. We look forward to working with you on this issue. If you have any questions, please contact me at (202) 208-3500, or Ms. Lyn Herdt, Chief, MMS Office of Congressional Affairs, at (202) 208-3502.

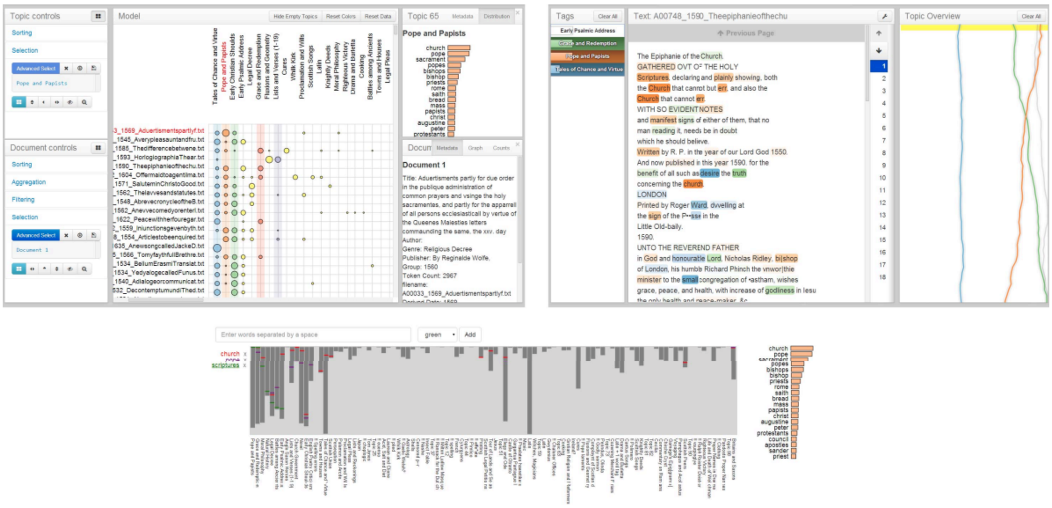
Sincerely,

Walter D. Cruickshank

Figure 2.22: Overview is a visual analytics tool for investigative journalism [Brehmer 2014]. The expandable hierarchical view on the right hand allows progressive disclosure of document clusters. Their labels are comprised of their N most frequent terms. On the right hand, a document viewer provides access to the original content.

Alexander et al. [Alexander 2014] have designed a tool to investigate a large text corpus, visible in Figure 2.23. Their work were motivated by literary scholars studing historical text corpora at the *Folger Shakespeare Library*. Based on *LDA*, they designed coordinated multiple views and different sorting strategies allowing both top-down and bottom-up analyses. In addition, they applied a set of principles fostering serendipitous discovery formulated in [Thudt 2012], i.e. providing multiple access points, highlighting adjacencies, offering flexible pathways for exploration and enticing curiosity and playfulness.

In our work, we identified focus and diversification processes followed by investigative journalists [L. Hunter 2011]. Lee Hunter et al. formulated diversification processes with principles similar to those formulated in [Thudt 2012]. Both processes also encompass the



Document 1

Title: Aduertisments partly for due order in the publique administration of common prayers and vsinge the holy sacramentes, and partly for the apparrell of all persons ecclesiasticall by vertue of the Queenes Maiesties letters commaunding the same, the xxv. day

Author:

Genre: Religious Decree

Publisher: By Reginalde Wolfe,

Group: 1560

Token Count: 2967

filename:

A00033_1569_Aduertismentspartlyf.txt

Text: A00748_1590_Theepiphanieofthechu

Previous Page

The Epiphanie of the Church.
GATHERED OVT O⁼ THE HOLY
Scriptures, declaring and plainly showing, both
the Church that cannot but err, and also the
Church that cannot err.
WITH SO EVIDENT NOTES
and manifest signs of either of them, that no
man reading it, needs be in doubt
which he should believe.
Written by R. P. in the year of our Lord God 1550.
And now published in this year 1590. for the
benefit of all such as desire the truth
concerning the church.
LONDON
Printed by Roger Ward, dvvelling at
the sign of the P••sse in the
Little Old-baily.
1590.
UNTO THE REVEREND FATHER
in God and honourable Lord, Nicholas Ridley, bi|shop
of London, his humble Richard Phinch the vnwor|thie
minister to the small congregation of •astham, wishes
grace, peace, and health, with increase of godliness in Iesu

Figure 2.23: Serendip [Alexander 2014] is built in collaboration with literary scholars. Resulting from a user-centered design this software allows top-down and bottom-up analyses and follows principles promoting serendipitous discoveries.

top-down and bottom-up analyses described in [Alexander 2014] as well as the hypothesis verification and generation tasks defined in [Brehmer 2014]. However we support such analysis through a multi-resolution exploration of text corpora to identify different angles or viewpoints with the precision and the exhaustiveness offered by the Bimax biclustering algorithm [Prelić 2006]. None of related work presented above proposes a solution to draw a broad overview while supporting detailed exploration of such *topic variants*, captured by subtle document relationships based on term co-occurrences. Exploring the plethora of overlapping biclusters extracted by Bimax requires however carefully designed visualizations. We therefore survey bicluster visualization approaches in the following section.

2.4 Bicluster Visualization

As described in section 1.4, bicluster structures fall in two main categories: hard biclusters where matrix rows and columns are allocated exclusively to one bicluster, and overlapping biclusters, where rows and columns can be assigned to multiple biclusters. However, overlapping biclusters lead to visualization issues as it complicates the creation of comprehensive overviews and the identification of the common and distinctive items of the biclusters. In their survey of bicluster visualization techniques, Sun et al. [Sun 2014] propose a design framework modeling 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 scheme, biclusters represent (n:m) relationships between two dimensions of the data.

Jigsaw [Stasko 2008] is a visual analytics tool devoted to the analysis of text corpora. The analyst can organize sets of named entities (i.e. by categories) in parallel coordinates view and can explore group relationships (1:n) as shown in Figure 2.24. Entities are connected if they co-occur in at least one documents, i.e. first order co-occurrences (see Figure 1.2 on page 15). By selecting one entity, this view shows group relationships (1:n) between two coordinates. Nevertheless, individual biclusters, i.e. (n:m) relationships, can be found interactivelly after alternating multiple times selection/reordering steps. Coordinated multiple views give multiple perspectives to the analyst. A graph view encodes multiple types of nodes using different graphical symbols: squares for documents and circles for entities. The entity-document memberships are shown as edges. This view depicts entity-document biclusters but they are not clearly delineated. Therefore, this tool is very powerful to identify group relationships (1:n) between entities but require multiple user manipulations to elicit biclusters individually.

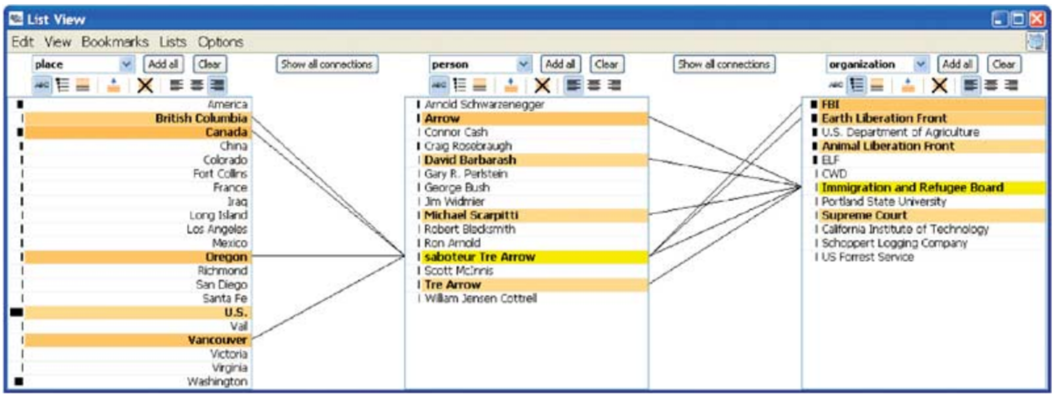


Figure 2.24: Jigsaw visual analytics approach based on named entities [Stasko 2008]. Parallel coordinates allow to analyze (1:n) relationships of entities through first-order co-occurrences.

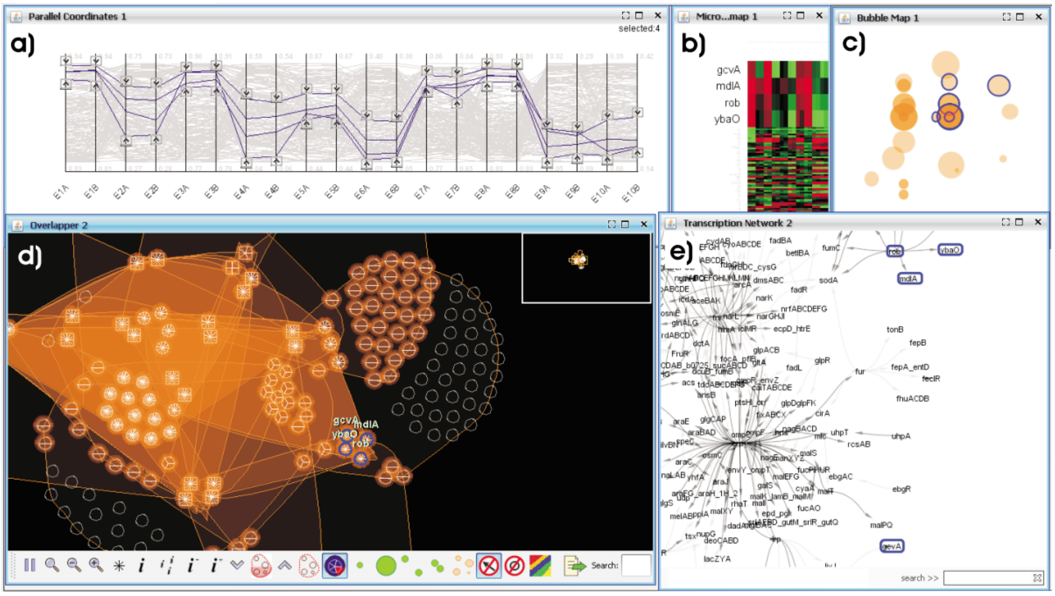


Figure 2.25: BicOverlapper [Santamaría 2008] proposes coordinated multiple views to explore overlapping biclusters in gene-expression data: a) Parallel coordinates reveal similar evolution of genes across coordinates. b) A matrix visualization reveals individual biclusters after reordering both dimensions. d) The node link diagram gives an overview of biclusters delineated by transparent hulls.

In bio-informatics, BicOverlapper [Santamaría 2008] visualizes biclusters extracted by the Bimax algorithm as a node-link diagram. The force-directed layout brings together the nodes from both dimensions and transparent hulls outline the biclusters and reveal their overlaps. However, in dense areas with numerous overlaps, node overlaps or edge crossing hinder the perception of common and distinctive parts of biclusters as shown in Figure 2.25.

Matrix visualizations and parallel coordinates are more effective than node-link diagrams at the bicluster level [Sun 2014]. And, even though matrix visualizations are less intuitive, they are often more readable for large and dense datasets [Ghoniem 2005]. Many solutions in bio-informatics propose coordinated multiple views with both a matrix heatmap and parallel coordinates [Barkow 2006, Santamaría 2008, Heinrich 2011]. They display the items of both dimensions of a matrix in well separated rows and columns that may be reordered to reveal biclusters. While individual biclusters can effectively be interpreted, the linear arrangement of each dimension fails to convey a clear overview of all overlapping biclusters without duplicating items [Heinrich 2011, Streit 2014] as shown in Figure 2.26. These approaches are not scalable with respect to the high dimensionality and the sparsity of term-document matrices. They also fail to support the comparison of biclusters and the identification of their common and distinctive items in the two dimensions.

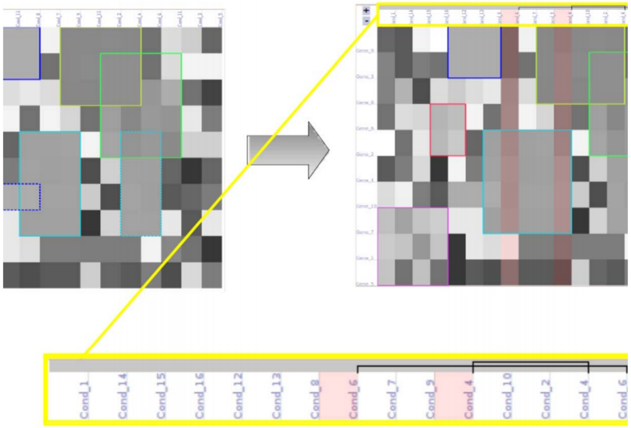


Figure 2.26: In a matrix visualization, Bicluster Viewer duplicates items in column to depict complete biclusters in a compact block [Heinrich 2011]. Common and distinctive parts of biclusters are hardly identifiable.

In response, Streit et al. [Streit 2014] propose a hybrid visualization where biclusters are represented by multiple matrices which are linked through their common columns/rows (see Figure 2.27). However the node-link representation cannot handle the large number of overlapping biclusters produced by Bimax. Bixplorer [Fiaux 2013] proposes a similar approach for text data to explore chained relationships between different categories of named entities (see Figure 2.28). This visualization works well as a workspace for the analyst to focus on a selection of biclusters of interest, chosen from a flat list. This bottom up approach does not support the prerequisite step consisting in exploring and identifying interesting biclusters within the whole corpus.

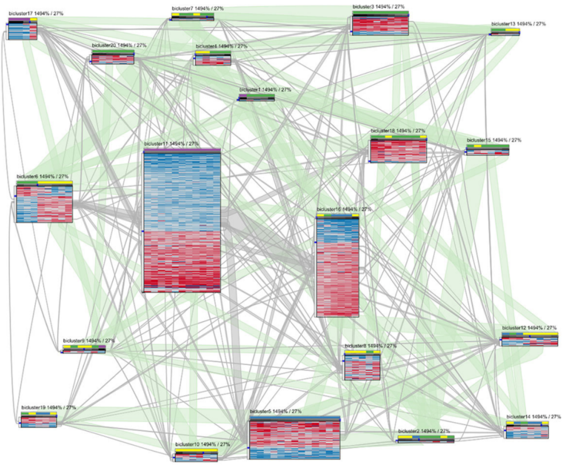
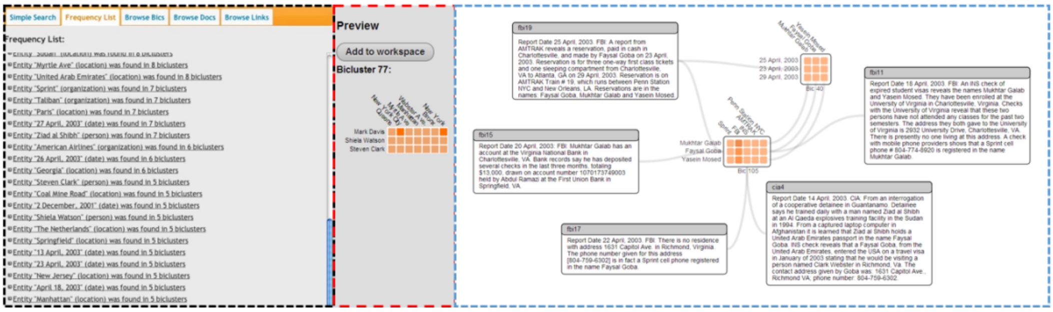


Figure 2.27: Furby: hybrid visualization with node-link diagrams and matrices [Streit 2014]. The multiple edge crossings induce visual clutter.



Frequency List:

- Entity "Sudan" (location) was found in 8 biclusters
- Entity "Myrtle Ave" (location) was found in 8 biclusters
- Entity "United Arab Emirates" (location) was found in 8 biclusters
- Entity "Sprint" (organization) was found in 7 biclusters
- Entity "Taliban" (organization) was found in 7 biclusters
- Entity "Paris" (location) was found in 7 biclusters
- Entity "27 April, 2003" (date) was found in 7 biclusters
- Entity "Ziad al Shibh" (person) was found in 7 biclusters
- Entity "American Airlines" (organization) was found in 6 biclusters
- Entity "26 April, 2003" (date) was found in 6 biclusters
- Entity "Georgia" (location) was found in 6 biclusters
- Entity "Steven Clark" (person) was found in 5 biclusters
- Entity "Coal Mine Road" (location) was found in 5 biclusters
- Entity "2 December, 2001" (date) was found in 5 biclusters
- Entity "Shiela Watson" (person) was found in 5 biclusters
- Entity "The Netherlands" (location) was found in 5 biclusters
- Entity "Springfield" (location) was found in 5 biclusters
- Entity "13 April, 2003" (date) was found in 5 biclusters
- Entity "23 April, 2003" (date) was found in 5 biclusters
- Entity "New Jersey" (location) was found in 5 biclusters
- Entity "April 18, 2003" (date) was found in 5 biclusters
- Entity "Manhattan" (location) was found in 5 biclusters

fbi19

Report Date 25 April, 2003. FBI: A report from AMTRAK reveals a reservation, paid in cash in Charlottesville, and made by Faysal Goba on 23 April, 2003. Reservation is for three one-way first class tickets and one sleeping compartment from Charlottesville, VA to Atlanta, GA on 29 April, 2003. Reservation is on AMTRAK Train # 19, which runs between Penn Station NYC and New Orleans, LA. Reservations are in the names: Faysal Goba, Mukhtar Galab and Yasein Mosed.

fbi15

Report Date 20 April, 2003: FBI: Mukhtar Galab has an account at the Virginia National Bank in Charlottesville, VA. Bank records say he has deposited several checks in the last three months, totaling $13,000, drawn on account number 1070173749003 held by Abdul Ramazi at the First Union Bank in Springfield, VA.

fbi17

Report Date 22 April, 2003. FBI: There is no residence with address 1631 Capitol Ave. in Richmond, Virginia. The phone number given for this address [804-759-6302] is in fact a Sprint cell phone registered in the name Faysal Goba.

fbi11

Report Date 18 April, 2003. FBI: An INS check of expired student visas reveals the names Mukhtar Galab and Yasein Mosed. They have been enrolled at the University of Virginia in Charlottesville, Virginia. Checks with the University of Virginia reveal that these two persons have not attended any classes for the past two semesters. The address they both gave to the University of Virginia is 2932 University Drive, Charlottesville, VA. There is presently no one living at this address. A check with mobile phone providers shows that a Sprint cell phone # 804-774-8920 is registered in the name Mukhtar Galab.

cia4

Report Date 14 April, 2003. CIA: From an interrogation of a cooperative detainee in Guantanamo. Detainee says he trained daily with a man named Ziad al Shibh at an Al Qaeda explosives training facility in the Sudan in 1994. From a captured laptop computer in Afghanistan it is learned that Ziad al Shibh holds a United Arab Emirates passport in the name Faysal Goba. INS check reveals that a Faysal Goba, from the United Arab Emirates, entered the USA on a travel visa in January of 2003 stating that he would be visiting a person named Clark Webster in Richmond, Va. The contact address given by Goba was: 1631 Capitol Ave., Richmond VA; phone number: 804-759-6302.

Figure 2.28: Bixplorer allows exploration of text corpus through the relationships between named entities [Fiaux 2013]. Such a bottom up approach does not convey a broad overview of biclusters.

More recently, BiSet [Sun 2015] was designed for top-down exploration of chained bicluster relationships (n:m:...:z) without duplicating items. Inspired by Jigsaw [Stasko 2008], the parallel coordinates in Figure 2.29 display named entities in multiple categories, and the chained relationships are shown through semantic edge bundles formed by the biclusters extracted from each pair of categories.

Through multiple interactions, BiSet enables a top-down exploration of chained relationships between multiple categories of named entities, without duplicating items. Our problem is different. Instead of exploring chained relationships of named entities, our goal is to support the exploration of thousands of biclusters extracted from the plain text contents of documents, i.e. many-to-many relationships between the terms and the documents. For this problem, BiSet's visualization becomes a bipartite graph where the thousands of edge bundles resulting from biclusters produce large number of intersecting curves.

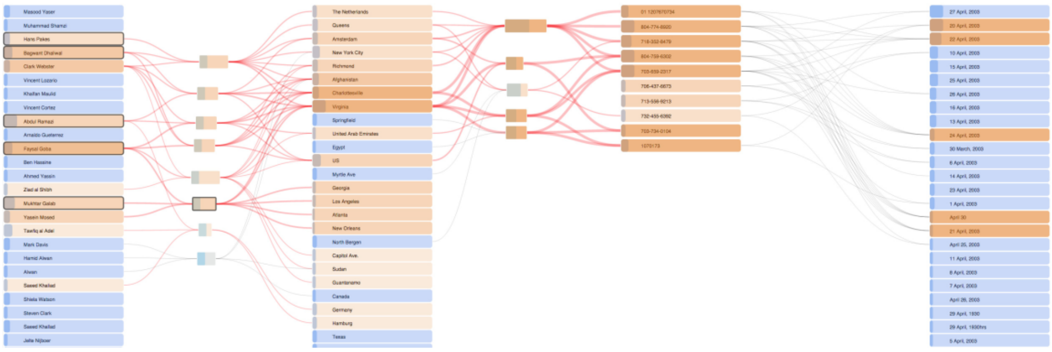
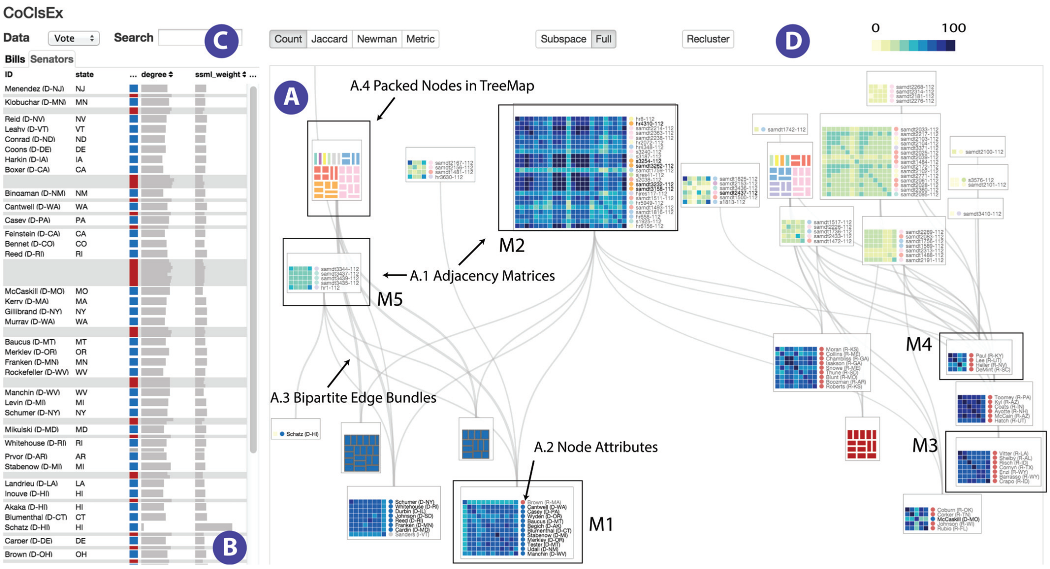


Figure 2.29: BiSet: parallel coordinates view with semantic edge bundles [Sun 2015]. BiSet allows top-down exploration of chained relationships between multiple categories of named entities. Bidirectional interactions through edge bundles and entities allow task-oriented rearrangements for sensemaking purposes.

For bipartite graphs, Xu et al. [Xu 2016] display separately one-dimensional projections of bisclusters either in an adjacency matrix or in a 2-level treemap as shown in Figure 2.30. The color intensity of cells in adjacency matrices informs the analyst about the consistency of nodes in the bicluster. The 2-level treemap organizes the nodes according to nominative and quantitative attributes coming from computed metrics or metadata. Both dimensions of biclusters are connected through biparite connections, bundled together to reduce visual clutter. Semantic interactions are proposed for cluster refinement by merging clusters with drag-and-drop interaction or by marking subsets of nodes with must-link or cannot-link constraints. However this approach does not scale to thousands of biclusters. Items are duplicated in clusters and for tasks consisting of searching *topic variants* characterizing alternate viewpoints about similar stories or events, it could be tedious to follow bipartite edge bundles to depict common and distinctive terms of biclusters.



Data: Vote
Search:
Bills | Senators

ID	state	... degree	ssml_weight ...
Menendez (D-NJ)	NJ		
Klobuchar (D-MN)	MN		
Reid (D-NV)	NV		
Leahy (D-VT)	VT		
Conrad (D-ND)	ND		
Coons (D-DE)	DE		
Harkin (D-IA)	IA		
Boxer (D-CA)	CA		
Bingaman (D-NM)	NM		
Cantwell (D-WA)	WA		
Casey (D-PA)	PA		
Feinstein (D-CA)	CA		
Bennet (D-CO)	CO		
Reed (D-RI)	RI		
McCaskill (D-MO)	MO		
Kerry (D-MA)	MA		
Gillibrand (D-NY)	NY		
Murray (D-WA)	WA		
Baucus (D-MT)	MT		
Merkley (D-OR)	OR		
Franken (D-MN)	MN		
Rockefeller (D-WV)	WV		
Manchin (D-WV)	WV		
Levin (D-MI)	MI		
Schumer (D-NY)	NY		
Mikulski (D-MD)	MD		
Whitehouse (D-RI)	RI		
Pryor (D-AR)	AR		
Stabenow (D-MI)	MI		
Landrieu (D-LA)	LA		
Inouye (D-HI)	HI		
Akaka (D-HI)	HI		
Blumenthal (D-CT)	CT		
Schatz (D-HI)	HI		
Carper (D-DE)	DE		
Brown (D-OH)	OH		

Figure 2.30: Visual co-cluster analysis of bipartite graphs [Xu 2016]. The space is divided horizontally to display each dimension of biclusters with either an adjacency matrix (A.1) or a 2-level treemap (A.4). The nodes (A.2) of both dimensions of the bipartite graph are connected through aggregated edge bundles (A.3).

Through a hierarchy of *topic variants* based on the overlap degree of terms, coordinated with a *Topic Variant Comparator* view, our solution proposes a trade-off to discern the common and distinctive parts of biclusters by avoiding node superposition and edge crossings while limiting the duplication of items. The directed labeling we propose allows to quickly explore thousands of biclusters in a radial tree visualization.

The next chapter describes the tasks taken into account and the nested bicluster structure leveraged by the visualizations.

Part II

Visual Analytics System

CHAPTER 3

Method and model

Contents

3.1 Introduction . . . 59

3.2 User centered approach . . . 62

3.2.1 Munzner's Nested Model . . . 62

3.2.2 Data and problem characterization . . . 63

3.2.3 Data and operation abstraction . . . 64

3.3 Multi-model and multi-resolution biclustering approach . . . 66

3.3.1 Text Processing . . . 66

3.3.2 Coarse-grained Topics . . . 66

3.3.3 Fine-grained Topic Variants . . . 71

3.4 Evaluation . . . 74

3.4.1 Protocol . . . 75

3.4.2 Dataset . . . 76

3.4.3 Analysis . . . 77

3.5 Discussion . . . 81

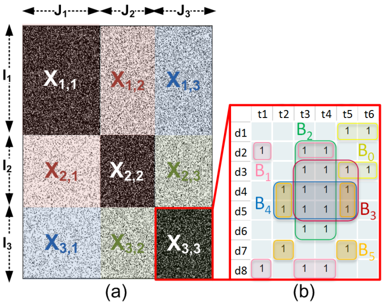
3.6 Conclusion . . . 82

3.1 Introduction

The design of a visual analytics software requires to understand the problems and tasks of application domain analysts. We focus in this thesis on the problems of investigative journalists when faced with face large text corpora. This chapter starts by a presentation of the user-centered method we followed for the design of our visual analytics software. The workflow and the tasks of the targeted users are described. In a nutshell, the journalists initiate their work by mapping the field under investigation. Then, they aim to analyze the relationships between documents, searching for multiple sources, i.e. fragments of text showing evidence related to their working hypotheses. Leveraging such a process in a large text corpus requires a well suited model supporting multi-resolution analyses.

Many recent visual text analytics solutions [Dou 2011, Lee 2012, Choo 2013, Alexander 2014] use variants of Latent Dirichlet Allocation (LDA) [Blei 2003] or Non-Negative Matrix Factorization [Lee 1999] to extract coarse-grained topics. These methods produce for each topic a weighted combination of terms and for each document a weighted combination of topics. Albeit less popular, biclustering methods [Govaert 2013, Madeira 2004] have also been used successfully to model coarse-grained topics. Applied on a term-document matrix, these methods consider simultaneously the rows and columns of the matrix and deliver homogeneous biclusters, each being a set of terms consistently grouped to describe a set of similar documents. We consider that a topic as understood by a journalist can be characterized by a bicluster from a data analytics perspective.

The nested biclustering approach we propose is depicted in Figure 3.1. It supports the analysis of both coarse-grained topics and fine-grained topic variants. The coarse-grained structure summarizes the corpus and gives an overview of the field under investigation. Then, for a topic of interest, the fine-grained structure allows the inspection of document relationships through their common and distinctive terms. This multi-resolution approach aims to spare the analyst the painstaking task of reading all documents in order to identify facts or viewpoints and the associated evidence.



	t1	t2	t3	t4	t5	t6
d1					1	1
d2	1		1	1		
d3			1	1	1	1
d4		1	1	1	1	
d5		1	1	1	1	
d6			1	1		
d7		1			1	
d8	1		1	1		

Figure 3.1: The proposed nested biclustering approach builds: (a) in black, diagonal biclusters (coarse-grained topics) and their pairwise confusion blocks; (b) the overlapping biclusters (topic variants) extracted from each topic e.g. $X_{3,3}$.

Our system is designed to accommodate any topic extraction method, as long as a bicluster structure can be derived from it. With topic models such as LDA, a bicluster structure can for example be extracted from the probability distributions with arbitrary probability thresholds to obtain overlapping topics, term-wise and document-wise. In particular, the proposed system deals with the fact that biclusters (and the associated topics) may come in various shapes [Madeira 2004], including their being disjoint or overlapping with respect to their term set and/or document set. For a given corpus, the shape of biclusters is however influenced by the chosen topic extraction method and/or its parametrization. Because the comprehension of a topic boils down to the journalist's ability to make sense of the related term set and explain the frontier/relationships between distinct topics, the present chapter aims to characterize in a systematic fashion the differences between two families of methods: disjoint biclustering and overlapping topic models. Conveying the relationships between topics of various shapes also requires the design of a new similarity metric that can handle both overlapping and disjoint biclusters.

As part of the journalistic work, the analysis of a coarse-grained topic also requires the identification and comparison of different viewpoints to ensure a well-balanced coverage of the topic. This is why our system extracts the fine-grained topics, called "topic variants" in this thesis, which lie within a given coarse-grained topic.

To assess the adequacy of the topic variants with regard to the user tasks, we propose an evaluation approach that does not depend on any labeled ground truth, making it applicable in practice on any real text corpus. This approach is based on multiple intrinsic bicluster metrics regarding both the terms and the documents of the topics as well as their variants. These metrics help to understand the structure of the topics, their quality and the shape of the underlying term co-occurrences. In this evaluation, we compare the disjoint topical structure given by `Coclus` [Ailem 2016], a diagonal biclustering method, with the overlapping topical structure yielded by the probabilistic Hierarchical Latent Dirichlet Allocation (`hLDA`).

The contributions of this work are the following:

1. we propose a new analysis approach of text corpora based on a nested bicluster structure supporting both overlapping and disjoint topics;
2. we propose a new similarity metric, suited for both overlapping and non-overlapping topics;
3. we set up and run an evaluation protocol based on multiple intrinsic bicluster metrics allowing
   (a) the systematic comparison of the topical structure of two topic extraction methods producing overlapping and disjoint topics and,
   (b) the appraisal of their suitability to user tasks on a real data set.

This last contribution aims to inform the design of the user study in section 4.5 investigating the influence of the choice of topic extraction method on the interpretability of topics by analysts.

In the rest of this chapter, section 3.2 describes the user-centered method and the definition of problems and tasks under consideration for this thesis. Next, the proposed nested biclustering approach and the related similarity metric are presented in section 3.3. Our evaluation method and results are described in section 3.4. Finally, we discuss the results in section 3.5.

3.2 User centered approach

Our work adopts a user-centred methodology where the final user is actively engaged at each level of the design, from data and tasks characterization [Brehmer 2013, Schulz 2013] until the software evaluation and validation [Isenberg 2013, Munzner 2009]. We are firmly committed to enforcing state-of-the-art methodology adapted to the design of visualization software. In this thesis, we follow the commonly used Munzner's nested model for visualization design presented below.

3.2.1 Munzner's Nested Model

Munzner et al. [Munzner 2009] proposes a methodology for visualization design based on the four nested levels presented in Figure 3.2.

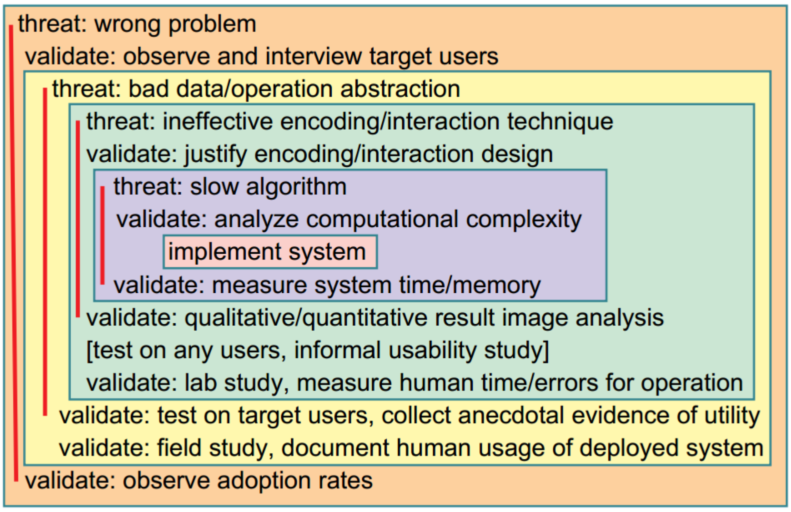

threat: wrong problem
validate: observe and interview target users
threat: bad data/operation abstraction
threat: ineffective encoding/interaction technique
validate: justify encoding/interaction design
threat: slow algorithm
validate: analyze computational complexity
implement system
validate: measure system time/memory
validate: qualitative/quantitative result image analysis
[test on any users, informal usability study]
validate: lab study, measure human time/errors for operation
validate: test on target users, collect anecdotal evidence of utility
validate: field study, document human usage of deployed system
validate: observe adoption rates

Figure 3.2: The four nested levels of the Munzner's model for visualization design (Figure from [Munzner 2009]).

This model provides a guidance to prevent threats to validity at each level of design and suggest appropriate evaluation methodologies [Isenberg 2013] to validate the design choices, upstream or downstream their implementation. The nesting order of levels does not impose an order in their achievement but indicates that the choices made at one level has a cascading impact in the lower levels. Each level may shed light on every other level during the design and the implementation. Therefore the development process is rather cyclic and each iteration can cover all or part of these different levels.

As seen in Figure 3.2, visualization design is guided by the user's data and tasks learned by the designer. Data and tasks can be defined at different levels of resolution. High-level tasks characterize the data and problems of the targeted users with its own domain vocabulary. Low-level tasks are drawn on data and operation abstraction. Firstly described in natural language with the domain expert, the tasks are next formalized with a typology of tasks [Brehmer 2013, Schulz 2013] that prepares the design of the visual encoding and the interactions. Finally, the system as a whole implements algorithms whose complexity in runtime and memory must be characterized. Specifically, the design of our visual analytics software, in the application domain of investigative journalism, is described in the next section, in compliance with the Munzner's Nested Model [Munzner 2009] and the typology of tasks for visualization proposed by Brehmer et al. [Brehmer 2013].

3.2.2 Data and problem characterization

In our work we are pointedly interested in the field of investigative journalism that is regularly facing large text corpora. In 2014 and 2015, we have met journalists in events such as "Les Assises du Journalisme"[1] in order to understand their tasks and their problems with text corpora investigation, and more generally the workflow they adopt during their inquiries. Next we started a partnership 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 started by conducting a semi-supervised interview. Globally, she expressed a sense of frustration about not having time to be exhaustive in her investigations. She is always forced to reduce the number of documents aggregated during her back-grounding work. A tool that could summarize document contents and extract all document relationships could clearly help her identify interesting facts without having to read every document. In addition, we completed our understanding of the problem through complementary documentation. In their manual for investigative journalism, Lee Hunter et al. [L. Hunter 2011] state that: "a hypothesis is a story and a method to test it". Indeed the core task of journalists is telling a story that can be "promoted, defended and remembered". Therefore the inquiry is initiated by the formulation of a hypothesis telling a story that has to be verified or disproved by the information found by the journalist. Hence, Lee Hunter et al. [L. Hunter 2011] propose a method to process inquiry based on hypothesis formulation and verification. But the investigation can also reveal facts that go against the initial hypothesis which require multiple reformulations.

From our preliminary analysis, we extracted a general workflow composed of three alternating processes summarized as follows:

- **Mapping the subject:** during this process, the journalist gets an overview of the subjects of inquiry.
- **Focus:** the journalist focuses the investigation on a specific aspect to identify facts, viewpoints that verify, refute or refine her hypothesis.

[1] http://www.journalisme.com/les-assises-presentation

- **Diversification:** the journalist looks one step farther from the targeted object, hoping to find unexpected information and new angles of analysis. This step is important to ensure that she does not miss essential information that could call into question the initial hypothesis.

While presented linearly, this workflow is actually discontinued, intertwined and repeated multiple times during the investigations. We took this workflow into account as we designed our system. In the next section we will describe in details the task abstraction.

3.2.3 Data and operation abstraction

During our preliminary analysis we have identified the need to analyze a large collection of texts with an exploratory tool. We defined three high-level tasks.

T1 summarize the corpus to identify topics of interest and aspects to investigate.

T2 find the documents that verify an aspect of the working hypothesis

T3 identify new angles or viewpoints that incite the journalist to refine or generate new hypothesis that better fit the facts discovered in the text content.

Each high-level task is divided into low-level tasks described in Table 3.2.

To summarize the corpus, our system leverages a bicluster structure to represent coarse-grained topics. Secondly, the nested bicluster structure reveals fine-grained *topic variants* lying within each topic. We consider this nested structure as a starting point for our task abstraction. We characterize the tasks by using the typology of Brehmer and Munzner [Brehmer 2013] formalizing the motivation of the tasks (Why) and the mean used to support them (How). For the *why* question Brehmer and Munzner suggest to specify the purpose of the visualization consumption, the type of search and the type of query concerned by the tasks. In this thesis, we focus only on the generic discovery purpose, and the two forms it takes, i.e. verify and generate hypothesis. Next, the type of search can be characterized by *locate*, *look-up*, *explore* or *browse* depending on whether the search target and its location are known or not, as showed in Table 3.1. The kinds of query can be *identify*, *compare* or *summarize*. Moreover, input and output data involved in each task must be considered. We also add a short description of how the tasks are achieved through a list of interactions. The description of the low-level tasks we identified with the journalist are described in Table 3.2.

	Target known	**Target unknown**
Location known	lookup	browse
Location unknown	locate	explore

Table 3.1: Different type of search depending on whether the search target and its location are known or not

ID	Short description	Description	Why					How
			Consume	Search	Input	Query	Output	
T1.1	Discovering the topics	Understand and locate topics of interest	Discover	Explore / Locate	All topics	Identify	Many topics	Encode, navigate, select
T1.2	Discovering the topic variants	Understand/Identify facts or view points related to events or stories	Discover	Explore	All topic variants	Identify	Many topic variants	Encode, derive, navigate, select
T2.1	Keyword-based search of topic variants	Identify/Select variant of interest related specific facts or view points	Verify	Locate	Terms	Identify	Many topic variants	Encode, filter, select
		Filter out others	Verify	Lookup	Many topic variants	Identify	Many topic variants	Encode, filter, change, select
T2.2	Compare topic variants	Common and distinctive terms/documents	Verify	Browse	Many topic variants	Identify	Terms/Documents of interest	Encode, select, change, arrange
T2.3	Show documents details	Identify documents of a topic variant / Show raw text for reading / Identify terms of variant in their context	Verify	Lookup	One topic variant	Identify	Raw text of the documents and the terms in their context	Encode, arrange, navigate, annotate
T3.1	Discovering topics relations	Identify the most similar topics	Generate	Browse	One topic	Identify	Many topics	Encode, navigate
T3.2	Suggesting terms for search query	The terms from the content that can be used for queries	Generate	Explore	All topic variants	Identify	Terms	Encode, navigate, select
T3.3	Suggesting new variants	Identify new variants sharing terms/documents with one variant of interest	Generate	Explore	Many terms/documents	Identify	Many topic variants	Encode, select

Table 3.2: The low-level tasks defined under the typology of Brehmer and Munzner [Brehmer 2013].

3.3 Multi-model and multi-resolution biclustering approach

Our tool relies on a nested topic structure built with two biclustering methods as presented in Figure 3.1. It is generic enough to support overlapping or disjoint topics, falling within the diagonal blocks and the confusion blocks of the term-document matrix. These blocks serve as a basis for the exploration of topics and their relationships. The complete processing pipeline, summarized in Figure 3.3, is described in details below.

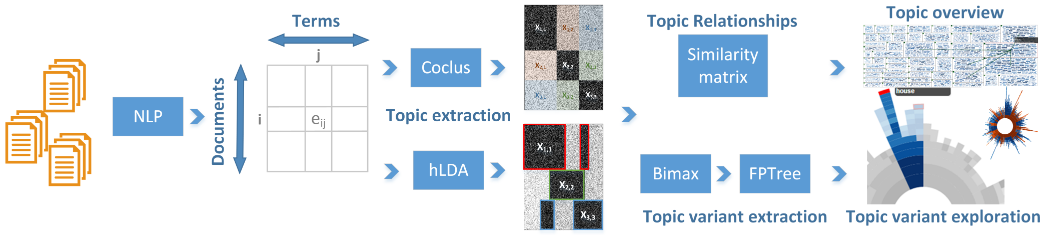


Figure 3.3: Our data processing pipeline: From left to right, NLP technology is used to select relevant text tokens. A term-document matrix is built. Various topic extraction methods, e.g. `Coclus` and `hLDA`, partition the TF-IDF matrix in topic blocks (in black) with various shapes. The latter are analyzed with Bimax to reveal topic variants. In parallel, topic relationships are extracted from their confusion blocks. Finally, interactive visualizations convey the topical structure and relationships to the analyst.

