Visual analytics of topics in Twitter in connection with political debates

Eder José de Carvalho

Eder José de Carvalho

Visual analytics of topics in Twitter in connection with political debates

Master dissertation submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP, in partial fulfillment of the requirements for the degree of the Master Program in Computer Science and Computational Mathematics. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Profa. Dra. Maria Cristina Ferreira de Oliveira

USP – São Carlos
June 2017

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados fornecidos pelo(a) autor(a)

C331v

Carvalho, Eder José de
Análise visual de tópicos no Twitter em conexão com debates políticos / Eder José de Carvalho; orientadora Maria Cristina Ferreira de Oliveira. - São Carlos - SP, 2017.
81 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2017.

1. Visualization, Social data analysis, Text clustering, Text segmentation. I. Oliveira, Maria Cristina Ferreira de, orient. II. Título.

Eder José de Carvalho

Análise visual de tópicos no Twitter em conexão com debates políticos

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Ciências – Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientadora: Profa. Dra. Maria Cristina Ferreira de Oliveira

USP – São Carlos
Junho de 2017

I dedicate this work to my father, José Maria de Carvalho, and to my mother, Rosmeri Terezinha de Carvalho.

ACKNOWLEDGEMENTS

Agradeço imensamente à minha família que sempre me apoiou em minha jornada acadêmica, acreditando muito mais em mim do que eu mesmo;

à minha orientadora Maria Cristina pela oportunidade, por todos os ensinamentos e, principalmente, pela paciência e compreensão durante todos esses anos. Sou também muito grato pela experiência internacional que tive graças ao seu apoio (e motivo pelo qual o último parágrafo está em inglês);

à Prof. Rosane que esteve presente desde o começo de meu mestrado e que direta ou indiretamente contribuiu para o desenvolvimento deste trabalho;

à banca por ter aceitado o convite para a defesa e por todos os valiosos comentários e sugestões após a mesma;

ao CNPq, CAPES e ao *Emerging Leaders in the Americas Program* (ELAP) pelo apoio financeiro;

I would like to thank Prof. Evangelos Milios, for receiving me as a visiting research student in Dalhousie University and for taking so much of your time to help my stay in Halifax. I also would like to thank Axel Soto, Raheleh Makki and Lulu Huang for all the hard work and for the good times.

“As invenções são, sobretudo,
o resultado de um trabalho de teimoso.”
(Santos Dumont)

RESUMO

CARVALHO, E. J. **Análise visual de tópicos no Twitter em conexão com debates políticos**. 2017. 81 p. Master dissertation (Master student Program in Computer Science and Computational Mathematics) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2017.

Mídias sociais como o Twitter e o Facebook atuam, em diversas situações, como canais de iniciativas que buscam ampliar as ações de cidadania. Por outro lado, certas ações e manifestações na mídia convencional por parte de instituições governamentais, ou de jornalistas e políticos como deputados e senadores, tendem a repercutir nas mídias sociais. Como resultado, gera-se uma enorme quantidade de dados em formato textual que podem ser muito informativos sobre ações e políticas governamentais. No entanto, o público-alvo continua carente de boas ferramentas que ajudem a levantar, correlacionar e interpretar as informações potencialmente úteis associadas a esses textos. Neste contexto, este trabalho apresenta dois sistemas orientados à análise de dados governamentais e de mídias sociais. Um dos sistemas introduz uma nova visualização, baseada na metáfora do rio, para análise temporal da evolução de tópicos no Twitter em conexão com debates políticos. Para tanto, o problema foi inicialmente modelado como um problema de clusterização e um método de segmentação de texto independente de domínio foi adaptado para associar (por clusterização) *tweets* com discursos parlamentares. Uma versão do algorimo MONIC para detecção de transições entre agrupamentos foi empregada para rastrear a evolução temporal de debates (ou agrupamentos) e produzir um conjunto de agrupamentos com informação de tempo. O outro sistema, chamado ATR-Vis, combina técnicas de visualização com estratégias de "recuperação ativa" para envolver o usuário na recuperação de *tweets* relacionados a debates políticos e associa-os ao debate correspondente. O arcabouço proposto introduz quatro estratégias de "recuperação ativa" que utilizam informação estrutural do Twitter melhorando a acurácia do processo de recuperação e simultaneamente minimizando o número de pedidos de rotulação apresentados ao usuário. Avaliações por meio de casos de uso e experimentos quantitativos, assim como uma análise qualitativa conduzida com três especialistas ilustram a efetividade do ATR-Vis na recuperação de tweets relevantes. Para a avaliação, foram coletados dois conjuntos de *tweets* relacionados a debates parlamentares ocorridos no Brasil e no Canadá, e outro formado por um conjunto de notícias que receberam grande atenção da mídia no período da coleta.

Palavras-chave: Visualização, Análise de dados sociais, Clusterização de texto, Segmentação de texto.

ABSTRACT

CARVALHO, E. J. **Visual analytics of topics in Twitter in connection with political debates**. 2017. 81 p. Master dissertation (Master student Program in Computer Science and Computational Mathematics) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2017.

Social media channels such as Twitter and Facebook often contribute to disseminate initiatives that seek to inform and empower citizens concerned with government actions. On the other hand, certain actions and statements by governmental institutions, or parliament members and political journalists that appear on the conventional media tend to reverberate on the social media. This scenario produces a lot of textual data that can reveal relevant information on governmental actions and policies. Nonetheless, the target audience still lacks appropriate tools capable of supporting the acquisition, correlation and interpretation of potentially useful information embedded in such text sources. In this scenario, this work presents two system for the analysis of government and social media data. One of the systems introduces a new visualization, based on the river metaphor, for the analysis of the temporal evolution of topics in Twitter in connection with political debates. For this purpose, the problem was initially modeled as a clustering problem and a domain-independent text segmentation method was adapted to associate (by clustering) Twitter content with parliamentary speeches. Moreover, a version of the MONIC framework for cluster transition detection was employed to track the temporal evolution of debates (or clusters) and to produce a set of time-stamped clusters. The other system, named ATR-Vis, combines visualization techniques with active retrieval strategies to involve the user in the retrieval of Twitter's posts related to political debates and associate them to the specific debate they refer to. The framework proposed introduces four active retrieval strategies that make use of the Twitter's structural information increasing retrieval accuracy while minimizing user involvement by keeping the number of labeling requests to a minimum. Evaluations through use cases and quantitative experiments, as well as qualitative analysis conducted with three domain experts, illustrates the effectiveness of ATR-Vis in the retrieval of relevant tweets. For the evaluation, two Twitter datasets were collected, related to parliamentary debates being held in Brazil and Canada, and a dataset comprising a set of top news stories that received great media attention at the time.