3.3.1 Text Processing

We use the popular *Vector Space Model* to build a term-document matrix. Each document is represented by a vector of distinct terms weighted by the *Term Frequency-Inverse Document Frequency* (TF-IDF). To this end, the raw text is processed using the Stanford CoreNLP library [Manning 2014] up to the Part-of-Speech tagging step. We keep only nouns and adjectives because they carry enough information to support the interpretation and exploration of topics and topic variants. We filter the resulting matrix keeping the 10,000 terms having the highest TF-IDF values.

3.3.2 Coarse-grained Topics

In order to spare the analyst the trouble of reading all documents, we summarize the corpus by extracting coarse-grained topics. On the one hand, the extracted topics must group documents that concern the same subject matter and also select the terms that are consistent for these documents. On the other hand, the topics must delimit meaningful subspaces, i.e. submatrices, for the second round of biclustering that extracts all topic variants.

3.3.2.1 A bicluster structure representing topics

Two topic extraction methods. In the survey of text mining in chapter 1 we have retained two families of methods suitable for the purpose of this thesis: biclustering methods and topic models. Biclustering methods exploit the duality of terms and documents to reveal consistent patterns. They group terms and documents in the same biclusters in such a way that the terms are representative to a set of similar documents. To extract such topics we can use `Coclus`, the diagonal biclustering algorithm proposed by Ailem et al. [Ailem 2016]. We anticipate that the non-repetition of the most common terms creates space for revealing the specificity of topics while giving consistent context to understand them. We hence expect that disjoint biclustering methods produce well separated biclusters and deliver useful patterns such as consolidated co-occurrences (see Figure 1.2), which helps find *topic variants*. `Coclus` does not depend on hyperparameters or arbitrary thresholds which cannot be defined by journalists. It must only be given the desired number of biclusters and can serve as a dimensionality reduction step prior to the fine-grained topic variant search.

To extract coarse-grained topics, we also retained topic models such as `hLDA`, providing a topic hierarchy where each topic is a probability distribution of terms (β) and each document is associated to one branch of the tree and is represented by a probability distribution of the topics (θ) found in its branch (see Figure 1.3). The high-order co-occurrence obtained in topics of `hLDA` are expected to enclose terms having more distant relationships. The fact that a term can be involved in multiple topics handles polysemy and reveals their multiple usage contexts in the corpus. However, the hyperparameters of the underlying probabilistic model (η, α and γ as described in section 1.2) need to be set and their impact on the topic shapes are not easy to anticipate by a lay user audience such as journalists. The probability distributions inherited from `LDA` (β & Θ) are also difficult to interpret and to manipulate [Alexander 2015]. Indeed, dealing with such distribution requires sophisticated visual analytics software [Dou 2011, Lee 2012, Alexander 2014]. For these reasons, the *crisp* membership of items offered by a bicluster structure allowing direct interpretation seems preferable.

Biclustering structure with a crisp membership. A bicluster structure can be formalized as follows. Given I a set of n documents and J a set of m distinct terms, a term-document matrix is defined by the sparse representation $X = \{e_{ij}, i \in [1..n], j \in [1..m]\}$ where $e_{ij} \neq 0$. With TF-IDF weights (see equation 1.1 p. 15) in the matrix cells, each entry e_{ij} measures how much a term is representative of the documents. These weights are exploited further along the data processing pipeline shown in Figure 3.3. A schema of two overlapping biclusters is presented in Figure 3.4 and we can define biclusters as follows. Given K the number of biclusters and X a matrix of size $n \times m$, bicluster $X_k, k \in [1..K]$

is a submatrix $I_k \times J_k$ such that $I_k \subseteq I$ and $J_k \subseteq J$. In disjoint biclustering, the hard allocation of rows and/or columns to one exclusive bicluster adds the following constraint: $\forall k, l \in [1..K]$ with $k \neq l$, $I_k \cap I_l = \emptyset$ and $J_k \cap J_l = \emptyset$. Such a bicluster structure can model any shape of topics, overlapping or disjoint, term-wise or document-wise, and can hence support a large number of topic extraction methods. In this thesis we investigated both biclustering and topic models through `Coclus` and `hLDA` respectively.

Configurations and implementation Although `hLDA` builds a hierarchy of topics, we consider only the topics placed at the leaf nodes as they contain more specific and interpretable terms. Indeed, the visual inspection of the topics enclosed at different depths of the `hLDA` hierarchy shows that internal nodes contain rather generic terms. In a way, they serve to categorize their children nodes rather than constitute real topics in their own right. By keeping the leaf nodes only, the bicluster structure created from the probability distribution of topics features overlapping term sets and disjoint document sets (see Figure 3.3). The number of iterations needed to learn the probability distributions is a parameter that must be set. The default value proposed in the Mallet toolkit [McCallum 2002] is 500 iterations.

Besides, `Coclus` provides a disjoint bicluster structure as direct output. It is computationally efficient with sparse matrices; its time complexity is $O(e \cdot p \cdot K)$, where $e = |X|$ is the number of non-zero entries in X, K is the number of clusters fixed as parameter and p is the number of iterations the algorithm runs until convergence to a local optimum. Empirically, less than 20 iterations were sufficient for the datasets we used (4,000 documents for 60,000 terms). Since the algorithm starts with a random initialization and converges to a local optimum, we resort to multiple tests to find the optimal partition for a given value of K, each test including the p iterations run by the algorithm until convergence. We perform 200 trials measuring the associated modularity score and, ultimately, we choose the partition yielding the best modularity. An alternative approach may consist in starting with a document clustering partition resulting from the efficient spherical k-means [Dhillon 2001b]. Then, `Coclus` refines the document partition while simultaneously finding the optimal term partition. For a given corpus, the number of topics to extract is an important parameter that must be chosen carefully, either by the analyst or trough a background process that sets the value of K empirically by varying its value in the range $[10..500]$. The direct mapping between a cocluster and a topic makes this task feasible for journalists.

While most evaluation approaches consist in analyzing the document partition through external metrics measuring the agreement between the partition and a ground truth, the structure of the topics is rarely considered in evaluations. Yet it is this structure that the analyst explores through the visualizations and that enables carrying out the tasks defined in the section 3.2.3.

On the one hand, the quality of the internal structure of topics is strongly related to the distributional hypothesis (Hypothesis 1 on page 14) and the lattent co-occurrence patterns constitute the *topic variants* expected to reveal the multiple angles and viewpoints related to facts or events. In section 1.5 we elicited the structural differences of both families of methods, from their theoretical foundation and mechanisms. To go further, the numerical experiment described below in section 3.4 achieves a statistical characterization of this topical structure based on real data sets and two representative methods, `Coclus` and `hLDA`. For instance, several metrics characterize the structure of the term co-occurrence patterns uncovered in each topic (i.e. *topic variants*), other metrics concern the topic size as well as the intra-cluster compactness and inter-cluster separability.

On the other hand, topic relationships help comprehend the frontiers of topics, highlight adjacency and create favorable conditions to serendipitous discovery. The next section explains how such relationships are built.

3.3.2.2 Topic relationships

We build topic relationships using a matrix-based measure of similarity between topics. The measure exploits the pairwise confusion blocks of topics shown in Figure 3.4. Indeed, these blocks inform about terms and documents two topics have in common. Previous work by Hanczar and Nadif [Hanczar 2011] uses the following Jaccard-based similarity measure:

$$Sim(X_k, X_l) = \frac{|X_k \cap X_l|}{|X_k \cup X_l|} = \frac{|X_k \cap X_l|_I + |X_k \cap X_l|_J}{|X_k \cup X_l|_I + |X_k \cup X_l|_J} \tag{3.1}$$

where $|.|$ denotes the cardinality of a set and, $|.|_I$ is the cardinality in the set of rows (i.e. documents) and $|.|_J$ is the cardinality in the set of columns (i.e. terms). But, for distinct diagonal biclusters (with $k \neq l$), we have $|X_k \cap X_l|_I = 0$ and $|X_k \cap X_l|_J = 0$. Thus this similarity measure needs to be adapted for diagonal biclusters. The diagonal partition divides the matrix X in $K \times K$ sub-matrices $X_{k,l}$ (Figure 3.4), $k \in [1..K]$ denotes one subset of the row partition and $l \in [1..K]$ denotes one subset of the column partition. Bicluster X_k is also noted $X_{k,k}$, its corresponding diagonal block in X. Given (X_k, X_l) a pair of diagonal biclusters, $X_{k,l}$ and $X_{l,k}$ constitute the confusion blocks that may share rows or columns with either diagonal bicluster of the pair. Hence, we consider that measuring the similarity between two diagonal biclusters consists in measuring to which extent they share rows or columns with their confusion blocks. So, we propose a new similarity measure $Sim(X_k, X_l)$ computing the extent of the intersection between the biclusters and their confusion blocks in rows and columns. Equation 3.1 is then transformed into:

$$\frac{|I_{k,k} \cap I_{k,l}| + |I_{l,l} \cap I_{l,k}| + |J_{k,k} \cap J_{l,k}| + |J_{l,l} \cap J_{k,l}|}{|I_k \cup I_l| + |J_k \cup J_l|} \tag{3.2}$$

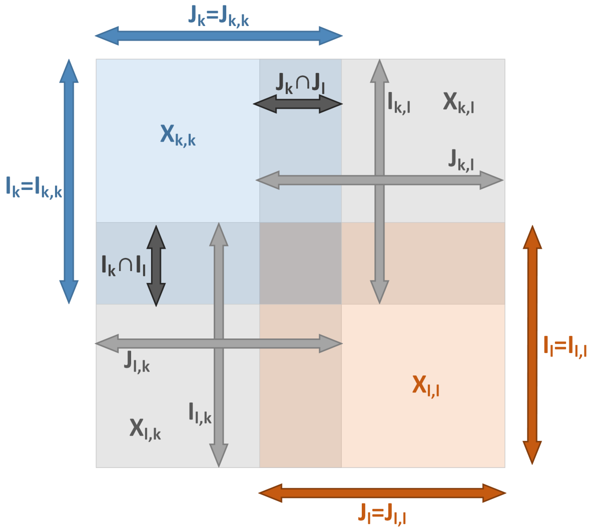


Figure 3.4: Two overlapping biclusters (in blue and orange): the columns of the confusion blocks show how many documents share common terms and the rows of the confusion blocks show how many terms occur in the same documents.

where $I_{k,l} = \{i \in I_k | \exists j \in J_l, e_{ij} \in X_{k,l}\}$ and $J_{k,l} = \{j \in J_l | \exists i \in I_k, e_{ij} \in X_{k,l}\}$ are respectively the set of rows and columns having non-empty vectors in block $X_{k,l}$.

We assume that any biclustering method building consistent topics yields a partition such that any term $j \in J_k$ occurs in at least one document $i \in I_k$. Similarly, any document $i \in I_k$ contains at least one term $j \in J_k$. Thus, we can say that $I_{k,k} = I_k$ and $J_{k,k} = J_k$ and then $I_{k,k} \cap I_{k,l} = I_{k,l}$ and $J_{k,k} \cap J_{l,k} = J_{l,k}$. Likewise, $I_{l,l} \cap I_{l,k} = I_{l,k}$ and $J_{l,l} \cap J_{k,l} = J_{k,l}$. Equation 3.2 can therefore be expressed:

$$\frac{|I_{k,l}| + |I_{l,k}| + |J_{l,k}| + |J_{k,l}|}{|I_k \cup I_l| + |J_k \cup J_l|}. \tag{3.3}$$

Equation 3.3 supports only disjoint biclusters. Indeed, in case of overlap, $I_k \cap I_l \neq \emptyset$ and $I_k \cap I_l = I_{k,l} \cap I_{l,k}$. Therefore, in Equation 3.3, the overlaps in lines are taken into account twice in $I_{k,l}$ and in $I_{l,k}$. Likewise, when $J_k \cap J_l \neq \emptyset$, $J_k \cap J_l = J_{k,l} \cap J_{l,k}$ and the overlaps in columns are taken into account in both $J_{l,k}$ and $J_{k,l}$

We can generalize Equation 3.3 for any pair of overlapping biclusters (X_k, X_l) by avoiding double counting of $I_k \cap I_l$ and $J_k \cap J_l$. We obtain the following equation:

$$\frac{|I_{k,l}| + |I_{l,k}| - |I_k \cap I_l| + |J_{l,k}| + |J_{k,l}| - |J_k \cap J_l|}{|I_k \cup I_l| + |J_k \cup J_l|}. \tag{3.4}$$

Note that for disjoint biclusters, Equation 3.4 is equivalent to Equation 3.3. When no similarity is observed between X_k and X_l, the confusion blocks $X_{k,l} = \emptyset$ and $X_{l,k} = \emptyset$.

We also have $I_k \cap I_l = \emptyset$ and $J_k \cap J_l = \emptyset$. So, $Sim(X_k, X_l) = 0$. In case of equality,

$$Sim(X_k, X_k) = \frac{|I_k| + |I_k| - |I_k| + |J_k| + |J_k| - |J_k|}{|I_k| + |J_k|} = 1.$$

We then have a similarity metric in $[0, 1]$ where 0 means no similarity and 1 means equality. We use Equation 3.4 to build a similarity matrix of topics. The next processing steps build the structure for the second level of detail devoted to *Topic Variants*.

3.3.3 Fine-grained Topic Variants

A journalist verifies his working hypothesis by finding multiple sources relating the same facts or stories. We work under the assumption that the aggregated corpus contains such repetitions. *Topic Variants* must, on the one hand, identify informative terms related to facts or stories, and on the other hand, retrieve multiple documents sharing them. If the multiplicity of documents strengthens the similarity of co-occuring terms through the distributional hypothesis, the multiplicity of terms also strengthens the similarity of documents. Hence, biclustering is a good candidate to extract such co-occurrence relationships. As terms and documents can be involved in multiple facts or stories, pattern-based overlapping biclustering such as `Bimax` [Prelić 2006], is appropriate to this purpose.

3.3.3.1 Topic Variants extracted by `Bimax`

Originally designed for gene expression data, *Bimax* [Prelić 2006] has never been used with textual data as far as we know. It takes a binarized term-document matrix as input and satisfies a constraint of maximal inclusion to extract all the maximal sets of terms that co-occur in maximal sets of documents. According to their study with gene expression data, this simple divide and conquer algorithm achieves similar results to those achieved by more complex methods. `Bimax` biclusters in a nested bicluster structure are formally defined as follows:

Given $X_k \subset X$, the bicluster or submatrix $I_k \times J_k$ of the topic $k \in K$, `Bimax` takes as input $\tilde{X}_k$, the corresponding binary matrix of size $\tilde{n}_k \times \tilde{m}_k$ such as

$$\tilde{X}_k = \{\tilde{e}_{ij} = \begin{cases} 1 & \text{if } e_{ij} \in X_k \text{ and } e_{ij} > \tau_k \\ 0 & \text{otherwise} \end{cases}, \forall i \in I_k, \forall j \in J_k\},$$

where $\tilde{n}_k \leq |I_k|$, $\tilde{m}_k \leq |J_k|$, and τ_k is the binarization threshold of the k^{th} topic.

Bicluster $\tilde{B}_{k,b}$ with $b \in [1..\beta_k]$ is the submatrix $\tilde{I}_{k,b} \times \tilde{J}_{k,b}$ where $\tilde{I}_{k,b} \subset I_k$ and $\tilde{J}_{k,b} \subset J_k$ and $\forall i \in \tilde{I}_{k,b}, \forall j \in \tilde{J}_{k,b}, \tilde{e}_{ij} = 1$.

In the nested structure, *Proposition 2* formalizing the maximal inclusion constraint expressed by Prelić et al. [Prelić 2006] becomes as follows:

Proposition 3. *For any $\tilde{B}_{k,b}$ described by the pair $(\tilde{I}_{k,b}, \tilde{J}_{k,b})$ the* maximal inclusion constraint *ensures that $\nexists b' \in [1..\beta_k]$ with $\tilde{I}_{k,b'} \subseteq I_k$ and $\tilde{J}_{k,b'} \subseteq J_k$ such that (1) $\forall i' \in \tilde{I}_{k,b'}, j' \in \tilde{J}_{k,b'} : \tilde{e}_{i'j'} = 1$ and (2) $\tilde{I}_{k,b} \subseteq \tilde{I}_{k,b'} \wedge \tilde{J}_{k,b} \subseteq \tilde{J}_{k,b'} \wedge (\tilde{I}_{k,b'}, \tilde{J}_{k,b'}) \neq (\tilde{I}_{k,b}, \tilde{J}_{k,b})$*

In essence, the maximal inclusion constraint ensures that no bicluster can be fully covered by another one. Instead, each bicluster is extended to its maximality. In a term-document matrix, Bimax extracts the maximal term co-occurrences shared by the maximal set of documents. In other words, *Bimax* finds all optimal consolidated co-occurrences patterns (see Figure 1.2) satisfying the exhaustiveness required by journalists. Such biclusters are called *topic variants* in this thesis.

We want to elicit all *Topic Variants* of each topic defined by the upper bicluster structure (see section 3.3.2). Since, for a given topic k, the complexity of Bimax is $O(\tilde{n}_k \tilde{m}_k \beta_k \min\{\tilde{n}_k, \tilde{m}_k\})$ [Prelić 2006], the number of Bimax biclusters β_k increases with the size and density of X_k, causing drops in runtime. While the extraction of coarse-grained topics acts as a dimensionality reduction mechanism, the diagonal blocks X_k corresponding to topics still come in various sizes and densities. Since Bimax works on a binary matrix, we define a configurable binarization threshold τ_k specific for each topic k. Entries of the submatrix X_k whose TF-IDF weight is lower than the threshold τ_k are set to 0; they are set to 1 otherwise. This reduces the density of the matrix ($|\tilde{X}_k| < |X_k|$), or even its dimensionality as zero vectors appear ($\tilde{n}_k \leq |I_k|$ and $\tilde{m}_k \leq |J_k|$). Moreover, Bimax takes three parameters: the minimum number of rows (*MinRows*) and the minimum number of columns (*MinCols*) per bicluster, and a maximum number of biclusters fixed to stop the algorithm (*MaxBC*). By definition of biclusters, $MinRows \geq 2$ and $MinCols \geq 2$. In section 4.3.3, we vary all these parameters individually to observe their effects on the resulting visualizations.

To extract *topic variants*, an efficient divide and conquer implementation of Bimax exists in BicAT [Barkow 2006]. Nevertheless we used our own java implementation of the incremental algorithm described in section 1.4.4 to prepare the dynamic approach we propose in chapter 6.

3.3.3.2 Term hierarchy

While Bimax fulfills the exhaustiveness required by journalists, it finds a large number of biclusters that overlap each other with respect to some of the terms and documents they contain, causing major interpretation issues. Since the meaning of a *Topic Variant* is mainly interpretable through its terms, we propose a hierarchy of terms that organizes *Topics Variants* (the Bimax biclusters) based on term overlaps. The example in Figure 3.5 is taken from the usage scenario unfolded in section 4.4.1. It concerns an event where Prim Minister Netanyahu met President Obama at the White House. In general, we observe that a term

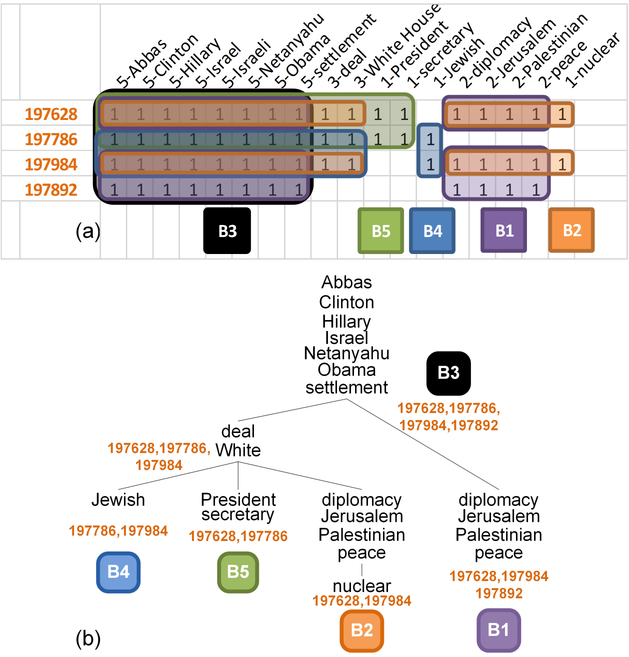


Figure 3.5: (a) Five `Bimax` biclusters (fine-grained topic variants) and (b) the related term hierarchy built with `FPTree`. The numbers in orange are document references. Each path from the root to a leaf describes a unique sequence of terms ordered by their overlap degree that represents one *topic variant*. They are grouped in the tree by their common prefix.

having a high overlapping degree (i.e. appearing in a large number of variants) tends to have a generic meaning associated to the event. Such a term, e.g. “Abbas”, “Clinton”, “Israël”, “Netanyahu”, “Obama” in Figure 3.5(b), also appears in all the documents gathered by the overlapping biclusters, e.g. documents #197628, #197786, #197984, #197892 in Figure 3.5(b). 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* (“diplomacy”, “Jerusalem”, “Palestinian”, “peace” in Figure 3.5(b)) or can be exclusive to one variant (“jewish”, “secretary” or “nuclear” in Figure 3.5(b)) and appear in fewer documents. Such terms are expected to reveal specific angles, viewpoints or facts (**T1.2**) and it makes sense to place them deeper in the branches of the hierarchy. For instance, B1 relates to the difficult diplomacy between Obama and Netanyahu and B2 concerns the nuclear deal with Iran. Both B5 and B6 raise the possible benefit for Hilary Clinton, the US Secretary of State and candidate for president elections, in case of successful meeting between Obama and Netanyahu.

To build such a term hierarchy, the *FPTree* algorithm [Han 2004] perfectly matches the needs described above. The lexicon of the topic is first sorted through 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 first term in the sequence of the inserted bicluster that does not match an existing path in the tree causes the creation of a new branch. A bicluster is placed at the path matching its complete term sequence. It can be placed on leaves (e.g. B1,B2,B4 or B5) or on internal nodes (e.g. B3).

Each path in the hierarchy is a unique sequence of terms describing one *topic variant*, starting with the most common terms at the root level and ending with the most distinctive terms in the leaves (**T2.2**). Moreover, each node in the hierarchy selects all the documents of the corresponding overlapping biclusters (**T2.3**). For instance in Figure 3.5, the term "abbas" appears in the five overlapping biclusters: B_1, B_2, B_3, B_4 and B_5. The union of the documents of these biclusters is the whole set. Next, the term "deal" appears in two overlapping biclusters: B_4 and B_5. The union of the documents of these biclusters is composed of documents #19768, #197786 and #197984. They constitute the maximal set of documents that share the term sequence defined by the path starting from "clinton" until "deal". As the analyst moves deeper along the branches towards the leaves, the document set under consideration melts down while the set of terms reaches its full extent.

Hence the nodes of the hierarchy correspond to the terms in the biclusters with different overlap degrees. They can be viewed as articulation points that gradually guide the analyst to specific *topic variants*. More generally, this hierarchical organization of biclusters aims to help the analyst understand the commonalities and specificities of fine-grained topic variants and possibly spot interesting facts or stories. While chapter 4 proposes coordinated views to explore the nested bicluster structure proposed in this section, the next section describes a numerical experiment that characterizes the structural properties of the topics and gives statistical evidence of the differences between a biclustering approach and a topic model approach.

3.4 Evaluation

The purpose of the journalists is to verify and generate hypotheses. This activity boils down to the exploration of term co-occurrences shared by multiple documents. While `Bimax` is leveraged to extract such co-occurrences at a granular level, the structure of the enclosing coarse-grained topics may have an important impact on the results. For example, it was found that topic overlaps harms the interpretation of topics [Sun 2014]. Before delving into the setup of a time-consuming user study, we seek presently to understand the influence of the chosen topic modeling technique on the characteristics of the resulting topics. This understanding will further guide us in the choice of independent variables to include in a controlled experiment comparing the influence of topic modeling techniques on the interpretability of topics by analysts. In this work, we focus on two methods, `Coclus` and `hLDA`.

3.4.1 Protocol

Our system is flexible enough to integrate any topic extraction method producing consistent sets of terms and documents, regardless of whether these sets are disjoint or have overlaps. It currently supports the extraction of overlapping topics using `hLDA` and disjoint topics using `Coclus`. In this section, we present an evaluation protocol allowing to understand in details topic shape and quality as well as the shape of the resulting `Bimax` biclusters and the corresponding term hierarchy built by the `FPTree` algorithm. We compare the `Coclus` method with `hLDA` in order to first elicit their differences and determine which method might better suit which user task. To this end, we have defined a set of bicluster metrics listed in Table 3.3 associated to each topic $k \in K$. The first three metrics inform about the size of the topics and the next three metrics about their quality in terms of compactness, separability and density. For intra-cluster compactness we used Root Mean Square Deviation. For the inter-cluster separability we used our own metric defined by the equation 3.4. In order to understand the co-occurrence structure of topics we propose several metrics concerning `Bimax` biclusters. Typically, we report their abundance (NbCCs) and how many terms/documents they cover in the topic. Finally, the term hierarchy given by `FPTree` is expected to be easier to understand than a flat list of sequences (topic variants). In such trees, deeper ramifications are expected to be more meaningful to the analyst than shallow ones. The global BranchingFactor metric measures the complexity of the internal structure of topic variants. However, we also aim to understand fan-out patterns with respect to the level/depth at which they occur. We then compute the branching factor at different levels of the term hierarchy (BrFact_L0-BrFact_L3). Wider initial fan-out patterns

Table 3.3: Intrinsic metrics computed for each topic.

Name	Description
NbTerms	log($\|J_k\|$), logarithm of the number of terms
NbDocs	log($\|I_k\|$), logarithm of the number of documents
Size	log($\|I_k\|.\|J_k\|$), logarithm of the topic size
Density	$\frac{\|X_k\|}{\|I_k\|.\|J_k\|}$, with $\forall e_{ij} \in X_k, e_{ij} \neq 0$
Compactness	$\frac{\sum_{i \in I_k, j \in J_k} (e_{ij} - \bar{e}_k)^2}{\|I_k\|.\|J_k\|}$, where $\bar{e}_k = \frac{\sum_{i \in I_k, j \in J_k} e_{ij}}{\|I_k\|.\|J_k\|}$, Root Mean Square (RMS)
Separability	$\frac{\min_{l \in [1..K]}(1 - sim(X_k, X_l))}{\max_{l,l' \in [1..K]}(1 - sim(X_l, X_{l'}))}$, where sim refers to Equation 3.4
NbCCs	Number of `Bimax` biclusters (β_k)
TermCoverage	Number of terms covered by at least one `Bimax` bicluster
DocCoverage	Number of documents covered by at least one `Bimax` bicluster
NbDocsByCC	(average) Number of documents in the `Bimax` biclusters
NbTermsByCC	(average) Number of terms in the `Bimax` biclusters
BranchingFactor	(average) Number of children of the internal FPTree nodes, the leaves being ignored
LastBranchingNode	(average) The level in each bicluster branch where the last FPTree node has multiple children
BrFact_L0-BrFact_L3	The branching factor at each level of the FPTree

suggest that the corresponding FPTree produces numerous linear sequences of terms with few shared terms between documents, which may be equated to more diversity of stories or events with short commonality. In contrast, narrower fan-out patterns close to the root produce term hierarchies that may be explored more easily. Indeed, the *topic variant* are grouped by longer common prefixes that can concern common stories or events.

For each metric we generated distribution histograms. The most informative ones are shown in Figure 3.6, Figure 3.7 and Figure 3.8. Some of them (e.g. the metrics related to the size) have a lognormal distribution. Moreover, we have identified several metrics that are sensitive to topic size. To avoid any bias due to the topic size distribution, we compared Coclus and hLDA through 14 fixed-size bins. We then applied a discretization of 14 equal intervals on the $\log(Size)$ domain, in which statistics are computed for each metric (Figure 3.7 and 3.8). The p-values generated using Student's T-test are displayed to assess the statistical significance of the samples. In the following analysis, we only consider the comparisons having a *p-value* < 0.05.

In order to fairly compare Coclus and hLDA, we harmonized as much as we could the pre-processing steps for both. We used a Python implementation of Coclus available online[2]. We ran 200 trials because Coclus converges to a local optimum with a random initialization step. We used the hLDA implementation shipped in Mallet [McCallum 2002] with the default values of the hyperparameters and the number of iterations ($\alpha = 10$, $\eta = 1$, $\gamma = 0.1$, $iter = 500$). We set a fixed hierarchy $depth = 5$. For both Coclus and hLDA, we used the tokenization provided by Mallet to build the TF-IDF matrix. Only the top 10,000 terms with the highest TF-IDF weights were kept.

Since Coclus needs the number of topics to be set a priori, we decided to use the number detected automatically by hLDA from the data. By giving both algorithms the same targeted number of topics, we make our best to present topics of comparable granularity to the analyst eventually. By considering only the leaf nodes in the topic hierarchy of hLDA and by keeping all the terms with a probability greater than 0, the obtained topics overlap with respect to their terms but are disjoint document-wise. In contrast, Coclus builds fully disjoint topics. To extract *topic variants*, we used our own implementation of the incremental algorithm (see section 1.4.4) with the following parameters: $MinRows = 2$, $MinCols = 2$, $MaxBC = 10,000$ and $\tau_k = 0, \forall k \in [1..K]$.

3.4.2 Dataset

We used a real dataset composed of news articles continuously aggregated in a database from multiple online news sources (BBC, CNN, Reuters, France24, Egypt Independent and Der Spiegel). We chose five time intervals lasting two weeks each as recapped in Table 3.4.

[2] https://pypi.python.org/pypi/coclust

Table 3.4: Characteristics of the dataset

Period start date	#Documents	#Terms	#Topics	Coclus	hLDA
				NbCCs>0	NbCCs>0
07/03/2016	3,738	55,983	99	92	98
04/04/2016	3,466	53,199	93	84	89
06/06/2016	3,467	52,775	109	95	109
04/07/2016	3,386	52,610	91	80	78
01/08/2016	3,498	53,118	105	93	99
05/09/2016	3,732	59,133	116	105	108

For each two-week period we ran both `Coclus` and `hLDA`, whose output is analyzed with `Bimax`. We chose non contiguous time periods to reduce the effect of long-lasting news stories that would result in very similar topics across two contiguous periods. The dataset has a total of 1,130 topics having a nonzero number of *Bimax* biclusters ($NbCCs > 0$), for which we also computed the metrics listed in Table 3.3.

3.4.3 Analysis

Our goal is to investigate the behavior of both `Coclus` and `hLDA` regarding the metrics enumerated in Table 3.3. We start our analysis with the size distribution in Figure 3.6. Firstly, we observe that the topic distribution of `Coclus` (in red) is right-skewed, with more topics of small to medium sizes. The topic distribution of `hLDA` has a bell shape and extracts more topics having a large size than `Coclus` does. Secondly, we aim to understand which of *NbTerms* or *NbDocs* has the greatest impact on topic size. For `hLDA` (in blue), the distribution of *NbDocs* is more spread and *NbTerms* shows a left-skewed distribution with more topics in the higher value ranges. We also aim to get more detailed patterns for topics of similar size in Figure 3.7(a). Unsurprisingly, *NbTerms* is higher for `hLDA`. This

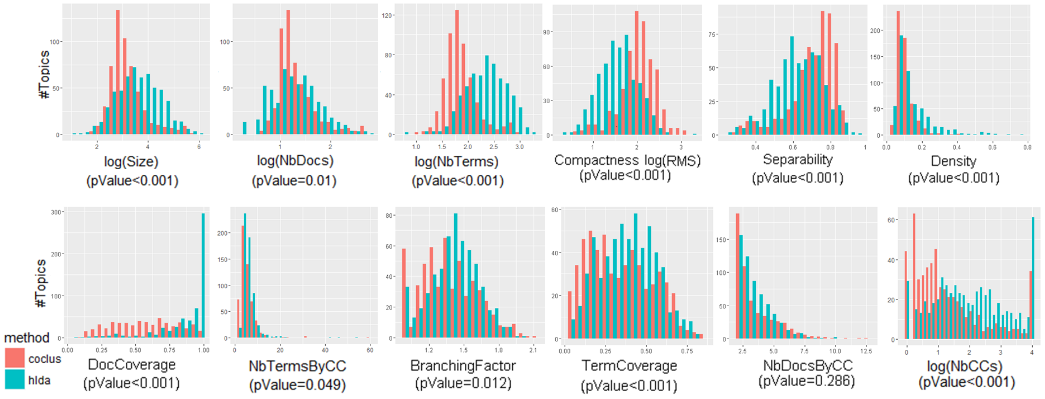
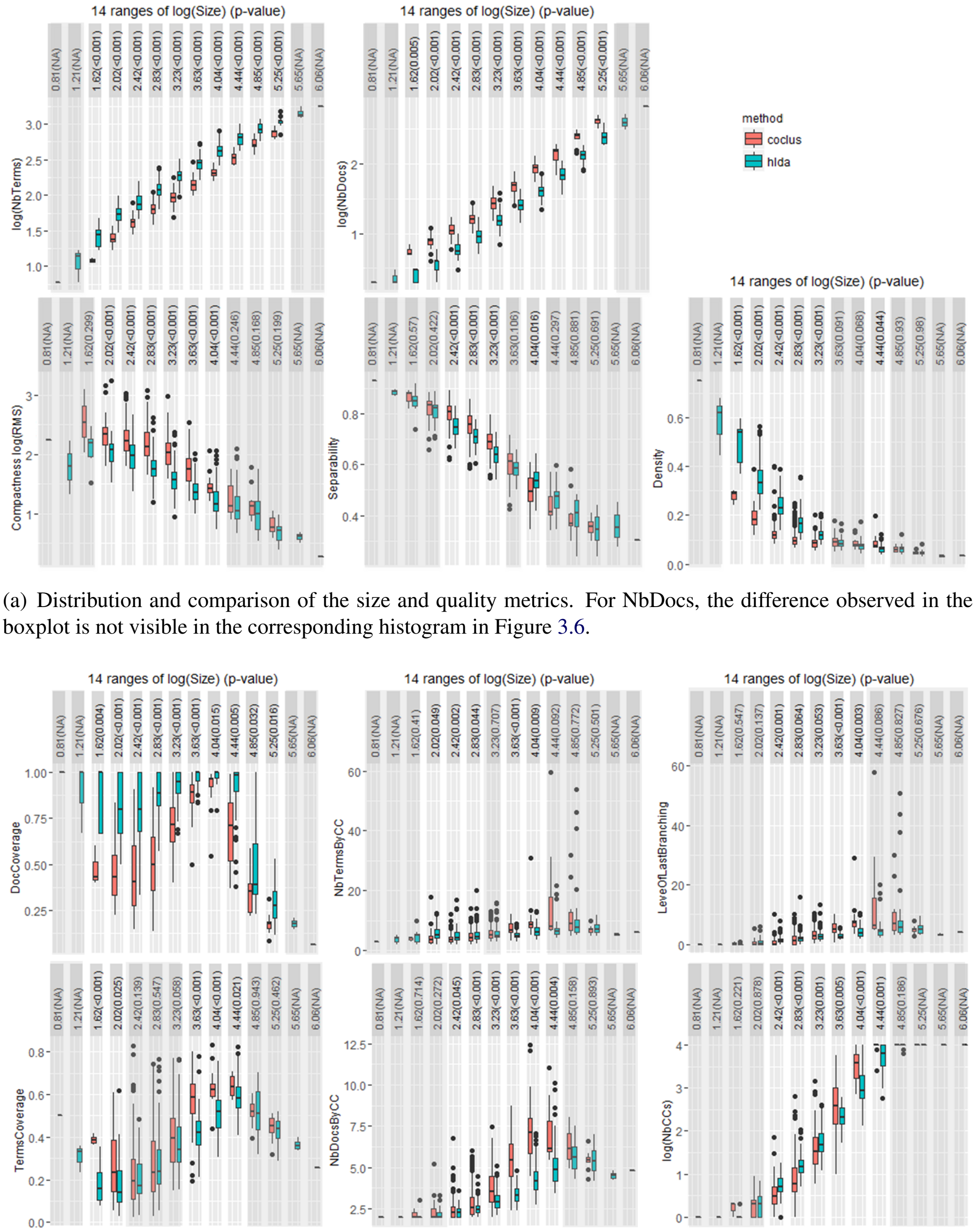


Figure 3.6: Distributions and comparison of different metrics for all topics produced by `Coclus` and `hLDA`. Note, for `hLDA`, the impact of its large topic size on the distributions of DocCoverage, BranchingFactor, TermCoverage, NbDocsByCC and NbCCs. These metrics need also to be compared with topics of similar size in Figure 3.7 and Figure 3.8.



(a) Distribution and comparison of the size and quality metrics. For NbDocs, the difference observed in the boxplot is not visible in the corresponding histogram in Figure 3.6.

(b) Distributions and comparison of metrics concerning `Bimax` biclusters discovered in each topic.

Figure 3.7: Comparison of `Coclus` and `hLDA` through different metrics (Shaded area for p-value>0.05).

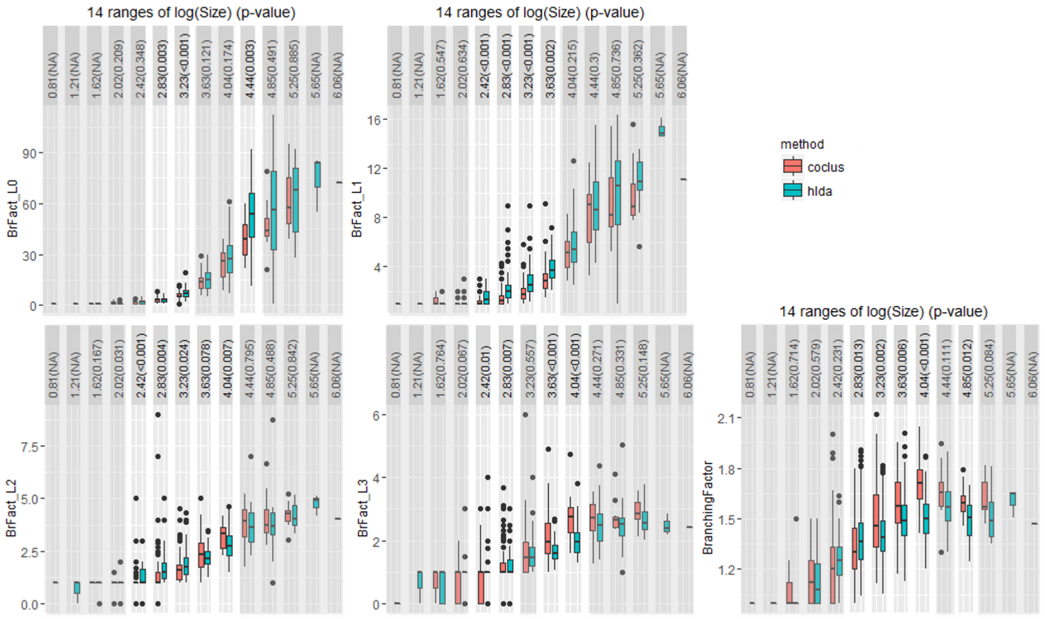


Figure 3.8: Comparison of `Coclus` and `hLDA` through metrics concerning the term hierarchy (Shaded area for p-value>0.05).

observation can be explained by the numerous overlapping terms generated by *hLDA* and its higher-order co-occurrences described in section 1.5 which in turn increase the topic vocabulary. However, for `Coclus` the number of documents (*NbDocs*) is higher, as shown in Figure 3.7(a). In this case, the simultaneous analysis of document and term vectors achieved by `Coclus` favors consolidated co-occurrences (see Figure 1.2 page 15). Thus, the topics extracted by `Coclus` tend to have more documents described by fewer, yet more specific terms.

Next, we would like to understand how the metrics influence one another and whether some of them explain the differences between `Coclus` and `hLDA`. To this end we used Principal Component Analysis (*PCA*). The topics and the metrics are respectively plotted through the first two components in Figure 3.9. Again, `hLDA` appears in blue and `Coclus` in red. We first analyze the metrics that better explain topic variance. In Figure 3.9(b), the X-axis sets apart two groups of metrics that are vertically centered (i.e they have no influence on the Y-axis). These are, on the right, the metrics related to the topic size (*Size*, *NbDocs*, *NbTerms*, *NbBCCs*) and, on the left, the quality metrics (the *Separability*, the *Compactness* with *RMS* and the *density*). In Figure 3.6, the histograms show that `Coclus` topics tend to be small, weakly compact (high *RMS*) and highly separable (high *Separability*). In contrast, `hLDA` topics tend to be large with high compactness (low *RMS*) and a low *Separability*. In Figure 3.9(a), the PCA projection of topics also reveals this pattern: on the left side, there

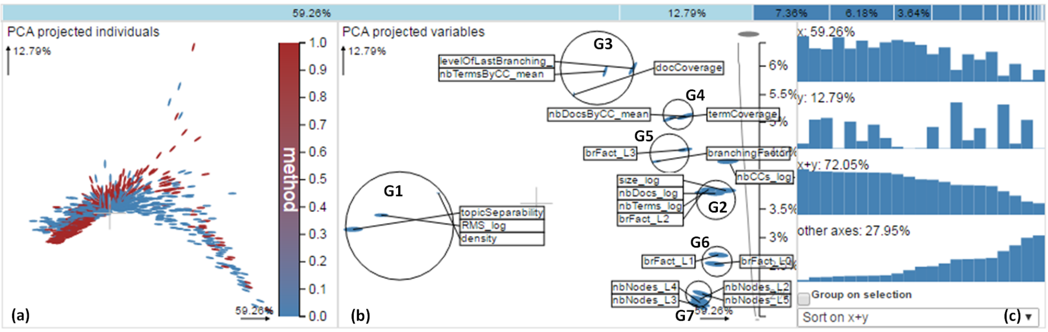


Figure 3.9: (a) PCA Projection of topics (blue for `hLDA` and red for `Coclus`). (b) PCA projection of metrics with strongly correlated groups (G1 to G7). (c) Interactive histograms to sort the metrics according to their contributions on the X- and/or Y-axis.

is a higher concentration of red topics (`Coclus`) and on the right side there is a higher concentration of blue topics (`hLDA`). We also observe in Figure 3.9(b) that a majority of metrics are correlated with the X-axis (they all appear on the extremities of the X-axis) and are thus sensitive to the size. This shows that the comparison of metrics through the histograms is biased by the preponderance of large topics for `hLDA`. A more qualified analysis of the metrics may be achieved by examining topics of similar size through the boxplots. Figure 3.7(a) confirms the differences described above, associating `hLDA` with higher compactness (lower *RMS*) and density values. This is an effect of the higher number of terms associated with `hLDA` topics for a given number of documents. But, `Coclus` makes for better separability due to its hard partitioning scheme.

So far, we analyzed the differences between `Coclus` and `hLDA` regarding topic shapes and quality. We also want to understand their differences with respect to the characteristics of the resulting `Bimax` biclusters, as this informs us on the properties of co-occurrence patterns. The boxplots of the *NbCCs* metric in Figure 3.7(b) reveal that, in the lower topic size ranges, `hLDA` leads to a higher number of biclusters than `Coclus` whereas the opposite trend is observed with large topics. Indeed, the comparatively larger number of terms provided by `hLDA` makes it easy to find common terms even in topics with a few documents. But for topics with a greater number of documents, `Coclus` selects better the documents sharing the terms related to the topics and surpasses `hLDA` regarding the number of extracted `Bimax` biclusters *NbCCs*, i.e. the number of distinct lists of terms shared by multiple documents.

In Figure 3.9(b), the *PCA* projection shows several groups of closely located metrics that are strongly correlated. We then try to explain this correlations below. Groups G1 and G2 concern the quality of topics (compactness, separability and density) and the size of topics (size, nbDocs, nbTerms) respectively. Group G3 captures three variables: *DocCoverage*, *NbTermsByCC* and *LevelOfLastBranching*. In Figure 3.7(b), we observe that hLDA achieves higher document coverage than Coclus. In addition, *NbTermsByCC* and *LevelOfLastBranching* follow the same crossing pattern as *NbCCs*. We discuss in detail this crossing pattern and the advantage of hLDA in terms of DocCoverage in section 3.5. In group G4, Coclus achieves higher *TermCoverage* and higher number of documents per bicluster (*NbDocsByCC*). We conclude that Coclus finds more documents sharing term co-occurrences. The superior term coverage suggests that the terms are more specific to the documents.

Finally, the last three groups (G5-G7 in Figure 3.9) concern the properties of the term hierarchy built by FPTree. In Figure 3.8, the global BranchingFactor shows larger values for Coclus. In the first hierarchy levels, hLDA topics have more branches, i.e. higher values for BrFact_L0 and BrFact_L1. This suggests that hLDA produces more linear sequences of terms in the tree, which may be equated to more noisy terms or more diversity of stories with short commonality. In the next levels (starting from BrFact_L3), Coclus topics are more ramified in the medium and large topic categories. We thus expect the Coclus term tree to be more easily explored. Below we discuss the most significant differences found between Coclus and hLDA and anticipate their impact on the journalist tasks.

3.5 Discussion

The first difference we found is that hLDA extracts larger topics, mainly due to their larger number of terms. This can be explained by the fact that hLDA identifies more high-order co-occurrences. This difference has an impact on most other metrics. Indeed, the higher the number of terms in topics (i.e. hLDA method), the more likely they will occur in its documents. This explains the advantage of hLDA in terms of compactness and density, but also the higher document coverage achieved by the Bimax biclusters. However, the higher the branching factor in the first levels of the term hierarchy the fewer the common terms found in Bimax biclusters. Nevertheless, we expect the higher-order co-occurrences fetched by hLDA to lead the journalist to examine a greater variety of documents and to explore more distant relationships through the term hierarchy. Therefore, we expect hLDA to be well suited for the diversification process of journalists as it would favor serendipitous findings, and hypothesis generation, in line with previous work by Alexander et al. [Alexander 2014] using LDA. This expectation comes with the caveat that large hLDA topics will generate an overwhelming quantity of linear sequences of terms that is difficult to explore, even with a tree visualization.

Secondly, a thorough analysis of the metrics for a given topic size shows that Coclus topics have comparably more documents. This observation is explained by the biclustering mechanism which promotes consolidated co-occurrences. As opposed to the previous observation about hLDA: the higher the number of similar documents in topics, the higher the chance to find documents sharing term co-occurrences. Hence, `Coclus` leads to `Bimax` biclusters with more documents.

In addition, we observed earlier that several metrics feature a crossing trend as the topic size grows, in particular the number of `Bimax` biclusters (NbCCs), the number of terms in each bicluster (NbTermsByCC) and the average of the deepest ramification level in the term hierarchy (LevelOfLastBranching). These metrics capture respectively document relationships as characterized by the abundance of term sequences, the length of these sequences and the depth of the ramifications (long vs. short commonalities). It appears that for these criteria, `hLDA` surpasses `Coclus` regarding small topics. But `hLDA` produces large topics when dealing with major news stories much more than `Coclus`. In this case, the chained relationships of the higher-order co-occurrences bring more noise with `hLDA`. In contrast, `Coclus` finds terms that are specific to more documents, surpassing `hLDA` for the same criteria when dealing with large topics. `Coclus` deals better with major news stories by locating more specific topic variants with slight differences. Moreover, in every topic size range, `Coclus` finds term co-occurrences shared by more documents, strengthening the evidence backing the working hypothesis of the journalist.

3.6 Conclusion

The proposed approach relies on a nested biclustering structure extracting coarse-grained topics and breaking them into fine-grained topic variants. This structure is exploited by every analytic component of the visual analytics software detailed in chapter 4. It supports multiple topic extraction methods delivering any top-level topic shapes. A new similarity metric developed for both overlapping and disjoint biclusters allows to depict topic frontiers and relationships. The numerical experiments examines multiple intrinsic metrics to characterize the properties of the coarse-grained topical structure of both `hLDA` and `Coclus`. We found that `hLDA` extracts more high-order co-occurrences of terms and identifies more distant document relationships. We expect this to be useful to refine a hypothesis or generate new ones, but it tends to generate more noise with major news stories. In contrast, `Coclus` spots more specific term co-occurrences shared by multiple documents. It handles better major news stories and is expected to better serve hypothesis verification. The result of this numerical experiment allowed us to express hypotheses for the design of the user study described in chapter 4. In addition, topic size must be considered as an independent variable to fairly assess the quality of the topic interpretation given by both methods.

The next chapter gives more details about the design rational of the proposed visualizations.

CHAPTER 4

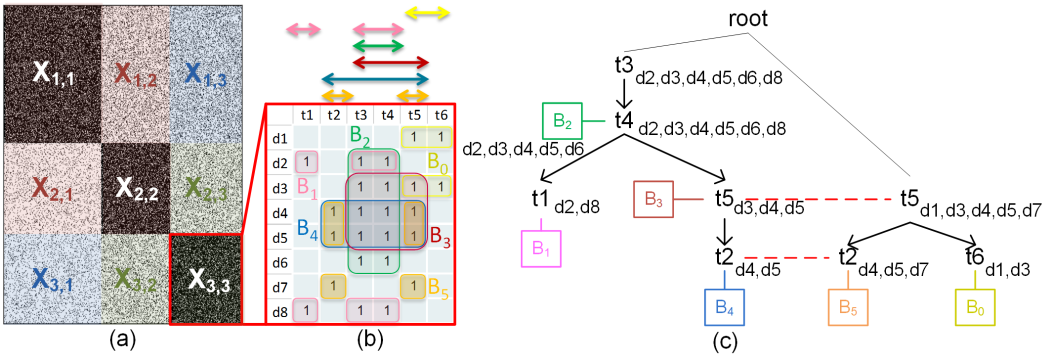
Multi-resolution Visual Analysis of Text Corpus

Contents

4.1 Introduction **83**
4.2 System overview **86**
4.2.1 Workflow and tasks abstraction 86
4.2.2 Four visual components 87
4.3 Visualization Design **89**
4.3.1 Topic Weighted Map 89
4.3.2 Topic Variant Overview 91
4.3.3 Human in the loop 94
4.3.4 Topic Variant Comparator 96
4.4 Evaluation **99**
4.4.1 Usage scenario 99
4.4.2 Qualitative evaluation with a domain expert 101
4.4.3 Discussion and future work 109
4.5 User study **110**
4.5.1 Properties of topic extraction methods 110
4.5.2 Experimental protocol 117
4.5.3 Results 119
4.5.4 Discussion 123
4.6 Conclusion **124**

4.1 Introduction

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



	t1	t2	t3	t4	t5	t6
d1					1	1
d2	1		1	1		
d3			1	1	1	1
d4		1	1	1	1	
d5		1	1	1	1	
d6			1	1		
d7		1			1	
d8	1		1	1		

Figure 4.1: The nested structure delivered by our hybrid biclustering approach. (a) The diagonal biclusters (topics) denoted $X_{k,k}, k \in \{1, 2, 3\}$ and the confusion blocks for each pair denoted $X_{k,l}, k \in \{1, 2, 3\}, l \in \{1, 2, 3\}$. (b) Bimax biclusters delivered for one topic. (c) The biclusters organized in a prefix-based term hierarchy and the documents retrieved at each node (the union of the documents sets in the biclusters of the subtree). If the terms of a bicluster do not contains an existing prefix a new branch is created which can involve redundancy, e.g. $\{t5, t2\} \subset \{t3, t4, t5, t2\}$.

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) search the documents that verify a given hypothesis and, 2) identify new angles or viewpoints that makes him/her refine or generate a new hypothesis that better fits the facts found in the data.

A first challenge is to avoid the analyst to read all documents for identifying such facts, angles or viewpoints. This challenge can be tackled by summarizing text corpora into multiple topics. Second, the journalists seek to identify multiple sources sharing common facts, angles or viewpoints concerning stories of interest. This can consist in analyzing all relationships between documents, and in identifying the terms they share. Ultimately the analyst aims to identify fragments of textual contents related to her working hypothesis.

A majority of visual analytics solutions are based on topic models derived from Latent Dirichlet Allocation [Blei 2003] and Non-Negative Matrix Factorization [Lee 1999, Choo 2013]. With these models, the semantic meaning of topics is often represented by the N most frequent terms. However, recent work [Alexander 2015] shows that “a topic is more than the top 10 words”. The authors argue that, to ensure that a topic model relates the semantic of the documents, it must reflect “the subtle patterns of co-occurrences”. The visual analytics tool described in this chapter leverages the advantages of the nested structures described in chapter 3 (see Figure 4.1) to provide multi-resolution analysis starting with coarse-grained topics and focusing on fine-grained *topic variants*.

While we can typically identify more than 50 coarse-grained topics in a corpus containing thousands of documents, the analyst must quickly locate topics of interest and understand their relationships. A major visualization issue is to organize topic terms in such a way that they give sufficient context to understand the meaning of the contents while showing structural patterns such as the importance of topics and their semantic relationships. To this end, we propose a novel visualization named *Topic Weighted Map* that depicts all extracted topics in multiple tag-clouds, while reflecting their relative size and similarity.

Another challenge has to do with the exploration and the interpretation of the numerous overlapping *Topic Variants* (*Bimax* biclusters) sharing many terms and documents. To address this issue we designed a hierarchical visualization showing term overlaps. This approach gives an overview of all topic variants while limiting their term redundancy and conveying their commonalities and specificities. The topic analysis through these variants focuses the work on a preselection of documents potentially useful for the journalist inquiry.

The latent topical structure shaping the *Topic Variants* found within each topic is determined by the upper-level topic extraction method. However, previous work does not provide clear recommendations regarding the choice of topic extraction methods to leverage analyst's tasks. The current work contributes to establish this kind of recommendation, according to the topic extraction method and the intrinsic properties of the extracted topics. Indeed, our system can constitute a common base to investigate the influence of the topical structures elicited by alternate topic models on the comprehension of topics and their variants by the analyst. First, we evaluated the usefulness of our system and the feasibility of investigative journalists' tasks through a usage scenario and a qualitative evaluation with an expert. Second, we conducted a controlled experiment that compares the interpretation quality of topics extracted by two methods; the first is Coclus [Ailem 2016], a diagonal biclustering method, and the second is the probabilistic Hierarchical Latent Dirichlet Allocation (hLDA) [Griffiths 2004]. Thereby we provide evidence of their differences from an analyst's perspective and identify the characteristics making either method suitable to tasks of investigative journalists.

Our contribution is manifold:

1. We designed a multi-resolution approach for the analysis of text corpora; it supports journalist's top-down and bottom-up analytical processes. We demonstrated the usefulness of our system through a usage scenario and a qualitative evaluation with an expert user.
2. We experiment *Bimax* with textual data and allow the analyst to control its parameters.
3. We propose a hierarchical model leveraged by coordinated multiple views allowing to explore numerous overlapping term-document biclusters, handling their term redundancies and conveying their commonalities and distinctive traits.

4. We propose the *Topic Weighted Map*, a novel topic visualization based on multiple tag clouds nested in a treemap reflecting their size and their relative similarity.
5. We conducted a controlled experiment investigating the impact of two topic extraction methods and their topical structure on the comprehension of topics and topic variants by analysts.

In the rest of this chapter, we recall the tasks under consideration and describe the visual components of the software system we implemented to meet this need in section 4.2. The visual encoding and the design rational are described in section 4.3. A usage scenario and a qualitative evaluation are reported and discussed in section 4.4. The controlled experiment, its results and the related discussion are described in section 4.5. Finally, we conclude our work in section 4.6.

4.2 System overview

4.2.1 Workflow and tasks abstraction

Our system has been designed following Munzner's four-level nested model [Munzner 2009] as described in section 3.2. We recap the process of investigative journalists [L. Hunter 2011]:

- **Mapping the subject:** during this process, the journalist gets an overview of the subjects of inquiry.
- **Focus:** the journalist focuses the investigation on a specific aspect to identify facts, viewpoints that validate, refute or refine her hypothesis.
- **Diversification:** the journalist looks one step farther from the targeted object, hoping to find unexpected information and new angles of analysis. This step is important to ensure that she does not miss essential information that could call into question the initial hypothesis.

While this workflow is presented linearly, it is actually intertwined and repeated during journalists' investigations. We took these processes into account as we designed our system. We also recall the tasks defined in section 3.2:

T1 Summarize the corpus to identify topics of interest and aspects to investigate.

T1.1 Understand and locate topics of interest

T1.2 Understand and identify *topic variants* to find facts and viewpoints related to stories or events.

T2 Find the documents that verify an aspect of the working hypothesis.

T2.1 Select *topic variants* of interest related to specific facts, angles or viewpoints / filter out the rest.

T2.2 Identify common and distinctive terms/documents of *topic variants*.

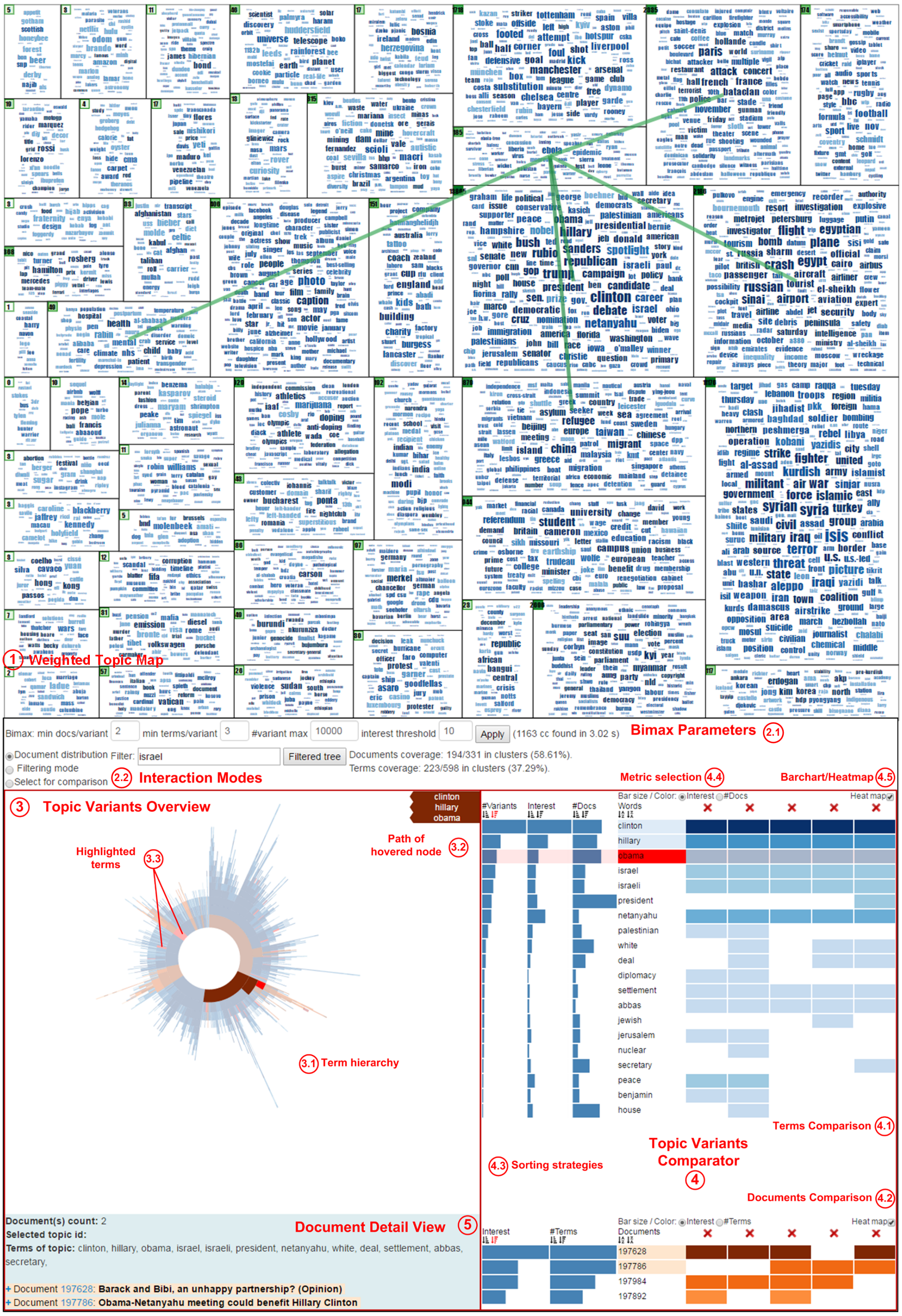
  - **T2.3** Retrieve the documents related to the *topic variants* / read raw original text / identify terms of a given topic variant in the text.

- **T3** Identify new angles or viewpoints to refine or generate new hypotheses that better fit the facts discovered in the text content.
  - **T3.1** Identify the most similar topics.
  - **T3.2** Identify terms of interest for queries.
  - **T3.3** Identify topic variants sharing documents/terms of interest.

4.2.2 Four visual components

The tool presented in Figure 4.2 supports topic exploration at multiple resolutions, from coarse-grained topics down to raw text within documents. It relies on the nested structure built with the hybrid biclustering approach described in chapter 3 and summarized in Figure 4.1. It comprises four visual components shown in Figure 4.2 that cover all the tasks described in section 4.2.1: the *Topic Weighted Map* (1) and the *Topic Variant Overview* (3), the *Topic Variant Comparator* (4) and the *Document Detail View* (5). The analyst starts his work by inspecting the *Topic Weighted Map* to get an overview of all the topics (**T1.1**) extracted from the corpus by 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 (3.1). The analyst can then explore all *Topic Variants* through a sunburst visualization (**T1.2**).This visualization is used to start the focus process aiming to verify specific aspects of the working hypotheses. The analyst can then filter the *Topic Variants* by keyword and hide uninteresting variants (**T2.1**). Next, a subset of *Topic Variants* can be chosen for further inspection in the *Topic Variant Comparator* (4). This view shows a matrix representation of the distribution of terms (4.1) and documents (4.2) within the selected topic variants (**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**), giving a precise semantic context to all upper level visualizations. Obviously, documents remain the ultimate material used by a journalist to evidence his hypotheses.

The diversification process is supported through multiple interactions combined to promote serendipity. Firstly, in the *Topic Weighted Map*, the analyst can browse the vicinity of a topic of interest and follow the suggested links with adjacent topics to deepen the investigation (**T3.1**). Secondly, the term hierarchy visible in the *Topic Variant Overview* reveals meaningful terms suggesting new keyword combinations for search queries **T3.2**. Thirdly, *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**).



Bimax: min docs/variant 2 min terms/variant 3 #variant max 10000 interest threshold 10 Apply (1163 cc found in 3.02 s)

● Document distribution Filter: israel Filtered tree Documents coverage: 194/331 in clusters (58.61%).
○ Filtering mode Terms coverage: 223/598 in clusters (37.29%).
○ Select for comparison

Document(s) count: 2
Selected topic id:
Terms of topic: clinton, hillary, obama, israel, israeli, president, netanyahu, white, deal, settlement, abbas, secretary,

+ Document 197628: **Barack and Bibi, an unhappy partnership? (Opinion)**
+ Document 197786: **Obama-Netanyahu meeting could benefit Hillary Clinton**

Figure 4.2: The 1^{st} version of our visual analytics tool and its components. The *Topic Weighted Map* visualization (1) shows 50 topics extracted from online news between Nov. 2 nd and Nov. 16 th , 2015. The topic selected in the *Topic Variant Overview* (3) is about the U.S. presidential elections. Five topic variants concerning Hillary Clinton have been sent to the *Comparator* (4) and the term “Obama” is hovered.

Through a Web architecture our tool is implemented in Java, Scala and Python for the backend, and in Javascript (*D3.js*) for the visualizations. The tool presented in Figure 4.2 is the first version we have developed. This version has been presented to an investigative journalist for a qualitative evaluation as described in the section 4.4.2. The collected feedback allowed us to identify limitations of the first version that have been addressed in the second version presented in Figure 4.8. In the rest of this chapter, we present the state of the second version of our tools. The next section describes in detail our visual encoding design.

4.3 Visualization Design

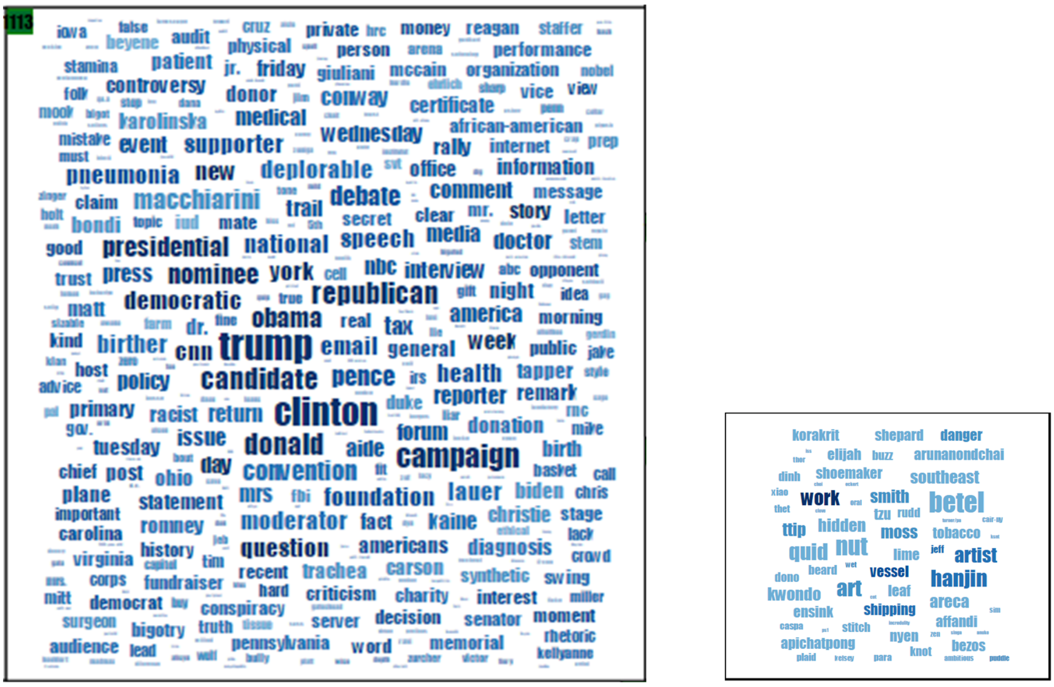
4.3.1 Topic Weighted Map

One of the challenges we address is to avoid the expert to read every document. We summarize the corpus by extracting coarse-grained topics as presented in chapter 3. On the one hand, the extracted topics must group documents that concern the same topics and also select the terms that are consistent for these documents. On the other hand, the topics must deliver a meaningful subspace for the second level of biclustering that extracts all *topic variants*.

In order to convey a bird's-eye view of the topics enclosed in the corpus, we define a novel visualization, called *Topic Weighted Map* (Figure 4.2(1)) which combines: 1) an MDS projection of topics in the 2D plane generating spatial coordinates used by 2) a Weighted Map visualization (a spatially consistent variant of treemaps) where each node is depicted by 3) a word cloud. For instance, Figure 4.2(1) represents 50 topics extracted from online news between Nov. 2^{nd} and Nov. 16^{th}, 2015. Topic size is proportional to the number of terms and the number of documents it involves, characterized as follows: $\sqrt{|I_k| \times |J_k|}$. Topic proximity in the 2D plane reflects topic similarity. We discuss the building blocks of the Topic Weighted Map visualization in detail below.

Topic word clouds. The diagonal biclustering step extracts topics, each grouping a set of terms that are consistent with respect to a set of documents. Rather than displaying the top N terms for each topic, we consider all terms and show their relative importance within the topic. To this end, we use a tag cloud visualization to depict each topic. The size of a term is mapped to an interest criterion (*Interest(j)*) based on the sum of its *TF-IDF* weights (see equation 1.1) across all documents belonging to the topic.

$$Interest(j) = \log{(1 + \sum_{i \in I_k} e_{ij})}, \forall j \in J_k \tag{4.1}$$


113
iowa false audit cruz private money reagan staffer
beyene physical person performance
stamina patient jr. friday giuliani mccain organization nobel
controversy donor conway certificate vice view
karolinska medical african-american
mistake event supporter wednesday rally internet prep
pneumonia new deplorable office information
claim macchiarini trail debate comment message
bondi secret clear mr. story letter
good presidential national speech media doctor stem
trust press nominee york nbc interview opponent
democratic republican night idea
matt dr. obama real tax america morning
kind birther cnn trump email general week public jake
host policy candidate pence health tapper
advice primary racist return clinton duke reporter remark
tuesday issue donald aide forum donation birth
chief post ohio day convention campaign basket call
plane statement mrs foundation lauer biden chris
important romney moderator fact kaine christie stage
carolina history question americans diagnosis lack
virginia tim recent trachea carson synthetic swing crowd
corps fundraiser hard criticism charity interest miller
mitt democrat conspiracy server decision senator moment
surgeon bigotry truth rhetoric
audience lead pennsylvania word memorial kellyanne
korakrit shepard danger
elijah buzz arunanondchai
dinh shoemaker southeast
work smith betel
tzu rudd
ttip hidden moss tobacco
quid nut lime artist
beard vessel hanjin
kwondo art leaf areca
ensink shipping affandi
caspa stitch
apichatpong nyen bezos
plaid para knot

(a) Large topic

(b) Small topic

Figure 4.3: (a) The large tag cloud concerns news about U.S. presidential campaign. (b) The small tag cloud contains two distinct stories. The first concerns articles about the danger of betel nuts consumption in Taiwan and the second concerns the bankruptcy of the South Korean firm Handjin Shipping.

The log transformation is applied to better distinguish the differences in the lower end of the range. The color intensity is mapped to the number of documents containing the term in the topic, the darker the more documents. Through this visual encoding, the analyst can quickly identify various patterns of interest such as: 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 like “betel” in Figure 4.3b); b) terms with a strong interest but appearing in many documents (large words with a dark color like “trump” in Figure 4.3a). We used the word cloud layout implemented by Davies in D3.js [Davies 2013].

Treemap of word clouds. In order to make sense of dozens of extracted topics, it is tempting to nest the individual word clouds described earlier in an overarching treemap structure. A more meaningful display can yet be obtained if topic similarity and topic relative weight are taken into account. Inspired from the underlying matrix model, we set topic weight to be the size of the corresponding bicluster (the product of the number of terms and the number of documents in our case). In turn, topic similarity can be captured by an MDS projection of the topic similarity matrix computed with our similarity metric defined by equation 3.4 in page 70. The resulting set of 2D coordinates and weights then serve to generate a Weighted Map layout, a variant of treemaps proposed by Ghoniem et al. [Ghoniem 2015] for georeferenced hierarchical data. In the resulting *Topic Weighted Map* visualization in Figure 4.2(1), each rectangle encloses a word cloud, rectangle size conveys topic importance and rectangle proximity conveys topic similarity.

Through this visualization, the analyst can get the broad picture of the topics covered in the corpus, locate topics of interest, and browse the vicinity of a topic to discover similar topics. In addition, topic relationships are also shown on demand by overlaying links [Fekete 2003] to the 5 most similar topics when the mouse hovers over a given topic. Link color encodes the strength of the relationship based on equation 3.4 in page 70. These links engage the analyst more actively into exploring related topics. Because larger topics tend to “attract” smaller ones when using our similarity measure, 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 can click on a topic in order to drill down and explore all document relationships through the *Topic Variant Overview*.

4.3.2 Topic Variant Overview

The *Topic Variant Overview* is designed to explore a large number of *topic variants*, i.e. biclusters uncovered by `Bimax` within a topic of interest. The purpose is to search for multiple sources that relate the same facts or the same stories. Under the *Maximal Inclusion* constraint (Proposition 2), *Bimax* results in all the optimal consolidated co-occurrence patterns (see Figure 1.2) satisfying the exhaustiveness required by journalists. However, 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.

In order to make sense of the large number of *Topic Variants* and their overlaps, we designed an interactive hierarchical visualization (see Figure 4.4). Indeed, the study of Dou et al. [Dou 2013] shows that the hierarchical exploration of topics is more effective than flat exploration based on lists. In addition, the hierarchy of clusters proved to be attractive for journalists in the Overview system [Brehmer 2014]. This is why we also adopt a hierarchical approach for the exploration of *Topic Variants*.

In fact, providing an overview of overlapping biclusters consists in finding a tradeoff between 1) representing repeated terms once, and having to draw explicit links between terms or contours around them, producing clutter (node-link views or parallel coordinates [Santamaría 2008], bipartite graphs [Sun 2015]), which complicates bicluster identification; 2) duplicating redundant terms when representing biclusters, e.g. in matrix views [Heinrich 2011, Streit 2014], which impedes the identification of common and distinctive elements of biclusters. The hierarchical visualization enhanced with the interactions proposed below embodies this trade-off and allows *topic variants* to be identified individually (**T1.2**, **T2.1**) as well as trough their common and distinctive terms (**T2.2**).

The semantic of a *Topic Variant* is mainly interpretable through its terms. So we propose a hierarchy of terms that organizes *Topics Variants* (the biclusters) based on their overlaps (schema (c) in Figure 4.1). The design rational of the term hierarchy and the adequacy of the FPTree algorithm [Han 2004] is already raised in section 3.3.3.2. We only describe the visual design and associated interactions leveraging the tasks defined in section 4.2.1.

We represent the resulting term hierarchy using a *Sunburst* visualization [Stasko 2000] implemented with D3.js [Bostock 2013]. The radial tree representation benefits from increasing space as we move away from the center, coinciding with an increased number of nodes due to branching. In Figure 4.4, each branch represents the complete term sequence of one *Topic Variant*. The hierarchy organizes the multiple term sequences shared by multiple documents and gives access on demand to the relevant documents (*Document Detail View* (5) in Figure 4.8) at different levels of topic granularity (**T2.3**). In the *Document*

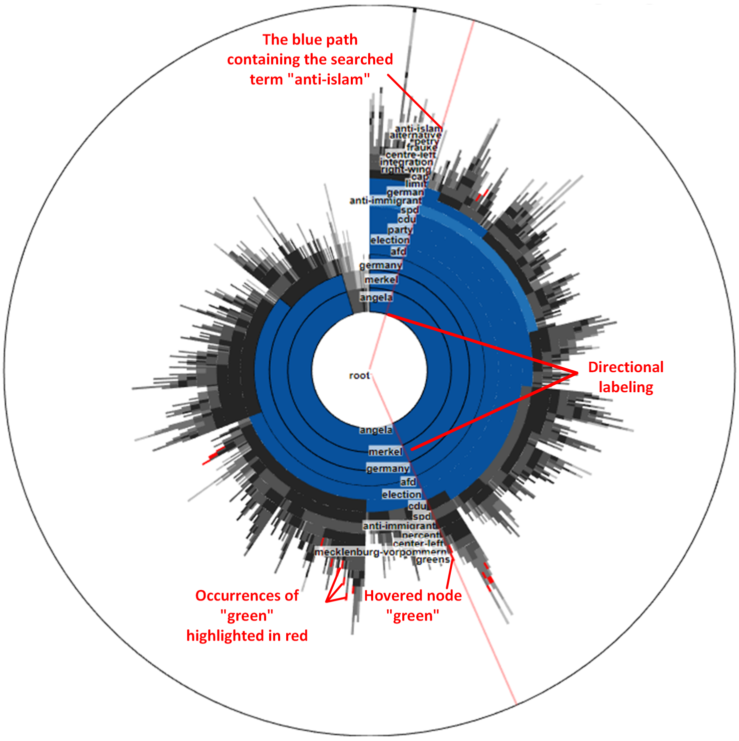


Figure 4.4: The *Topic Variant Overview* resulting from a topic about local elections in Germany in September 2016. Two variants are shown concerning the event.

Detail View the content can be expanded to understand the specific meaning of terms from their context and the terms are highlighted in distinct colors to quickly locate them in the text. A first goal is to find informative terms and associated documents that can confirm or disprove the working hypothesis (**T2**). A second goal is to suggest combinations of terms to help the analyst to find unexpected viewpoints or to express queries (**T3.2**).

As the analyst hovers over a node, the related root-to-leaf path of the *topic variant* is materialized by a directional labeling, an overlay red radial line like in Figure 4.4 starting from the center and extending outwards. Node labels are placed alongside, through different layouts chosen by the user: on the left side, on the right side or alternatively on both sides of the red line (see Figure 4.8(3b)). While the latter layout is less efficient for reading sequences of terms, it optimizes their placement at their node position and reduces the label overlaps. Label orientation is adjusted to always obtain comfortable reading angles as described in Figure 4.5. For easier comparison across *topic variants* (**T3.3**), hovering over a term highlights it in red in all branches where they occur throughout the Sunburst (see Figure 4.4).

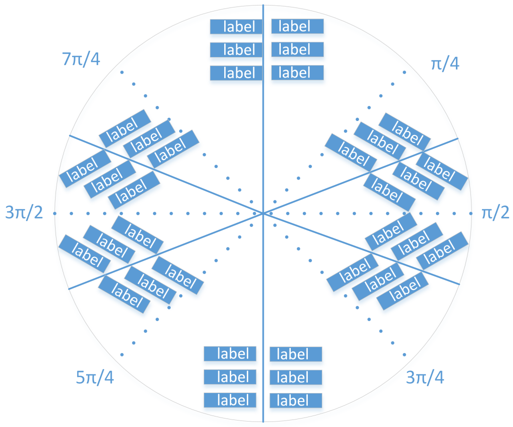


Figure 4.5: Label orientation changes automatically according to the angle of the ray and the label alignment.

We adopted this on demand labeling strategy to avoid the visual clutter that occurs when all the terms in the Sunburst are simultaneously displayed. In the first version we used for the qualitative evaluation with a journalist (see section 4.4.2), we labeled the hovered node only. The sequence from the root until the hovered node was displayed on the right hand of the Sunburst as a breadcrumbs trail. Through this first approach, the journalist felt lost and disorientated in the sunburst. The directional labeling described above addresses this issue and have been developed in a second version of our tool used for the controlled experiment presented in section 4.5. None of the participants to this study was disturbed by the directional labeling. Indeed, it displays better the context of the whole term sequence at the hovered position during the exploration. The analyst can quickly scan through many *topic variants* to spot interesting events or stories (**T1.2**) as well as common and distinctive sub-sequences of terms (**T2.2**). For example, Figure 4.4 shows two scanned *topic variants* both sharing the terms (“angela”, “merkel”, “germany”, “afd”, “election”), which tells that the topic is about elections in Germany. The titles of the documents informs that they concern a specific story: the situation of the political party of Angela Merkel (CDU) in view of the local election held in Germany in September 2016. By analyzing the short differences between the scanned *topic variants*, we found two angles, among others, concerning the event. One discusses the defeat of Merkel’s party in the Mecklenburg-Vorpommern state (Merkel’s district), and the second relates a debate about the responsibility of Merkel’s immigration policy for the defeat.

We propose three interaction modes to further support tasks targeted in section 4.2.1. These modes can be chosen as shown in Figure 4.8(3d).

1. The *Document distribution* mode allows the analyst to highlight in shades of orange all paths containing at least one of the documents associated with the clicked node (see Figure 4.6a). This points the analyst to new *topic variants* sharing the selected documents (**T3.3**). These new *Topics Variants* can bring new documents together with ones already known, revealing new angles, new facts or viewpoints.

2. The *Filtering* mode shows the paths that contain all the terms entered in the search field. The paths from the root to the deepest term of the query are colored in shades of blue in Figure 4.6b. All other nodes remain shaded in gray (**T2.1**). At any time, the *topic variants* that are grayed out can be interactively hidden/unhidden.

3. The *Select for comparison* mode sends the paths of the clicked nodes to the *Topic Variant Comparator* visible in Figure 4.6c. The nodes belonging to these paths/variants appear in shades of blue in the *Topic Variant Comparator* view while other nodes remain in gray shades. The *Topic Variant Comparator* gives the analyst a detailed view of the selected *topic variants* to better identify common and distinctive terms/documents (**T2.2**,**T3.3**).

4.3.3 Human in the loop

As described in section 3.3.3, the number of *Bimax* biclusters increases when the size and the density of the matrix grow, causing drops in execution time. We tackle such issues in the *Topic Variant Overview* through an iterative process where the analyst can modify the parameters of *Bimax* (see (1) Figure 4.8) and adapt the results depending on the shape of each topic. Following the visual analytics principle, we then enable the analyst to steer the `Bimax` algorithm until useful results are found.

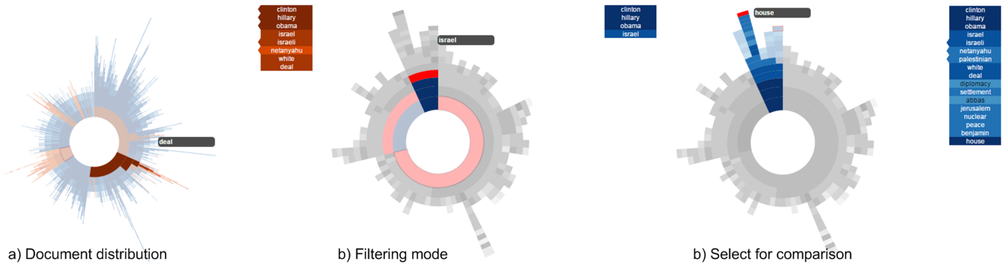


Figure 4.6: Interaction Modes. a) The orange biclusters contain any document selected by the clicked node “Israel”. b) The biclusters not matching the term “Israel” are filtered. c) The bicluster colored in blue is sent to the topic variant comparator.

The parameters that the user can tweak for each topic are recalled below. A configurable binarization threshold can be defined on the TF-IDF matrix to reduce the density of the matrix, and its dimensionality. In addition `Bimax` takes three parameters: the minimum number of rows (*MinRows*) and the minimum number of columns (*MinCols*) per bicluster, and a maximum number of biclusters fixed to stop the algorithm (*MaxBC*). Increasing *MinCols* ignores co-occurrence patterns with too few terms, which could be considered irrelevant. Increasing *MinRows* ignores the term co-occurrences found in too few documents. By default, our system considers biclusters having at least 3 terms co-occurring in at least 2 documents. For all coarse-grained topics, `Bimax` is run with the same default parameter values.

We observed empirically using text datasets that the time performance drops drastically as the number of biclusters reaches 10,000, which we use as the default value of *MaxBC*. This parameter is used to stop `Bimax` when the density or the size of the matrix are too large, hence leading to a huge number of biclusters and significant drops in runtime. In this case, the user cannot by any means exploit the overwhelming results. The density of the topic matrix must then be reduced by raising the binarization threshold to keep only the most meaningful entries given by the TF-IDF weights. For the sake of clarity, we call this threshold “Interest threshold”. The analyst may easily understand that it must be raised to keep the most interesting variants only. Conversely, if the number of *Topic Variants* is low, the threshold must be lowered to include more topic variants. We note the importance of using a binary matrix at this step. The binarization component is greatly flexible and can accommodate any weighting scheme (not only TF-IDF) combined with any query that captures user needs and prior knowledge in order to retain the matrix entries to be processed by `Bimax`.

In Figure 4.7, we visualize the effect of varying each parameter separately on the term hierarchy built from the U.S. presidential elections topic. After each parameter variation, the root node “Obama” is clicked to highlight in orange the distribution of the selected documents. With the default parameters ($MinT = 3$, $MinD = 4$, $\tau_k = 5$), only the first levels of the 13,000 biclusters are visible in the sunburst visualization. Increasing both τ_k and *MinT* reduces the dispersion of the documents concerning "Obama" (see the path in shade of orange in the Figure 4.7), but the changes of τ maintain the variety regarding the number of terms per variant. As *MinD* increases, the number of terms per variant tends to be reduced but the documents selected by the node "Obama" remain largely dispersed in the biclusters until the node disappears.