Keywords: Visualization, Social data analysis, Text clustering, Text segmentation.

LIST OF FIGURES

CONTENTS

CHAPTER

1

INTRODUCTION

In recent years several government organizations in Brazil and the world have been committed to implement open data policies (Open Government Data, OGD) BREITMAN *et al.* (2012) in order to make data generated by such institutions available on the Web. The motivation is to increase the transparency of governmental actions and the capacity of citizens to accompany and interfere in government policies that affect their daily life. In 2011 Brazil became a member of the *Open Government Partnership*, a multinational initiative to promote the widespread adoption of OGD.

One of several actions resulting from this effort is the *Portal Brasileiro de Dados Abertos*[1] (Brazilian Open Data Portal), which maintains a catalog of files from diverse sources. The Brazilian Senate also maintains a *Portal de Atividade Legislativa*[2] (Portal of Legislative Activity), with many records on parliamentary activity. Therefore, there is a favorable scenario for expanding the quantity and diversity of data for public access and creating an "ecosystem" of users of such data. The range of potential "consumer" profiles of OGD is wide: application developers; journalists; scientists and scholars; officials of government organizations, private corporations that provide services to the government, or the government itself in its various instances; or even "ordinary" people interested in exercising their citizenship.

Research questions associated with open data have roots mainly in the scientific community working on themes related to the Semantic Web LOPEZ *et al.* (2013); GLORIA *et al.* (2013); KALAMPOKIS; TAMBOURIS; TARABANIS (2013), with focus on processing data from multiple sources in order to make them accessible for automatic processing by algorithms – with a well-defined access interface and a standard format – or by humans – through browsers, with the display of interpretable representations.

The standardization of data formats, for example, is critical, particularly to enable broad

[1] <http://dados.gov.br/> [Accessed 1 June 2017]
[2] <http://www25.senado.leg.br/web/atividade> [Accessed 1 June 2017]

access, interoperability and reuse of solutions. It is also true that the mere existence of OGD repositories does not guarantee the available data will be used in its full potential, which has motivated initiatives to stimulate the development of applications and tools[3], or the learning of use strategies and access tools[4]. Although such initiatives may be considered relatively successful, there is a concern that a significant portion of potential OGD users will not benefit from them because they can not perform the essential operations necessary to collect, process, integrate, and interpret the data GRAVES; HENDLER (2013). Computational Visualization techniques have great potential as facilitators in this scenario.

The discipline known as Computational Visualization came about with the dissemination of computers capable of generating interactive graphics CARD; MACKILAY; SHNEIDERMAN (1999); OLIVEIRA; LEVKOWITZ (2003); CHEN (2006b); WONG; THOMAS (2004). This research area has two aspects: Scientific Visualization techniques generally include the creation of graphical models of data from the physical world, obtained from the human body, the planet Earth, molecular models, among others. Such data embody spatial information that is determinant in the generation of visual representations. Data of a non-physical nature, without an inherent spatial mapping – such as financial data, document collections, and other abstract conceptions – are, on the other hand, typically mapped by Information Visualization techniques. It should be noted that these strands are not always clearly distinguishable, for example, geographical data are often associated with one or the other, depending on the context.

The motivation for Visualization is to facilitate the interpretation of data, whether simple or complex, and therefore visualization techniques and tools can be exploited to support the dissemination of OGD to end users. In this context, research questions concern not only storage formats or access patterns, but also the search for techniques to create representations that facilitate tasks that require identifying relevant information in the data and its interpretation.

In particular, we are interested in visualization applied to the analysis of textual contents. Previous contributions from the *Visualization, Imaging and Computer Graphics* (VICG) group in this area PAULOVICH *et al.* (2008); SALAZAR *et al.* (2013); GOMEZ-NIETO *et al.* (2014) motivated a collaborative project with the Research Group on *Machine Learning and Networked Information Spaces* (MALNIS), led by Prof. Evangelos Milios, from the University of Dalhousie, Canada, in "Visual Text Analytics", a topic of interest to both groups (FAPESP Grant 2013/50380-4). Both groups also collaborated on a project funded by the *Canada-Latin American and the Caribbean Research Exchange Grants* (LACREG), entitled "Visual Text Analytics for Open Government Data". This master's project has been developed in the context of these cooperation initiatives.

We propose investigating the application of visualization techniques and visual representations to support the analysis of open data provided in textual format, in particular parliamentary

[3] <http://wiki.dados.gov.br/II-Encontro-Nacional-de-Dados-Abertos.ashx> [Accessed 1 June 2017]
[4] <https://br.okfn.org/tag/escola-de-dados/> [Accessed 1 June 2017]

discourses and their relation with texts published in social media, such as *Twitter*.

1.1 Research Problem

Social media channels like *Twitter* and *Facebook* attract much interest from researchers in journalism and social sciences, due to their scope and impact. In several situations, they act as channels of initiatives that seek to broaden citizenship actions, possibly stimulated or triggered by political or social organizations. Professionals and volunteers from such organizations tend to be potential consumers of OGD, seeking support to base their actions and arguments.

On the other hand, certain actions and manifestations of politicians such as deputies and senators tend to reverberate in social media, and vice versa. The repercussion may also occur in conventional legislative channels, such as parliamentary speeches, proposed amendments, specific legislation, etc. These different media reflect discussions stemming from the multiple views, opinions and roles of the social actors involved.

In this scenario, it becomes particularly interesting to investigate how the interaction between these multiple manifestations occurs, and to try to understand how they affect the unfolding of events. However, this requires a significant effort in collecting and organizing the data - in this case, the texts associated with the various manifestations in multiple dissemination channels, and the information associated with such texts, i.e., metadata that identifies subject, place, publication vehicle, person, date, etc. Collection and organization are necessary activities to enable the subsequent interpretation and analysis of data, and all these steps require extensive effort and specific knowledge. The availability of adequate computational tools to support such tasks is essential to enable more studies and actions in this area.

In this project, our focus is precisely on this aspect of the problem: we believe that there is a lack of adequate tools for the potential analysts involved in this type of activity, and our working hypothesis is that text visualization solutions can contribute to overcoming this gap, facilitating access to data and interpretation. This effort requires the ability to correlate the content of speeches with what is published on social media, such as *Twitter*. Visualizations can highlight the topics, actors and attributes relevant to particular events, and allow the association of the occurrence of those topics with data from diverse sources.

1.2 Goals

The objective of this project is to provide empirical evidence on the applicability of visualization techniques in general and text visualization in particular to support analysts interested in collecting, correlating and interpreting textual data from governmental data portals (specifically, the Chamber and/or Senate) and social media publications such as Twitter. Several text visualization techniques were considered. As discussed in Chapter 2, most of them required

some kind of text preprocessing to generate a representation of the collection that can be manipulated by computational algorithms, typically a vector model obtained, for instance, from a frequency count of relevant terms SALTON; WONG; YANG (1975), or from some probabilistic topic extraction technique BLEI; NG; JORDAN (2003).