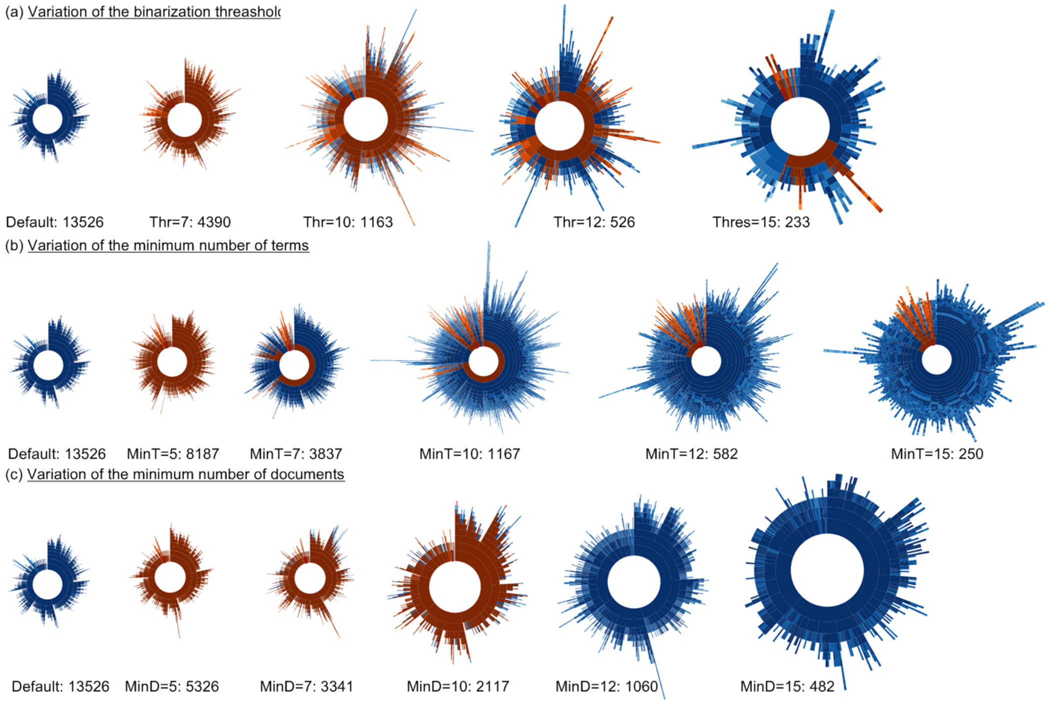
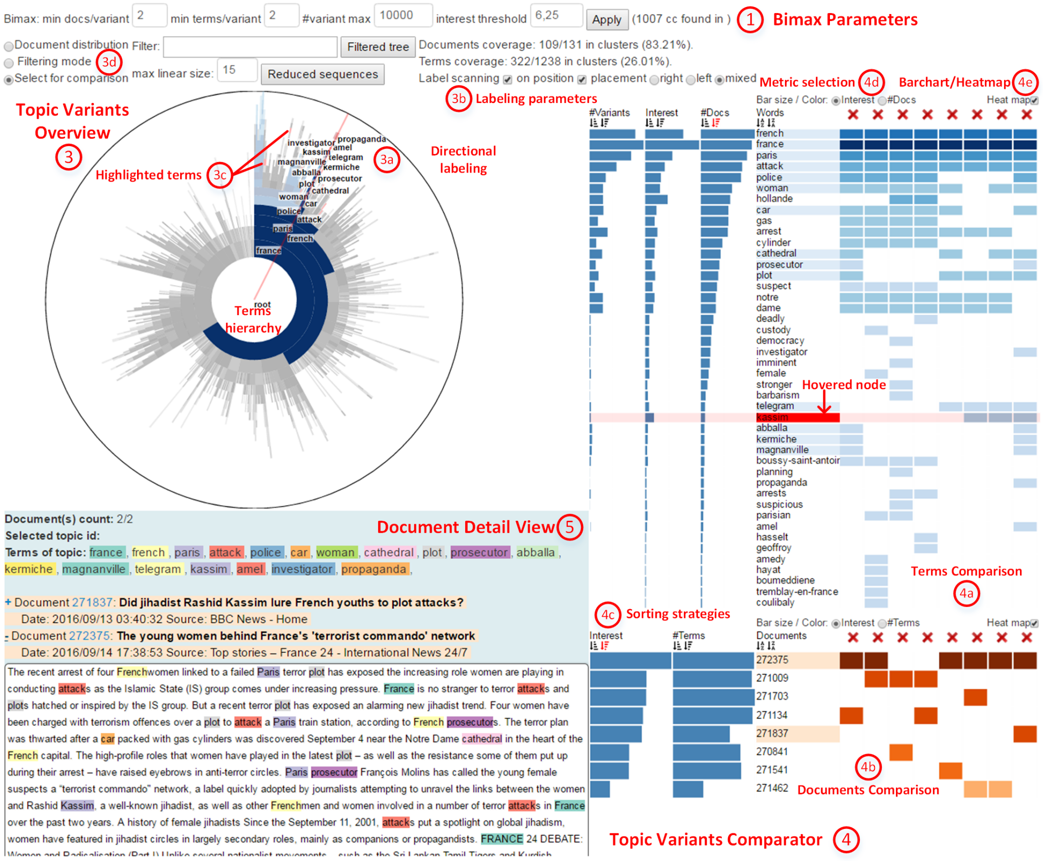


Figure 4.7: Number of biclusters as the parameters of `Bimax` vary.

This set of parameters enables the analyst to drive *Bimax* until interesting insights are found. The nested structure obtained in this fashion allows the analyst to explore all document relationships within topics, which could potentially reveal interesting facts or stories. The next section describes the visualization designed to compare selected topic variants.

4.3.4 Topic Variant Comparator

The *Topic Variant Comparator* (Figure 4.8 (4)) is a workspace where the analyst can store and remove *Topic Variants* of his choice. It provides multiple interactions enabling the user to find meaningful terms documents. Both the *Topic Weighted Map* and the *Topic Variant Overview* reveal informative terms at their respective level of detail. The first reveals the most interesting ones at the granularity of coarse-grained topics. The second proposes a hierarchy of terms that drives the user toward *Topic Variants* of interest. Discovering thereby sets of terms shared by multiple documents is a first means to identify meaningful terms related to stories or facts. However, the hierarchy shown in the *Topic Variant Overview* organizes the terms of *topic variants* based on the prefix commonality of term sequences in the related `Bimax` biclusters, the order of terms in the sequences being determined by the overlap degree of terms in the biclusters of the topic. The *Topic Variant Comparator* completes this view by supplying multiple sorting strategies offering multiple perspectives. The analyst can hence better understand the *Topic Variants* and identify the most informative terms as shown in Figure 4.8.



Bimax: min docs/variant 2 min terms/variant 2 #variant max 10000 interest threshold 6,25 Apply (1007 cc found in)

○ Document distribution Filter: Filtered tree
○ Filtering mode max linear size: 15 Reduced sequences
● Select for comparison

Documents coverage: 109/131 in clusters (83.21%).
Terms coverage: 322/1238 in clusters (26.01%).
Label scanning ☑ on position ☑ placement ○ right ○ left ● mixed

Bar size / Color: ● Interest ○ #Docs Heat map ☑

Bar size / Color: ● Interest ○ #Terms Heat map ☑

Document(s) count: 2/2
Selected topic id:
Terms of topic: france , french , paris , attack , police , car , woman , cathedral , plot , prosecutor , abballa , kermiche , magnanville , telegram , kassim , amel , investigator , propaganda ,

+ Document 271837: **Did jihadist Rashid Kassim lure French youths to plot attacks?**
Date: 2016/09/13 03:40:32 Source: BBC News - Home
- Document 272375: **The young women behind France's 'terrorist commando' network**
Date: 2016/09/14 17:38:53 Source: Top stories – France 24 - International News 24/7

The recent arrest of four Frenchwomen linked to a failed Paris terror plot has exposed the increasing role women are playing in conducting attacks as the Islamic State (IS) group comes under increasing pressure. France is no stranger to terror attacks and plots hatched or inspired by the IS group. But a recent terror plot has exposed an alarming new jihadist trend. Four women have been charged with terrorism offences over a plot to attack a Paris train station, according to French prosecutors. The terror plan was thwarted after a car packed with gas cylinders was discovered September 4 near the Notre Dame cathedral in the heart of the French capital. The high-profile roles that women have played in the latest plot – as well as the resistance some of them put up during their arrest – have raised eyebrows in anti-terror circles. Paris prosecutor François Molins has called the young female suspects a "terrorist commando" network, a label quickly adopted by journalists attempting to unravel the links between the women and Rashid Kassim, a well-known jihadist, as well as other Frenchmen and women involved in a number of terror attacks in France over the past two years. A history of female jihadists Since the September 11, 2001, attacks put a spotlight on global jihadism, women have featured in jihadist circles in largely secondary roles, mainly as companions or propagandists. FRANCE 24 DEBATE: Women and Radicalisation (Part I) Unlike several nationalist movements – such as the Sri Lankan Tamil Tigers and Kurdish

Figure 4.8: The 2^{nd} version of our tool. The selected topic concerning an event in France : four women have been arrested after they fail a terrorist attack in Paris near Notre Dame Cathedral. Multiple *topic variants* have been sent to *Topic Variant Comparator*. In the matrix view, the terms are sorted by the number of documents in which they appear and the color of cells are mapped to the TF-IDF metric. Thereby, we hence identify the terms that are representative to few documents. We found "Kassim" (hovered term colored in red) being the name of the instigator of multiple attacks in France. He communicated with the women through "telegram" (the term just before "Kassim"), a messaging application.

In Figure 4.8 (4), the *Topic Variant Comparator* displays *Topic Variants* as columns in a matrix visualization. The matrix is split horizontally in two parts. In the upper part (4a), rows correspond to the terms occurring in the *Topic Variants*; the cells are colored in shades of blue. In the lower part (4b), rows correspond to documents containing the*Topic Variants*; the cells are colored in shades of orange. Obviously, 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) arranges the terms in the same order as in the branches of the hierarchy,
2. the degree of interest of terms is based on the *TF-IDF* weights,
3. the last metric is the number of documents where the term occurs within the topic.

More details of the last two metrics are given in section 4.3.1. These metrics, shown on the left hand side as barcharts (4c), enable multiple term sorting strategies. We also provide an alphabetical sorting for fast term lookups. In addition, the user can choose through a checkbox either of these metrics to display in the matrix cells (4d), either in a barchart mode or in heatmap mode (4e). The different combinations of sorting strategies and displayed metrics provide multiple perspectives on the data. This allows the analyst to focus the investigation on the most informative terms. For instance, the journalist can sort the terms by interest and display the number of documents to identify meaningful terms occurring in few documents within the topic. In Figure 4.8, the term “kassim” has been found by this mean. When she sorts the terms by the bicluster overlap degree, and shows the degree of interest metric, the journalist can identify meaningful terms that are specific to few *Topic Variants*.

The hover interaction in the *Topic Variant Comparator* is synchronized with the *Topic Variant Overview*. The hovered terms are colored in red in both and their path is highlighted from the root. The hovering is also synchronized in the list of terms in the *Topic Variant Comparator*. The hovered terms are colored in red and their path in blue. By clicking on a matrix cell, the selected documents are listed in the *Document View* (5).

The bottom part of the matrix (4b) shows the document distribution. Only two metrics are proposed: an interest measure based on the *TF-IDF* aggregated row-wise and the number of terms in the topic variant for each document. In this part of the matrix, the interactions take place at the column/variant level. The hover interaction highlights the complete path in the *Topic Variant Overview* as well as in the *Topic Variant Comparator*. By clicking on any (orange) cell of a column, all documents of the corresponding *Topic Variant* are listed in the *Document View*. The labels of the selected documents are colored in orange in the list.

These interactions are designed to help the user compare the *Topic Variants* (**T2.2**) and identify terms and documents being shared or specific (**T3.2, T3.3**). We can also support the diversification process. For instance the journalist can hover over an interesting term for further investigation, and identify new variants to add in the workspace (**T3.3**). For instance, in Figure 4.8, “kassim” has been hovered (red term in the *Topic Variant Comparator*). The highlighted terms in the Sunburst (3c) allows to locate new *topic variants* potentially interesting to add in the comparator.

4.4 Evaluation

The first version of our tool (Figure 4.2) has been evaluated in two ways. Firstly we present in the next section a usage scenario that explores a corpus of 3,992 news articles. Secondly we conducted a qualitative evaluation with an analytic journalist. We describe our method and our results in the section 4.4.2.

4.4.1 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. A video related to this scenario is provided with comments in french (`https://youtu.be/rj9YrTMPClQ`). The data set is composed of 3,992 news articles aggregated from multiple online news sources (BBC, CNN, Reuters, France24, Egypt Independent and Der Spiegel) between the 2^{nd} of November 2015 and the 16^{th} of November 2015. We extracted the topics with `Coclus` a diagonal biclustering algorithm [Ailem 2016]. We performed multiple tests and found the best modularity value to be 50 biclusters or topics. In Figure 4.9, the *Topic Weighted Map* shows these 50 topics. Despite the variety of topic sizes and the broad range of topics, we observe that similar topics are located close together. For example, there are two football related topics close to one another, and two astronomy topics at the top of the figure. On the right side, the large topics depict the most treated events in the news for that period (the crash of a Russian airplane in Egypt, the American presidential elections, the war in the Middle East, the immigrant crisis, and the terrorist attack in Paris). The topic related to the American elections attracts 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”. Recall that the lighter terms are found in less documents in this topic than the darker ones.

By clicking on the topic, the *Topic Variant Overview* displays a sunburst (Figure 4.2) containing 1,163 *Topics Variants* (leaf nodes). By hovering over the nodes at the center of the hierarchy we discover the most shared terms among the *Topic Variants*: “republican”, “clinton”, “rubio”, “israel”, “trump”, “debate”, “candidate”. The terms are all clearly related to the US presidential campaign except “israel”. We hypothesize that Israel is a

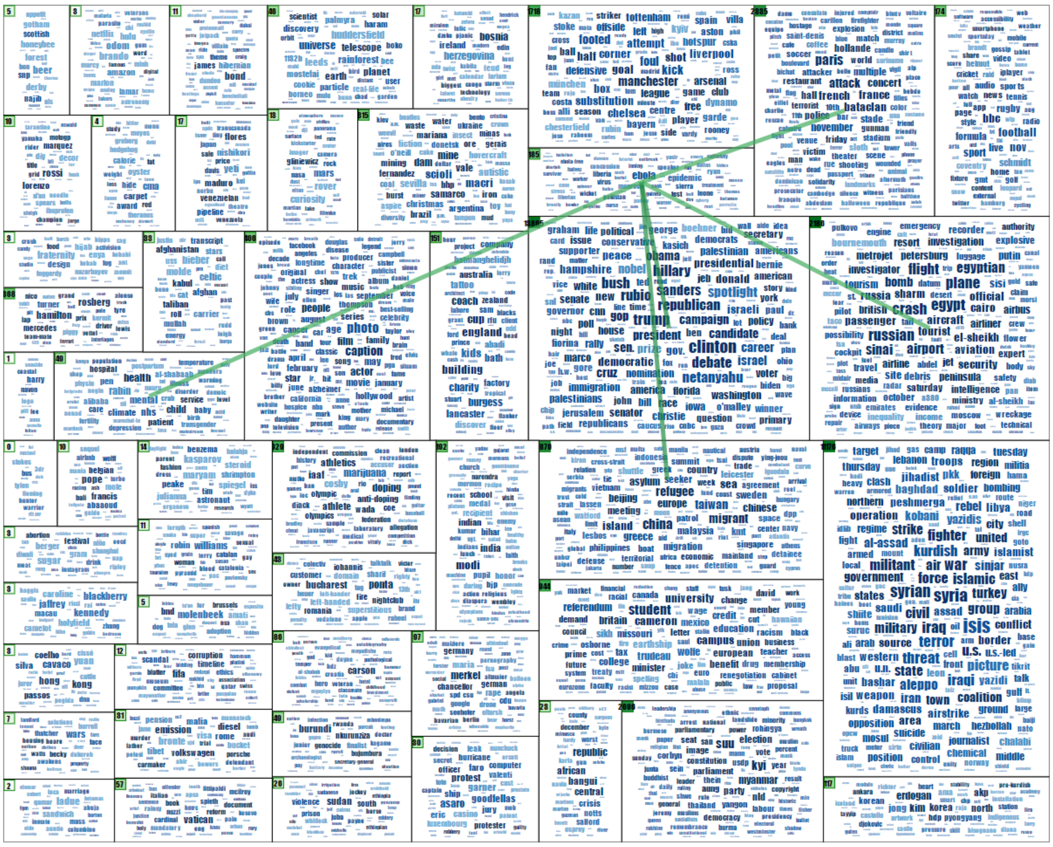


Figure 4.9: The *Weighted Topic Map* shows 50 topics extracted from 3,992 news articles aggregated from multiple online news sources between the 2^{nd} November 2015 and the 16^{th} November 2015. One of the links highlighted from the topic about Ebola targets another topic about health being far away.

topic debated between candidates. We then explore the relationships between documents to explain why Israel may be related the US election topic. We first use the *Document distribution* mode to know if the documents linked to the *Topic Variants* matching the keyword "israel" also appear in the *Topic Variants* related to the US elections. By clicking on the central node "israel", certain *Topic Variants* (branches) are selected in orange. Some of them have a common path "clinton", "hillary", "obama", "israel", "israeli", "netanyahu". We can hence quickly distinguish the *Topic Variants* specific to Israel from those linking the US elections to Israel.

Next we use the *Filtering* mode and hide *Topic Variants* of the elections that don't contain "israel". The result is shown in Figure 4.10(a). Then we decide to compare all *Topic Variants* containing "clinton". We switch to the mode *Selection for comparison* and add all the variants with "clinton" in the *Topic Variant Comparator* (Figure 4.10(b)). We chose to sort the terms by the number of enclosing variants and to display the degree of interest metric in the matrix cells. Some patterns emerge from the matrix. First of all, the terms shared by all variants helps focus on the topic: “clinton”, “hillary”, “obama”, “israel”, “israeli”, “netanyahu”, “settlement” and “abbas”. In the middle, one *Topic Variant* groups all these shared terms and all documents. By quickly reading the titles of the documents, we identify an event : “Netanyahu meets Obama at the White House”.

In Figure 4.10(b), the two *Topic Variants* on the left have specific terms: “palestinian”, “diplomacy” and “peace”. One of them contains the term “nuclear”. These variants group three documents evoking difficult diplomacy between Obama and Netanyahu. The term “nuclear” refers to their disagreement about the nuclear deal with Iran. On the right, two variants have other specific terms: “president”, “secretary” and “jewish”. These two variants bring in a new document 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 for the benefit of 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”.

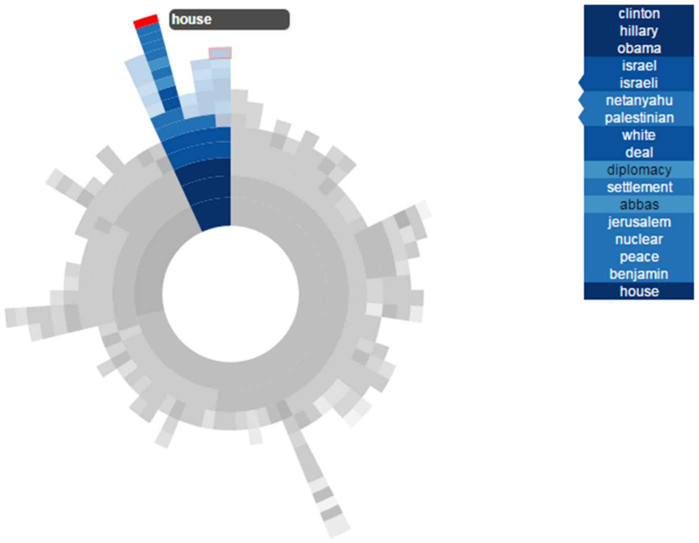
This usage scenario shows the ability of our tool to drill down into a topic, generate and refine hypothesis, identify document relationships that reveal facts and stories while distinguishing multiple angles or viewpoints.

4.4.2 Qualitative evaluation with a domain expert

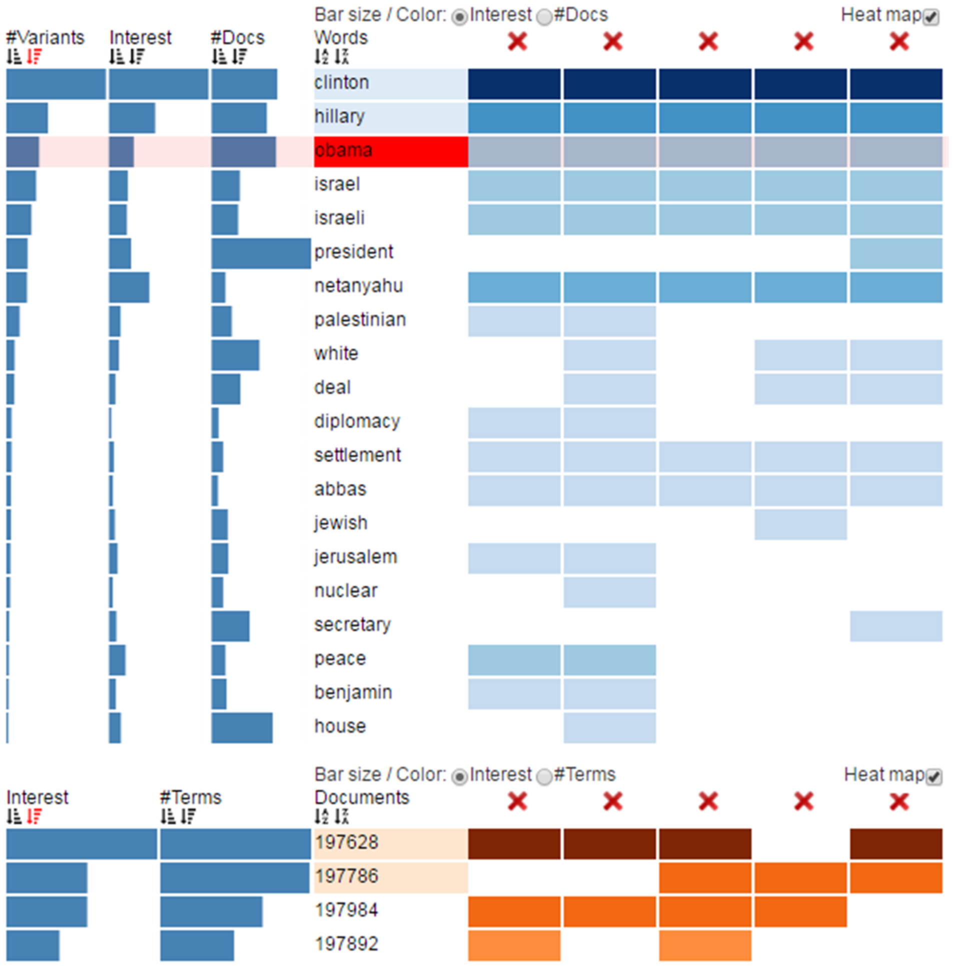
Our Visual Analytics tool has been designed in collaboration with Warda Mohamed, a professional analytic journalist and editor at Orient XXI. She also writes for a number of French media, including Le Monde diplomatique and Mediapart. As part of her daily work, she needs to verify, confirm and refine hypotheses by confronting them to available evidence in document collections.

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 as well as collect her feedback about our system.

In the third meeting we carried out a qualitative evaluation split in two parts. The evaluation lasted around 3 hours. We made a voice recording of the entire meeting to thoroughly analyze expert feedback. In the first part, we demonstrated the usage of the tool on a small set of 9 documents previously supplied to us by the expert, corresponding to the handpicked material she used to prepare a previously published paper. The aim was to analyze the corpus on our own and confront our findings with hers without prior



(a) The *Topic Variant Overview* where *Topic Variants* of the elections that don't contain "israel" are hidden. The *Topic Variants* in shade of blue are sent to the *Topic Variant Comparator* in Figure 4.10(b)


#Variants
Interest
#Docs
Bar size / Color: ● Interest ○ #Docs
Heat map ☑

Words	✖	✖	✖	✖	✖
clinton	■	■	■	■	■
hillary	■	■	■	■	■
obama	■	■	■	■	■
israel	■	■	■	■	■
israeli	■	■	■	■	■
president					■
netanyahu	■	■	■	■	■
palestinian	■	■			
white		■		■	■
deal		■		■	■
diplomacy	■	■			
settlement	■	■	■	■	■
abbas	■	■	■	■	■
jewish				■	
jerusalem	■	■			
nuclear		■			
secretary					■
peace	■	■			
benjamin	■	■			
house		■			

Interest
#Terms
Bar size / Color: ● Interest ○ #Terms
Heat map ☑

Documents	✖	✖	✖	✖	✖
197628	■	■	■		■
197786			■	■	■
197984	■	■	■	■	
197892	■		■		

(b) Five *topic variants* selected in the *Topic Variant Comparator* to compare different aspects of the event. We chose those of the subtree containing "Clinton"

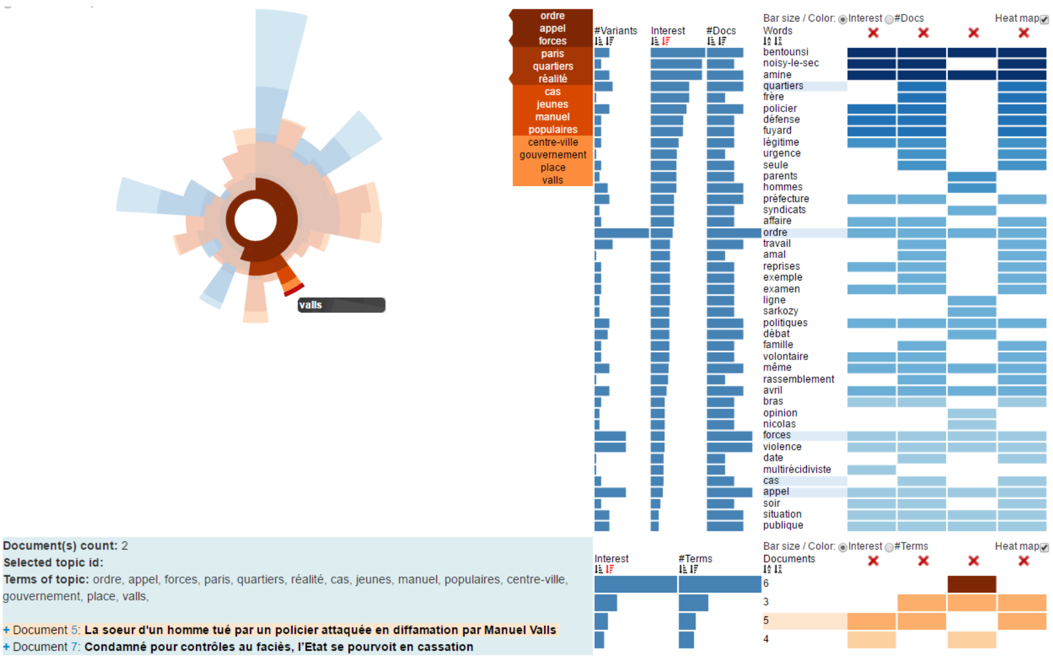
Figure 4.10: The Sunburst and the *Topic Variant Comparator* used for the usage scenario.

knowledge of her conclusions. She was investigating the blunders made by the French police. Her starting hypothesis was three-fold: (1) The main victims of police blunders are of African or Arab descent. (2) They are in majority innocent of any offense, but are killed. (3) These killings remain unpunished and have been covered by the French State for decades. We investigated the dataset with our tool to find facts that verify, refute or refine the three aspects of the hypothesis. The comparison of our findings to hers on this small corpus aimed to validate the consistency of our findings with a form of ground truth that was known to the expert and build trust in the tool. In the same time, using a familiar dataset allowed her to appreciate the functionality offered by the tool as we operated it.

During 30 minutes we gave a live demo of the system while explaining what the tool reveals from her small collection. During this phase, she could ask questions to understand how the visualizations and the interactions work. She also could 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* and the *Topic Variant Comparator* in Figure 4.11. After a quick glance, the expert validated the relevance of the terms shown in the *Topic Variants* view. She said: "*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, she manipulated the visualizations with a larger corpus, i.e. the one used for our usage scenario in section 4.4.1. This part lasted two hours and a half. She manipulated the tool in order to answer our questions covering all the tasks listed in Table 3.2. 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 in the *Weighted Topic Map* shown in Figure 4.9 two topics respectively about: football, astronomy, health/medicine and Asia'. For three of these themes she found the two topics in a few seconds. 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 (see Figure 4.12). For each of these four themes, we asked her to explain the difference between the two related topics she spotted, which she did. Globally she enjoyed the *Topic Weighted Map* view: *"I like this tool, because 3,000 documents it's very large 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



Document(s) count: 2
Selected topic id:
Terms of topic: ordre, appel, forces, paris, quartiers, réalité, cas, jeunes, manuel, populaires, centre-ville, gouvernement, place, valls,

+ Document 5: **La soeur d'un homme tué par un policier attaquée en diffamation par Manuel Valls**
+ Document 7: **Condamné pour contrôles au faciès, l'Etat se pourvoit en cassation**

Figure 4.11: The *Topic Variants Overview* shows lists of co-occurring terms that constitutes relationships between 9 documents used by the investigative journalist to write an article. This small set of documents served us as a ground truth to validate with the expert the effectiveness of the tool and to build trust in the system. In the *Topic Variant Comparator*, the terms are sorted by degree of interest. In the Top-10 terms in the list we identify "bentounsi","fuyard", "légitime", "défense". The journalist validated that even if she didn't know the subject matter, she could investigate closely to the top-10 terms for her enquiry.

largest topics were stacked on the right hand side.

Next she wondered why the refugees and China were grouped in one topic. As this was the purpose of our next question, we proposed her to select this topic and explore it in the *Topic Variant Overview*. She first hovered over the *Topic Variants*. We had one question at this point: "Are the terms in the center of the visualization consistent with the general meaning of the upper level topic captured from the Topic Weighted Map ?" After scanning different *Topic Variants* she agreed. But she also explored the terms deeper in each branch revealing various European countries, but also Eritrea as shown in Figure 4.13. 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 both the countries and the debated questions such as the Shengen Agreement, and asylum claims*". Next, she commented that she may have trouble using this visualization if she were on her own, even though she understood its benefits. One reason might be the on-demand labeling strategy we adopted at the time of the qualitative evaluation, which labeled a single word at a time. Even if the complete sequence of terms was displayed on

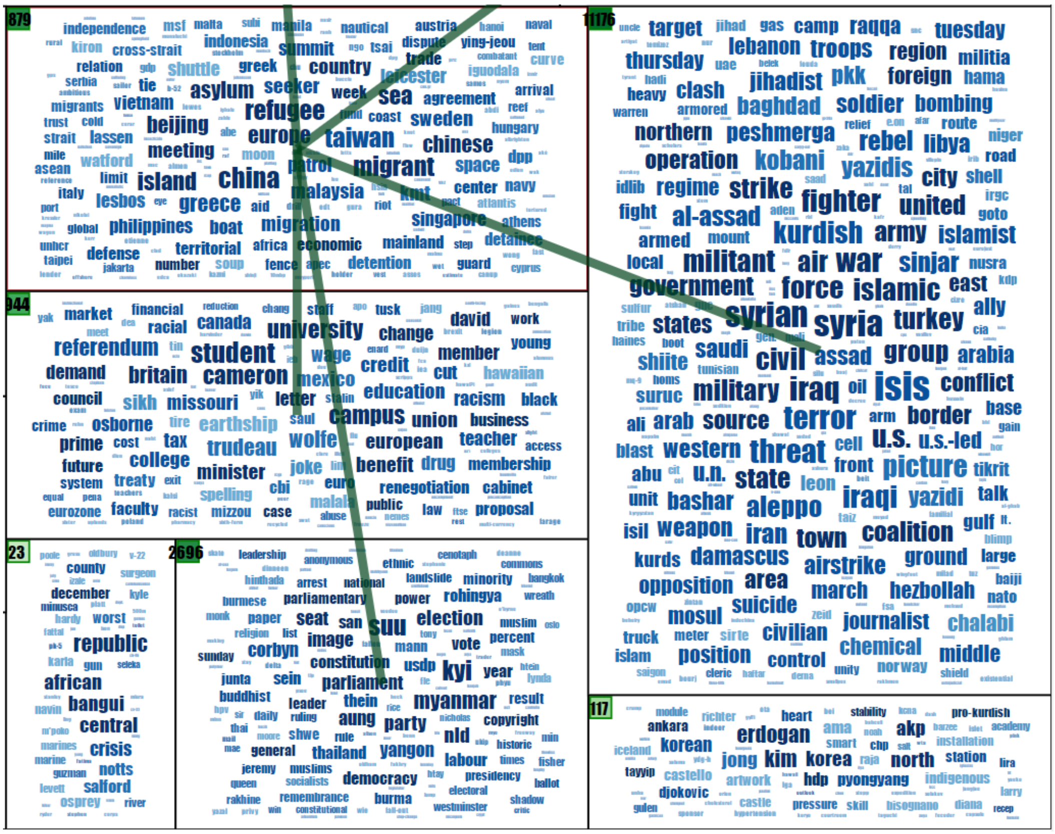
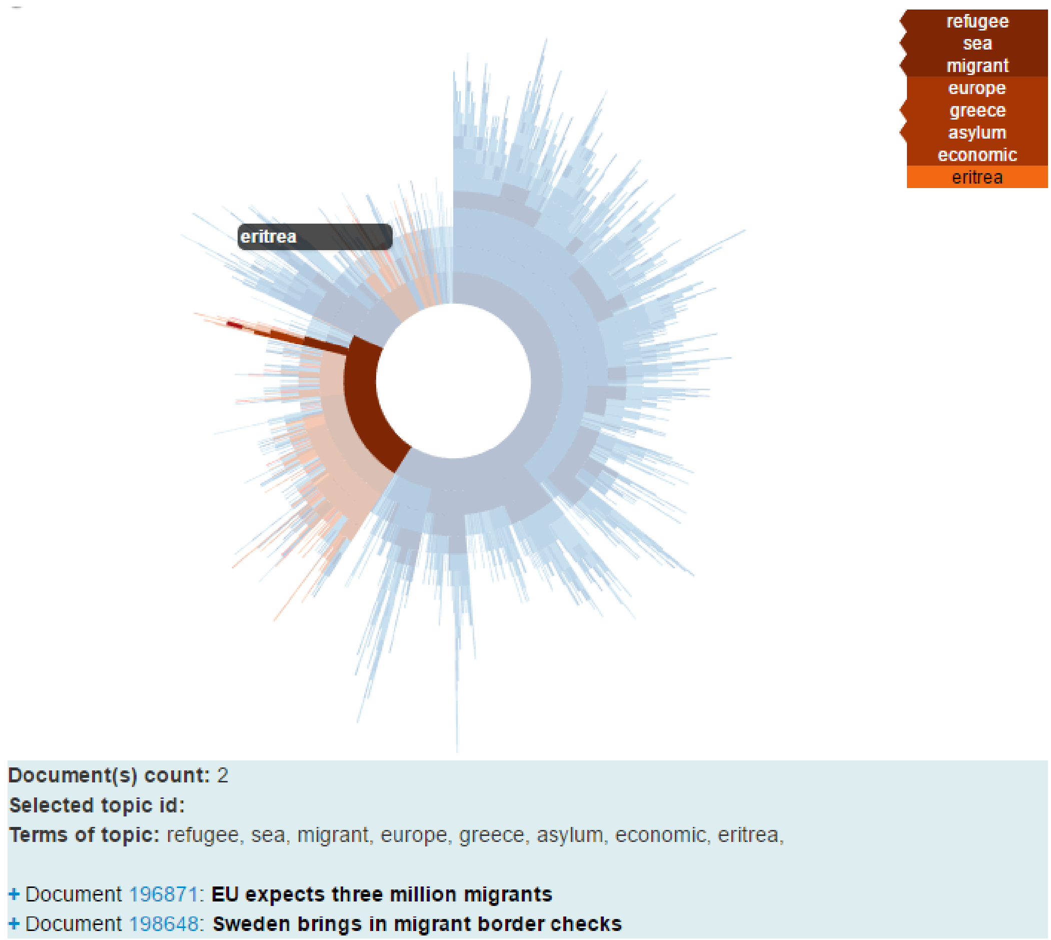

879
independence msf malta subi manila nautical austria hanoi naval
kiron cross-strait indonesia summit tsai dispute ying-jeou tent
relation shuttle greek country trade leicester iguodala curve
serbia tie asylum seeker week sea agreement arrival
migrants vietnam refugee coast sweden reef
trust cold beijing europe taiwan chinese hungary
strait lassen meeting moon patrol migrant space dpp
mile asean watford limit island china malaysia center navy
italy lesbos greece aid riot singapore atlantis
port global philippines boat migration athens
unhcr defense territorial africa economic mainland step detainee
taipei jakarta number soup fence detention guard cyprus
1176
target jihad gas camp raqqa tuesday
thursday lebanon troops region militia
hadi clash jihadist pkk foreign hama
heavy armored baghdad soldier bombing
warren northern peshmerga relief route niger
operation kobani rebel libya road
idlib regime strike yazidis city shell
fight al-assad aden fighter united irgc goto
armed mount kurdish army islamist
local militant air war sinjar nusra
government force islamic east kdp
sulfur states syrian syria turkey ally
tribe haines shiite saudi civil assad group arabia cia
suruc military iraq oil isis conflict
ali arab source terror arm border base gain
blast western threat cell u.s. u.s.-led
abu u.n. state leon front picture tikrit
unit bashar aleppo iraqi yazidi talk
isil weapon iran town coalition gulf
kurds damascus airstrike ground large
opposition area march hezbollah baiji
opcw mosul suicide journalist nato
truck meter sirte civilian zeid chalabi
islam position control chemical middle
saigon cleric unity norway shield
944
market financial chang staff tusk jang
yak racial canada university change david work
referendum student wage member young
demand britain cameron mexico credit cut hawaiian
council sikh missouri letter education racism black
crime osborne earthship saul campus union business
prime cost tax wolfe european teacher access
future college trudeau joke benefit drug membership
system treaty exit minister euro renegotiation cabinet
spelling cbi public
eurozone faculty racist mizzou case abuse law proposal
23
county surgeon
december kyle
worst
republic
gun
african
bangui
central
crisis
notts
salford
osprey river
2696
leadership anonymous ethnic cenotaph commons
arrest national landslide minority bangkok
burmese parliamentary power rohingya
paper seat san suu election muslim
religion image vote percent
corbyn mann
sunday constitution usdp kyi year
junta sein parliament
buddhist leader thein myanmar result
daily aung party copyright
thai shwe rule nld historic
general thailand yangon labour
jeremy muslims democracy presidency
queen socialists electoral ballot
rakhine remembrance burma shadow
constitutional westminster
117
module richter heart stability pro-kurdish
ankara erdogan ama akp academy
iceland korean smart chp installation
tayyip jong kim korea north station lira
castello artwork hdp pyongyang indigenous
djokovic larry
gulen castle pressure skill bisognano diana recep

Figure 4.12: Two topics concerning Asia are close together. The topic about China includes however news about refugee crisis.



Document(s) count: 2
Selected topic id:
Terms of topic: refugee, sea, migrant, europe, greece, asylum, economic, eritrea,

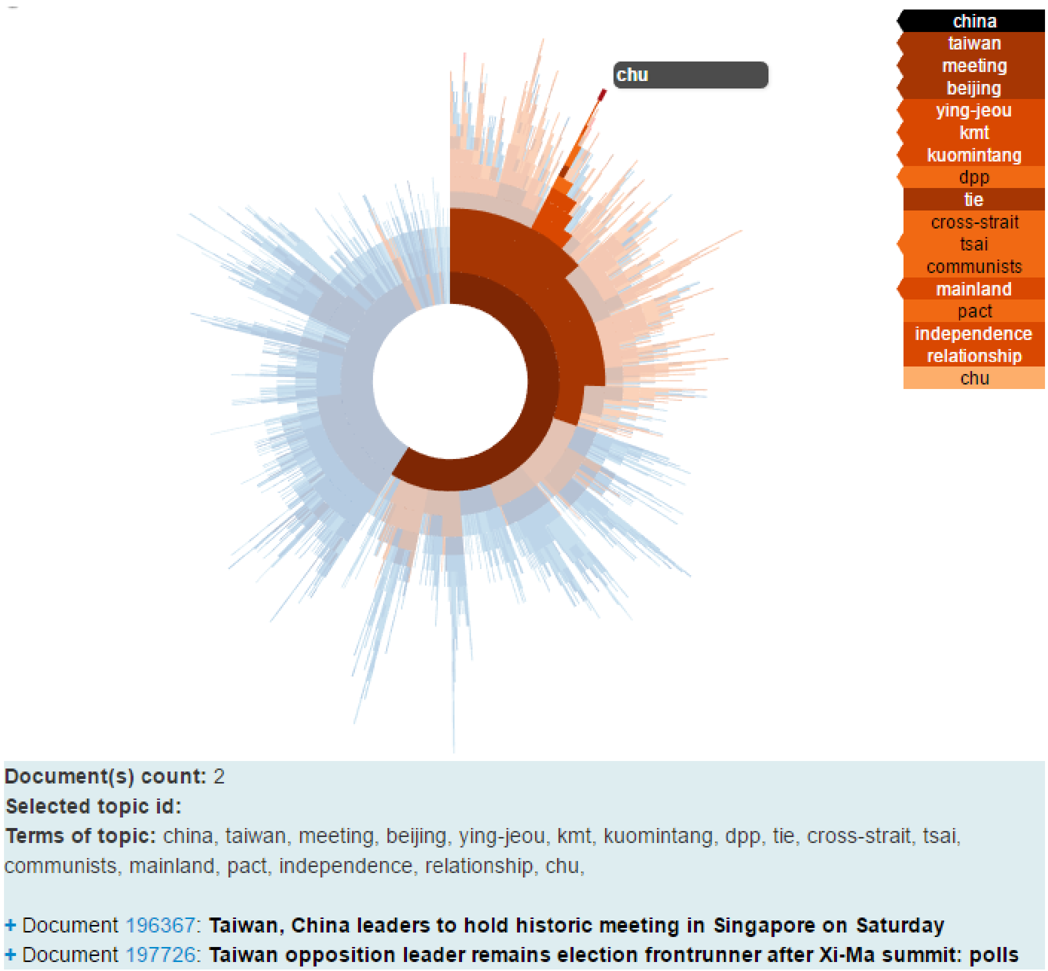
+ Document 196871: **EU expects three million migrants**
+ Document 198648: **Sweden brings in migrant border checks**

Figure 4.13: The large orange node on the left of the first ring corresponds to the term "Refugee" and groups many related topic variants. The highlighted one mentions "eritrea", and the complete term sequence is shown on the right hand.

the right hand as a breadcrumbs trail, the user's focus is often near the mouse pointer. This complicated the whole interpretation of term sequences. In the second version of our tool, we have developed the directional labeling shown in Figure 4.4 that better suits exploration of multiple term sequences.

We then proposed to use the *Document distribution* mode to explain the relationship between China and the European refugee crisis. She clicked on multiple central nodes. Clicking alternatively on the roots "China" then "refugee", distinct parts of the hierarchy turn orange (see Figure 4.13 and 4.14 for comparison). However, clicking on the term "island" turns orange a large part of the hierarchy, covering subset of both "China" and "refugee" branches as depicted in Figure 4.15. 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 brought together 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 *Topic Variants* of interest and add them in the *Topic Variant Comparator* to identify the most meaningful ones. We noticed that she followed a repetitive analytic scheme. After manipulating the sorting strategies of the *Topic Variant Comparator*, she checked the meaning of the 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*." In the first version of the *Document View*, we observed that when all the terms are highlighted in the same color this hinders locating terms of interest in the document body.



Document(s) count: 2
Selected topic id:
Terms of topic: china, taiwan, meeting, beijing, ying-jeou, kmt, kuomintang, dpp, tie, cross-strait, tsai, communists, mainland, pact, independence, relationship, chu,

+ Document 196367: **Taiwan, China leaders to hold historic meeting in Singapore on Saturday**
+ Document 197726: **Taiwan opposition leader remains election frontrunner after Xi-Ma summit: polls**

Figure 4.14: The large orange node on the right of the first ring corresponds to the term "China". The highlighted topic variant concerns a "meeting" between "China" and "Taiwan".

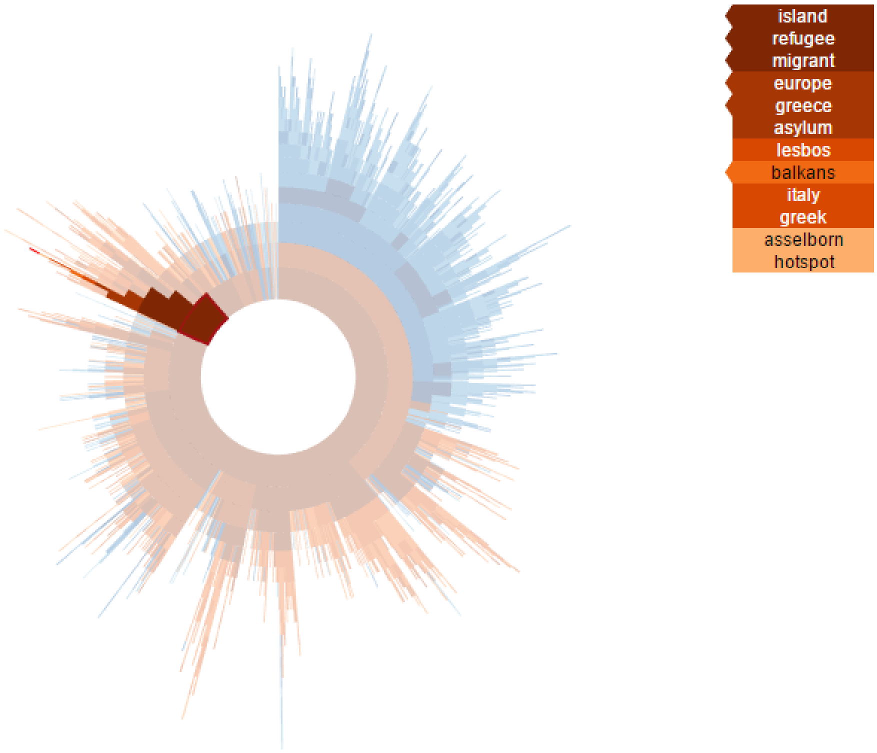


Figure 4.15: The small orange node highlighted on the left of the first ring corresponds to the term “Island”. This term brings documents that also appear in the paths colored in oranges. We can see that the documents containing “Island” cover the branch starting from “China” and the branch starting from “Refugee”. Therefore, “Island” contributes to the grouping of a mixture of documents about both themes in the same topic.

In the second version proposed in Figure 4.8, the terms are highlighted with the 12 categorical color scheme of Brewer [Brewer 2017]. When the *topic variant* contains more than 12 terms the color scheme is repeated consecutively in the sequence. In general, a maximum of two distinct terms are highlighted with the same color, which does not impede the search.

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 a 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 to analyze the corpus according to different angles and reopen the previous ones. Next, the exclusive assignment of terms to topics (due to hard partitioning) is troublesome. If certain terms in the documents are ignored because they are already used in another topic, someone who is unfamiliar with the subject could miss important aspects. We will discuss this issue and the solution we envision in the next section.

4.4.3 Discussion and future work

The usage scenario presented above shows that our visual analytics tool allows to explore the multiplicity of angles and view points shared by documents, and give evidence of the feasibility of the tasks **T1** to **T3** described in section 4.2.1. The participation of the expert at the begining of the design was a first guarantee that the problem and the tasks were well characterized [Munzner 2009]. If the qualitative evaluation gave us an appreciation of the usefulness of our tool, it also allowed us to identify some limitations we addressed in a second version.

However, the qualitative evaluation relying only on one expert and a semi-directed exploration remains preliminary. Indeed, the expert is 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 are facing the issue that a majority of journalists limit their corpora to a small number of documents. Without large corpora holding a known ground truth, we have to observe the long-term adoption of the tool by several journalists through their new enquiries, while avoiding the bias of the semi-directed approach. This is one of our concerns for future work. We also aim to conduct a controlled experiment to compare the effectiveness of the sunburst compared to other implicit hierarchy visualizations [Schulz 2011] or explicit hierarchies such as word tree [Wattenberg 2008], in order to leverage the tasks listed in section 4.2.1. Our qualitative evaluation showed us that the Sunburst view might be difficult to use by journalists due to labeling issues. Our own experience with the directional labeling developed in the second version of our tool is promising. We aim in future work to compare the effectiveness of this approach with other labeling approaches such as excentric labeling [Fekete 1999].

We have seen earlier in this chapter that the hard partitioning of data is problematic since both terms and documents can logically belong to multiple topics. We have alleviated this problem in the *Topic Weighted Map* view by computing and overlaying topic relationships in support of the diversification process in analytical journalism. In future work, we would like to further exploit information about topic overlap as captured by the confusion blocks described in section 3.3.2. Typically, we will exploit them to help the user import terms and/or documents from related topics into the topic of interest. This could result in a partial merge or a complete merge between two or multiple topics, which allows the analyst to shape new topics depending on her needs. This requires the integration of user-defined grouping constraints into the biclustering algorithm while rebuilding the topics for the rest of the matrix based on modularity.

Another limitation of most biclustering algorithms is that they require the number of biclusters to be set in advance. However, we observed that certain topics actually contain multiple themes. This could be resolved by increasing the number of biclusters. However, the diagonal biclustering algorithm starts with a random initialization that may result in a completely different partition. To preserve the mental map of the analyst, an incremental approach must be adopted for both the biclustering algorithms and the visualizations. This incremental approach is likely to be suitable for the analysis of text streams, which we aim to handle in future work. An incremental version of Bimax is already available and we propose a dynamic version in chapter 6. More work is required to make modularity-based diagonal biclustering dynamic too. Moreover, all the visualizations presented in this work must be adapted to gracefully handle incremental analyses.

4.5 User study

In order to deal with large text corpora, the visual analytics tool proposed in the current chapter allows top-down navigation following the visual information seeking mantra "Overview first, zoom and filter, details on demand" [Shneiderman 1996], and avoids the painstaking tasks of reading numerous documents to search for facts, and the related viewpoints that validate or disprove working hypotheses. The usefulness of this approach for investigative journalism has been showed through a qualitative evaluation in the previous section. However, previous work does not provide clear recommendations regarding the choice of topic extraction methods to leverage analyst's tasks. Indeed, the standard clustering evaluation approaches consist in assessing the quality of a document partition against labeled benchmarks and global metrics such as Accuracy, Normalized Mutual Information (NMI) [Strehl 2002], and Adjusted Rand Index (ARI) [Rand 1971]. While these metrics tell how close the partition is to a known ground truth, they give little information about the underlying topical structure and they do not focus on topic comprehension by the analyst.

In the present work we would like to understand how the topical structures extracted by different methods may impact topic comprehension through tasks we characterized with journalists. To answer this question from an analyst perspective, we used the multi-model capability of our system to conduct a controlled experiment in which subjects have to interpret topics extracted from a real news corpus. In this experiment, we compare `Coclus`, a disjoint biclustering method, to `hLDA`, a probabilistic topic model that yields overlapping topics. We start in the next section by eliciting the properties of both methods.

4.5.1 Properties of topic extraction methods

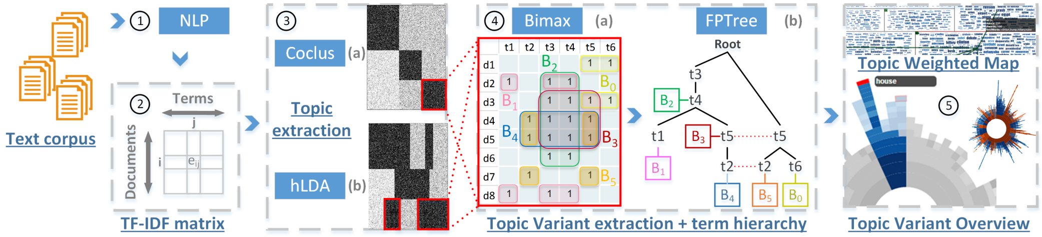


Figure 4.16: Our data processing pipeline: From left to right, NLP technology is used to select relevant text tokens. A term-document matrix is built. Various topic extraction methods can partition the TF-IDF matrix in topic blocks (in black) with various shapes. The blocks are analyzed by Bimax to reveal topic variants and FPTree hierarchizes their term sets. Finally, interactive visualizations convey the topical structure to the analyst.

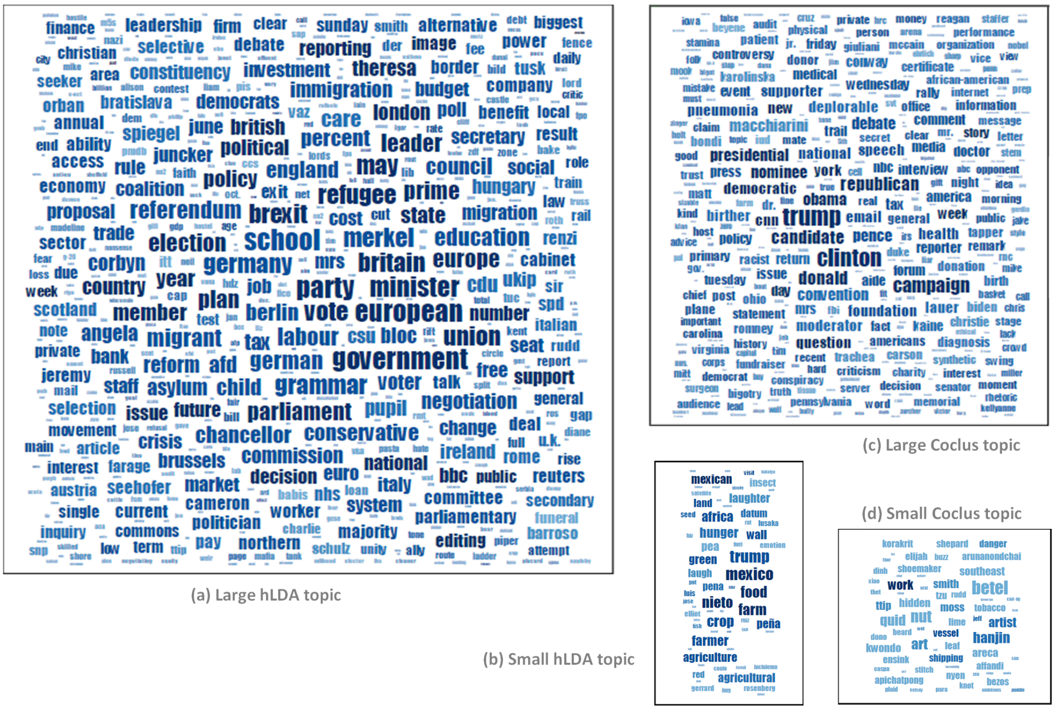
Text Processing. To model the corpus, we adopt the widely used *Vector Space Model* (*VSM*) that provides a term-document matrix. First, the raw text goes through tokenization, part-of-speech tagging and lemmatization steps using Stanford CoreNLP [Manning 2014] (Figure 4.16.1). We keep only nouns and adjectives because they carry enough information to support the interpretation and exploration of topics and *topic variants*. As shown in Figure 4.16.2, each document i is represented by a vector of distinct terms j, weighted by the *Term Frequency-Inverse Document Frequency* (*TF-IDF*) score. This weighting scheme measures how much a term is representative of the documents and is exploited both by the biclustering methods and the visualizations.

Topic extraction. Biclustering methods exploit the duality of terms and documents to obtain high-order co-occurrences within biclusters while promoting the appearance of consolidated term co-occurrences. To extract such biclusters as coarse-grained topics, we use the diagonal biclustering algorithm proposed by Ailem et al. [Ailem 2015]. This algorithm extracts disjoint biclusters as shown in Figure 4.16.3a. The number of topics to extract is an important parameter of this algorithm and must be chosen carefully. Hierarchical LDA (`hLDA`) builds a topic hierarchy where the number of topics is discovered automatically from the data. Like `LDA`, `hLDA` does not take into account the duality of term-document relationships and captures only higher-order co-occurrences. The bicluster structure derived from the probability distribution of `hLDA` we obtain topics with overlapping terms sets and disjoint documents sets.

Beyond the structural differences between both methods summarized above and thoroughly detailed in chapter 1 and 3, their impact on topic comprehension is not well understood. In addition, the *Bimax* biclusters revealed within each topic, a.k.a *topic variants*, can be representative of facts, angles or viewpoints shared by multiple documents. Nonetheless, the quality of the *topic variants* supplied by Bimax essentially depends on the quality of the topical structure offered by the coarse-grained topics, e.g. the terms can represent specific stories or general themes; term co-occurrences can be first-order, high-order or consolidated; the overlapping terms can inform the stories and facts treated in the documents or can be generic; distinctive terms can represent distinct stories or interesting angles and viewpoints of the same story, but sometimes also uninteresting noise. This incited us to conduct a controlled experiment. We define in the next section the independent variables and the hypotheses taken into account for this experiment.

4.5.1.1 Independent variables and hypotheses

To design the experiment, we systematically examined the tag clouds and the *Topic Variant Overview* (Sunburst) obtained with both methods. Figure 4.17 shows large and small tag clouds resulting from both methods. If large tag clouds are interpretable without prior knowledge for both extraction methods, for small tag clouds, we observe that the vocabulary of Coclus topics is very specific to the events at hand and requires specific knowledge to understand the story. For instance, “betel”, “nuts”, “tobacco”, “danger”, concern articles about the danger of betel nuts consumption in Taiwan. Likewise, “handjin”, “vessel”, “shipping”, “art”, “artist”, “moss” concern the specific story related to the bankruptcy of



(a) Large hLDA topic

(b) Small hLDA topic

(c) Large Coclus topic

(d) Small Coclus topic

Figure 4.17: Examples of treemap tiles representing large and small coarse-grained topics extracted by hLDA (a,b) and Coclus (c,d) respectively.

the South Korean firm Handjin Shipping. 3,000 sailors and the British artist Rebecca Moss, whose residency aboard the Handjin Geneva vessel took part of an artistic project, were retained at sea because Handjin's vessels were barred from international ports due to the firm bankruptcy. Such a tag cloud relates two independent stories whose the terms cannot be associated to a story or event without prior knowledge. In contrast, the small hLDA tag cloud concerns one story and the terms give more context to interpret the topic: it concerns agriculture in Mexico and the consequences of the Trump's declared intention to build a wall along the U.S. border with Mexico.

In addition, Figure 4.18(a) shows two scanned *topic variants* both sharing the terms ("film","actor"), which tells that the topic is about cinema. The large differences in the remaining terms of these topic variants reveal two distinct stories. The first concerns a review of the film "Sully" directed by Clint Eastwood. The second discusses the place of female actors in movies. In Figure 4.18(b), the larger commonality of the two scanned *topic variants* concerns a specific story: the situation of the political party of Angela Merkel (CDU) in view of the local election held in Germany in September 2016. By analyzing the short differences between the scanned *topic variants*, we found two angles concerning the event. One discusses the defeat of Merkel's party in the Mecklenburg-Vorpommern state (Merkel's district), and the second relates a debate about the responsibility of Merkel's immigration policy for the defeat. Figure 4.18(b) shows a small Coclus topic with nine unrelated stories separated in distinct subtrees (delineated by dashed lines). These stories are represented each by few *topic variants* yielding few documents.

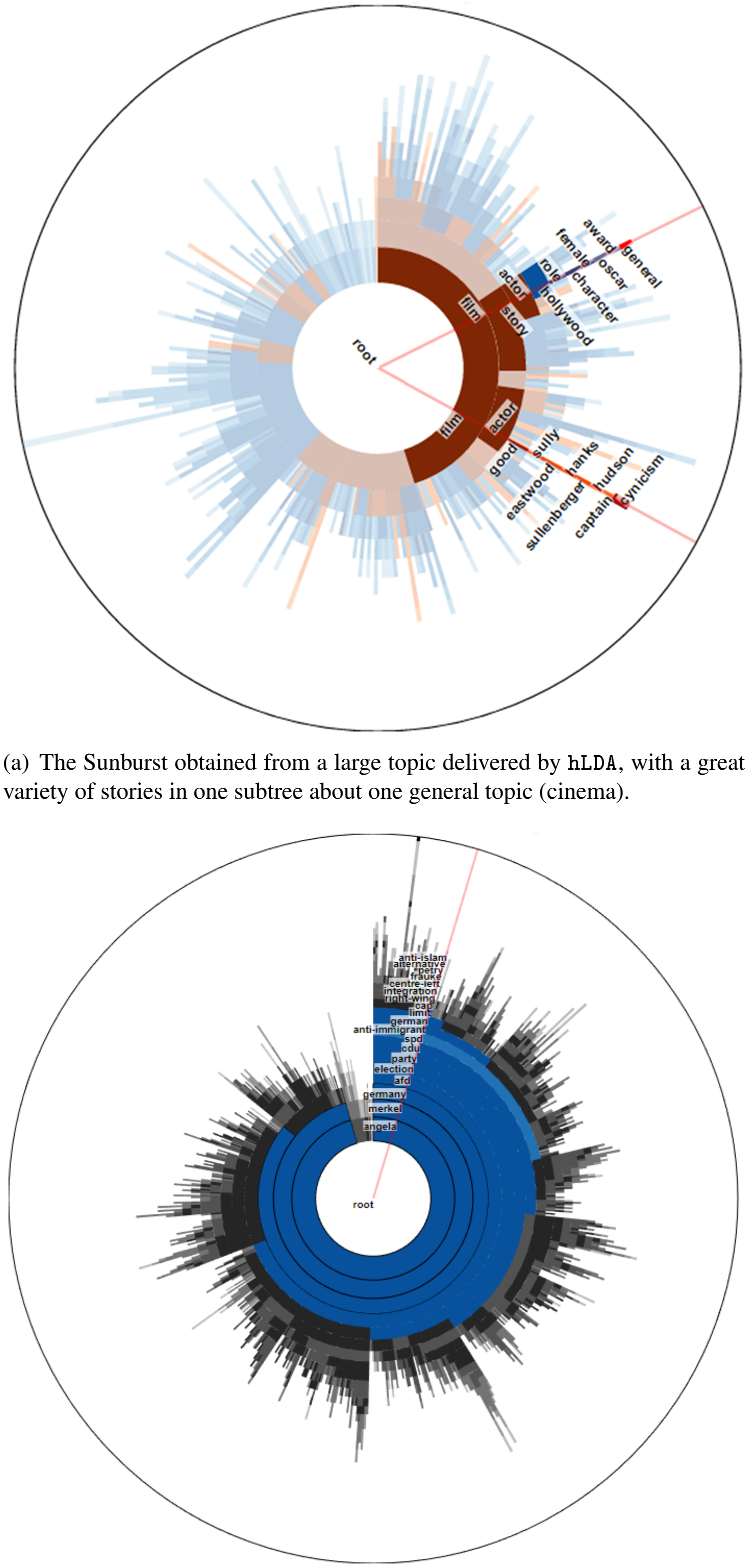
For each method we observed hence specific traits which can be summarized as follows:

O1 in Coclus topics, topic variants are often grouped by common terms describing specific events or stories (Figure 4.18(b));

O2 some small Coclus topics aggregate unrelated stories or events, that end up in distinct subtrees in the term hierarchy (Figure 4.19);

O3 within hLDA topics, especially the large ones, generic shared terms correspond more to a general theme and result in *topic variants* from mixed events or stories (Figure 4.18(a)).

Seeking to identify which topic attributes qualify as independent variables for the present controlled experiment on topic comprehension, we ran a preliminary computational experiment described in section 3.4 looking for statistically significant structural differences between the topics extracted by Coclus and the ones extracted by hLDA. Some metrics captured the size of the topics and their quality in terms of separability, compactness and density. Other metrics measured the abundance of Bimax biclusters and the structure of the resulting term hierarchy such as the branching depth. While a hierarchical structure is expected to be easier to explore than a flat list of term sequences [Dou 2013],



(a) The Sunburst obtained from a large topic delivered by `hLDA`, with a great variety of stories in one subtree about one general topic (cinema).

(b) The Sunburst obtained from a large topic delivered by `Coclus`, with a common term sequence about one story (Germany local election of September 2016) and grouping many *topic variants* with short variations.

Figure 4.18: The Sunburst obtained from two large topics of `hLDA` and `Coclus` respectively.

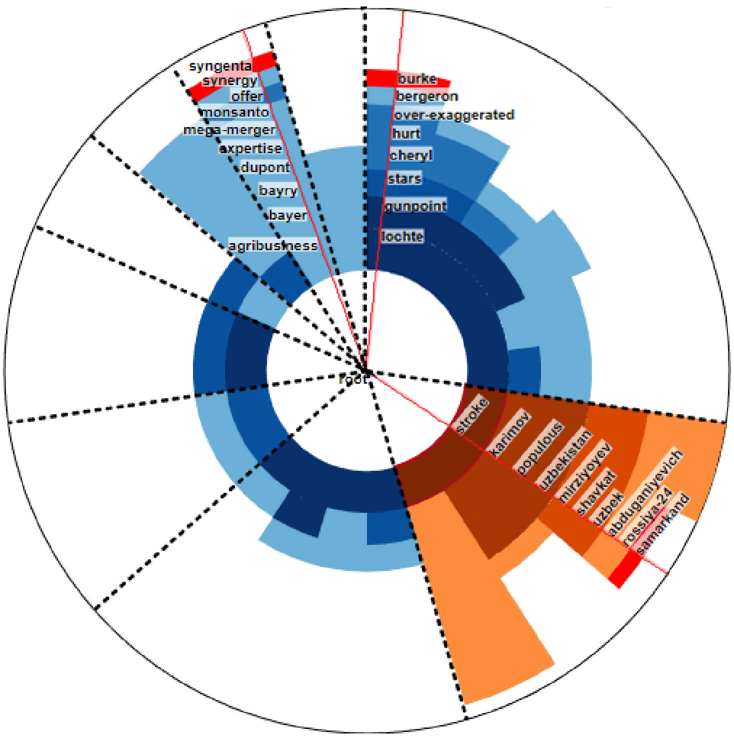


Figure 4.19: Nine unrelated low frequency events separated in distinct subtrees (delineated by dashed lines) in a small Coclus topic.

we also anticipated that late (deeper) branching of term sequences is more meaningful for the analyst than early branching. We reasoned that longer subsequences of shared terms are more likely to group *topic variants* concerning the same events or stories, and to better reveal differences among the related viewpoints or angles (Figure 4.18(b)).

For the sake of brevity, we found that hLDA extracts more often large topics with a comparatively larger number of terms. Many metrics appeared to be correlated with topic size, suggesting that topic size was a suitable independent variable for the present controlled experiment. In contrast, Coclus topics have more documents and fewer terms comparatively. This can be explained by the biclustering mechanism which produces a disjoint term partition and promotes consolidated co-occurrences (Figure 1.2). We infered that the terms in Coclus topics are more document-specific, which may explain observation **O1**. In addition, we anticipated that Coclus deals better with major news topics by locating more specific *topic variants* with more commonalities and slight specificities as shown in Figure 4.18(b). Again, this may explain observation **O1**. However, the shorter commonalities in term sequences within large hLDA topics (Figure 4.18(a)) is consistent with observation **O3**. Indeed the high-order co-occurrences fetched by hLDA yields a greater variety of documents and more distant relationships to explore through the term hierarchy. Finally, since Coclus emphasizes consolidated co-occurences, one can expect it to reveal low frequency events by grouping documents and specific terms placed in well separated topic subtrees, as observed in Figure 4.19 (**O2**). The previous analysis leads to elaborate the following hypothesis we would like to verify through this controlled experiment:

H1.1 Small topics produced by `hLDA` are easier to interpret using a word cloud than those produced by `Coclus` (Figure 4.17b Vs. Figure 4.17d).

H1.2 Small topics produced by `hLDA` lead to *Topic variants* that allow a richer analysis and a more precise interpretation than the ones based on `Coclus`.

H2.1 Large `Coclus` topics are more focused and allow a more precise comprehension than large `hLDA` topics, when both are depicted as word clouds.

H2.2 With less variety and more commonality between *topic variants* in the subtrees, the term hierarchies of large topics are easier to explore with `Coclus` (Figure 4.18(a) Vs. Figure 4.18(b)).

Before describing the design of our experiment in section 4.5.2, we give more details about the dataset and relevant data processing parameters.

4.5.1.2 Dataset and data processing parameters

We used a real dataset composed of news articles continuously aggregated in a database from multiple online news sources (BBC, CNN, Reuters, France24, Egypt Independent and Der Spiegel). We ran both `Coclus` and `hLDA` on a corpus of 3,732 documents and 59,133 terms, covering the period from September 5th to September 18th, 2016. For a fair comparison of the methods, we harmonized as much as we could the pre-processing steps for both. We used a Python implementation of `Coclus` available online[1]. We used the `hLDA` implementation shipped in Mallet [McCallum 2002] with the default values of the hyperparameters and the number of iterations ($\alpha = 10$, $\eta = 1$, $\gamma = 0.1$, $iter = 500$). We set a fixed hierarchy $depth = 3$. For both `Coclus` and `hLDA`, we used the text preprocessing pipeline described in section 3.3 except that we did not filter the first 10,000 terms with the highest *TF-IDF* weight in the matrix. All nouns and adjectives were kept.

Since `Coclus` needs the number of topics to be set a priori, we decided to use the number detected automatically by `hLDA` from the data. By giving both algorithms the same targeted number of topics, we made our best effort to present topics of comparable granularity to the analyst eventually. From the topic hierarchy of `hLDA`, we consider only the leaf nodes as they contain more specific and interpretable terms as described in section 4.5.1. In the corpus described above, `hLDA` found 83 topics whose variants are extracted with our own implementation of the incremental algorithm of `Bimax`, configured with the following parameters: $MinRows = 2$, $MinCols = 2$, $MaxBC = 10,000$ and $\tau_k = 0, \forall k \in [1..K]$. In the sunburst visualization, the minimum angle to display is fixed to 0.3 degrees. We can

[1] `https://pypi.python.org/pypi/coclust`

hence display up to $1,200$ term sequences (or Bimax biclusters). Therefore, the optimal τ_k parameter is searched automatically to obtain a maximum of $1,200$ biclusters. The coarse topics are presented to the participants as individual word clouds extruded from the *Topic Weighted Map* and topic variants are presented in the *Topic Variant Overview*.

4.5.2 Experimental protocol

19 volunteers participated in the experiment. 16 have a background in computer science and 3 in other domains. We consider two independent variables: the topic extraction method (Coclus and hLDA) and topic size. The 166 topics (83 topics from each method) have been distributed in 3 classes based on topic size using the following quantile split: 25% of small, 50% of medium and 25% of large. For each combination of topic size and method, two topics were presented to each participant, giving a total of $2 \times 3 \times 2 = 12$ topics per participant. In order to avoid the same participant examining two semantically related topics (extracted with different methods), we adopted a between-subject approach. To assess topic similarity we inspected the 83 topics of each method using the corresponding *Topic Weighted Map* overview and identified the most salient pairs of similar topics. Five pairs of topics were chosen and separated in two user groups. In each group, the topic list was randomly completed with non-paired topics until obtaining two topics for each size and method. For each user group we defined two series of 6 distinct topics presented in a random order to mitigate the learning effect. Participants had also the choice to analyze one or two topic series. This between-subject approach increases the variety of topics by having 4 distinct topics per method and per size. The experiment started with a 15-minute demo followed by 15 minutes of training with two topics. We dedicated 30 minutes for each series of topics (5 minutes for each topic). The experiment time was not strictly limited, but the elapsed time was shown to encourage the participants to enter a response. 13 participants explored two series of topics and 6 participants explored only one. We obtained 32 interpretations per topic extraction method and per topic size.

The purpose of the study was first to understand how different topic extraction methods influence topic comprehension. From the word clouds displayed for each topic, we asked the participants to report the stories they found by entering up to 5 terms and by proposing a title that describes their understanding. In order to evaluate the interpretation of each reported story, we tried to be as objective as possible by defining the scores described in Table 4.1. Hence a better score was obtained for correct and precise interpretations. To characterize a global interpretation quality for each of the word clouds, we considered the average score of every story found. By doing so, we meant to evaluate the feasibility of task **T1.1** per se using the word clouds.

Secondly, we aimed to understand the influence of the methods on the exploration and

Table 4.1: Scores and criteria to evaluate topic interpretations through tag clouds.

0	No story is found or the chosen terms are not semantically related.
1	Terms are semantically related but the title is not given or is not related to the terms.
2	Terms are semantically related and the proposed title does not correspond to a theme in any document of the topic.
3	Terms are semantically related and the title corresponds to a theme treated in at least one document of the topic.
4	Terms are semantically related and the title describes precisely an event or a story treated in at least one document of the topic.

the interpretation of the *topic variants*. To do so, we presented the *Topic Variant Overview* to the participant with a preselected *topic variant*. The participant was asked to explain his interpretation in one or two sentences, and to find two other *topic variants* revealing different angles or viewpoints related to the same story. The rationale was to evaluate the following aspects:

- the interpretability of the *topic variants* in terms of precision and correctness;
- the ability to explore and find topic variants and to understand their commonalities and specificities;
- the difficulty to deal with the amount, the variety and the heterogeneity of *topic variants*;
- the difficulty of dealing with term redundancies in the hierarchy;
- the ability to deal with different topic sizes (small vs. large);

The precision and the correctness received scores between 0 and 4 in the same spirit of the scores in table 4.1, as detailed in Table 4.2.

Table 4.2: Scores and criteria to evaluate *topic variants* through the Sunburst.

0	No *topic variant* is found.
1	One *topic variant* is found but no explanation is given or, if any, it is not related to the terms.
2	One *topic variant* is found. The explanation is consistent with the term sequence but corresponds to an hypothetical interpretation the documents do not hold.
3	One *topic variant* is found and the explanation corresponds to a general description of an event treated in at least one of the documents but it does not describe a specific angle, viewpoint or fact.
4	One *topic variant* is found and the explanation describes with precision a specific angle, viewpoint or fact treated in at least one of the documents. The concerned story must be the one designated by the initial *topic variant*.

However, the more numerous the alternative viewpoints or angles found in the view, the better the exploration and understanding of commonalities and specificities. So, instead of averaging the scores of topic variants, we considered as a global topic score the sum of scores obtained by every *topic variant* found. As the participant had to interpret a maximum of 3 *topic variants*, the best score could not exceed 12 for each topic. Thereby we also evaluate the feasibility of the tasks **T1.2**, **T2.2** and **T3.3** through the *Topic Variant Overview*. The interaction giving access to document titles and raw contents was enabled for the word clouds as well as the *Topic Variant Overview*. Generally, a quick check on the list of document titles was enough for the users to validate the hypothetical interpretations they elaborated from the term sequences. By lack of time, the users rarely looked at document contents, and if so, only one or two sentences around a specific term of interest helped them understand the precise meaning of the term in its context. Finally we also asked the participant to judge the usefulness of the *Topic Variant Overview* (Sunburst) and to quantify the difficulty of exploring them by answering the following questions on a Likert scale:

Q1 The subtle differences between term sequences help identify precise and interesting angles or viewpoints.

Q2 Term sequences reveal a great variety of stories interesting to explore together.

Q3 Term redundancy across sequences complicates the analysis.

Q4 The number and the variety of the sequences are too large, which complicates the analysis.

4.5.3 Results

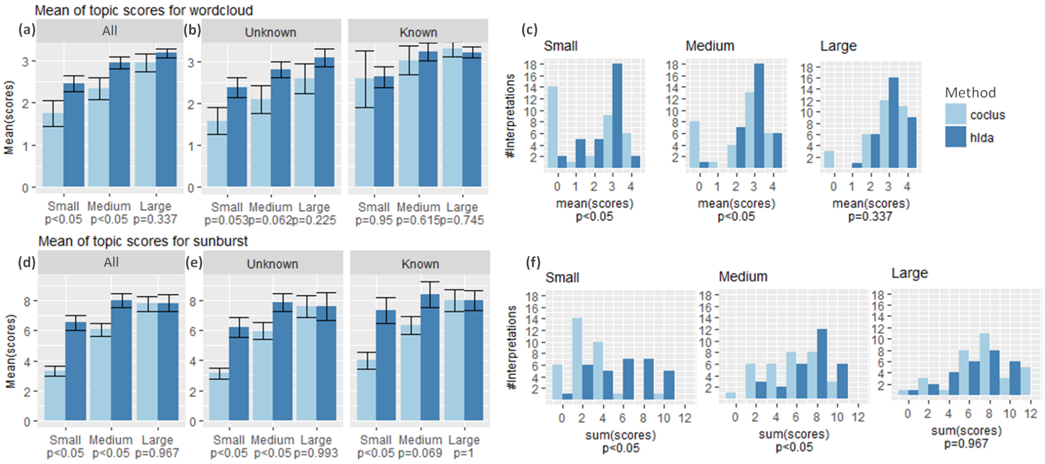


Figure 4.20: The qualitative scores obtained for `Coclus` and `hLDA` topics. The bar charts show the means and the standard errors of topic scores by size (a,d) or by size and by prior knowledge (b,e). Histograms (c,f) depict the distribution of scores for each size.

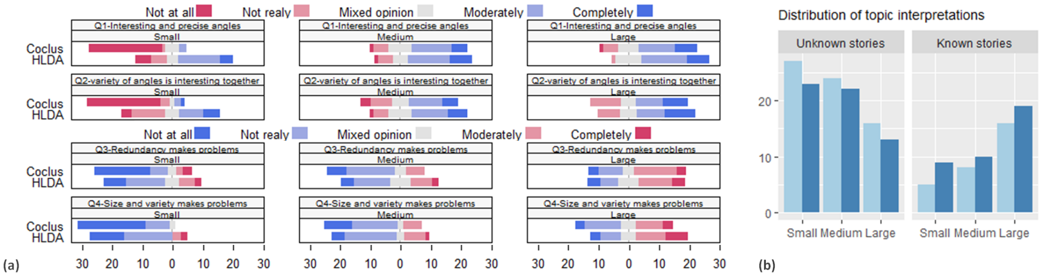


Figure 4.21: (a) Histograms showing user ratings of the quality of the *Topic variant Overview* on a Likert scale (built with R package `HH` [Heiberger 2014]). (b) The distribution of topics by size and by prior knowledge.

Interpretation scores: precision and correctness. For each size and method we computed the mean and standard error of the global score received by every topic. We present the results in Figure 4.20. A p-value less than 0.05 indicates that the difference is statistically significant, according to the Student's T-test. In Figure 4.20a the small and medium-sized tag clouds get significantly better interpretations based on *hLDA* topics than *Colcus* topics (**H1.1**). We essentially explain this difference by the higher number of terms in small `hLDA` topics, which retains more semantic context in the word clouds. However, we note in Figure 4.20c that for small topics, `Coclus` got more interpretations with the highest score=4, and as many as `hLDA` for medium and large topics. The differences for small and medium sizes are mainly due to a large number of interpretations with a score 0 obtained by `Coclus`, that decreases as the size of topics increases. This contrasted result for `Coclus` can be explained by the lower number and the specificity of `Coclus` terms (**O2**) that require prior knowledge to successfully interpret smaller word clouds. The large topics yield no significant differences and obtain an average score around 3. While **H2.1** was not verified, we can conclude that for both methods the word clouds showing large topics allow a correct interpretation of a theme (**T1.1**) but not always of precise events or stories. Figure 4.20d shows that the participants better explore the *topic variants* for `hLDA` with small and medium topics, by reporting more precise and correct interpretations (**H1.2**). However for large topics the average score is close to 8 for both `hLDA` and `Coclus`. It is interesting to note in Figure 4.20f that for large topics, `hLDA` interpretation never get the highest score of 12 while `Coclus` has 6 interpretations with the highest score (**H2.2**). For both methods the score increases with topic size showing that our *Topic Variant Overview* supports effectively the search for interesting angles and viewpoints among more than a thousand *Topic Variants* (**T1.2, T2.2, T3.3**).

Does prior knowledge matter? The familiarity of participants with the topics can have an impact on the interpretation score. To separate the analysis of known and unknown topics, we asked the participant to rate their familiarity with each topic on a 5-level Likert scale. We considered that the topics were familiar if the scores were between 3 and 5. Figure 4.21b shows the distribution of topics in the two groups. Unsurprisingly, we can see that familiar topics are mainly large and unknown topics are small. In Figures 4.20b and 4.20e, we observe that the familiarity of topics has an impact only on the interpretations based on word clouds. Indeed, in Figure 4.20b, the differences between `Coclus` and `hLDA` are not significant when the story is known. The differences appear more convincing when the stories are unknown.

Judgment of usefulness. We also collected qualitative judgments of *topic variants* by participants through multiple questions formulated in section 4.5.2 and reported in Figure 4.21a. For the first question (**Q1**), both techniques have a good appreciation for medium and large topics with a slight advantage for `hLDA`. However `Coclus` has a poor appreciation with small topics (**H1.2**). The responses to question **Q2** follow exactly the same pattern as for question **Q1**. For the next two questions, the color of the bar chart is reversed because the responses “Not at all” and “Not really” correspond to better performance. For both methods the two kinds of difficulties increase with topic size. For question **Q3**, the participants report greater difficulty with `hLDA` in small and medium topics while for large topic the difficulty is the same for both methods. However through question **Q4**, the variety and the amount of *topic variants* make the exploration more difficult with `hLDA`, and more significantly for large topics (**H2.2**).

Participant observations. We also analyzed the observations entered by the participant themselves through a field available under each topic to enter free comments, as well as notes we took while observing the participants. We labeled these observations in different categories reported in Figure 4.22. The categories mainly concern problems observed during the analysis of certain topics. We can hence discern the variety of limitations reported by participants and establish profiles for each size and method. First, we notice the presence of the category “KNOWLEDGE_DEPENDENT” for every size and method but it appears more importantly for the topics built by `Coclus` with a medium size. This remains consistent with the large number of unknown stories observed for small and medium topics in Figure 4.21b, but the more specific terms given by `Coclus` can also be an explanation (**H1.2**). However, for both methods alike, the lack of data (“NEED_MORE_DATA”) seems to be a more important limitation for small topics.

Moreover, we took note of each topic for which the participants needed to access document content (“NEED_DOCS”). This need appears for every size and method except for

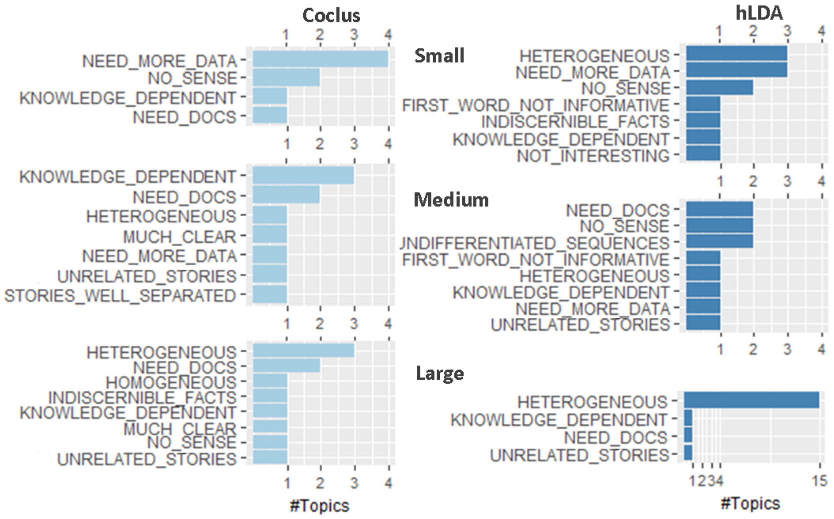


Figure 4.22: Observations of participants collected during the user study.

small hLDA topics. This observation is certainly related to the lack of familiarity with the concerned topics. Sometimes the participants reported that in some *topic variants* the terms were not semantically related, or even the documents concerned different stories. We grouped these observations under the category “NO_SENSE”. Again, both methods and all sizes are concerned. This situation is mainly due to term polysemy or the usage of the same concepts in different contexts.

In addition, for some medium and large topics from both methods, unrelated stories have been forced together in one topic (“UNRELATED_STORIES”). However, one participant reported for a topic of Coclus that unrelated stories were well separated in different subtrees of the hierarchy (“STORIES_WELL_SEPARATED”). In this case, the variety of stories does not hinder the analysis because uninteresting *topic variants* can easily be ignored or filtered out, as we also observed in **O2**. Yet concerning Coclus, two topics appeared to be very clear compared to other topics seen until then (“MUCH_CLEAR”) and one large topic has been qualified as “HOMOGENEOUS” (**H2.2**). However Coclus shows some limitations concerning more generally the difficulty of dealing with noise. Typically, for medium and large topics, participants reported problems concerning the heterogeneity of the stories revealed by topic variants grouped in one subtree (“HETEROGENEOUS”). One participant reported also the difficulty for a large topic to discern interesting facts (“INDISCERNIBLE_FACTS”). Nonetheless, for hLDA, the comments concerning the difficulty to deal with noise are made for all topic sizes and are more diverse. For instance, some participants report that the first word is rather generic and not informative (“FIRST_WORD_NOT_INFORMATIVE”) or that the differences between term sequences are difficult to discern (“UNDIFFERENTIATED_SEQUENCES”).