Achieving this goal required creating representative textual corpora. This corpora was built from the collection of parliamentary speeches available in the open data portals of the Chamber/Senate, as well as texts published by parties, parliamentarians and others (such as journalists, NGOs) on Twitter, on topics linked to proposals formally debated in Parliament and known to have generated discussion or controversy. Such texts are typically associated with complementary metadata (date, name and party of the parliamentarian, etc.), which can also be considered in creating visualizations. Using such corpora and complementary data as a departing point, visualization techniques applicable in this scenario were investigated, which are adequate to support data cleansing and organization as well as exploratory data analysis in an integrated manner. The objective was to favor analysis aimed at identifying and evidencing correlations between texts originating from different sources (the legislative house and the social network).

Along the investigation we aimed at identifying the typical exploratory tasks performed by potential data analysts, with a view to implementing the visualization techniques to support their execution. Validation was conducted considering possible case studies and usage scenarios, relying, whenever possible, on the involvement of target users.

Our collaborators have been performing similar processes on data from the Canadian Parliament, but applying traditional analytical techniques of text mining and social network analysis. Thus, we could conduct the research simultaneously in datasets/texts from both countries, interacting with our partners from MALNIS. The preprocessing and visualization tools developed are, in principle, independent of the language in which the texts are written (Portuguese or English).

1.3 Outline

This text is organized as follows. Chapter 2 presents a brief overview of recent related work in text visualization. Chapter 3 describes a visual interface proposed and developed as part of this work and in collaboration with the Research Group MALNIS. The research project and the applied methodology are presented in Chapter 4. Finally, concluding remarks on the contribution and its limitations are provided in Chapter 5.

CHAPTER

2

RELATED WORK

2.1 Initial Remarks

This chapter presents the main concepts, tools and related work relevant to the problem addressed in this project that of employing visualization techniques to analyze the temporal evolution of topics in Twitter in connection with political debates, and it is organized as follows. Section 2.2 provides an overview of recent contributions on visualization applied to text and social media data. A brief review of visualization techniques and tools for handling collections of textual documents and data of public interest in general, including posts from social media, is presented in Section 2.3. Finally, a few state-of-the-art visualization techniques for text which rely on the usage of multidimensional projections are described in Section 2.4.

2.2 Visualization for Social Media Data

Social media are the means by which ordinary people and communities can share, create, discuss and modify Internet content KIETZMANN *et al.* (2011). Despite the simple concept, social media have been acting nowadays as a powerful tool for democratization of information and knowledge, turning readers into authors. This phenomenon, leveraged by the recent creation of a wide variety of web-based technologies and services such as weblogs (blogs) and micro-blogging (e.g. Twitter), social and professional network sites (e.g. Facebook and LinkedIn), and media sharing sites (e.g. YouTube and Instagram), has raised the interest not only from social science researchers but also from professionals involved with companies and political issues.

Although visual representations have already been employed by social scientists to understand human relationships since the 1930s, as shown in the review by FREEMAN (2000), the way that social media extend these relationships by employing a wide variety of technologies, in addition to its popularity, demands a focus on the most recent contributions. Moreover, while the content of social media can comprise different types of data (e.g. videos and photos), in this

work we address textual content only.

From the perspective of data mining techniques, such textual content can been viewed as a large collection of documents. It is possible to apply statistical models on the collections in order to extract topics/themes for summarization. In general, topics provide the basis to create visual representations of textual collections. DOU *et al.* (2013), for example, proposed the system HierarchicalTopics to analyze large collections of text based on topic extraction. The first step in the process consists in collecting textual data from various sources, including social media sites, research publications, news, etc. The data then goes through a pre-processing stage to remove stopwords and emojis (the same as emoticons) and to prepare it for topic extraction. The topic extraction itself could be performed by a model that best fits the data, such as Latent Dirichlet Allocation (LDA) BLEI; NG; JORDAN (2003) and Author Topic Model (ATM) ROSEN-ZVI *et al.* (2004), the two models considered in their experiments. As the document collection grows, more topics may be required for summarization, and therefore more difficult it will be to generate clean and legible visualizations. One possibility to overcome this problem is to organize the topics in a hierarchical structure. With this in mind, DOU *et al.* (2013) designed their own algorithm called Topic Rose Tree (TRT) to construct a multilevel hierarchical structure with any given number of topics. Finally, two visualizations were designed to present the results. The first shows the hierarchical structure of topics as a tree, in which leaf nodes represent topics, while the remaining nodes denote group of topics (shown on the left side of Figure 1). The second visualization employs the well-known river flow metaphor, by extending the ThemeRiver HAVRE; HETZLER; NOWELL (2000) technique to allow the temporal visualization of single topics and group of topics (shown on the center and right side of Figure 1). An overview of the visualization system is shown in Figure 1, which depicts the hierarchical topic structure, a topic flow of multiple topics and detailed flow views of specific

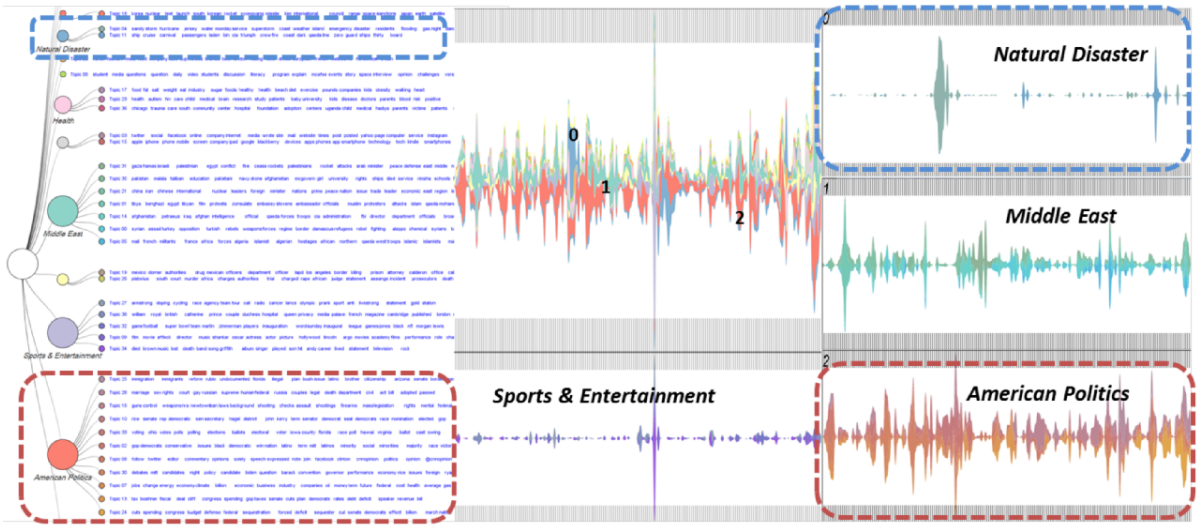


Figure 1 – An overview of the interface of the *HierarchicalTopics* system. The hierarchical structure is shown on the left and the temporal evolution of the topics on the central area and to the right. Figure extracted from DOU *et al.* (2013).

topics.

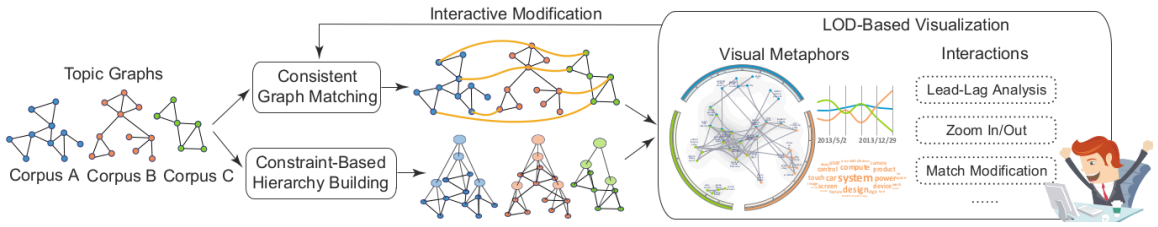


Figure 2 – An overview of the *TopicPanorama* system. Figure extracted from LIU *et al.* (2014).

Another interesting approach for topic visualization is introduced by LIU *et al.* (2014). As in the previous case of DOU *et al.* (2013), they propose a system for analyzing large collections of documents also based on topic extraction. Their system, called TopicPanonama, was designed under the assumption that a full picture of the topics discussed in multiple sources, such as news and blogs, is able to provide new insights for decision-making. In order to develop their hypothesis, the first step in the implementation consists in applying topic graphs construction models to the data. As the data comprises multiple sources, and different text corpora may exhibit distinguishing characteristics (a more in-depth discussion on this matter can be found in Chapter 4), a single graph could not fit each corpus properly, and therefore a different graph is built for each of the sources. Then, a consistent graph matching algorithm, extended to solve issues of inconsistency, is employed to find correspondences among the multiple topic graphs. Figure 2 shows an overview of the system. Finally, a density map is combined with a node-link diagram to visually represent the topic graph. Due to the potentially large number of topics, LIU *et al.* (2014) also employ a hierarchical clustering algorithm, the Bayesian Rose Tree (BRT) in their case, to organize the topics. The hierarchical structure is then shown as a radial stacked tree and combined with the density-based graph to create the final visualization, which is displayed in a circular layout, as illustrated in Figure 3 with label (a). The visualization only shows the representative nodes at each level of the stacked graph hierarchy, each of which is represented as a circular glyph, which conveys the degree of uncertainty of the matching results. The other non-representative nodes are shown as a density map. Moreover, an incremental algorithm is employed in order to integrate user feedback into the graph matching algorithm; it allows users to modify the mapping results. Additionally, the lead-lag relationships across multiple corpora, which show which source (lead) is followed by the others (lag) in regards to a selected topic, is visually represented as a twisted-line-like visualization, as indicated in Figure 3 with label (d). Figure 3 shows the visual interface of the TopicParonama system.

In some scenarios the extraction of topics alone may not be sufficient to meet the different information needs of analysts. For instance, questions such as what or who influences the emergence of specific topics, or to what extent these topics draw people's attention, are some possible research questions. XU *et al.* (2013), for instance, proposed a framework to study the agenda-setting theory and topic competition effects on social media. The agenda-

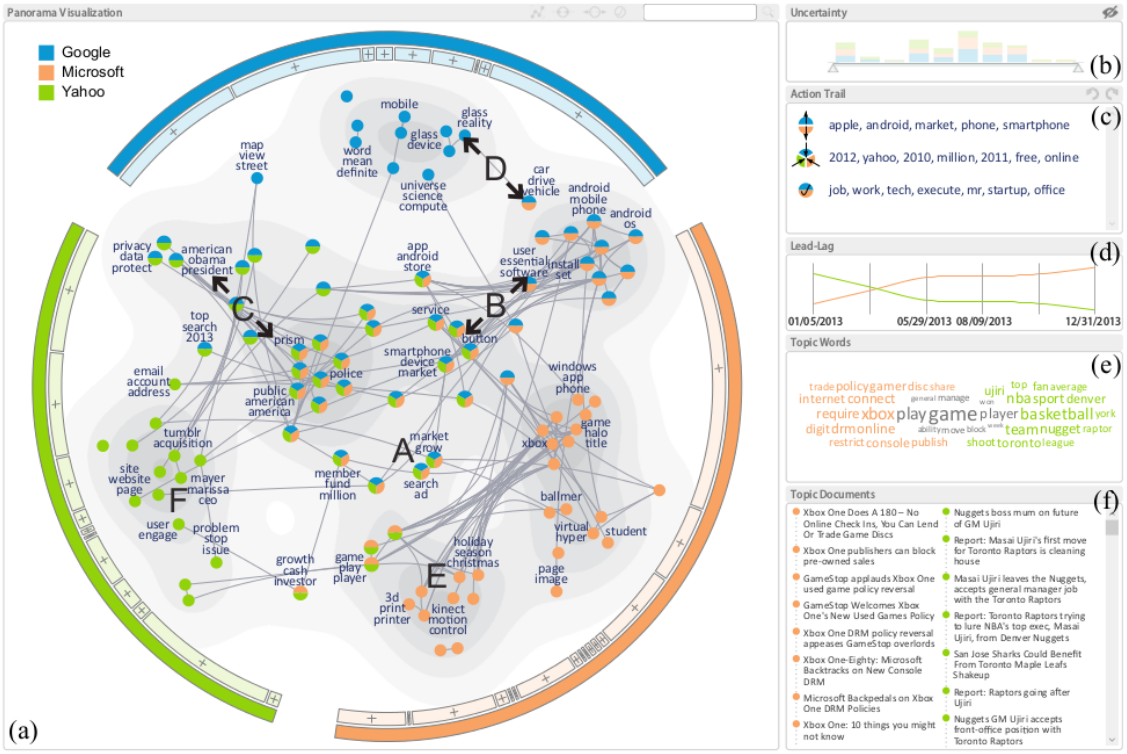


Figure 3 – Visual interface of the *TopicPanorama* system showing a full picture of topics related to Google, Microsoft, and Yahoo on multiple media: (a) the final visualization to examine the matched graph; (b) widget used to filter unwanted uncertainty glyphs (the representative nodes); (c) match modification trail; (d) lead-lag analysis; (e) topic word; (f) document. Figure extracted from LIU *et al.* (2014).

setting theory, introduced by Maxwell McCombs and Donald Shaw in the 1970s, states that the emphasis placed on certain topics by the traditional media determines their saliency as perceived by the general public. The democratization of information, as mentioned earlier, fostered by the growth of social media technologies, has enabled ordinary people to disseminate information over the Internet, which in turn influences the saliency of a particular topic in the public agenda. The framework by XU *et al.* (2013), as illustrated in Figure 4, consists of three major components: *data storage* and *pre-processing*, *data analysis* and *interactive visualizations*.