As for Coclus, one participant found “INDISCERNIBLE_FACTS” for one small topic, and more importantly, the “HETEROGENEOUS” rating appears much more frequently for large hLDA topics (15) than for any Coclus topic (4), which strengthens **H2.2**.

4.5.4 Discussion

The goal of our experiment was to study how different topic extraction methods impact topic comprehension by analysts and eventually to assess which method is better for which tasks. First, the better scores obtained by hLDA with small and medium topic sizes (Figure 4.20) validate both **H1.1** and **H1.2**. The responses to questions **Q1** and **Q2** (Figure 4.21a) also confirm these hypotheses. However, the contrasted scores for small and medium Coclus topics (Figure 4.20c) could be due to the lack of familiarity observed in Figure 4.21b and 4.22; the terms being too few and specific, under a disjoint partition scheme, to give sufficient context to understand the topics without prior knowledge. Secondly, **H2.1** and **H2.2** are not directly validated by the scores of large topics in Figure 4.20 by lack of statistical significance. But the responses to question **Q4** (Figure 4.21a) show that with hLDA, it is more difficult to cope with the variety of *topic variants*. Moreover the analysis of user observations in Figure 4.22 reveals that the participants identified more issues with hLDA concerning noise, even with small topics to some extent, but much more for large ones. While the responses to questions **Q1** and **Q2** (Figure 4.21a) are slightly better for hLDA, Coclus received good appreciations too. Coclus is also the only method that obtained positive observations (Figure 4.22) and the highest score in Figures 4.20c and 4.20f.

This controlled experiment reveals a trade-off to find between the two topical structures of Coclus and hLDA. On the one hand, Coclus produces less interpretable small and medium topics, but more specific large topics (Figure 4.18(b)) with sometime low-frequency unrelated stories that are well separated in distinct subtrees (Figure 4.19). Therefore, Coclus seems interesting to verify hypotheses. We also envision improving user analysis by supporting interactive topic editing. To this end, interactions such as reallocating, removing or merging *topic variants* are useful, but are easily done when the stories are well separated in distinct subtrees in the hierarchy, as Coclus does. On the other hand, hLDA produces small and medium topics with more context (Figure 4.17a) allowing better interpretations, but the large topics are less specific and sometimes produce intertwined stories in the same subtree (Figure 4.18(a)). While low-frequency events can be scattered or hidden by the hLDA topical structure, the large variety of its topics remains interesting and reveals precise and interesting viewpoints. It seems to better promote serendipitous discoveries and hypothesis generation.

The process of inquiry such as journalistic investigation alternates both verification and generation of hypothesis. It is hence useful to deal with the limitations of both methods in order to better support both processes. Applying visual analytics principles fits this goal by giving the possibility to the expert to steer both methods until interesting results are found. By starting with hLDA the expert can better analyze the variety of topics and their relationships. In a second analysis, the expert can adopt Coclus in a subset of hLDA topics to better reveal the specificities of the stories. Thereby, we expect that *topic variants* extracted by Bimax could be better organized in the *Topic Variant Overview* in well discernible stories to facilitate topic exploration and transformation. This will be our main concern for future work.

Another issue we aim to deal with is term redundancy in the *Topic Variant Overview*. While Bimax searches for optimal biclusters, they contain lots of overlapping terms with slight differences. To build a term hierarchy, we consider biclusters as sequences of terms ordered by their overlap degree. Then, FPTree hierarchizes the term sequences by maximizing the common subsequences starting with the first terms. In this way, when differences are observed in prefix terms (e.g. the sequences of B4 and B5 in Figure 4.16(4b)), the sequences are separated in distinct subtrees, even if they have large commonalities in their remaining terms (t5, t2 aligned through the red dashed line in Figure 4.16(4b)). In future work, we aim to tackle this issue by considering the whole set of bicluster terms to group *topic variants*. We expect that such an approach better brings out events, while quickly identifying the commonalities and the specificities of *topic variants* while searching for interesting viewpoints, angles and facts during the investigation. The preservation of a hierarchical representation in order to ease the exploration of large number of biclusters remains a challenging issue.

4.6 Conclusion

In this chapter, we described a visual analytics tool supporting analytic journalists in dealing with large corpora. With our system analysts can access their data at multiple levels of detail. For topic overview, we designed a new visualization, the *Topic Weighted Map*, that combines an MDS projection, a Weighted Map visualization and multiple word clouds. We experimented a hybrid biclustering approach, leveraging hard biclustering and overlapping biclustering algorithms. We demonstrated that through this hybrid structure, *Bimax* biclusters reveal meaningful *Topic Variants* that help the analyst understand document relationships. We proposed a new approach to explore overlapping term-document biclusters based on a hierarchy of associated terms. A qualitative evaluation showed that this term hierarchy linked to the *Topic Variant Comparator* view allows the analyst to explore a large number of *Topic Variants* and find useful facts or viewpoints. Overall, the system supports the analysis of corpora that are significantly larger than journalists may be used to. The system lets the journalist initiate a focus process on specific aspects to verify hypotheses, and then engage in a diversification process to discover new aspects and refine hypothesis.

We conducted a comparative study showing the impact on topic comprehension of topic size and of the topical structure elicited by two different topic extraction methods, `Colclus` and `hLDA`. The study reveals that the two methods have opposite behaviors with respect to topic size. Small `Coclus` topics hold specific vocabulary and are more difficult to interpret without prior knowledge, whereas small `hLDA` topics contain richer vocabulary giving more context for the analyst's comprehension. When `hLDA` large topics yield themes with heterogeneous *topic variants*, large `Coclus` topics uncover more specific and more separable events in spite of slight differences between *topic variants*. We conclude that `Coclus` is more suitable for hypothesis verification while `hLDA` is more relevant for hypothesis generation.

The current chapter was devoted to a visual analytics system for static text corpora. The next chapter propose dynamic visualizations dealing with text streams.

Part III

Towards Visual Analytics of text streams

CHAPTER 5

Situation Awareness and Real Time Exploration of Text Streams

Contents

5.1 Introduction . . . 129

5.2 Related Work . . . 130

5.3 Architecture and Models . . . 131

5.3.1 Event-driven Web-oriented Architecture . . . 131

5.3.2 Text Processing . . . 133

5.3.3 Model for Multiple Dynamic Frequency Matrices . . . 133

5.4 Visualization Design . . . 134

5.4.1 Temporal View . . . 134

5.4.2 Map View . . . 136

5.5 Case Study . . . 136

5.5.1 Data and Tasks . . . 136

5.5.2 The Case-Study . . . 138

5.6 Discussion and future work . . . 142

5.6.1 Discussion . . . 142

5.6.2 Challenges and Future Work . . . 144

5.7 Conclusion . . . 145

5.1 Introduction

The wide adoption of portable devices has led to an exponential increase in information sharing in social media. Nowadays everyone can produce information everywhere at any time. In this context, many professions use microblogs such as Twitter to find relevant information. For example, emergency services use them to target more efficiently their interventions in a disaster area [Chae 2014]. The law enforcement forces also exploit them to identify and track criminals [Denef 2013]. Last but not least, investigative journalists follow targeted user accounts to find sources of information and documents related to their ongoing investigations [Marcus 2011]. In this regard, visual analytics tools may be used to provide such users with valuable help to carry out situation awareness and exploration tasks dealing with text streams.

In this chapter, we present a work in-progress that aims eventually to propose a novel visual analytics approach for text stream analysis. The visual analytics tool we describe is a preliminary step designed to take the IEEE VAST Challenge 2014 [1] (mini-challenge 3) [Médoc 2014]. Based on a user-centered method, our design is driven by two generic tasks: situation awareness and exploratory analysis of text streams. Through an efficient web-oriented architecture, the backend exploits in real time the metadata and named-entities of the messages, while the frontend offers dynamic and interactive visualizations supporting flexible analyses of Twitter-like text streams.

To evaluate the tool, we conducted a case-study through our investigation of streaming text messages in the context of the VAST Challenge questions. This case-study led us to identify, regarding the targeted tasks, the advantages and the limitations of the proposed visualizations, the needs in terms of analytic processing, and more generally the challenges we aim to address in future work.

The remainder of this chapter is organized as follows. In *Section* 5.2 we present related work that has inspired our approach, our data models, and our visualizations. We describe briefly in *Section* 5.3 our software architecture and our data model, then our visualization design in *Section* 5.4. In *Section* 5.5 we develop a case-study illustrating how our prototype may be used to explore text streams. Finally we reflect on the issues that we encountered during this work in *Section* 5.6.

5.2 Related Work

The design of our tool is inspired by four main pieces of research. Firstly, we based our design on the following tasks listed by Rohrdantz et al. [Rohrdantz 2011] to achieve situation awareness:

1. *monitoring the current situation* by a short-term analysis of data;
2. an *exploration* task consisting in navigating through all dimensions of data using interactive visualization, allowing the user to fully understand multiple aspects of data;
3. *event tracking* as well as *change and trend detection* tasks allow the user to follow the temporal evolution of data in real time;

[1] http://hcil2.cs.umd.edu/newvarepository/benchmarks.php

4. a *historical retrieval* task allows the user to analyze the historical development of a selected subset of data;

5. finally, a *temporal context* task seeks to compare each data item relatively to the others at any time point, for example by using ranking [Krstajić 2013].

Secondly, one of the most used techniques to model text corpora is the *Vector Space Model* (VSM), also known as *Bag Of Words*, where each document is represented by a word vector. The entries of the vector are the words of the entire corpus and the weights are the frequency of these words in the document. Based on the *VSM*, Turney and Pantel [Turney 2010] identified three models to handle three kinds of matrices: *Term-Document* matrices, *Word-Context* matrices and *Pair-Pattern* matrices. We propose a more general framework described in *Section 5.3* defining general spaces that can handle these three kinds of matrices for streaming text messages. We refer to these spaces as *Object Space* and *Context Space*; they support the creation of multiple *Object-Context* matrices.

Thirdly, a Theme River visualization [Havre 2002] conveys an overview of the temporal evolution of multiple items of interest depicted as layers in a stacked area chart (e.g. word counts, author counts etc.). The evolution of each item can be analyzed individually, or globally through its relative importance compared to other items of the same kind. In this way the *temporal context* and *change and trends detection* tasks can be achieved. We will see further in this chapter that, by adding interaction mechanisms and parsimonious animations, most of the remaining tasks can be achieved too.

Lastly, Kandogan et al. [Kandogan 2013] propose a reference web architecture for real-time visual analytics on large streaming data. For the needs of the VAST Challenge, we implemented a subset of their architectural patterns using Esper Streaming Engine (see *section 5.3.1*). However, we had to implement our own model of multiple dynamic frequency matrices presented in *section 5.3.3*.

5.3 Architecture and Models

5.3.1 Event-driven Web-oriented Architecture

We adopted a web-oriented architecture (see *figure 5.1*) comprising a Java backend running on a Tomcat Web Server and a Javascript web frontend based on the D3.js and AngularJS libraries.

During our initial requirements study, we identified two main tasks: monitoring the current situation and exploring the history. These tasks rely on two different data scopes for which flexible expiration strategies can be defined depending on various considerations such as elapsed time or a fixed number of items. Monitoring the current situation bears on

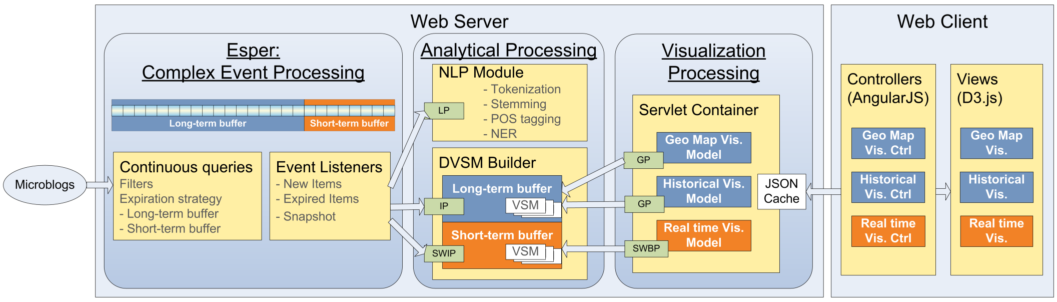


Figure 5.1: Event-driven web-oriented architecture

a sliding short-term period up to the current time. We call this scope the *short-term buffer*. Exploring the history requires a longer period in which the historical data is collected. We call this scope the *long-term buffer*.

For exploration and monitoring tasks, the visualizations do not exploit directly raw texts. At regular time intervals, the Analytical Processing in Figure 5.1 applies the NLP module on the streaming messages to extract named entities. The designer must define beforehand pairs of features chosen from categories of named entity or metadata. Given these pair-wise combinations the Dynamic Vector Space Model Builder (DVSM Builder) computes multiple dynamic matrices. The server-side Visualization Processing component transforms the multiple DVSM in a JSON data structure proper to the visualizations. At regular intervals the frontend calls the server to retrieve the new state of data and renders the visualizations. The last callback responses are cached in JSON on the server and can be retrieved by the client call signature. This improves the interaction efficiency in a multi-user mode. The cached data is marked as obsolete as soon as the Visualization Processing component updates the structure.

The design of our Analytical Processing and Visualization Processing components was guided by the processing approach for streaming data proposed by Kandogan et al. [Kandogan 2013]:

1. The *Local Processing* (LP) is stateless and applies local transformation on each incoming item.

2. The stateful *Incremental Processing* (IP) contains a common internal model that is updated for each incoming item.

3. The *Sliding Window Processing* (SWP) requires an expiration strategy defining the selection criterion of the items that represent the window. The common internal model is maintained at regular intervals.

4. Finally the *Global Processing* (GP) computes in batch mode all the items collected

from the stream.

In the case of SWP we privilege the use of incremental algorithms. But, since most data mining algorithms compute the data in batch mode, we may want to apply them on a static snapshot retrieved from the sliding window at regular intervals. Unlike Kandogan et al. [Kandogan 2013] we distinguish two types of SWP: the *Sliding Window Batch Processing* (SWBP) and the *Sliding Window Incremental Processing*(SWIP). Each component in Figure 5.1 is identified by one of these processing approaches in the green rectangles.

All these processing approaches as well as the architectural patterns *Queue of Observers* and *Sliding Window Repository* [Kandogan 2013] were implemented using Esper, an opensource Java component for Complex Event Processing. Each streaming item is wrapped in an event handled by the Esper Engine. Based on continuous queries that filter items and define expiration strategies, events are sent regularly to registered components, or each time an item flows in or expires. The listeners receive either a full snapshot of events, or only the newly added/expired events. We define two continuous queries respectively for long-term and short-term buffers.

This architecture is designed in order to minimize server-side processing of client calls, hence supporting more efficient interactions. Although this architecture induces a large usage of memory, it scales up well in a cluster through a distributed in-memory computing engine such as Apache Spark if needed.

5.3.2 Text Processing

Besides the textual content, microblog messages carry exploitable metadata such as date and author information, hashtags and mentioned users. We use Twitter's Text Processing Library to extract such information from the messages. We leverage Named-Entity Recognition using the Stanford CoreNLP API to extract meaningful structured data in real time such as person names, locations and organizations present in messages. We refer to all such metadata and structured contents as '*aspects*'.

5.3.3 Model for Multiple Dynamic Frequency Matrices

Based on the VSM, our model is designed in order to ensure the following requirements. The model must compute multiple dynamic frequency matrices by combining each desired *aspect* of the messages. Each matrix must be updated as messages stream in/out and the state of matrices must represent the messages belonging to the related long-term or short-term buffers. Hence the number of rows and columns in the matrices can vary over time, increasing when new messages stream in, or decreasing when old ones expire.

Since libraries such as Apache Lucene or S-Space, don't handle DVSM with a flexible combination of spaces, we implemented our own model. We defined two kinds of spaces:

Object Spaces and *Context Spaces*. In *Object Spaces*, entries can be any aspect extracted from message contents. In *Context Spaces*, entries correspond to aspects for which items of *Object Space* can occur many times. Various *Object-Context* matrices can then be built by combining instances of both *Spaces*, the weight being the number of occurrences of *Object* items in each *Context* items. For example, hashtags can serve as *Object Space*, while time in minutes is used for *Context Space* to analyze the temporal evolution of hashtags.

The entries of each *Space* as well as the related matrix weights are maintained as messages stream in. When a message expires, various removal strategies may apply depending on how the matrices are impacted and how the effects have to be propagated to the spaces. For instance, if the vector corresponding to a space entry is empty (all the weights are equal to 0) we propose two choices: either keeping this entry in the space to retain the information that it has existed earlier, or removing/forgetting it until a new message reintroduces it in the buffer.

Moreover, the model is flexible enough to define several instances of both *Spaces* and combine them in several matrices. An application of this model is the analysis of the relationships between all aspects of the messages. For example, by combining author as *Context Space* and hashtags or extracted named-entities as *Object Space*, one can know how much one author talks about something, about someone or about a place. These relationships are monitored in real time in the short-term buffer while their historical development is traced in the long-term buffer. This model sets the stage for event detection.

Matrix combinations can also be used to achieve multi-resolution temporal analyses. One instance of *Object Space* can be combined with different *Context Space* representing different time granules. For example, defining the two combinations Author-TimeInMinute, Author-TimeInSecond lets the user switch on demand from a per-minute to a per-second time resolution without rebuilding the whole matrix: both matrices are kept up-to-date and available in memory.

5.4 Visualization Design

Our tool is composed of two windows *Figure 5.2(a,b)* that can be visualized simultaneously on a dual-monitor setup.

5.4.1 Temporal View

In *Figure 5.2(a)*, two dynamic and interactive temporal views are displayed. Both comprise four visual components: a configuration bar (1), a Theme River in the Detail view (2) with its Legend (3) and its Focus+Context lens (4). In addition, the Message view displays full text messages (5). At the top of the window (a), the Historical Theme River (HTR) view

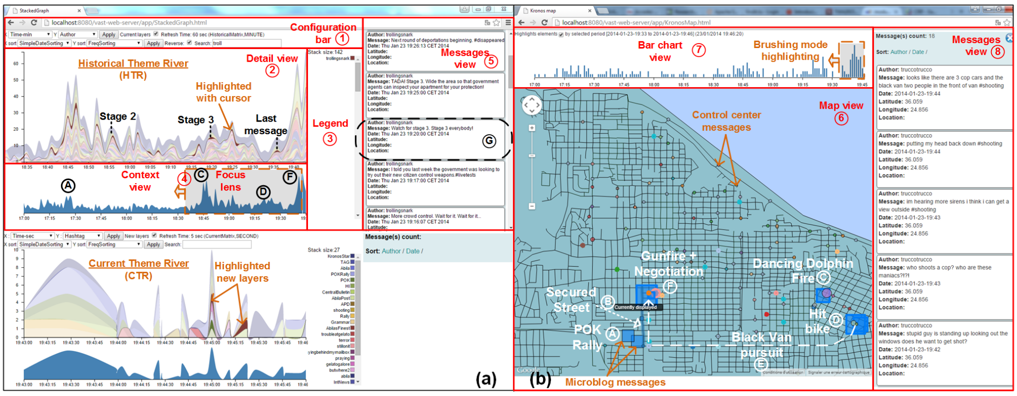


Figure 5.2: **(a)** Two dynamic Theme Rivers visualize the temporal evolution of different aspects of streaming messages from a microblog. **(b)** A map of Abila City is displayed using the GoogleMap API. Capital letters correspond to the multiple events found in the stream: (A) the POK Rally with (B) the secured street, (C) the "Dancing Dolphin" fire, (D) the black van hit a bike, (E) the black van pursuit with the police, (F) the gun fire and negotiations with the police and (G) the last messages of "trollingsnark" launching the "stage 3" of operations.

is dedicated to historical analysis. It shows the evolution of text streams, based on the long-term buffer. On the bottom, the Current Theme River (CTR) view provides situation awareness of text streams based on the short-term buffer. From the configuration bar (1) the analyst chooses which *Context Space* to encode on the time axis and which *Object Space* to encode as aspect (i.e. authors, hashtags or extracted named-entities) on the Y axis. At regular intervals, the corresponding frequency matrix is retrieved from the server. In the Detail view (2) a dynamic Theme River depicts the temporal evolution of the chosen object-context martix where weights are encoded as layer thickness. Therefore, the analyst gets a global overview of the temporal development of the selected aspect and achieves the *temporal context* task.

To support the *exploration* task, we added a *Focus + Context* lens (4) allowing the analyst to zoom in and explore the Detail view (2) by dragging a selected time-window along the whole history. This lens is defined interactively and resized at will. The layers/items are identified by their color mapped to the labels in the Legend (3). Furthermore, when the mouse hovers over the Theme Rivers or the legend, the targeted layer is highlighted showing its temporal patterns and a tooltip provides more information. The legend can be filtered according to a text query entered in the search box, and the user can access the actual messages on demand (5) by clicking on a layer or on a legend item.

Finally, through the two dynamic Theme Rivers the analyst can *monitor the current situation* and *detect changes*. To support the *historical retrieval* task, the items of the HTR view are automatically highlighted when they match those in the short-term buffer. In contrast, the CTR view highlights the items never encountered in the long-term buffer to support the *trend detection* task. To avoid overwhelming the user when too many items are highlighted, this behavior can be deactivated using a check box in (1).

5.4.2 Map View

In *Figure 5.2(b)*, the Map view (6) encodes with circle shapes the messages coming from the law enforcement Control Center (CC), i.e. alerts issued by the police or fire department provided in the steaming data of the IEEE VAST Challenge 2014. The square shapes represent the MicroBlog (MB) messages supplying spatial coordinates. Shape size is proportional to the number of messages at the given position. While the Map view gives a cumulative representation of all the historical events, the bar Chart view (7) shows their temporal distribution across the long-term buffer.

Moreover, each message can be highlighted on the map with a thick border to achieve two objectives. Firstly, the *monitoring mode* helps *monitor the current situation* by observing the items belonging to a sticky time window capturing the last few minutes of the stream. Secondly, the *brushing mode* consists for the user to select interactively a period in the bar chart and drag it at will on the timeline. The messages belonging to the active time period are highlighted accordingly; by clicking on a shape the corresponding messages are displayed (8). This helps the user carry out *exploration* and *change and trends detection* tasks.

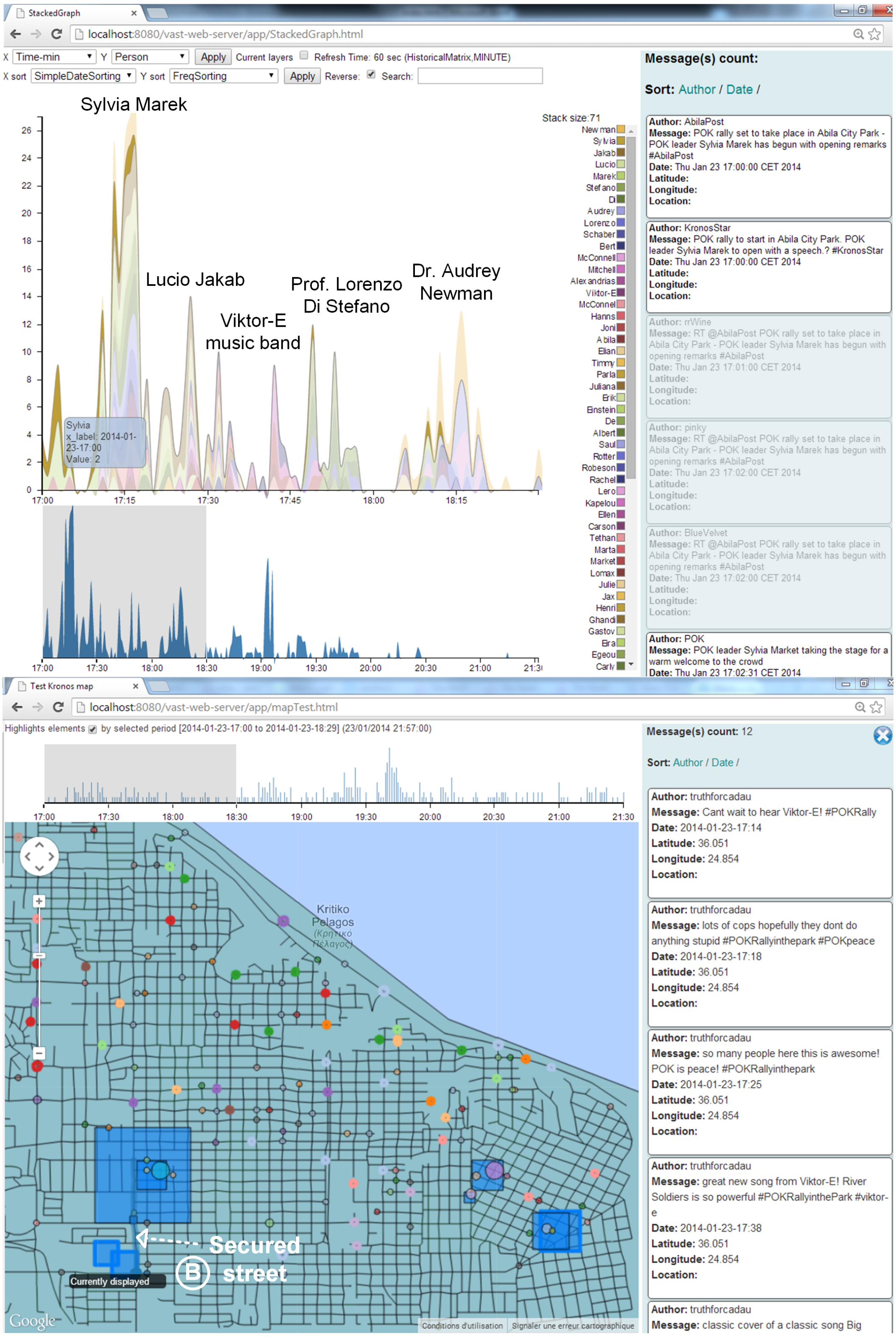
5.5 Case Study

The present case study describes our own analysis of a text stream as provided by the IEEE VAST Challenge 2014.

5.5.1 Data and Tasks

A fictive scenario is drawn up: several employees of the GASTech company have disappeared; the Protector of Kronos (POK) is an organization known to be hostile to the activities of the company on the account of pollution and possible ties to health problems in the city of Abila.

The mission of the analyst consists in investigating the disappearance of GasTech employees and answer general questions such as where the employees are? who is behind a possible abduction crime? To answer these questions, the analyst has access to two kinds of streaming messages. Firstly, the MB messages are formatted like Twitter messages. Some



X Time-min ▾ Y Person ▾ Apply Current layers □ Refresh Time: 60 sec (HistoricalMatrix,MINUTE)

X sort SimpleDateSorting ▾ Y sort FreqSorting ▾ Apply Reverse: ☑ Search:

Message(s) count:

Sort: Author / Date /

Author: AbilaPost
Message: POK rally set to take place in Abila City Park - POK leader Sylvia Marek has begun with opening remarks #AbilaPost
Date: Thu Jan 23 17:00:00 CET 2014
Latitude:
Longitude:
Location:

Author: KronosStar
Message: POK rally to start in Abila City Park. POK leader Sylvia Marek to open with a speech.? #KronosStar
Date: Thu Jan 23 17:00:00 CET 2014
Latitude:
Longitude:
Location:

Author: rrWine
Message: RT @AbilaPost POK rally set to take place in Abila City Park - POK leader Sylvia Marek has begun with opening remarks #AbilaPost
Date: Thu Jan 23 17:01:00 CET 2014
Latitude:
Longitude:
Location:

Author: pinky
Message: RT @AbilaPost POK rally set to take place in Abila City Park - POK leader Sylvia Marek has begun with opening remarks #AbilaPost
Date: Thu Jan 23 17:02:00 CET 2014
Latitude:
Longitude:
Location:

Author: BlueVelvet
Message: RT @AbilaPost POK rally set to take place in Abila City Park - POK leader Sylvia Marek has begun with opening remarks #AbilaPost
Date: Thu Jan 23 17:02:00 CET 2014
Latitude:
Longitude:
Location:

Author: POK
Message: POK leader Sylvia Market taking the stage for a warm welcome to the crowd
Date: Thu Jan 23 17:02:31 CET 2014

Highlights elements ☑ by selected period [2014-01-23-17:00 to 2014-01-23-18:29] (23/01/2014 21:57:00)

Message(s) count: 12

Sort: Author / Date /

Author: truthforcadau
Message: Cant wait to hear Viktor-E! #POKRally
Date: 2014-01-23-17:14
Latitude: 36.051
Longitude: 24.854
Location:

Author: truthforcadau
Message: lots of cops hopefully they dont do anything stupid #POKRallyinthepark #POKpeace
Date: 2014-01-23-17:18
Latitude: 36.051
Longitude: 24.854
Location:

Author: truthforcadau
Message: so many people here this is awesome! POK is peace! #POKRallyinthepark
Date: 2014-01-23-17:25
Latitude: 36.051
Longitude: 24.854
Location:

Author: truthforcadau
Message: great new song from Viktor-E! River Soldiers is so powerful #POKRallyinthePark #viktor-e
Date: 2014-01-23-17:38
Latitude: 36.051
Longitude: 24.854
Location:

Author: truthforcadau
Message: classic cover of a classic song Big

Figure 5.3: Each peak in the Historical Theme River concerns talks delivered by members of the POK organization in a park. The highlighted layer corresponds to Sylvia Marek. In the map, a street is secured by the police. The highlighted alerts are those appearing in the time window selected in the timeline.

of them contain geolocation data to locate them in Abila City. Secondly, the CC messages correspond to geolocated alerts sent by the police and fire departments.

Hence the first objective is to discover in real time important events through thousands of messages. A second objective is to identify the main persons related to the events, the witnesses of a criminal scene, potential hostages or even the criminals themselves. A third objective is to track in real time these events on a map and on a time line.

The VAST challenge dataset is split in two segments. The first is provided as a static file to allow tool design and serves as an initial history for the plot. The second is provided as a live stream through a socket connection made available by the challenge organizers a few hours before the closure of the challenge.

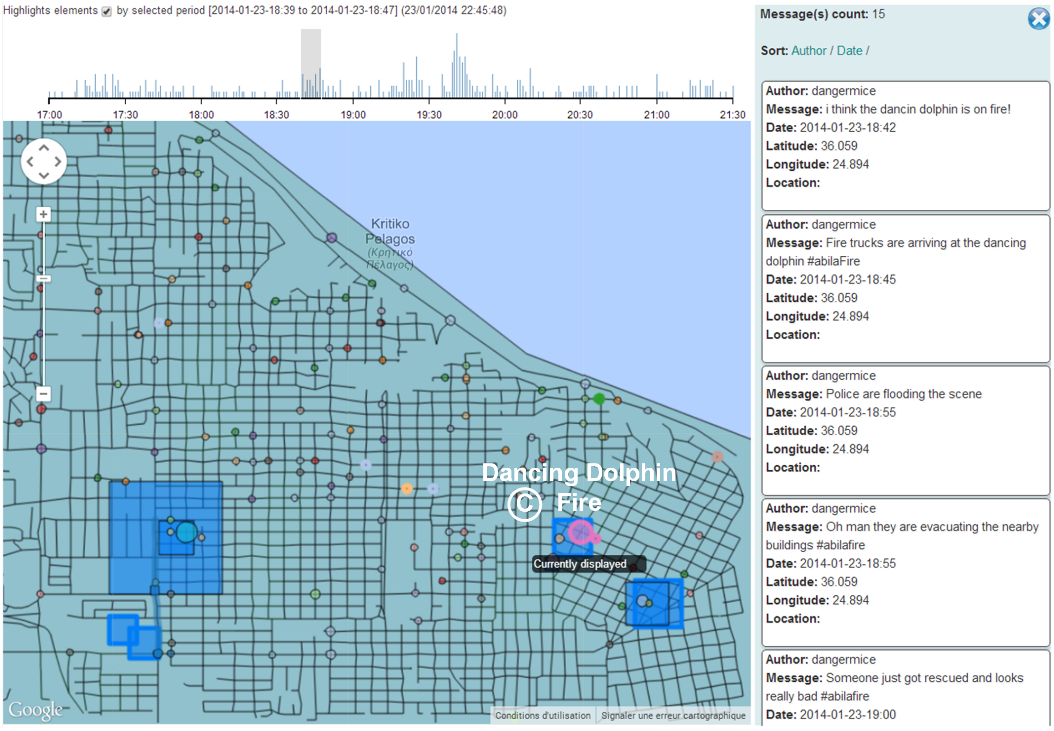
5.5.2 The Case-Study

The stream starts at 18:30. In the CTR view, we choose to monitor author activity. At 18:34, a new author starts texting; his layer is highlighted as a new author in the CTR view. Clicking on the layer to read his messages, we find his first message to be quite ironic. We decide to note his name '*trollingsnark*' to track him.

In the HTR, we explore alternately all aspects of data. Through both the author and hashtag aspects we notice, in the Context view, a peak located between 17:23 and 17:29 (see Figure 5.2 (A)). Zooming on this period in Figure 5.3, we inspect the layers using tooltip information and browse through the message view in order to explain the peak. Amidst the jabber, we learn that many people are attending a rally organized by the POK organization in a park. Analyzing the person aspect (i.e. named-entities) in Figure 5.3 gives us more insight about the events occurring during the rally: "Sylvia Marek" speaks at 17:17 followed by "Lucio Jakab" at 17:26, the music band Viktor-E, Prof Lorenzo Di Stefano and Dr. Audrey Newman. In the Map view we locate a street secured by the police in Figure 5.3(B). A large square shape near this street depicts MB messages sent by a single author commenting the rally events and sheds light on the peak observed earlier.

In the Map view, at 18:40 a CC message (circle) is highlighted (see Figure 5.4 (C)), which draws our attention to the message "POSSIBLE FIRE-REPORT". Next, around 18:42 in Figure 5.4, an MB message is highlighted saying "*i think the dancing dolphin is on fire!*". Reverting to the temporal views we switch to the hashtag aspect and filter the CTR with the query '*fire*'. In the HTR view a peak appears around 18:47. Both views confirm that a building called '*Dancing Dolphin*' is on fire.

During idle periods we track regularly the messages of our ironic author. In the HTR view, we switch to the author aspect, we filter the author list with the query '*trollingsnark*' to see his temporal trends and read his messages. At 18:56, a message makes this author suspicious: "*There we go. False flag operation entering stage 2.*" (see Figure 5.5). We



Highlights elements ☑ by selected period [2014-01-23-18:39 to 2014-01-23-18:47] (23/01/2014 22:45:48)

Message(s) count: 15

Sort: Author / Date /

Author: dangermice
Message: i think the dancin dolphin is on fire!
Date: 2014-01-23-18:42
Latitude: 36.059
Longitude: 24.894
Location:

Author: dangermice
Message: Fire trucks are arriving at the dancing dolphin #abilaFire
Date: 2014-01-23-18:45
Latitude: 36.059
Longitude: 24.894
Location:

Author: dangermice
Message: Police are flooding the scene
Date: 2014-01-23-18:55
Latitude: 36.059
Longitude: 24.894
Location:

Author: dangermice
Message: Oh man they are evacuating the nearby buildings #abilafire
Date: 2014-01-23-18:55
Latitude: 36.059
Longitude: 24.894
Location:

Author: dangermice
Message: Someone just got rescued and looks really bad #abilafire
Date: 2014-01-23-19:00

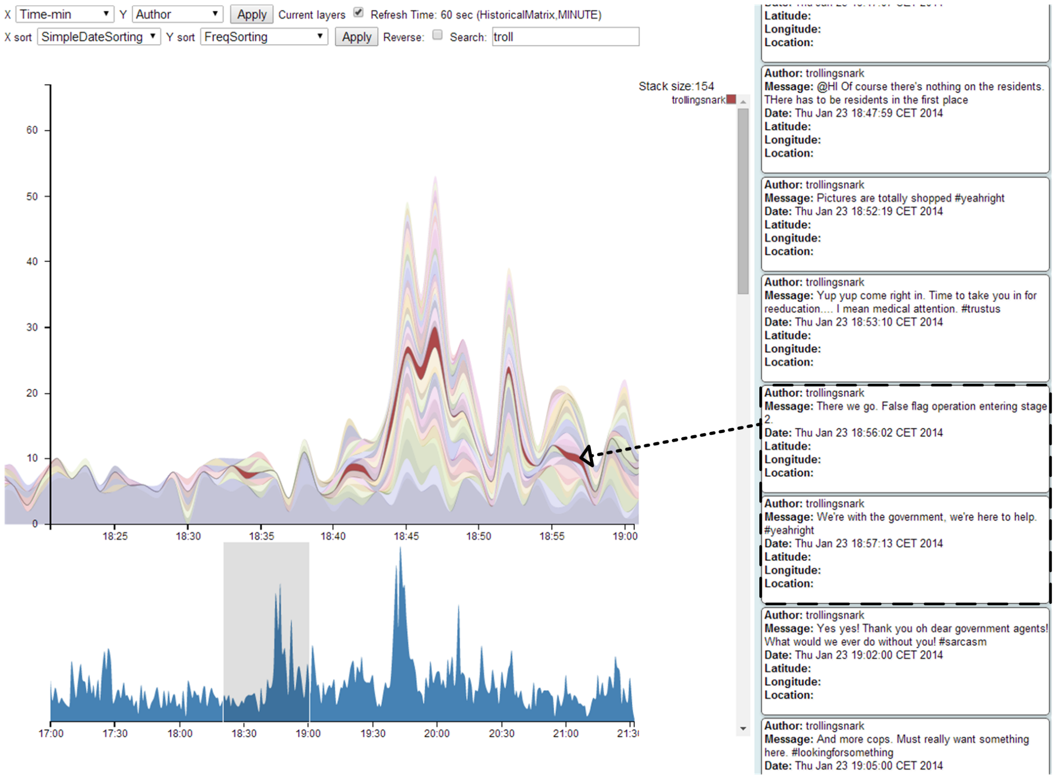
Figure 5.4: The pink circle alert concerns a fire started in the '*Dancing Dolphin*' building. A witness describes the event through a set of messages.

assume that the "false flag operation" is the fire at the *Dancing Dolphin*, that is, a diversion created by terrorists.

On the map, we observe at 19:20 a new highlighted square shape that grows in a few minutes. The corresponding messages report that a "Black Van has hit a bike" as shown in Figure 5.6(D). At 19:39 we observe in Figure 5.6(E) six highlighted CC alerts forming a path along the street. These alerts identify a Black Van and report a high-speed pursuit with the police. Using the *brushing mode* on the time line, we highlight successively the six CC alerts over the path reported also in Figure 5.2(E). We hence track the Black Van pursuit that stops on the map close to a highlighted square shape in Figure 5.6(F). This corresponds to messages posted from a restaurant. A witness tweets that "*some crazy black van just got pulled over in the parking*" and that two persons exchange gunfire with the police.

During another idle period, we go back to the messages of our suspicious author in Figure 5.6(G). At 19:20 he writes: "*Watch for stage 3. Stage 3 everybody!*" and at 19:26 "*Next round of deportations beginning. #disappeared*". These messages strengthen our suspicions about him.

We continue our analysis with the live stream. Moving back and forth between the map and the CTR view, we track the events and verify the information by filtering and drilling down into the hashtag aspect in the HTR view. We learn that the van carries hostages, and



X Time-min ▾ Y Author ▾ Apply Current layers ☑ Refresh Time: 60 sec (HistoricalMatrix,MINUTE)
X sort SimpleDateSorting ▾ Y sort FreqSorting ▾ Apply Reverse: □ Search: troll

Latitude:
Longitude:
Location:

Author: trollingsnark
Message: @HI Of course there's nothing on the residents. THere has to be residents in the first place
Date: Thu Jan 23 18:47:59 CET 2014
Latitude:
Longitude:
Location:

Author: trollingsnark
Message: Pictures are totally shopped #yeahright
Date: Thu Jan 23 18:52:19 CET 2014
Latitude:
Longitude:
Location:

Author: trollingsnark
Message: Yup yup come right in. Time to take you in for reeducation.... I mean medical attention. #trustus
Date: Thu Jan 23 18:53:10 CET 2014
Latitude:
Longitude:
Location:

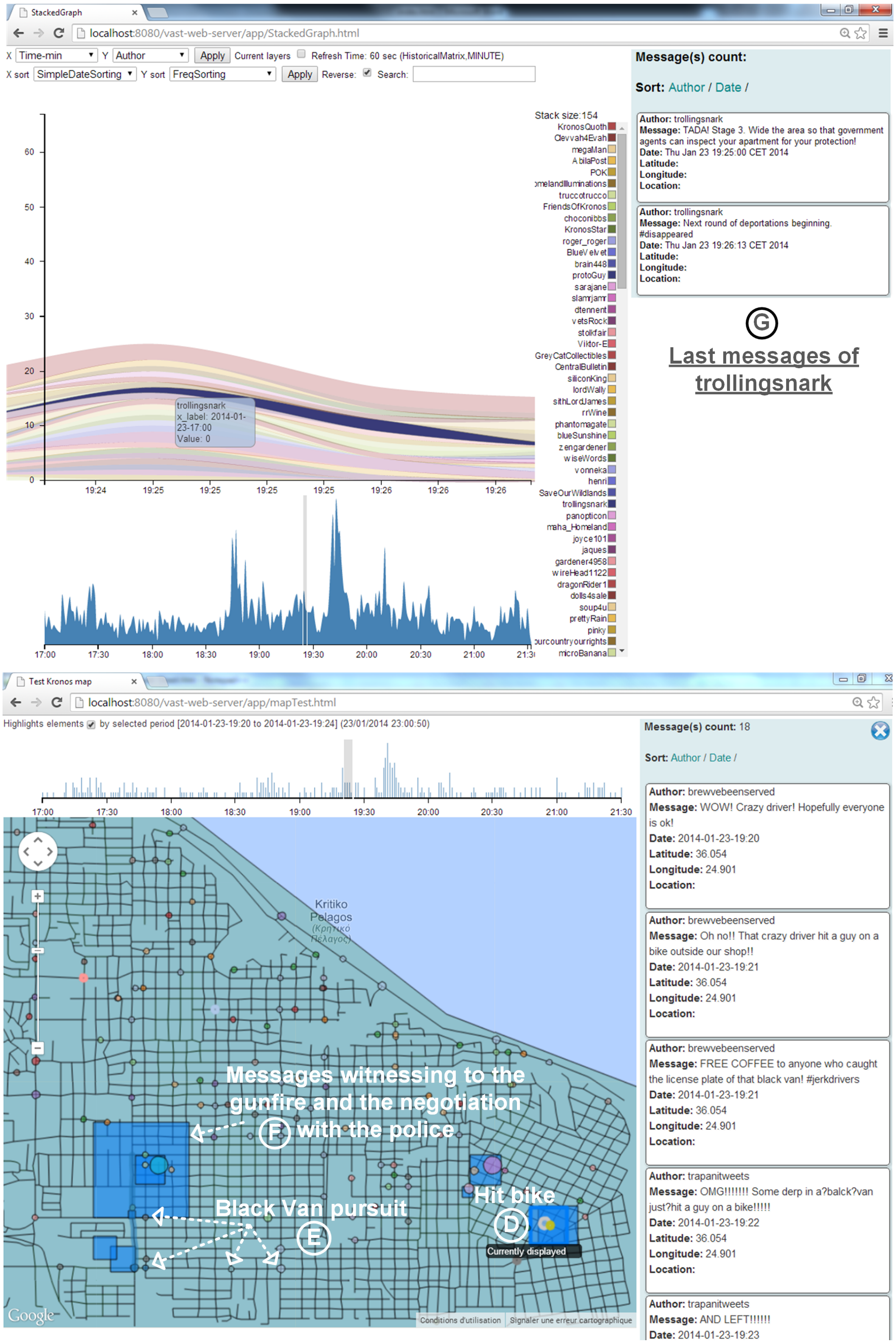
Author: trollingsnark
Message: There we go. False flag operation entering stage 2.
Date: Thu Jan 23 18:56:02 CET 2014
Latitude:
Longitude:
Location:

Author: trollingsnark
Message: We're with the government, we're here to help. #yeahright
Date: Thu Jan 23 18:57:13 CET 2014
Latitude:
Longitude:
Location:

Author: trollingsnark
Message: Yes yes! Thank you oh dear government agents! What would we ever do without you! #sarcasm
Date: Thu Jan 23 19:02:00 CET 2014
Latitude:
Longitude:
Location:

Author: trollingsnark
Message: And more cops. Must really want something here. #lookingforsomething
Date: Thu Jan 23 19:05:00 CET 2014

Figure 5.5: Ironic messages of ‘*trollingsnark*’ among which the message “*False flag operation entering stage 2*”. We assume that the “operation” refers to the fire caused by the terrorist as a diversion.