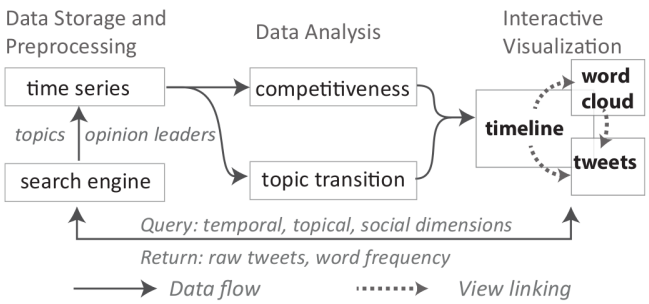


Figure 4 – Illustration of the framework introduced by XU *et al.* (2013). The framework consists of three major components: *data storage* and *pre-processing*, *data analysis*, and *interactive visualizations*. Figure extracted from the corresponding work.

The first component employs Apache Lucene[1] for text indexing and searching, and allows the data analysis component to extract time-series data efficiently by inputting keywords and time range queries. The results are then input to the topic competition model, expanded in order to contemplate multiple influence sources, which models the co-evolutionary relations between topics and opinion leaders – authors manually identify the set of topics and opinion leaders in their experiments. In the interactive visualizations component, they also employed a variation of the ThemeRiver technique to compose the visualization, as illustrated in Figure 5. The layers of the stacked graph (i.e. ThemeRiver) correspond to the topics and the opinion leaders are represented as threads overlapped with the layers to depict their co-evolutionary relationship, as indicated in Figure 5 with labels (a) and (b), respectively. The width of a thread is proportional to the contribution to the topic of the corresponding opinion leader group. Moreover, the threads can switch layers as the leaders change their topical focus. Additionally, multiple secondary visualizations are displayed on demand as users interact with the system, such as a word cloud that provides a visual summary of the textual content of tweets and a radial view that displays pairwise competition among topics, as illustrated in Figure 5 with labels (c) and (d), respectively.

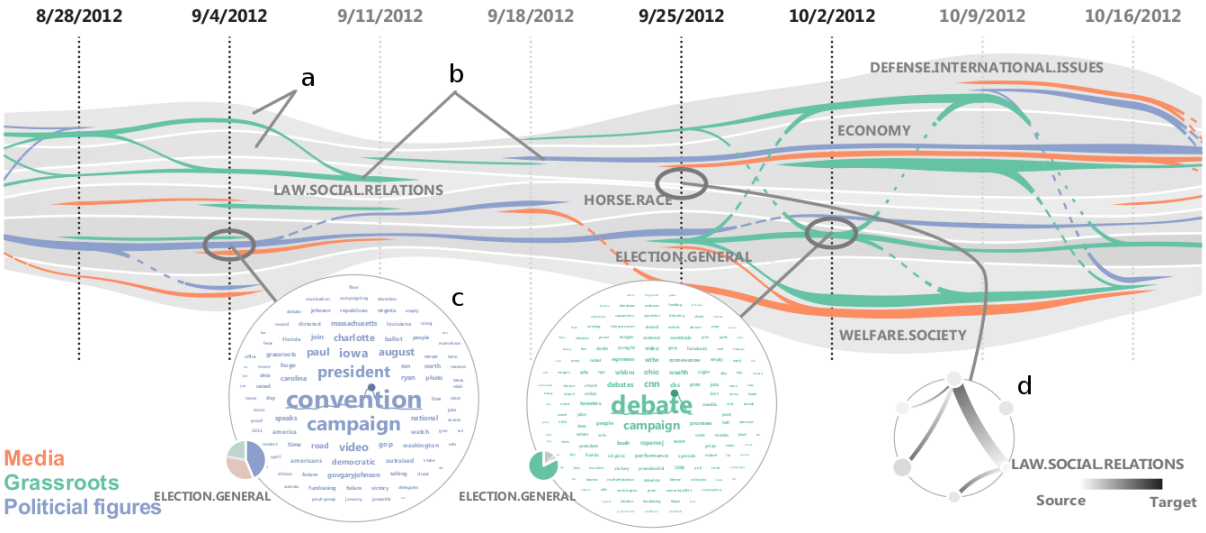


Figure 5 – Visual interface of the system by XU *et al.* (2013) showing the co-evolutionary relations between the topics and the opinion leader groups in a dataset related to the 2012 US presidential election. Figure extracted from the corresponding work.

While topic based approaches are suitable to capture popular trends and/or the overall picture of social media content, the study of information that does not receive enough attention requires a different treatment, such as the adoption of algorithms for modeling the process of information dissemination. In this context, ZHAO *et al.* (2014) developed a visual analytics system, named FluxFlow, to analyze the spreading of anomalous information in social media. Similarly to the work by XU *et al.* (2013), FluxFlow also consists of three major components: a *data pre-processing* and *storage* module, a *data analysis* module, and a *visualization* module. The

[1] <https://lucene.apache.org/> [Accessed 1 June 2017]

first component employs Apache Hadoop[2] for data filtering, retweeting thread reconstruction, and thread feature extraction. The data analysis module then applies the one-class conditional random fields (OCCRF) model to detect sequential anomalous retweeting threads based on the set of features extracted in the previous step. Specifically, an anomaly score is computed for each retweeting thread, which is built based on the interactions between Twitter users (e.g. retweet and mention), which are then ranked as a list of abnormal threads in a non-increasing order. Additionally, some contextual information, such as how Twitter users in these threads interact with each other, are computed to further support a better understanding of the abnormality. The visualization module presents the anomalous threads and the contextual information through multiple visualizations.