X Time-min ▾ Y Author ▾ Apply Current layers □ Refresh Time: 60 sec (HistoricalMatrix,MINUTE)

X sort SimpleDateSorting ▾ Y sort FreqSorting ▾ Apply Reverse: ☑ Search:

Message(s) count:

Sort: Author / Date /

Author: trollingsnark
Message: TADA! Stage 3. Wide the area so that government agents can inspect your apartment for your protection!
Date: Thu Jan 23 19:25:00 CET 2014
Latitude:
Longitude:
Location:

Author: trollingsnark
Message: Next round of deportations beginning. #disappeared
Date: Thu Jan 23 19:26:13 CET 2014
Latitude:
Longitude:
Location:

G
Last messages of trollingsnark
Test Kronos map
localhost:8080/vast-web-server/app/mapTest.html

Highlights elements ☑ by selected period [2014-01-23-19:20 to 2014-01-23-19:24] (23/01/2014 23:00:50)

Message(s) count: 18

Sort: Author / Date /

Author: brewvebeenserved
Message: WOW! Crazy driver! Hopefully everyone is ok!
Date: 2014-01-23-19:20
Latitude: 36.054
Longitude: 24.901
Location:

Author: brewvebeenserved
Message: Oh no!! That crazy driver hit a guy on a bike outside our shop!!
Date: 2014-01-23-19:21
Latitude: 36.054
Longitude: 24.901
Location:

Author: brewvebeenserved
Message: FREE COFFEE to anyone who caught the license plate of that black van! #jerkdrivers
Date: 2014-01-23-19:21
Latitude: 36.054
Longitude: 24.901
Location:

Author: trapanitweets
Message: OMG!!!!!!! Some derp in a?balck?van just?hit a guy on a bike!!!!!
Date: 2014-01-23-19:22
Latitude: 36.054
Longitude: 24.901
Location:

Author: trapanitweets
Message: AND LEFT!!!!!!
Date: 2014-01-23-19:23

Figure 5.6: The last messages of '*trollingsnark*' launching the stage 3 of operations, i.e. the deportation of hostages (G), are posted five minutes after the "Black Van has hit a bike" (D) and one minute before the pursuit and the gunfire exchange with the police.

that the police is negotiating with the terrorists.

We decide to analyze the messages of our suspect in the temporal context of all the events discovered in the city. In HTR, we brush a long period (see *Figure 5.2*). We observe that his activity starts at the beginning of the fire event (C) and stops at the beginning of the Black Van pursuit (E). The message announcing "*stage 2*" appears fifteen minutes after the fire starts and the one about "*Stage 3*" five minutes after the bike accident. His last message comes one minute after the pursuit starts. As a result we have found converging clues tying this suspicious author to the criminals in the van and making him the potential head of the operations. The stream ends while the terrorists surrender and the hostages are rescued.

5.6 Discussion and future work

In the previous section we described what we learned in real time from the stream using our dynamic visual analytics tool. In the next section we discuss the lessons learned concerning user issues, the solutions we provide as well as their advantages and limitations.

5.6.1 Discussion

5.6.1.1 Support for Real-time Analysis

Firstly, in a real-time analysis the user is faced with a dilemma: is it better to explore past history or to monitor the current situation? If the throughput of messages is low, the user may start with a quick exploration of the CTR view and take advantage of idle periods to explore the HTR view. For higher throughput, a collaborative approach may be adopted. Instead of opposing the CTR and the HTR views, we propose to show relationships between them through the auto-highlighting feature in both. In the CTR the highlighted layers emphasize what is new. During the historical exploration, the highlighted layers of the HTR show that these items are currently active in the CTR. We noticed that, when a large number of items occurs in the CTR, everything gets highlighted in the HTR, which may hinder the analysis. In this situation, the user can disable this feature using the checkbox ('Current layers').

A second issue for the analyst is to handle the pressure of real-time analyses. The user stops frequently his HTR exploration to see what is happening in the CTR. To minimize the cost of context switching, our tool helps the user preserve the current analysis by keeping its state including the active filters, the time window selection and the list of messages selected for reading. Even if the underlying data changes, the newly displayed items correspond to what the user expects to see when he continues his analysis.

5.6.1.2 Support for Multi-aspect Analysis

The data model proposed in *section 5.3.3* supports the analysis of the temporal evolution of multiple aspects of the stream. We observed that certain message aspects are more valuable for event analysis or for certain tasks. For example, author names in the HTR view do not bring enough semantics of the messages to orient the analysis during the exploration task. Even if the hashtag aspect brings stronger semantics, both show thin layers due to the high number of items they display. This hinders the detection of their individual temporal patterns. In contrast, the aspects resulting from the named-entity recognition (persons, organization, location) contain fewer items with stronger semantics. Moreover, these items bring a mixture of witness statements from different authors which, put in their temporal order, allow to tell a story with greater confidence. As a result, the person aspect leads to faster and more exhaustive analysis than other aspects to discover the rally-related events. But, for the Dancing Dolphin fire and the black van pursuit events, no named-entities are found; the hashtags and authors aspects are more useful in this case. Finally the author aspect turns out to be essential to track suspects or witnesses (e.g. tracking 'trollinsnark') while hashtags help verify hypotheses.

This multi-aspect analysis incites the user to often switch between aspects during his investigation. Supporting this interaction is crucial in real-time exploration. Our proposed architecture and model, handling multiple matrices maintained incrementally and available at any time in the server-side memory, has successfully handled the 4,058 messages provided in the IEEE VAST Challenge 2014.

5.6.1.3 Location of Events on the Map

The map visualization is useful to locate and monitor the area in which events occur in real time. The displayed messages are aggregated throughout time. The time dimension is flatened to give a historical overview in the map while monitoring the current situation via the synchronized highlighting mode. The low number of geolocated *MB* messages and the stability of their position helps preserve the mental map of the user and identify the changes quickly (growing shapes or new ones). Moreover, the list of messages ordered by time allows to easily understand the sequence of events. However, for *CC* alerts, a lot of scattered items appear and disappear which overwhelms the user. The brushing mode allows hence to control the extent and the speed of the sliding window to observe trajectories of alerts, as we did to identify the trajectory of the "Black Van".

5.6.1.4 Coordinated Multiple Views

This study shows that the Theme River views with the interactions we propose allows to obtain an overview, to zoom in on a period of interest and drill-down the messages to detect and understand events while monitoring the current situation to detect new items. These views are suitable for a low number of items, each carrying a strong semantic value and therefore referencing a mixture of meaningful messages. This observation will guide the choice of analytical processes such as clustering and dimension reduction techniques we aim to include in future work.

This study shows also that the Map view is useful to locate and monitor the area in which interesting events occur in real time. Although MB messages located on the map correspond often to meaningful witness statements reported by a single author, they must be verified through other aspects in the temporal views. The dual monitor setup supports this complementarity between both views.

5.6.2 Challenges and Future Work

In this work we have identified many limitations recapped below. In the HTR, the adoption of a global ranking of layers does not match their local ranking in user selected time windows and complicates the analysis. By filtering the corresponding matrix on the period of interest and applying the sorting strategy in this subset, we may offer a more accurate *temporal context*. We may also benefit from the temporal multi-resolution capability of our model to propose a finer time granularity for this selected period.

Scalability issues may arise as we support *historical retrieval* through an in-memory data structure. To cater for long time periods or high throughput streaming, more efficient data storage and information retrieval techniques must be considered to keep the message history accessible by the analyst. We are currently testing state-of-the-art distributed in-memory structures.

To monitor the current situation, the CTR represents explicitly the time as an axis. On the one hand, this representation lets the user estimate when items will expire from the CTR (e.g. when items approaches the end of the sliding window) and become scattered in the HTR. The user can hence assess the priority they have for his ongoing analysis. On the other hand, real-time analyses require the user to get a quick understanding of all aspects of data. This is not possible with a river showing one aspect at a time. Addressing the limitation of single-aspect river views, we may consider abstracting the time dimension, recovering one visual dimension to represent one more aspect to characterize events (ie: what, where, who). One option may be to use a matrix visualization to show the relationships between multiple aspects and leverage the multiple dynamic matrices available in-memory. Time is then implicitly apprehended through animations by brushing an auxiliary timeline.

Moreover, discovering the relationships between these aspects (groups of authors that talk about the same hashtags, the same places, the same persons, etc.) may be useful to better understand events. This may consist in analyzing the co-occurence of *Object Space* items in *Context Space* items and reciprocally. We aim to discover these co-occurences by applying co-clustering techniques to the frequency matrices we already build. We expect that the resulting co-clusters reduce the item size to display and contain a mixture of meaningful messages, as discussed in *section 5.6.1.4*. In this perspective the remaining challenges are first to propose efficient ways to run these analytic techniques in real time. For instance, we plan to use incremental clustering algorithms or static approaches based on consecutive snapshots.

5.7 Conclusion

Our Visual Analytics tool allows the user to achieve situation awareness and exploration tasks for text streams. Through a case study, we show the user's ability to discover, locate and explain different events in their temporal context. In this activity, the monitoring and exploration of all message aspects appears to be crucial. These tasks are supported by a flexible and dynamic VSM underlying the web-oriented architecture we propose. Nonetheless, we have highlighted several limitations in this solution, among which the lack of accurate temporal context for a selected period, the limitation of single-aspect river views and the large use of memory by our model and architecture. This preliminary work has paved the way for us to start experimenting with novel visualizations and analytical processing techniques for text streams, in order to tackle these issues. To this end, the next chapter proposes `DynBimax`, a dynamic version of the `Bimax` algorithm.

CHAPTER 6

Dynamic Bimax

Contents

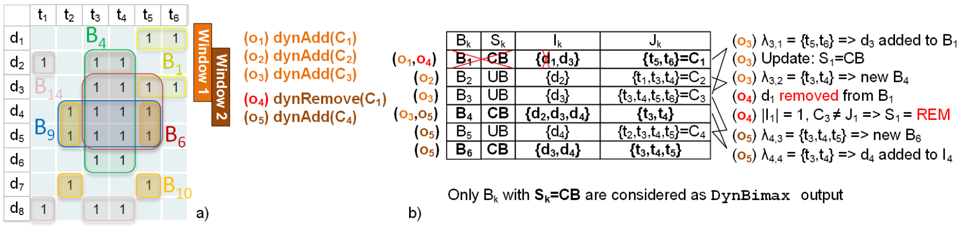
6.1 Introduction 147
6.2 Dynamic Bimax Algorithm 147
6.3 Proof of condition validity after and before remove 150
6.4 Performance evaluation 151
6.5 Conclusion 153

6.1 Introduction

Dealing with news streams, investigative journalists need to verify or refine their hypotheses by identifying viewpoints or facts shared by multiple sources. To support these tasks for static text corpora, `Bimax` [Prelić 2006] can be combined with a diagonal biclustering method [Ailem 2015]. The latter provides coarse-grained topics from which `Bimax` extracts fine-grained topic variants. To obtain good insights, the analyst has yet to refine the model by reassigning documents and terms within the partition. Therefore, the `Bimax` results must be updated dynamically in order to better capture user knowledge. To deal also with dynamic news streams, sliding windows are commonly used. While most biclustering methods are static [Govaert 2013, Madeira 2004], other solutions process consecutive time slices by using the prior partition to smooth the next one [Greene 2010]. In contrast, dynamic approaches are proposed for one-way overlapping clustering [Pérez-Suárez 2013] and consist in updating the partition when objects are added, removed or modified. While these solutions deal with data streams they provide coarse-grained clusters or biclusters. We propose in this chapter a dynamic algorithm, `DynBimax`, that extends the incremental version of `Bimax` [Prelić 2006] to provide dynamic topic variants.

6.2 Dynamic Bimax Algorithm

After recalling the general notations used in this chapter, we present below the necessary changes to the incremental `Bimax` algorithm [Prelić 2006] (simply referred to as `Bimax` for



	t_1	t_2	t_3	t_4	t_5	t_6
d_1					1	1
d_2	1		1	1		
d_3			1	1	1	1
d_4		1	1	1	1	
d_5		1	1	1	1	
d_6			1	1		
d_7		1			1	
d_8	1		1	1		

(o_1) dynAdd(C_1)
(o_2) dynAdd(C_2)
(o_3) dynAdd(C_3)
(o_4) dynRemove(C_1)
(o_5) dynAdd(C_4)

	B_k	S_k	I_k	J_k
(o_1,o_4)	~~B_1~~	~~CB~~	$\{\cancel{d}_1,d_3\}$	$\{t_5,t_6\}=C_1$
(o_2)	B_2	UB	$\{d_2\}$	$\{t_1,t_3,t_4\}=C_2$
(o_3)	B_3	UB	$\{d_3\}$	$\{t_3,t_4,t_5,t_6\}=C_3$
(o_3,o_5)	B_4	CB	$\{d_2,d_3,d_4\}$	$\{t_3,t_4\}$
(o_5)	B_5	UB	$\{d_4\}$	$\{t_2,t_3,t_4,t_5\}=C_4$
(o_5)	B_6	CB	$\{d_3,d_4\}$	$\{t_3,t_4,t_5\}$

(o_3) $\lambda_{3,1} = \{t_5,t_6\}$ => d_3 added to B_1
(o_3) Update: S_1=CB
(o_3) $\lambda_{3,2} = \{t_3,t_4\}$ => new B_4
(o_4) d_1 removed from B_1
(o_4) $|I_1| = 1$, $C_3 \neq J_1$ => S_1 = REM
(o_5) $\lambda_{4,3} = \{t_3,t_4,t_5\}$ => new B_6
(o_5) $\lambda_{4,4} = \{t_3,t_4\}$ => d_4 added to I_4

Only B_k with **S_k=CB** are considered as `DynBimax` output

Figure 6.1: a) The Bimax biclusters of the whole static binary matrix. b) The state of B after 5 operations of DynBimax through a sliding window (size=3 and shift=1)

brevity in the sequel) needed to obtain DynBimax.

Given I a set of n documents in rows and J a set of m terms in columns, a binary matrix is defined as $X = \{e_{ij} \in \{0,1\}, \forall i \in [1..n], \forall j \in [1..m]\}$. A bicluster $B_k, k \in [1..\beta]$ is a submatrix $I_k \times J_k$ where $I_k \subset I$ and $J_k \subset J$ and for which all entries $e_{ij} = 1, \forall i \in I_k, \forall j \in J_k$. Bimax verifies the maximal inclusion constraint (*MIC*) reported in Proposition 2 page 26, according to Prelić et al. [Prelić 2006]. The *MIC* ensures that no bicluster is completely included in another one (Figure 6.1a).

Bimax [Prelić 2006] is described in details in section 1.4.4 and in Algorithm 2 page 26. For the sake of clarity in this chapter we recall the principle of Bimax algorithm with the help of dynAdd in algorithm 3 and the Figure 6.1. For any new document $i \in I$ described by the set of terms $C_i = \{j \in J | e_{ij} = 1\}$, all biclusters $B_k \in B$ are scanned to look for any intersection $\lambda_{i,k}$ between C_i and the terms J_k. $\lambda_{i,k}$ is the maximal set of terms shared between i and I_k. The condition in line 6 extends $B_{k'}$ to its maximality by adding i in $I_{k'}$ (in Figure 6.1b, o_5 adds d_4 to B_4 due to $\lambda_{4,4} = J_4 = \{t_3, t_4\}$). Otherwise, a new bicluster is added, with the documents $I_k \cup \{i\}$ (B_6 during o_5 with $J_6 = \{t_3, t_4, t_5\}$). Finally, after updating B, the condition in line 13 creates a new unitary bicluster (UB) with $I_k = \{i\}$ and $J_k = C_i$ (B_5 during o_5). Note that $\forall i \in I$, $\exists B_k \in B | J_k = C_i$. This ensures that any new row will also be compared with all previously added rows. For the result, only the biclusters with a minimum number of rows (MinR) and columns (MinC) are kept.

To handle dynamic data in Algorithm 3, we defined two distinct methods: 1) dynAdd corresponds to Bimax with several changes highlighted in red and 2) dynRemove to remove any row i from each bicluster $B_k \in B$ containing it. If the *MIC* is no longer met, B_k must be removed too. To do so, B_k is compared to all other biclusters $B_{k'} \in B$, leading to a lengthy search in B. We first reduce the size of B, by integrating the MinC constraint in line 5 of dynAdd. For easy check of *MIC*, we introduce a state S_k taking four possible values for each bicluster: (1) UB for unitary biclusters with $I_k = \{i\}$ and $J_k = C_i$, (2) NB for non-constrained biclusters when $|I_k| < MinR$, (3) CB for constrained biclusters when $|I_k| > MinR$ and (4) REM when the bicluster is marked for removal ($|.|$ is the set cardinality). Only constrained biclusters (BC) are kept for the DynBimax result. The other

Table 6.1: Conditions changing S_k in $changeStateOnRemoving(k)$. $\perp$ is a contradiction.

Prior S_k	$\lvert I_k \rvert = 0$	$I_k = \{r\} \wedge C_r = J_k$	$I_k = \{r\} \wedge C_r \neq J_k$	$1 < \lvert I_k \rvert < MinR$	$\lvert I_k \rvert \geqslant MinR$
UB	**REM**	$\perp$	$\perp$	$\perp$	$\perp$
NB	$\perp$	**UB**	**REM**	NB	$\perp$
CB	$\perp$	**UB**	**REM**	**NB**	CB

states are used internally by the algorithm. This state is updated in `dynRemove` and `dynAdd` according to Tables 6.1 and Table 6.2. In Table 6.1, $|I_k| = 0$ results in a simple removal of B_k. But when one row $I_k = \{r\}$ remains, B_k must be kept as UB if $C_r = J_k$ and removed otherwise ($S_1 = REM$ during o_4 in Figure 6.1). The remaining conditions, i.e. $I_k > 1$, results in $S_k \in \{NB, CB\}$, depending on $MinR$. In this case, we have to check if B_k still verifies the *MIC*.

In Table 6.3 we analyze all the situations of inclusion between B_k and any $B_{k'} \in B, k' \neq k$. Only one condition (marked by *REM*) captures the *MIC* violation and removes B_k (line 10 of `dynRemove`). We demonstrate in section 6.3 that all other situations can be ignored. Finally, the modification of a row is handled by removing the outdated row, followed by adding the up-to-date version.

Algorithm 3 DynBimax

```
1:  dynAdd($C_i$)
2:  var:$B$
3:  for $k \in [1..|B|]$ do
4:      $\lambda_{i,k} := J_k \cap C_i$
5:      if $|\lambda_{i,k}| > MinC$ then
6:          if $\exists k' \in [1..|B|]$ with $J_{k'} = \lambda_{i,k}$ then
7:              $I_{k'} := I_{k'} \cup \{i\}$
8:              $changeStateOnAdding(C_i, k')$
9:          else
10:             $B_{|B|+1} = (I_k \cup \{i\}, \lambda_{i,k})$
11:             $B := B \cup \{B_{|B|+1}\}$
12:             $changeStateOnAdding(C_i, |B|)$
13: if $\nexists (I_{k'}, J_{k'}) \in B$ with $J_{k'} = C_i$ then
14:     $B_{|B|+1} = (\{i\}, C_i); B := B \cup \{B_{|B|+1}\}$
15:     $changeStateOnAdding(C_i, |B|)$
```

```
1:  global:$MinC$                                  ▷ Minimum #columns
2:  global:$MinR$                                  ▷ Minimum #rows
3:
4:  dynRemove($C_i$)
5:  var:$B$
6:  for all $k \in [1..|B|]$ with $i \in I_k$ do
7:      $I_k := I_k \setminus \{i\}$
8:      $changeStateOnRemoving(k)$
9:      if $S_k = NB \vee S_k = CB$ then
10:         if $\exists k' \in [1..|B|]$ with $I_k = I_{k'} \wedge J_k \subset J_{k'}$
    then
11:             $S_k := REM$
12:     if $S_k = REM$ then
13:         $B := B \setminus \{B_k\}$
14:
15: DynBimax()
16: output:$\{B_k \in B | S_k = CB\}$
17:
18:
19:                         ▷ The changes in dynAdd for DynBimax
```

Table 6.2: Changing S_k in $changeStateOnAdding(k)$.

$\lvert I_k \rvert = 1$	$1 < \lvert I_k \rvert < MinR$	$\lvert I_k \rvert \geqslant MinR$
UB	NB	CB

Table 6.3: Only one condition must be checked between B_k and any $B_{k'} \in B$ for which MIC is not met after a row removal from B_k.

	$J_k \subset J_{k'}$	$J_k = J_{k'}$	$J_{k'} \subset J_k$	otherwise
$I_k \subset I_{k'}$	$\perp$	$\perp$	$\perp$	ignore
$I_k = I_{k'}$	**REM**	$\perp$	$\perp$	$\perp$
$I_{k'} \subset I_k$	keep	$\perp$	$\perp$	$\perp$
otherwise	$\perp$	$\perp$	$\perp$	keep

6.3 Proof of condition validity after and before remove

We assume that, before removing any row i from B_k, the prior biclustering B verifies the *MIC*. Axiom 1, ensured by line 6 of `dynAdd` in Algorithm 3, guarantees that each bicluster is extended by the rows to its maximality. Axiom 2, guaranteeing the exhaustiveness, is ensured by the lines 6 to 12.

Axiom 1. $\forall B_k \in B, \forall i \in [1..n] : C_i \cap J_k = J_k \Rightarrow i \in I_k$

Axiom 2. $\forall B_k \in B, \forall i \in [1..n] : C_i \cap J_k \neq \emptyset \Rightarrow \exists B_{k'} \in B : i \in I_{k'} \wedge J_{k'} = C_i \cap J_k$

Lemma 1. *Given two biclusters B_k and $B_{k'}$ with $k \neq k'$, $J_k \subset J_{k'} \Rightarrow I_{k'} \subset I_k$*

Proof. By definition, $\forall i' \in I_{k'}, \forall j' \in J_{k'}$: $j' \in C_{i'}$. In addition, $J_k \subset J_{k'} \Rightarrow \forall j \in J_k$: $j \in J_{k'}$ and thus, $j \in C_{i'}$. This implies that, $C_{i'} \cap J_k = J_k$. Thus, according to the Axiom 1, $\forall i' \in I_{k'}$: $i' \in I_k$, and $I_{k'} \subset I_k$. □

The symetric lemma is true:

Lemma 2. *Given two biclusters B_k and $B_{k'}$ with $k \neq k'$, $I_k \subset I_{k'} \Rightarrow J_{k'} \subset J_k$.*

For all the cases in Table 6.3, if the prior state does not verify the *MIC*, the condition can be ignored for the search.

1. $J_k = J_{k'}$ $(\neg MIC)$
   Prior: $J_k = J_{k'}$ $(\neg MIC \Rightarrow \bot)$
   $J_k = J_{k'}$ implies $\forall i \in I_k, C_i \cap J_{k'} = J_{k'}$ and $\forall i' \in I_{k'}, C_{i'} \cap J_k = J_k$. Axiom 1 implies $i \in I_{k'}$ and $i' \in I_k$ and thus $I_k = I_{k'}$ $(\neg MIC)$.

2. $I_k \subset I_{k'} \wedge J_k \subset J_{k'}$ $(\neg MIC)$
   Prior: $I_k \cup \{i\} \not\subseteq\not\supseteq I_{k'} \wedge J_k \subset J_{k'}$ $(\bot)$[1]
   Lemma 1 results in a contradiction: $I_{k'} \subset I_k \cup \{i\}$.

3. $I_k = I_{k'} \wedge J_k \subset J_{k'}$ $(\neg MIC \Rightarrow REM)$
   Prior: $I_{k'} \subset I_k \cup \{i\} \wedge J_k \subset J_{k'}$ (MIC)

4. $I_{k'} \subset I_k \wedge J_k \subset J_{k'}$ $(MIC \Rightarrow keep)$
   Prior: $I_{k'} \subset I_k \cup \{i\} \wedge J_k \subset J_{k'}$ (MIC)

5. $I_k \not\subseteq\not\supseteq I_{k'} \wedge J_k \subset J_{k'}$ $(\bot)$
   Prior: $I_k \cup \{i\} \not\subseteq\not\supseteq I_{k'} \wedge J_k \subset J_{k'}$ $(\bot)$
   Lemma 1 results in a contradiction: $I'_k \subset I_k \cup \{i\}$.

6. $I_k \subset I_{k'} \wedge J_{k'} \subset J_k$ (MIC)
   Prior: $I_k \cup \{i\} \not\subseteq\not\supseteq I_{k'} \wedge J_{k'} \subset J_k$ $(\bot)$
   Lemma 1 results in a contradiction: $I_k \cup \{i\} \subset I_{k'}$.

7. $I_k = I_{k'} \wedge J_{k'} \subset J_k$ $(\neg MIC)$
   Prior: $I_{k'} \subset I_k \cup \{i\} \wedge J_{k'} \subset J_k$ $(\neg MIC \Rightarrow \bot)$

8. $I_{k'} \subset I_k \wedge J_{k'} \subset J_k$ $(\neg MIC)$
   Prior: $I_{k'} \subset I_k \cup \{i\} \wedge J_{k'} \subset J_k$ $(\neg MIC \Rightarrow \bot)$

9. $I_k \not\subseteq\not\supseteq I_{k'} \wedge J_{k'} \subset J_k$ $(\bot)$
   Prior: $I_k \cup \{i\} \not\subseteq\not\supseteq I_{k'} \wedge J_{k'} \subset J_k$ $(\bot)$
   Lemma 1 results in a contradiction: $I_k \cup \{i\} \subset I'_k$.

10. $I_k \subset I_{k'} \wedge J_{k'} \not\subseteq\not\supseteq J_k$ ($\bot \Rightarrow$ can be ignored)
    Prior: $I_k \cup \{i\} \not\subseteq\not\supseteq I_{k'} \wedge J_{k'} \not\subseteq\not\supseteq J_k$ (MIC)
    We can show with Axiome 2 that $\exists B_{k"} : I_{k"} = I_k \wedge J_{k"} = J_k \cup J_{k'}$, that meets the condition to remove B_k. But, Lemma 2 results in a contradiction: $J_{k'} \subset J_k$. This case can hence be ignored.

11. $I_k = I_{k'} \wedge J_{k'} \not\subseteq\not\supseteq J_k$ $(\bot)$
    Prior: $I_{k'} \subset I_k \cup \{i\} \wedge J_{k'} \not\subseteq\not\supseteq J_k$ $(\bot)$
    Lemma 2 results in a contradiction: $J_k \subset J_{k'}$.

12. $I_{k'} \subset I_k \wedge J_{k'} \not\subseteq\not\supseteq J_k$ $(\bot)$
    Prior: $I_{k'} \subset I_k \cup \{i\} \wedge J_{k'} \not\subseteq\not\supseteq J_k$ $(\bot)$
    Lemma 2 results in a contradiction: $J_k \subset J_{k'}$.

13. $I_k \not\subseteq\not\supseteq I_{k'} \wedge J_{k'} \not\subseteq\not\supseteq J_k$ $(MIC \Rightarrow keep)$
    Prior: $I_k \cup \{i\} \not\subseteq\not\supseteq I_{k'} \wedge J_{k'} \not\subseteq\not\supseteq J_k$ (MIC)

[1] $A \not\subseteq\not\supseteq B \Leftrightarrow A \not\subseteq B \wedge B \not\subseteq A$

We observe that except for conditions 3, 4, 10 and 13, the prior state is proved to be $\neg MIC$ or to result in a contradiction due to lemma 1 or lemma 2. However, B_k must be removed if prior state is MIC and the current state is $\neg MIC$. Only condition 3 leads to this situation.

6.4 Performance evaluation

we evaluate the computation time DynBimax compared to Bimax by varying the sliding windows. Our experiment runs on a real dataset comprising 3,992 news articles aggregated from multiple online news sources (BBC, CNN, Reuters, France24, Egypt Independent and Der Spiegel) from November 2 to November 16, 2015. The corpus is modeled by a *TF-IDF* matrix where the top 10,000 nouns and adjectives are kept. The matrix is binarized with a threshold $\texttt{Thr} = 5$. The sliding windows are defined with increasing sizes (from 100 to $1,000$ documents) and increasing shift ratios (5% to 20% of the window size). In this way, each window is moved over all documents of the corpus. At each step, we run both algorithms with MinR=2 and MinC=2. For Bimax, we measure the computation time of dynAdd over the entire window. For DynBimax, we cumulate the time of both dynRemove and dynAdd for the expiring documents and for the incoming ones. We aim to identify the shift ratio at which DynBimax surpasses Bimax. Figure 6.2 shows, for each shift size, how the computation time evolves as the window size grows. We observe that for a shift of less than 20%, DynBimax is faster than Bimax and the difference increases with the window size. However for a 20% shift, DynBimax reaches its limit because dynAdd is faster than dynRemove. For a given window size, the boxplots show a larger variation for Bimax. We account for it by the event burstiness typical of breaking news, e.g. Paris attacks in Step 13 of Figure 6.3. Such a major event commented by an increasing share of articles yields more biclusters and takes longer to process. However, for DynBimax we observe lower variations, indicating a better stability against news burstiness.

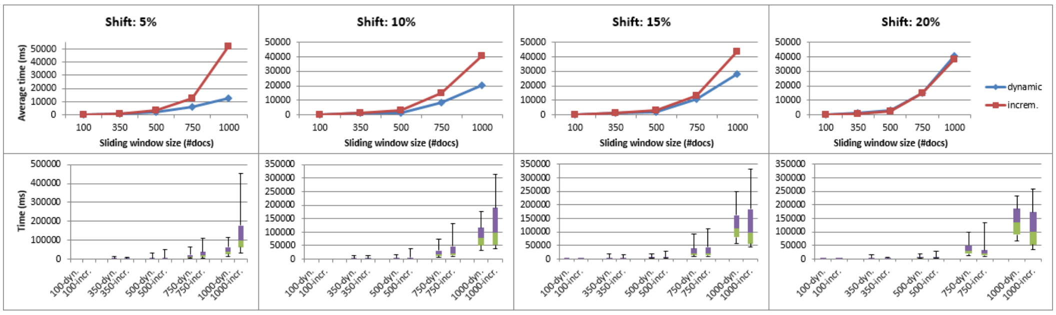


Figure 6.2: computation time of DynBimax vs. Bimax for different sizes and shifts of sliding widows.

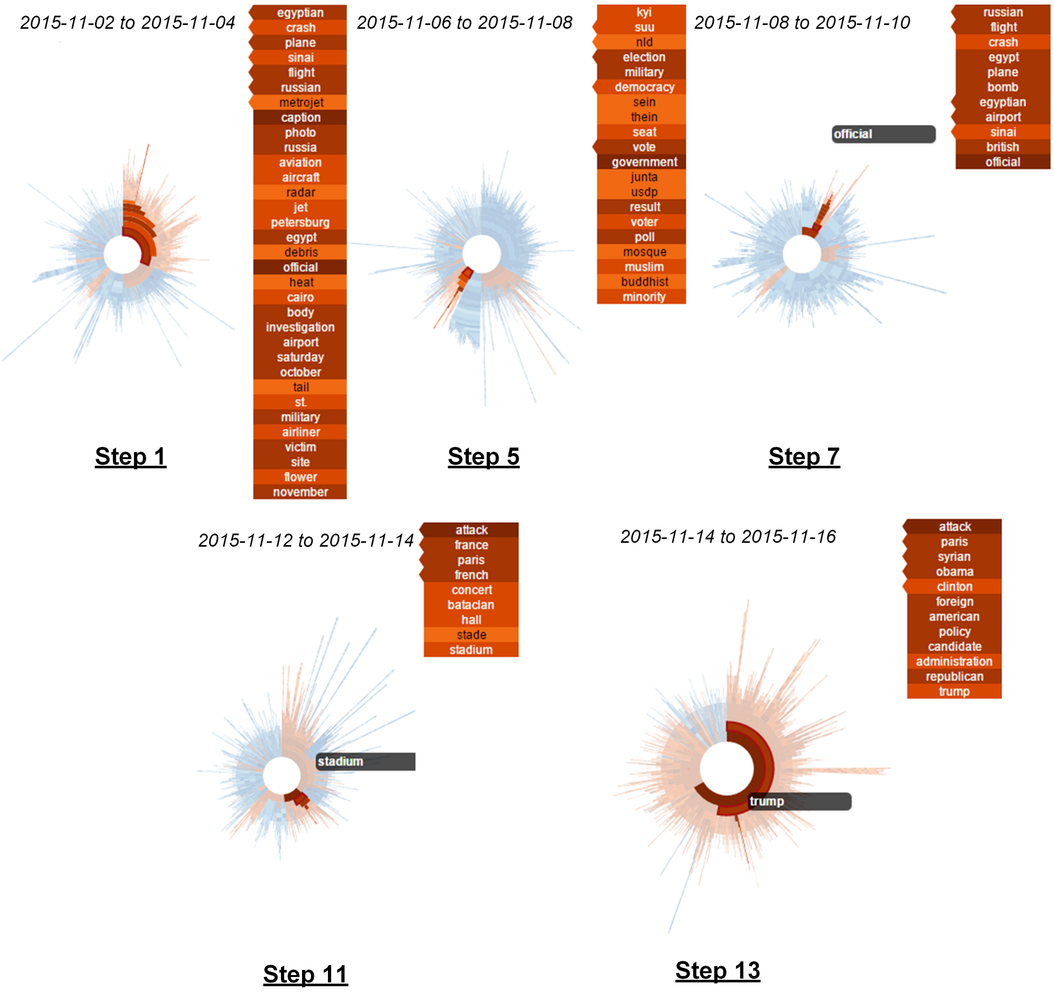


Figure 6.3: Visualization through a term hierarchy of DynBimax applied on news streams.

For each window step, Figure 6.3 shows the topic variants extracted by DynBimax and organized in a term hierarchy with the FPTree algorithm [Han 2004]. One variant is selected whose the term sequence is displayed by the breadcrumbs trail. Each path, from root to leaf, depicts one bicluster (B_k) described by its unique sequence of terms (J_k). The first terms of a sequence are closer to the root and group the biclusters and all the documents sharing them. They describe events or stories treated in the news. Less frequent terms are placed further away and describe a topic variant. Hovering over a node displays its term in a tooltip and its path on a breadcrumbs trail. Clicking a node retrieves all the documents sharing its path. We color in orange all the biclusters containing at least one of these documents in order to show their dispersion. After shifting the window, we systematically choose a topic by clicking the representative parent node and we highlight one topic variant. We can see that from step 1 to 7 the topic about a plane crash in Egypt is fading. Steps 11 and 13 show the surge of Paris attacks news and a topic variant about its impact on the U.S. presidential election campaign.

6.5 Conclusion

We presented DynBimax a dynamic overlapping biclustering, allowing to add, modify and remove objects from prior partitions of streaming news articles. Following a sliding window strategy, we showed through visualization that DynBimax enables the user to track the evolution of topics over time, while identifying specific viewpoints or facts. Our numerical experiment shows that DynBimax surpasses Bimax for a window shift of less than 20% of the window size. But, DynBimax trades computation-time stability to better cope with the surge of breaking news. Finally, we identified the necessary and sufficient condition for which the maximal inclusion constraint is not satisfied after removing a row. In future work, we envision to define a structure pruning the search space to faster target biclusters meeting this condition.

Conclusion and perspectives

This thesis presents a novel visual analytics approach for the exploration of text corpora. In particular, the developed software supports the process carried out by investigative journalists for hypothesis verification, refining and generation. Through the user-centered methodology outlined in chapter 3, we designed a multi-resolution system based on two nested bicluster structures. The top-level biclusters summarize the corpus by grouping similar documents within coarse-grained topics described by subsets of representative terms. The low-level biclusters reveal fine-grained *topic variants* within each topic, i.e. all optimal co-occurrences of terms shared by multiple documents. The visual analytics software is composed of two main visual components. The *Topic Weighted Map* gives an overview of the field under investigation. The *Topic Variant Overview* allows drilling down *topic variants* for searching multiple facts, angles or viewpoints related to stories found in the corpus.

The survey of text mining techniques in Chapter 1 elicits the different assumptions underlying the topic models and biclustering mechanisms. Both methods consider different kinds of co-occurrence patterns which inform about how they deal with the distributional hypothesis. In particular, topic models consider only high-order co-occurrences while biclustering methods additionally promote the consolidated co-occurrences strengthening context similarity involved in the distributional hypothesis. This difference has an impact on the topical structure delivered by both approaches and explains a lot of observations and results obtained from our experiments in chapters 3 and 4.

Chapter 3 describes the nested biclustering approach supporting the muti-resolution analysis of text corpora. This approach is flexible enough to handle any top-level topic shapes, i.e. biclusters overlapping or being disjoint, document-wise and/or term-wise. To analyze the relationships between coarse-grained topics, we developed a new similarity metric for both overlapping and disjoint biclusters. We defined an evaluation protocol based on multiple intrinsic metrics characterizing the properties of the coarse topical structure. We used `hLDA` as a baseline to assess the strengths and weaknesses of `Coclus`, the diagonal biclustering method we leverage for topic extraction. We found that `hLDA`, which extracts more high-order co-occurrences, identifies more often large topics with more distant document relationships and a comparatively larger number of terms. In contrast, `Coclus` spots additionally consolidated co-occurrences and thus the terms of topics are more document-specific. Multiple metrics show that large topics of `Coclus` enclose more specific *topic variants* with more commonalities and slight specificities. We hence anticipate that `Coclus` deals better with major news topics. The result of this numerical experiment guides the design of the user study presented in chapter 4 in order assess the impact of these differences on the comprehension of topics. Typically, we need to consider the topic size as an independent variable, and hence fairly observe the quality of the topic interpretation given by both methods. Future work would consist in extending our analysis to encompass full-overlapping topic extraction methods such as LDA, as well as overlapping biclustering methods [Shafiei 2006a].

The visual design of every component of our visual analytics software is presented in chapter 4. Through the proposed visualizations, the analyst can access the corpus at multiple levels of detail. For high-level topic overview, we designed a new visualization, the *Topic Weighted Map*, that combines an MDS projection, a *Weighted Map* visualization and multiple tag clouds. The plethora of overlapping term-document biclusters representing *topic variants* are organized in the term hierarchy of the *Topic Variant Overview*. The latter alleviates term redundancies and exposes the commonalities and specificities of *topic variants*. We also enable the domain expert to steer the *Bimax* algorithm until finding interesting results. Multiple interactions and coordinated views let the journalist initiate a focus process on specific aspects to verify hypotheses, and then engage in a diversification process to discover new aspects and refine her working hypothesis. We experimented with a nested biclustering approach, leveraging hard biclustering and overlapping biclustering algorithms. A usage scenario and a qualitative evaluation showed that, through the *Weighted Topic Map* the journalist can quickly identify topics of interest, and through the *Topic Variant Overview*, coordinated with the *Topic Variant Comparator*, the analyst can explore a large number of *Topic Variants* and find useful facts or viewpoints. Overall, the system supports the analysis of corpora that are significantly larger than journalists may be used to. Future work includes needs to improve the understanding of topic frontiers and to provide interactions to change them. We also aim to conduct a user study to compare our directional labeling with excentric labeling in the *Topic Variant Overview*.

The *Topic Variant Overview* and the tag clouds of *Topic Weighted Map* provide both common base to analyze the impact on topic comprehension of the size and structure of topic elicited by two different topic extraction methods. In chapter 4, we conducted a user study comparing Colclus and hLDA. Based on the results of our numerical experiment outlined in chapter 3, we expressed hypotheses, confirmed in part by the user study. We expected small topics to be easier to interprete with hLDA whereas Coclus might surpasses hLDA for large topics. The study reveals that small Coclus topics hold specific vocabulary and are more difficult to interpret without prior knowledge, whereas small hLDA topics contain richer vocabulary giving more context for the analyst's comprehension. The superiority of Coclus with large topics is not clearly established but several independent results converge towards the following trends. When large topics of hLDA yield themes with heterogeneous *topic variants*, large Coclus topics uncover more specific and more separable events with slight differences between *topic variants*. We infer that Coclus is more suitable for hypotheses verification while hLDA is more relevant for hypotheses generation. Short-term extensions of this work include a user study with investigative journalists in order to confirm our assumptions about the suitability of the Coclus and hLDA methods for high-level tasks such as verifying, refining and generating hypotheses.