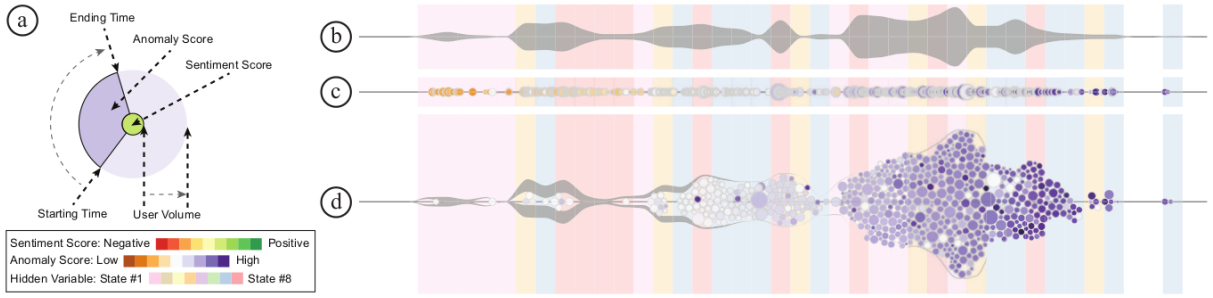


Figure 6 – Illustration of the three main views of the *FluxFlow* system: (b) a user volume chart; (c) a linear circle view that encodes each Twitter user as a circle; (d) a volume circle view that combines the previous two views – the "gray ribbon" aggregates users with low anomaly scores. (a) shows a thread glyph for aggregating the main information of a retweeting thread in a compact form. Figure extracted from ZHAO *et al.* (2014).

A temporal evolution of the threads is presented in three main views, as illustrated in Figure 6. Once again, the temporal evolution (or temporal trends) of the data under analysis is conveyed with a river flow metaphor. The first view, shown in Figure 6 with label (b), shows the volume of users participating in a retweeting thread over time. The second view, shown in Figure 6 with label (c), encodes each Twitter user as a small circle displayed over the time axis, where circle size indicates the number of followers and its color corresponds to the anomaly score of a particular user. Finally, the first two visualizations are combined in order to create a composite one, as indicated in Figure 6 with label(d). Moreover, a thread glyph, as illustrated in Figure 6 with label (a), summarizes a number of variables associated with a thread, such as the tweet sentiment score and the thread anomaly score, in order to provide a compact form for the user to understand key characteristics of the abnormality. In addition to these three main views, the system allows the user to visualize the hierarchical structure of the threads, which is displayed as a tree (shown in Figure 7 with label (a)), and provides some contextual information such as the distribution of the threads in the anomaly feature space, which is accomplished by a multidimensional projection algorithm, and the actual content of the tweets, as indicated in

[2] <http://hadoop.apache.org/> [Accessed 1 June 2017]

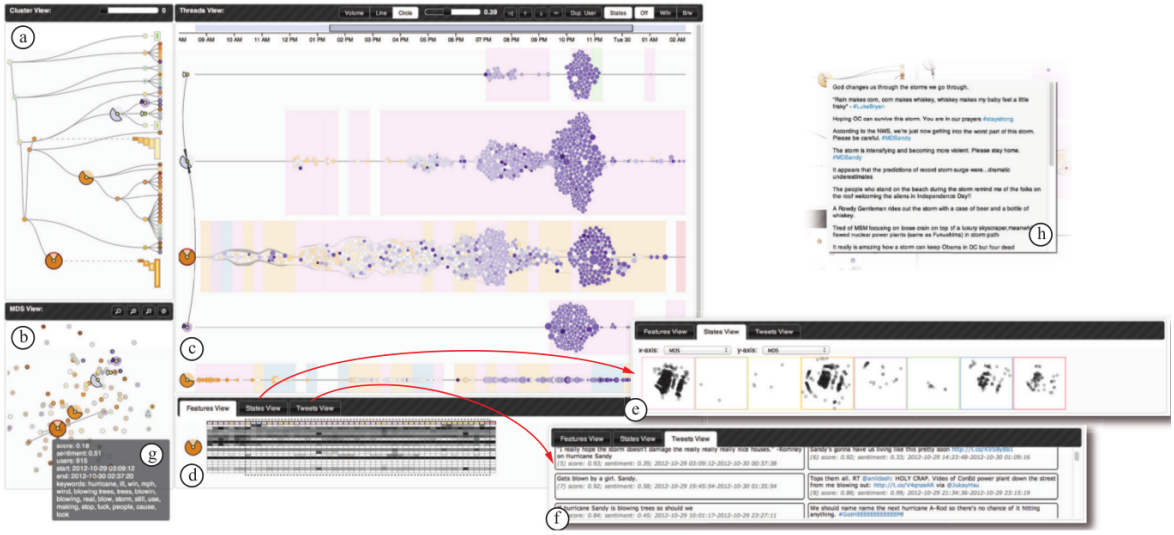


Figure 7 – Visual interface of the *FluxFlow* system: (a) the hierarchical structure of a thread; (b) the thread distribution in the anomaly feature space; (c) a threads view, which shows the composite view, and a detail information panel; (d) a feature view; (e) a states view; (f) a tweets view; (g) informative tooltips and (h) context menus. Figure extracted from ZHAO *et al.* (2014).

Figure 7 with labels (b) and (f), respectively. Figure 7 shows the visual interface of the FluxFlow system.

Interesting initiatives that explore the use of visual representations to facilitate the interpretation of social media data have been carried out at the *Laboratório de Estudos Sobre Imagens e Cibercultura* (LABIC)[3], from the Federal University of *Espírito Santo* (UFES). The LABIC researchers collect data from social media and use classic visualization techniques to "understand" the behavior of social network users on a particular subject. Analyses made by the LABIC coordinator Dr. Fábio Malini, have been cited in press media, such as New York Times, El Diario, O Globo, and television news programs such as Globo News. From the several subjects analyzed, one in particular has made LABIC visible in the world media, the 2013 protests in Brazil[4]. In this particular case, a script was used to collect tweets with recurrence of keywords such as "Dilma" and "MarcoCivil", and the data was plotted using Gephi[5], an open-source network analysis and visualization software.