Chapters 5 is an ongoing work for twitter-like text stream investigation, initiated during 2014 VAST Challenge. The visual analytics tool we proposed for the contest allows the user to achieve situation awareness and exploration tasks for text streams. Through a case study, we show the user's ability to discover, locate and explain different events in their temporal context. In this activity, the monitoring and exploration of all message aspects appear to be crucial. These tasks are supported by a flexible and dynamic *VSM* underlying the web-oriented architecture we propose. Nonetheless, we have highlighted several limitations in this solution, among which the lack of accurate temporal context for a selected period, the limitation of single-aspect river views and the large use of memory by our model and architecture.

Since in VAST challenge MC3, the Microblogs messages contain semi-structured content, our visualizations exploit them directly to explore different aspects of twitter-like data. Only fast analytic processing is done such as tokenization, part-of-speech tagging and named entity extraction. However, uncovering the relationships between these aspects (groups of authors that talk about the same hashtags, the same places, the same persons, etc.) may be useful to better understand events. This may consist in taking two aspects of data as the two dimensions of a dynamic *Object-Context* frequency matrix and in applying biclustering to reveal their relationships. In this perspective a remaining challenge is to propose efficient ways to execute these analytic techniques in real time. Moreover, the large volume of data requires a multi-resolution approach relying on a dynamic nested structure.

We started such work with the lower level structure of our system, i.e. the *topic variants*. Indeed, in chapter 6 we propose DynBimax, an extension of Bimax algorithm that support dynamic text streams involving adding, modifying and removing documents. Both chapters 5 and 6 are early initiatives to deal with text streams. For dynamic data, a remaining challenge has to do with the design of comprehensible visual representations of complex relationships that do not overwhelm the user. A combination of matrix-based visualizations, multi-modal graphs and tree visulizations, with parsimonious animations that preserve the analyst's mental map, may help monitor multiple aspects of text streams and uncover insightful structural patterns. Finally, supplying interactions to put the user in the loop by steering the underlying algorithms and refining the results has the potential to lead to a novel and useful visual analytics tool to investigate text streams.

List of publications

Accepted papers

- Analyse exploratoire de corpus textuels pour le journalisme d'investigation, Nicolas Médoc, Mohammad Ghoniem and Mohamed Nadif. In EGC 2017, vol. RNTI-E-33, pp. 477-480.

- Exploration visuelle de variantes de sujet par une approche hybride de biclustering, Nicolas Médoc, Mohammad Ghoniem and Mohamed Nadif. In Actes de la 28ième conférence francophone sur l'Interaction Homme-Machine (IHM '16), ACM, pp. 103-114, DOI: 10.1145/3004107.3004116.

- Exploratory Analysis of Text Collections Through Visualization and Hybrid Biclustering, Nicolas Médoc, Mohammad Ghoniem and Mohamed Nadif. In Berendt B. et al. (eds) Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2016, Lecture Notes in Computer Science, vol 9853, Springer, Cham, pp. 59-62, DOI: 10.1007/978-3-319-46131-1_13.

- Vers une approche Visual analytics pour explorer les variantes de sujet d'un corpus, Nicolas Médoc, Mohammad Ghoniem and Mohamed Nadif. In EGC 2016, vol. RNTI-E-30, pp. 539-540.

- Visual Analytics of Text Streams Through Multiple Dynamic Frequency Matrices, Nicolas Médoc, Mickaël Stefas, Mohammad Ghoniem, Mohamed Nadif. In 2014 IEEE Conference on Visual Analytics Science and Technology (VAST), pp. 381-382, DOI: 10.1109/VAST.2014.7042576.

Under revision papers

- Things that Matter for Topic Comprehension: Size and Topic Model, Nicolas Médoc, Mohammad Ghoniem and Mohamed Nadif. Selected for the fast track review process of IEEE Transaction on Visualization and Computer Graphics (TVCG) Journal.

Bibliography

[Aggarwal 2012a] C.C. Aggarwal and C.X. Zhai. Mining text data. Springer US, 2012. (Cited on pages 3, 13 and 15.)

[Aggarwal 2012b] C.C. Aggarwal and C.X. Zhai. *A Survey of Text Clustering Algorithms*. In Mining Text Data, pages 77–128. Springer US, 2012. DOI: 10.1007/978-1-4614-3223-4_4. (Cited on pages 3, 18 and 21.)

[Aggarwal 2013] C.C. Aggarwal and C.K. Reddy. Data Clustering: Algorithms and Applications. CRC Press, August 2013. (Cited on page 18.)

[Ailem 2015] 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. (Cited on pages 23, 111 and 147.)

[Ailem 2016] M. Ailem, F. Role and M. Nadif. *Graph modularity maximization as an effective method for co-clustering text data*. Knowledge-Based Systems, vol. 109, pages 160–173, 2016. (Cited on pages 6, 22, 23, 24, 28, 29, 61, 67, 85 and 99.)

[Aldous 1985] D.J. Aldous. *Exchangeability and related topics*. In École d'été de Probabilités de Saint-Flour XIII-1983, pages 1–198. Springer, 1985. (Cited on page 17.)

[Alencar 2012] A.B. Alencar, M.C. de Oliveira and F.V. Paulovich. *Seeing beyond reading: a survey on visual text analytics*. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 2, no. 6, pages 476–492, November 2012. (Cited on page 32.)

[Aletras 2013] N. Aletras and M. Stevenson. *Evaluating topic coherence using distributional semantics*. In Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers, pages 13–22, 2013. (Cited on page 30.)

[Alexander 2014] E. Alexander, J. Kohlmann, R. Valenza, M. Witmore and M. Gleisher. *Serendip: Topic model-driven visual exploration of text corpora*. In 2014 IEEE Conference on Visual Analytics Science and Technology (VAST), pages 173–182, October 2014. (Cited on pages vii, 48, 50, 51, 60, 67 and 81.)

[Alexander 2015] E. Alexander and M. Gleicher. *Task-Driven Comparison of Topic Models*. IEEE TVCG, vol. PP, no. 99, pages 1–1, 2015. (Cited on pages 6, 30, 48, 49, 67 and 84.)

[Archambault 2013] D. Archambault, D. Greene and P. Cunningham. *TwitterCrowds: Techniques for Exploring Topic and Sentiment in Microblogging Data*. CoRR, vol. abs/1306.3839, 2013. (Cited on pages vi, 34, 44 and 45.)

[Archambault 2016] D. Archambault and H.C. Purchase. *On the effective visualisation of dynamic attribute cascades*. Information Visualization, vol. 15, no. 1, pages 51–63, 2016. (Cited on page 42.)

[Bach 2014] B. Bach, E. Pietriga and J. D. Fekete. *GraphDiaries: Animated Transitions andTemporal Navigation for Dynamic Networks*. IEEE Transactions on Visualization and Computer Graphics, vol. 20, no. 5, pages 740–754, May 2014. (Cited on page 42.)

[Bach 2016] B. Bach, P. Dragicevic, D. Archambault, C. Hurter and S. Carpendale. *A Descriptive Framework for Temporal Data Visualizations Based on Generalized Space-Time Cubes*. Computer Graphics Forum, pages n/a–n/a, April 2016. (Cited on pages 37 and 41.)

[Barkow 2006] S. Barkow, S. Bleuler, A. Prelić, P. Zimmermann and E. Zitzler. *BicAT: a biclustering analysis toolbox*. Bioinformatics, vol. 22, no. 10, pages 1282–1283, May 2006. (Cited on pages 53 and 72.)

[Bertin 1977] J. Bertin and J. Bertin. La graphique et le traitement graphique de l'information. Number 91 (084.21) BER. 1977. (Cited on page 31.)

[Blei 2003] D.M. Blei, A.Y. Ng and M.I. Jordan. *Latent dirichlet allocation*. the Journal of machine Learning research, vol. 3, pages 993–1022, 2003. (Cited on pages 3, 16, 46, 48, 60 and 84.)

[Bostock 2013] M. Bostock. *D3-Sunburst*, 2013. (Cited on page 92.)

[Bradley 1997] P.S. Bradley, O.L. Mangasarian and W.N. Street. *Clustering via concave minimization*. Advances in neural information processing systems, pages 368–374, 1997. (Cited on page 19.)

[Brehmer 2013] M. Brehmer and T. Munzner. *A Multi-Level Typology of Abstract Visualization Tasks*. IEEE TVCG, vol. 19, no. 12, pages 2376–2385, December 2013. (Cited on pages 8, 50, 62, 63, 64 and 65.)

[Brehmer 2014] 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 TVCG, vol. 20, no. 12, pages 2271–2280, December 2014. (Cited on pages vii, 1, 21, 33, 49, 50, 51 and 91.)

[Brewer 2017] Brewer and Cynthia A. *ColorBrewer. http://www.ColorBrewer.org*, 2017. Online; accessed 01-jun-2017. (Cited on page 108.)

[Broniatowski 2013] D.A. Broniatowski, M.J. Paul and M. Dredze. *National and local influenza surveillance through Twitter: an analysis of the 2012-2013 influenza epidemic*. PloS one, vol. 8, no. 12, page e83672, 2013. (Cited on page 1.)

[Cambria 2014] E. Cambria and B. White. *Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article]*. IEEE Computational Intelligence Magazine, vol. 9, no. 2, pages 48–57, May 2014. (Cited on pages 2 and 13.)

[Celeux 1992] G. Celeux and G. Govaert. *A classification EM algorithm for clustering and two stochastic versions*. Computational statistics & Data analysis, vol. 14, no. 3, pages 315–332, 1992. (Cited on pages 20 and 25.)

[Chae 2014] J. Chae, D. Thom, Y. Jang, S.Y. Kim, T. Ertl and D.S. Ebert. *Public behavior response analysis in disaster events utilizing visual analytics of microblog data*. Computers & Graphics, vol. 38, pages 51–60, February 2014. (Cited on page 129.)

[Chang 2009] J. Chang, J. BoydGraber, S. Gerrish, C. Wang and D. Blei. *Reading Tea Leaves: How Humans Interpret Topic Models*. In Y. Bengio, D. Schuurmans and J.D. Lafferty, editors, Advances in Neural Information Processing Systems 22, pages 288–296. 2009. (Cited on page 29.)

[Cho 2004] H. Cho, I.S. Dhillon, Y. Guan and S. Sra. *Minimum sum-squared residue co-clustering of gene expression data*. In Proceedings of the 2004 SIAM International Conference on Data Mining, pages 114–125. SIAM, 2004. (Cited on page 22.)

[Choo 2013] J. Choo, C. Lee, C. K. Reddy and H. Park. *UTOPIAN: User-Driven Topic Modeling Based on Interactive Nonnegative Matrix Factorization*. IEEE TVCG, vol. 19, no. 12, pages 1992–2001, December 2013. (Cited on pages 46, 48, 60 and 84.)

[Chuang 2012] J. Chuang, D. Ramage, C. Manning and J. Heer. *Interpretation and trust: designing model-driven visualizations for text analysis*. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI ’12, pages 443–452, New York, NY, USA, 2012. ACM. (Cited on pages 46 and 48.)

[Church 1990] K.W. Church and P. Hanks. *Word Association Norms, Mutual Information, and Lexicography*. Comput. Linguist., vol. 16, no. 1, pages 22–29, March 1990. (Cited on page 15.)

[Collins 2009] C. Collins, F.B. 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. (Cited on pages v, 35 and 36.)

[Crain 2012] S.P. Crain, K. Zhou, S-H Yang and H. Zha. *Dimensionality Reduction and Topic Modeling: From Latent Semantic Indexing to Latent Dirichlet Allocation and Beyond*. In C.C. Aggarwal and C.X. Zhai, editors, Mining Text Data, pages 129–161. Springer US, 2012. (Cited on pages 3 and 16.)

[Cui 2011] Weiwei Cui, Shixia Liu, Li Tan, Conglei Shi, Yangqiu Song, Zekai Gao, Huamin Qu and Xin Tong. *TextFlow: Towards Better Understanding of Evolving Topics in Text*. IEEE TVCG, vol. 17, no. 12, pages 2412–2421, December 2011. (Cited on pages vi, 37 and 40.)

[Davies 2013] J. Davies. *D3-cloud. https://github.com/jasondavies/d3-cloud*, 2013. (Cited on page 90.)

[Deborah 2010] L.J. Deborah, R. Baskaran and A. Kannan. *A Survey on Internal Validity Measure for Cluster Validation*. International Journal of Computer Science and Engineering Survey, vol. 1, no. 2, pages 85–102, December 2010. (Cited on page 29.)

[Deerwester 1990] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer and Richard Harshman. *Indexing by latent semantic analysis*. Journal of the American Society for Information Science, vol. 41, no. 6, pages 391–407, September 1990. (Cited on page 16.)

[Dempster 1977] A.P. Dempster, N.M. Laird and D.B. Rubin. *Maximum likelihood from incomplete data via the EM algorithm*. Journal of the royal statistical society. Series B (methodological), pages 1–38, 1977. (Cited on pages 20 and 25.)

[Denef 2013] S. Denef, P.S. Bayerl and N.A. Kaptein. *Social Media and the Police: Tweeting Practices of British Police Forces During the August 2011 Riots*. In Proc. CHI'13, pages 3471–3480, New York, NY, USA, 2013. ACM. (Cited on pages 1 and 129.)

[Dhillon 2001a] I.S. Dhillon. *Co-clustering Documents and Words Using Bipartite Spectral Graph Partitioning*. In Proc. of the 7th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, KDD ’01, pages 269–274, New York, NY, USA, 2001. ACM. (Cited on page 22.)

[Dhillon 2001b] I.S. Dhillon and D.S. Modha. *Concept Decompositions for Large Sparse Text Data Using Clustering*. Machine Learning, vol. 42, no. 1, pages 143–175, 2001. (Cited on pages 19 and 68.)

[Dhillon 2003] Inderjit S. Dhillon, Subramanyam Mallela and Dharmendra S. Modha. *Information-theoretic Co-clustering*. In Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’03, pages 89–98, New York, NY, USA, 2003. ACM. (Cited on page 22.)

[Ding 2001] C.HQ Ding, X. He, H. Zha, M. Gu and H.D. Simon. *A min-max cut algorithm for graph partitioning and data clustering*. In Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on, pages 107–114. IEEE, 2001. (Cited on page 20.)

[Don 2007] Anthony Don, Elena Zheleva, Machon Gregory, Sureyya Tarkan, Loretta Auvil, Tanya Clement, Ben Shneiderman and Catherine 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. (Cited on pages vii and 49.)

[Dou 2011] W. Dou, X. Wang, R. Chang and W. Ribarsky. *ParallelTopics: A probabilistic approach to exploring document collections*. In 2011 IEEE VAST, pages 231–240, 2011. (Cited on pages vii, 4, 37, 46, 47, 48, 60 and 67.)

[Dou 2012] W. Dou, X Wang, D. Skau, W. Ribarsky and M.X. Zhou. *LeadLine: Interactive visual analysis of text data through event identification and exploration*. In 2012 IEEE VAST, pages 93–102, 2012. (Cited on pages vi and 40.)

[Dou 2013] Wenwen Dou, Li Yu, Xiaoyu Wang, Zhiqiang Ma and William Ribarsky. *HierarchicalTopics: Visually Exploring Large Text Collections Using Topic Hierarchies*. IEEE TVCG, vol. 19, no. 12, pages 2002–2011, 2013. (Cited on pages vi, 19, 41, 91 and 113.)

[Dumais 1992] Susan T. Dumais and Jakob Nielsen. *Automating the Assignment of Submitted Manuscripts to Reviewers*. In Proceedings of the 15th Annual International ACM SIGIR, SIGIR ’92, pages 233–244, New York, NY, USA, 1992. ACM. (Cited on page 16.)

[Endert 2012] A. Endert, P. Fiaux and C. North. *Semantic Interaction for Sensemaking: Inferring Analytical Reasoning for Model Steering*. IEEE Transactions on Visualization and Computer Graphics, vol. 18, no. 12, pages 2879–2888, December 2012. (Cited on pages 32 and 33.)

[Ester 1996] M. Ester, H-P Kriegel, J. Sander, X. Xu*et al. A density-based algorithm for discovering clusters in large spatial databases with noise.* In Kdd, volume 96, pages 226–231, 1996. (Cited on page 19.)

[Everitt 2011] B.S. Everitt, S. Landau, M. Leese and D. Stahl. Cluster Analysis. John Wiley & Sons, 2011. (Cited on page 18.)

[Fekete 1999] J-D Fekete and C. Plaisant. *Excentric Labeling: Dynamic Neighborhood Labeling for Data Visualization*. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '99, pages 512–519, New York, NY, USA, 1999. ACM. (Cited on pages 37 and 109.)

[Fekete 2003] J Fekete, D Wang, Niem Dang, Aleks Aris and Catherine Plaisant. *Interactive poster: Overlaying graph links on treemaps*. In Proceedings of the IEEE Symposium on Information Visualization Conference Compendium (InfoVisŠ 03), pages 82–83, 2003. (Cited on page 91.)

[Fiaux 2013] P. Fiaux, M. Sun, L. Bradel, C. North, N. Ramakrishnan and A. Endert. *Bixplorer: Visual Analytics with Biclusters*. Computer, vol. 46, no. 8, pages 90–94, August 2013. (Cited on pages viii, 54 and 55.)

[Fulda 2015] J. Fulda, M. Brehmer and T. Munzner. *TimeLineCurator: Interactive Authoring of Visual Timelines from Unstructured Text*. IEEE Transactions on Visualization and Computer Graphics, vol. PP, no. 99, pages 1–1, 2015. (Cited on page 35.)

[Gambette 2010] Philippe Gambette and Jean Véronis. *Visualising a Text with a Tree Cloud*. In Hermann Locarek-Junge and Claus Weihs, editors, Classification as a Tool for Research, pages 561–569. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010. (Cited on pages v, 35 and 36.)

[Gan 2014] Q. Gan, M. Zhu, M. Li, T. Liang, Y. Cao and B. Zhou. *Document visualization: an overview of current research*. Wiley Interdisciplinary Reviews: Computational Statistics, vol. 6, no. 1, pages 19–36, January 2014. (Cited on page 32.)

[Gerrish 2010] S. Gerrish and D.M. Blei. *A language-based approach to measuring scholarly impact*. In Proceedings of the 27th International Conference on Machine Learning (ICML-10), pages 375–382, 2010. (Cited on page 1.)

[Ghoniem 2005] Mohammad Ghoniem, Jean-Daniel Fekete and Philippe Castagliola. *On the Readability of Graphs Using Node-Link and Matrix-Based Representations: A Controlled Experiment and Statistical Analysis*. Information Visualization, vol. 4, no. 2, pages 114–135, June 2005. (Cited on page 53.)

[Ghoniem 2007] M. Ghoniem, D. Luo, J. Yang and W. Ribarsky. *NewsLab: Exploratory Broadcast News Video Analysis*. In 2007 IEEE VAST, pages 123–130, October 2007. (Cited on pages vi, 37, 38 and 39.)

[Ghoniem 2015] Mohammad Ghoniem, Maël Cornil, Bertjan Broeksema, Mickaël Stefas and Benoît 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. (Cited on page 91.)

[Govaert 1995] G. Govaert. *Simultaneous clustering of rows and columns*. Control and Cybernetics, vol. 24, pages 437–458, 1995. (Cited on page 22.)

[Govaert 2009] G. Govaert. Data analysis. ISTE & Wiley, series editor édition, 2009. (Cited on page 21.)

[Govaert 2013] G. Govaert and M. Nadif. Co-Clustering. ISTE & Wiley, series editor édition, 2013. (Cited on pages 3, 21, 24, 25, 28, 60 and 147.)

[Greene 2010] D. Greene and P. Cunningham. *Spectral co-clustering for dynamic bipartite graphs*. In Paper presented at the Workshop on Dynamic Networks and Knowledge Discovery (DyNAK 2010) at the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2010), Barcelona, September 24th 2010. Sun SITE Central Europe (CEUR), 2010. (Cited on page 147.)

[Gretarsson 2012] B. Gretarsson, J. O'Donovan, S. Bostandjiev, T. Höllerer, A. Asuncion, D. Newman and P. Smyth. *TopicNets: Visual Analysis of Large Text Corpora with Topic Modeling*. ACM Trans. Intell. Syst. Technol., vol. 3, no. 2, pages 23:1–23:26, February 2012. (Cited on pages vi and 42.)

[Griffiths 2004] T.L. Griffiths, M.I. Jordan, J.B. Tenenbaum and D.M. Blei. *Hierarchical Topic Models and the Nested Chinese Restaurant Process*. In Advances in Neural Information Processing Systems 16, pages 17–24. MIT Press, 2004. (Cited on pages 6, 17, 28 and 85.)

[Halvey 2007] M. Halvey and M. Keane. *An assessment of tag presentation techniques*. In Proceedings of the 16th international conference on World Wide Web, pages 1313–1314. ACM, 2007. (Cited on page 34.)

[Han 2004] 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, vol. 8, no. 1, pages 53–87, 2004. (Cited on pages 73, 92 and 152.)

[Hanczar 2011] B. Hanczar and M. Nadif. *Using the bagging approach for biclustering of gene expression data*. vol. 74, no. 10, pages 1595–1605, 2011. (Cited on page 69.)

[Harris 1954] Z.S. Harris. *Distributional structure*. Word, vol. 10, no. 2-3, pages 146–162, 1954. (Cited on pages 3 and 14.)

[Havre 2002] S. Havre. *ThemeRiver: visualizing thematic changes in large document collections*. IEEE TVCG, vol. 8, no. 1, pages 9–20, 2002. (Cited on pages v, 38, 39 and 131.)

[Heiberger 2014] R.M. Heiberger and N.B. Robbins. *Design of Diverging Stacked Bar Charts for Likert Scales and Other Applications*. Journal of Statistical Software, vol. 57, no. 5, pages 1–32, 2014. (Cited on pages xi and 120.)

[Heinrich 2011] Julian Heinrich, Robert Seifert, Michael Burch and Daniel Weiskopf. *BiCluster Viewer: A Visualization Tool for Analyzing Gene Expression Data*. In George Bebis, Richard Boyle, Bahram Parvin, Darko Koracin, Song Wang, Kim Kyungnam, Bedrich Benes, Kenneth Moreland, Christoph Borst, Stephen DiVerdi, Chiang Yi-Jen and Jiang Ming, editors, Advances in Visual Computing, number 6938 de Lecture Notes in Computer Science, pages 641–652. Springer Berlin Heidelberg, September 2011. (Cited on pages viii, 53, 54 and 92.)

[Hofmann 1999] T. Hofmann. *Probabilistic Latent Semantic Indexing*. In Proc. of the 22Nd Annual International ACM SIGIR, SIGIR '99, pages 50–57, New York, NY, USA, 1999. ACM. (Cited on page 16.)

[Horta 2014] D. Horta and R. J. G. B. Campello. *Similarity Measures for Comparing Biclusterings*. IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 11, no. 5, pages 942–954, September 2014. (Cited on pages 23 and 29.)

[Hotho 2005] A. Hotho, A. Nürnberger and G. Paaß. *A brief survey of text mining*. In Ldv Forum, volume 20, pages 19–62, 2005. (Cited on pages 3 and 14.)

[Huang 2008] A. Huang. *Similarity measures for text document clustering*. In Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand, pages 49–56, 2008. (Cited on pages 18 and 19.)

[Isenberg 2013] T. Isenberg, P. Isenberg, J. Chen, M. Sedlmair and T. Moller. *A Systematic Review on the Practice of Evaluating Visualization*. IEEE Transactions on Visualization and Computer Graphics, vol. 19, no. 12, pages 2818–2827, December 2013. (Cited on pages 29 and 62.)

[Kandogan 2013] E. Kandogan, D. Soroker, S. Rohall, P. Bak, F. van Ham, J. Lu, H-J Ship, C-F Wang and J. Lai. *A reference web architecture and patterns for real-time visual analytics on large streaming data*. In Proc. SPIE, VDA 2014, volume 901708. IS&P, 2013. (Cited on pages 131, 132 and 133.)

[Kaufman 1987] L. Kaufman and P. Rousseeuw. Clustering by means of medoids. North-Holland, 1987. (Cited on page 19.)

[Kaufman 2009] L. Kaufman and P.J. Rousseeuw. Finding groups in data: an introduction to cluster analysis, volume 344. John Wiley & Sons, 2009. (Cited on page 19.)

[Keim 2010] D.A. Keim, F. Mansmann and J. Thomas. *Visual Analytics: How Much Visualization and How Much Analytics?* SIGKDD Explor. Newsl., vol. 11, no. 2, pages 5–8, May 2010. (Cited on pages 2 and 32.)

[Kim 2011] K. Kim, S. Ko, N. Elmqvist and D. S. Ebert. *WordBridge: Using Composite Tag Clouds in Node-Link Diagrams for Visualizing Content and Relations in Text Corpora*. In 2011 44th Hawaii International Conference on System Sciences (HICSS), pages 1–8, January 2011. (Cited on pages vi, 43 and 44.)

[Krstajić 2013] M. Krstajić, M. Najm-Araghi, F. Mansmann and D.A. Keim. *Story Tracker: Incremental visual text analytics of news story development*. Information Visualization, vol. 12, no. 3-4, pages 308–323, July 2013. (Cited on pages 21 and 131.)

[Kucher 2015] K. Kucher and A. Kerren. *Text visualization techniques: Taxonomy, visual survey, and community insights*. In 2015 IEEE Pacific Visualization Symposium (PacificVis), pages 117–121, April 2015. (Cited on page 32.)

[L. Hunter 2011] Mark L. Hunter, Nils Hanson, Sabbagh Rana, Luuk Sengers, Drew Sullivan and Pia Thordsen. Story-Based Inquiry: A manual for investigative journalists. http://markleehunter.free.fr/documents/SBI_english.pdf. 2011. (Cited on pages 50, 63 and 86.)

[Labiod 2011] L. Labiod and M. Nadif. *Co-clustering for binary and categorical data with maximum modularity*. In Data Mining (ICDM), 2011 IEEE 11th International Conference on, pages 1140–1145. IEEE, 2011. (Cited on page 22.)

[Lee 1999] Daniel D Lee and H Sebastian Seung. *Learning the parts of objects by non-negative matrix factorization*. Nature, vol. 401, no. 6755, pages 788–791, 1999. (Cited on pages 48, 60 and 84.)

[Lee 2010] Bongshin Lee, N.H. Riche, A.K. Karlson and S. Carpendale. *SparkClouds: Visualizing Trends in Tag Clouds*. IEEE TVCG, vol. 16, no. 6, pages 1182–1189, 2010. (Cited on pages v, 34 and 35.)

[Lee 2012] H. Lee, J. Kihm, J. Choo, J. Stasko and H. Park. *iVisClustering: An Interactive Visual Document Clustering via Topic Modeling*. Computer Graphics Forum, vol. 31, no. 3pt3, pages 1155–1164, 2012. (Cited on pages vii, 4, 46, 47, 48, 60 and 67.)

[Leighton 1988] T. Leighton and S. Rao. *An approximate max-flow min-cut theorem for uniform multicommodity flow problems with applications to approximation algorithms*. In Foundations of Computer Science, 1988., 29th Annual Symposium on, pages 422–431. IEEE, 1988. (Cited on page 20.)

[Liu 2009] S. Liu, M.X. Zhou, S. Pan, W. Qian, W. Cai and X. Lian. *Interactive, Topic-based Visual Text Summarization and Analysis*. In Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM '09, pages 543–552, New York, NY, USA, 2009. ACM. (Cited on page 39.)

[Liu 2010] Y. Liu, Z. Li, H. Xiong, X. Gao and J. Wu. *Understanding of Internal Clustering Validation Measures*. In 2010 IEEE 10th International Conference on Data Mining (ICDM), pages 911–916, December 2010. (Cited on pages 18 and 29.)

[Liu 2014] Shixia Liu, Xiting Wang, Jianfei Chen, Jun Zhu and Baining Guo. *Topic-Panorama: A full picture of relevant topics*. In 2014 IEEE Conference on Visual Analytics Science and Technology (VAST), pages 183–192, October 2014. (Cited on pages vi and 45.)

[Luo 2012] D. Luo, J. Yang, M. Krstajic, W. Ribarsky and D. Keim. *EventRiver: Visually Exploring Text Collections with Temporal References*. IEEE Transactions on Visualization and Computer Graphics, vol. 18, no. 1, pages 93–105, January 2012. (Cited on pages vi and 40.)

[MacQueen 1967] J. MacQueen*et al.* *Some methods for classification and analysis of multivariate observations*. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, volume 1, pages 281–297. Oakland, CA, USA., 1967. (Cited on page 19.)

[Madeira 2004] S.C. Madeira and A.L. Oliveira. *Biclustering Algorithms for Biological Data Analysis: A Survey*. IEEE/ACM Trans. Comput. Biol. Bioinformatics, vol. 1, no. 1, pages 24–45, January 2004. (Cited on pages 3, 21, 60 and 147.)

[Manning 2014] C.D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S.J. Bethard and D. McClosky. *The Stanford CoreNLP Natural Language Processing Toolkit*. In Association for Computational Linguistics (ACL) System Demonstrations, pages 55–60, 2014. (Cited on pages 14, 66 and 111.)

[Marcus 2011] A. Marcus, M.S. Bernstein, O. Badar, D.R. Karger, S. Madden and R.C. Miller. *Twitinfo: Aggregating and Visualizing Microblogs for Event Exploration*. In Proc. CHI'11, pages 227–236. ACM, 2011. (Cited on pages 5 and 130.)

[McCallum 2002] A.K. McCallum. *MALLET: A Machine Learning for Language Toolkit.*, 2002. (Cited on pages 68, 76 and 116.)

[McLachlan 2004] G. McLachlan and D. Peel. Finite mixture models. John Wiley & Sons, 2004. (Cited on page 21.)

[Médoc 2014] Nicolas Médoc, Mickaël Stefas, Mohammad Ghoniem and Mohamed Nadif. *Visual Analytics of Text Streams through Multiple Dynamic Frequency Matrices*. In IEEE VAST 2014, 2014. (Cited on pages 7 and 130.)

[Mimno 2011] D. Mimno, H.M. Wallach, E. Talley, M. Leenders and A. McCallum. *Optimizing semantic coherence in topic models*. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 262–272. Association for Computational Linguistics, 2011. (Cited on page 30.)

[Munzner 2009] T. Munzner. *A Nested Model for Visualization Design and Validation*. IEEE TVCG, vol. 15, no. 6, pages 921–928, November 2009. (Cited on pages viii, 50, 62, 63, 86 and 109.)

[Murtagh 2012] F. Murtagh and P. Contreras. *Algorithms for hierarchical clustering: an overview*. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 2, no. 1, pages 86–97, 2012. (Cited on page 19.)

[Navigli 2012] R. Navigli. *A Quick Tour of Word Sense Disambiguation, Induction and Related Approaches*. In SOFSEM 2012: Theory and Practice of Computer Science, number 7147 de Lecture Notes in Computer Science, pages 115–129. Springer Berlin Heidelberg, January 2012. (Cited on pages 3 and 17.)

[Newman 2004] M.EJ Newman and M. Girvan. *Finding and evaluating community structure in networks*. Physical review E, vol. 69, no. 2, page 026113, 2004. (Cited on page 20.)

[Newman 2010] D. Newman, Y. Noh, E. Talley, S. Karimi and T. Baldwin. *Evaluating Topic Models for Digital Libraries*. In Proceedings of the 10th Annual Joint Conference on Digital Libraries, JCDL '10, pages 215–224, New York, NY, USA, 2010. ACM. (Cited on page 29.)

[Ng 2001] A.Y. Ng, M.I. Jordan, Y. Weiss*et al. On spectral clustering: Analysis and an algorithm*. In NIPS, volume 14, pages 849–856, 2001. (Cited on page 20.)

[O'Connor 2010] B. O'Connor, R. Balasubramanyan, B.R. Routledge and N.A. Smith. *From tweets to polls: Linking text sentiment to public opinion time series*. ICWSM, vol. 11, no. 122-129, pages 1–2, 2010. (Cited on page 1.)

[Paatero 1994] P. Paatero and U. Tapper. *Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values*. Environmetrics, vol. 5, no. 2, pages 111–126, 1994. (Cited on page 16.)

[Pérez-Suárez 2013] A. Pérez-Suárez, J. Martínez-Trinidad, J.A. Carrasco-Ochoa and J.E. Medina-Pagola. *An algorithm based on density and compactness for dynamic overlapping clustering*. Pattern Recognition, vol. 46, no. 11, pages 3040–3055, November 2013. (Cited on page 147.)

[Prelić 2006] A. Prelić, S. Bleuler, P. Zimmermann, A. Wille, P. Bühlmann, W. Gruissem, L. Hennig, L. Thiele and E. Zitzler. *A systematic comparison and evaluation of biclustering methods for gene expression data*. Bioinformatics, vol. 22, no. 9, pages 1122–1129, May 2006. (Cited on pages 3, 21, 25, 26, 51, 71, 72, 147 and 148.)

[Rand 1971] W.M. Rand. *Objective Criteria for the Evaluation of Clustering Methods*. Journal of the American Statistical Association, vol. 66, no. 336, pages 846–850, 1971. (Cited on pages 23, 29 and 110.)

[Rivadeneira 2007] A. Rivadeneira, D. Gruen, M. Muller and D. Millen. *Getting our head in the clouds: toward evaluation studies of tagclouds*. In Proceedings of the SIGCHI conference on Human factors in computing systems, pages 995–998. ACM, 2007. (Cited on page 34.)

[Rohrdantz 2011] C. Rohrdantz, D. Oelke, M. Krstajić and F. Fischer. *Real-time visualization of streaming text data: Tasks and challenges*. In Proc. IEEE Workshop on Interactive Visual Text Analytics for Decision Making, 2011. (Cited on pages 4 and 130.)

[Role 2011] F. Role and M. Nadif. *Handling the Impact of Low Frequency Events on Co-occurrence Based Measures of Word Similarity - A Case Study of Pointwise Mutual Information*. pages 218–223, October 2011. (Cited on page 15.)

[Role 2014] F. Role and M. Nadif. *Beyond cluster labeling: Semantic interpretation of clusters contents using a graph representation*. Knowledge-Based Systems, vol. 56, pages 141–155, January 2014. (Cited on page 21.)

[Rufiange 2013] S. Rufiange and M.J. McGuffin. *DiffAni: Visualizing Dynamic Graphs with a Hybrid of Difference Maps and Animation*. IEEE Transactions on Visualization and Computer Graphics, vol. 19, no. 12, pages 2556–2565, December 2013. (Cited on page 42.)

[Saber 2015] H.B. Saber and M. Elloumi. *A New Study on Biclustering Tools, Biclusters Validation and Evaluation Functions*. International Journal of Computer Science & Engineering Survey, vol. 6, no. 1, pages 01–13, February 2015. (Cited on page 29.)

[Sacha 2016] D. Sacha, M. Sedlmair, L. Zhang, J. A. Lee, D. Weiskopf, S. North and D. Keim. *Human-centered machine learning through interactive visualization*. Bruges, Belgium, May 2016. (Cited on page 37.)

[Salton 1975] G. Salton, A. Wong and C. S. Yang. *A Vector Space Model for Automatic Indexing*. Commun. ACM, vol. 18, no. 11, pages 613–620, November 1975. (Cited on pages 3 and 13.)

[Santamaría 2008] R. Santamaría, R. Therón and L. Quintales. *A visual analytics approach for understanding biclustering results from microarray data*. BMC Bioinformatics, vol. 9, no. 1, page 247, May 2008. (Cited on pages viii, 53 and 91.)

[Schaeffer 2007] S.E. Schaeffer. *Graph clustering*. Computer science review, vol. 1, no. 1, pages 27–64, 2007. (Cited on page 20.)

[Schulz 2011] H. Schulz, S. Hadlak and H. Schumann. *The Design Space of Implicit Hierarchy Visualization: A Survey*. IEEE Transactions on Visualization and Computer Graphics, vol. 17, no. 4, pages 393–411, April 2011. (Cited on page 109.)

[Schulz 2013] H.J. Schulz, T. Nocke, M. Heitzler and H. Schumann. *A Design Space of Visualization Tasks*. IEEE Transactions on Visualization and Computer Graphics, vol. 19, no. 12, pages 2366–2375, December 2013. (Cited on pages 62 and 63.)

[Shafiei 2006a] M. Shafiei and E. Milios. *Model-based overlapping co-clustering*. In Proceeding of SIAM Conference on Data Mining, 2006. (Cited on pages 4, 21 and 156.)

[Shafiei 2006b] M.M. Shafiei and E.E. Milios. *Latent Dirichlet Co-Clustering*. In Sixth International Conference on Data Mining, 2006. ICDM '06, pages 542–551, December 2006. (Cited on page 28.)

[Shneiderman 1996] B. Shneiderman. *The eyes have it: a task by data type taxonomy for information visualizations*. In IEEE Symposium on Visual Languages, pages 336–343, September 1996. (Cited on pages 31, 35 and 110.)

[Stasko 2000] John Stasko, Richard Catrambone, Mark Guzdial and Kevin McDonald. *An evaluation of space-filling information visualizations for depicting hierarchical structures*. International Journal of Human-Computer Studies, vol. 53, no. 5, pages 663–694, November 2000. (Cited on page 92.)

[Stasko 2008] John Stasko, Carsten Görg and Zhicheng Liu. *Jigsaw: Supporting Investigative Analysis through Interactive Visualization*. Information Visualization, vol. 7, no. 2, pages 118–132, June 2008. (Cited on pages vii, 52 and 55.)

[Steinbach 2000] M. Steinbach, G. Karypis and V. Kumar. *A comparison of document clustering techniques*. In KDD workshop on Text Mining, 2000. (Cited on pages 3 and 19.)

[Strehl 2002] A. Strehl and J. Ghosh. *Cluster Ensembles:A Knowledge Reuse Framework for Combining Multiple Partitions*. Machine Learning Research, pages 583–617, 2002. (Cited on pages 23, 29 and 110.)

[Streit 2014] Marc Streit, Samuel Gratzl, Michael Gillhofer, Andreas Mayr, Andreas Mitterecker and Sepp Hochreiter. *Furby: fuzzy force-directed bicluster visualization*. BMC Bioinformatics, vol. 15, no. Suppl 6, page S4, May 2014. (Cited on pages viii, 53, 54 and 92.)

[Sun 2014] M. Sun, C. North and N. Ramakrishnan. *A Five-Level Design Framework for Bicluster Visualizations*. IEEE TVCG, vol. 20, no. 12, pages 1713–1722, December 2014. (Cited on pages 4, 22, 52, 53 and 74.)

[Sun 2015] M. Sun, P. Mi, C. North and N. Ramakrishnan. *BiSet: Semantic Edge Bundling with Biclusters for Sensemaking*. IEEE TVCG, vol. PP, no. 99, pages 1–1, 2015. (Cited on pages viii, 55 and 91.)

[Thomas 2005] J. Thomas and K. Cook. Illuminating the path: the research and development agenda for visual analytics. IEEE Press, 2005. (Cited on pages 2 and 32.)

[Thudt 2012] A. Thudt, U. Hinrichs and S. Carpendale. *The Bohemian Bookshelf: Supporting Serendipitous Book Discoveries Through Information Visualization*. In

Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '12, pages 1461–1470, New York, NY, USA, 2012. ACM. (Cited on page 50.)

[Tufte 2006] Edward R Tufte. *Beautiful evidence*. New York, 2006. (Cited on page 35.)

[Turney 2010] P.D. Turney and P. Pantel. *From Frequency to Meaning: Vector Space Models of Semantics*. J. Artif. Int. Res., vol. 37, no. 1, pages 141–188, January 2010. (Cited on pages 2, 3, 15, 29 and 131.)

[Vendramin 2010] L. Vendramin, R.J.G.B. Campello and E.R. Hruschka. *Relative clustering validity criteria: A comparative overview*. Statistical Analysis and Data Mining, vol. 3, no. 4, pages 209–235, August 2010. (Cited on page 29.)

[Viegas 2009] F. B. Viegas, M. Wattenberg and J. Feinberg. *Participatory Visualization with Wordle*. IEEE TVCG, vol. 15, no. 6, pages 1137–1144, November 2009. (Cited on pages v and 34.)

[Šilić 2010] A. Šilić and B.D. Bašić. *Visualization of Text Streams: A Survey*. In Knowledge-Based and Intelligent Information and Engineering Systems, pages 31–43. Springer, Berlin, Heidelberg, September 2010. (Cited on page 32.)

[Wallach 2009] H.M. Wallach, I. Murray, R. Salakhutdinov and D. Mimno. *Evaluation Methods for Topic Models*. In Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09, pages 1105–1112, New York, NY, USA, 2009. ACM. (Cited on page 29.)

[Wang 2009] P. Wang, C. Domeniconi and K.B. Laskey. *Latent dirichlet bayesian co-clustering*. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pages 522–537. Springer, 2009. (Cited on pages 18 and 28.)

[Wanner 2014] F. Wanner, A. Stoffel, D. Jäckle, B.C. Kwon, A. Weiler and D.A. Keim. *State-of-the-Art Report of Visual Analysis for Event Detection in Text Data Streams*. In R. Borgo, R. Maciejewski and I. Viola, editors, EuroVis - STARs. The Eurographics Association, 2014. (Cited on pages 4 and 32.)

[Wattenberg 2008] M. Wattenberg and F. B. Viégas. *The Word Tree, an Interactive Visual Concordance*. IEEE Transactions on Visualization and Computer Graphics, vol. 14, no. 6, pages 1221–1228, November 2008. (Cited on pages v, 35 and 109.)

[Wei 2010] Furu Wei, Shixia Liu, Yangqiu Song, Shimei Pan, Michelle X. Zhou, Weihong Qian, Lei Shi, Li Tan and Qiang 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. (Cited on pages vi, 37 and 39.)

[Weskamp 2004] Marcos Weskamp. *News Map. http://newsmap.jp/*, 2004. (Cited on pages vi and 43.)

[Wise 1995] J.A. Wise, J.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*. pages 51–58, October 1995. (Cited on pages v and 33.)

[Xu 2003] W. Xu, X. Liu and Y. Gong. *Document Clustering Based on Non-negative Matrix Factorization*. In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, SIGIR ’03, pages 267–273, New York, NY, USA, 2003. ACM. (Cited on page 16.)

[Xu 2016] P. Xu, N. Cao, H. Qu and J. Stasko. *Interactive visual co-cluster analysis of bipartite graphs*. In 2016 IEEE Pacific Visualization Symposium (PacificVis), pages 32–39, April 2016. (Cited on pages viii and 56.)