2.3 Visualization applied to Public and Textual Data

Visualizations can be applied in a wide variety of contexts, from textual corpora, such as news, academic texts and literary books, to scientific data such as meteorological conditions and physico-chemical characteristics of plants. Furthermore, research in this field is not restricted to the academic world, given that several non-academic organizations are now recognizing its

[3] <http://www.labic.net/> [Accessed 1 June 2017]

[4] <https://goo.gl/KCy3Y1>. URL shortened by Google URL Shortener at <goo.gl> [Accessed 1 June 2017]

[5] <https://gephi.org> [Accessed 1 June 2017]

value. The Brazilian newspaper "Estadão", for example, has made available a portal blog with several interactive visualizations of data on Brazilian politics, elections and social demographics[6], created using data from several sources, not necessarily public. These visualization range from simple bar charts to more sophisticated ones such as the "Mapa dos incêndios de ônibus em SP, mês a mês, desde 2013" (a map of buses set on fire during the protest riots in 2013)[7], as illustrated in Figure 8. Although most visualizations and tools available in the portal are tailored to show a specific dataset, they do allow journalists and readers to carry out substantial analysis. In particular, it is worth mentioning the section referred to as "Projetos Especiais" (special projects), which aggregates the most relevant projects published by the blog, such as the "Basômetro" (basometer), which has been well received not only by journalists but also by the academic community. It was the first project developed for the blog, and it was designed to reveal the level of support to government bills by parliament members in the Brazilian congress.

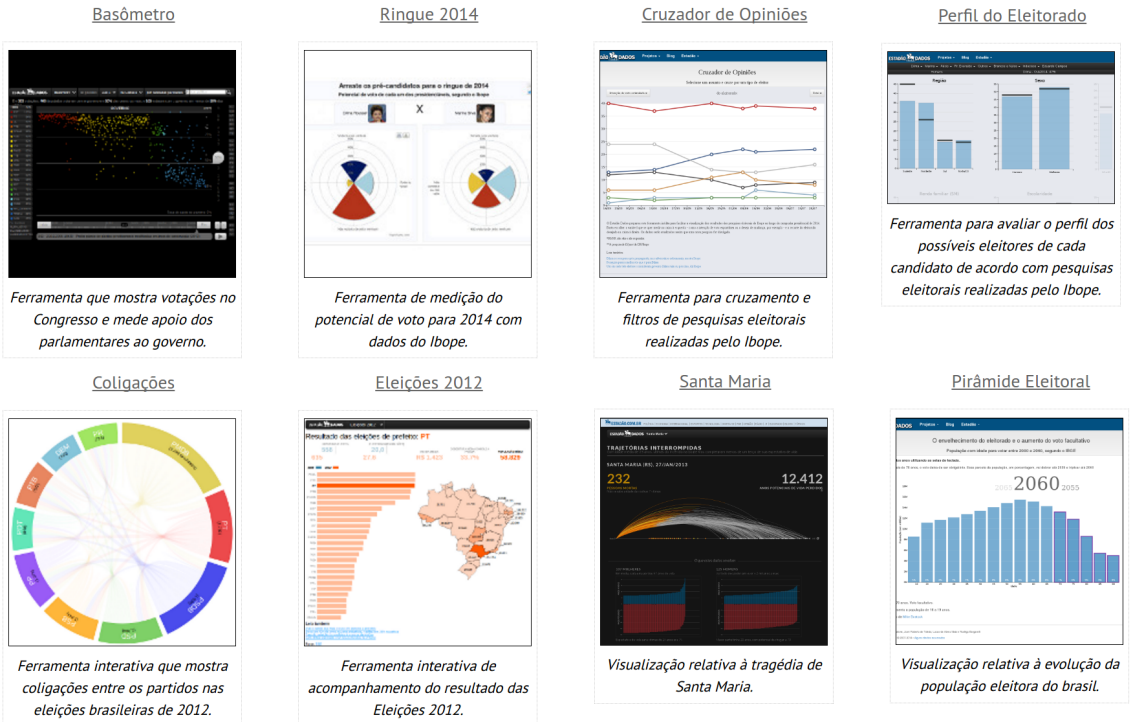


Figure 8 – Examples of visualizations available at the portal blog "Estadão" .

Another very appealing solution widely known is provided by Gapminder.org[8], founded by Ola Rosling, Anna Rosling Rönnlund and Hans Rosling in February 25 2005, which relies mostly on data made available by international organizations such as the World Bank, UNESCO, and journals such as The Lancet[9]. Gapminder's mission is to leverage initiatives that advance the United Nations Millennium Development Goals[10], and promote a better understanding of the

6 <http://blog.estadaodados.com/> [Accessed 1 June 2017]
7 <http://blog.estadaodados.com/mapa-incendios-onibus-sp/> [Accessed 1 June 2017]
8 <https://www.gapminder.org/> [Accessed 1 June 2017]
9 <http://www.thelancet.com/> [Accessed 1 June 2017]
10 <http://www.un.org> [Accessed 1 June 2017]

world and the society by dismantling misconceptions with reliable statistics. They have developed a number of data visualization tools, of which Trendalyzer is the most famous. Trendalyzer lets people explore global statistics by converting data into interactive visualizations. The success of the tool comes from the fact that the time dimension stays in the background, which is associated to a temporal axis and exhibited via animation (shown in Figure 9 with labels (a) and (b)), whereas variables of interest are mapped to the x-y axes as in conventional scatter plots. Given a set of countries, for instance, one could map life expectancy in years to the x-axis, the per capita income to the y-axis and the year referred by the data to the temporal background axis. Each country is represented as a circle, with circle size and color typically indicating population and country, respectively. A very appealing presentation given by Hans Rosling at TED Talks conference[11] in 2006, with title "The best stats you've ever seen", has already had over 11 million views. Figure 9 shows the visual interface of the Trendalyzer system, which shows how long people live and how much money they earn[12].

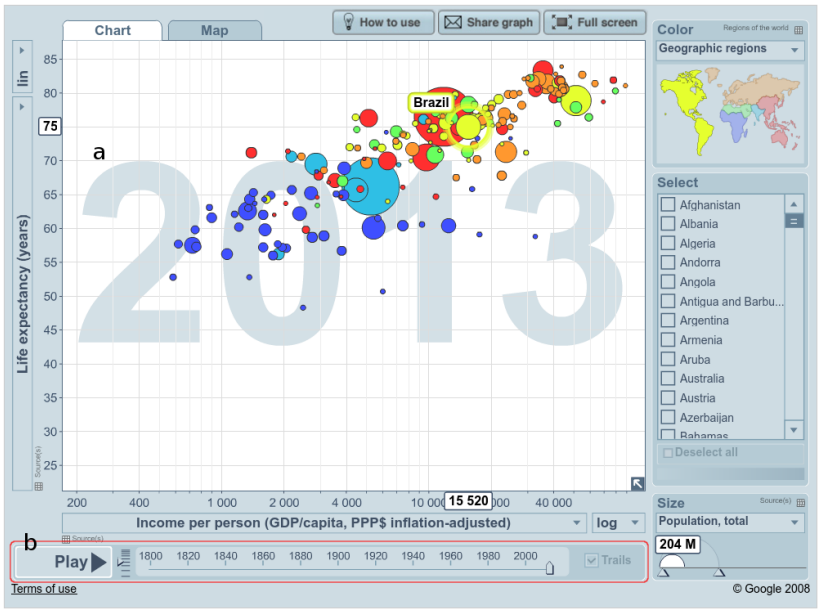


Figure 9 – Visual interface of the *Trendalyzer* system showing how long people live and how much money they earn: (a) the time dimension and (b) the associated temporal axis.

Cooperation between the private sector and academia has also produced interesting visualization results. BREHMER *et al.* (2014) with the assistance of the Associated Press journalists[13], developed the Overview, a system for systematic analysis of large collections of documents based on clustering, visualization, and tagging. According to the authors, due to the involvement of collaborators and users who were domain experts, a major contribution of the work were the lessons learned during the development of the tool, since 4 versions of the tool were created and 6 case studies conducted. Nevertheless, the basis the system remained

[11] <https://www.ted.com/talks> [Accessed 1 June 2017]
[12] <https://goo.gl/F9itG8>. URL shortened by Google URL Shortener at <goo.gl> [Accessed 1 June 2017]
[13] <https://www.ap.org/en-us/> [Accessed 1 June 2017]

stable during all versions, which consists in following the common practice of converting each document into a vector of word weights by the formula Term Frequency-Inverse Document Frequency (TF-IDF), and computing similarities between documents measured by the cosine distance. A hierarchical structure is then created based on those similarities. The visualization itself consists of showing this hierarchical structure as a tree, as illustrated in Figure 10, where each node represents a group of similar documents. Each node is labeled with keywords extracted via TF-IDF scores, which provides an overview of the topics discussed in those documents. Additionally, the system allows the user to access the actual document, as well as tagging of both the documents and the tree nodes in order to keep track of any findings and/or discoveries during the analysis.

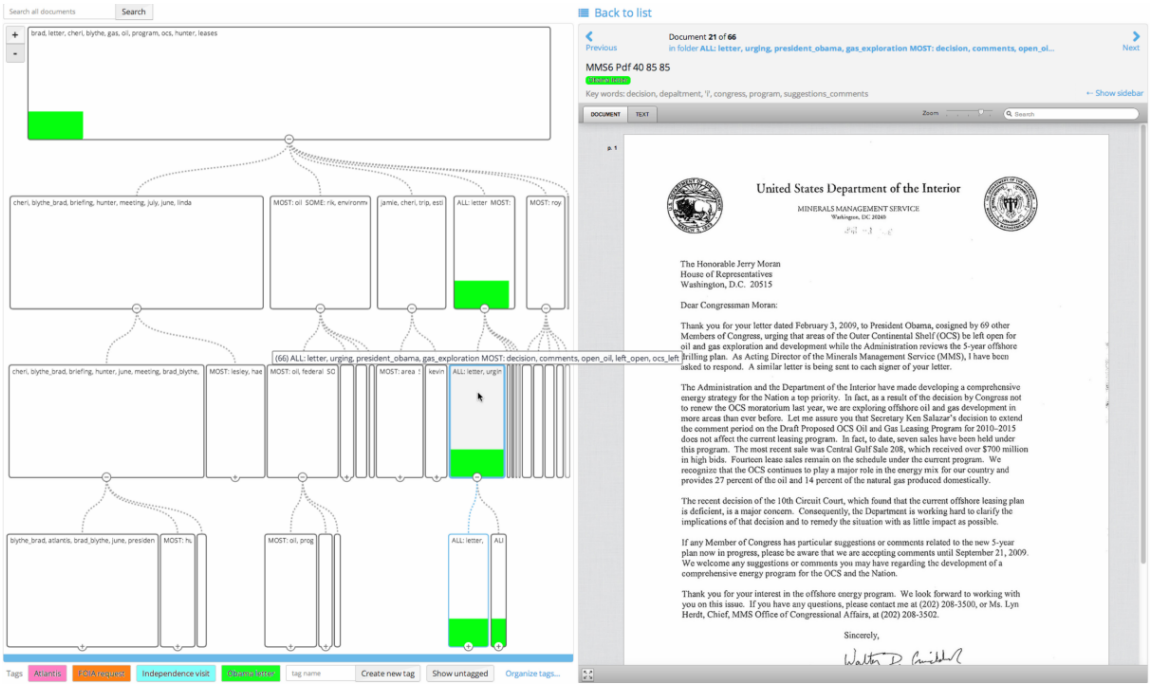


Figure 10 – Main screen of the Overview system, showing groups of similar documents as a tree (left) and a document selected from the highlighted group (right). The document collection analyzed comprises a collection of White House email messages concerning drilling in the Gulf of Mexico prior to the 2010 BP oil spill. Figure extracted from BREHMER *et al.* (2014).

When the goal is to create visualizations applied specifically to data made available by government institutions, an important question that arises is how such data are published and how to consume them. In general, this data is available thanks to initiatives from organizations and communities around the world that seek to promote open/free knowledge. In Brazil, the *Information Access Law*[14] has been recently enacted, under which all public government institutions are required to supply information requested by citizens, and institutions are preparing and reviewing procedures to handle the upcoming demands.

This is in parallel with several initiatives to make available data relative to government decisions at the executive, legislative and judiciary levels, such as partnerships between the

[14] <goo.gl/sjSdSQ>. URL shortened by Google URL Shortener at <goo.gl> [Accessed 1 June 2017]

Open Knowledge Foundation, the Transparency Hacker Community and the W3C. The "Portal Brasileiro de Dados Abertos" (Brazilian Open Data Portal), for example, was proposed by the government as a large catalog of public data and follows the W3C recommendations, which state that the publication and dissemination on the web of information from the public sector should be shared in open and raw formats, and be logically understandable to allow their use by socially developed applications. GRAVES (2013), for instance, developed the VisualBox, designed specifically for handling data published as Linked Data BIZER; HEATH; BERNERS-LEE (2009), Linked Data is a term used to describe a method of publishing structured data so that it can be interlinked and become more useful through semantic queries[15].

The system is based on LODSPeaKr, a framework for creating applications based on Linked Data. It is worth noting that this application requires minimal user knowledge of Web and SPARQL technologies. Knowing the basics, a user will be able to create visualizations using the two system components: *models* and *templates*. The first component supports creating SPARQL queries to retrieve data published in RDF, and the second component allows creating a number of visualizations of the retrieved data. The visualizations are provided by libraries such as Google Maps and D3.js, as illustrated in Figure 11.

When the task requires the analysis of text document collections, methods based on topic modeling are once again present. However, there are approaches that can not only present the topics but also allow users to introduce their knowledge directly into the model in order to achieve better results. UTOPIAN (User-driven Topic modeling based on Interactive Nonnegative

[15] <https://en.wikipedia.org/wiki/Linked_data> [Accessed 1 June 2017]

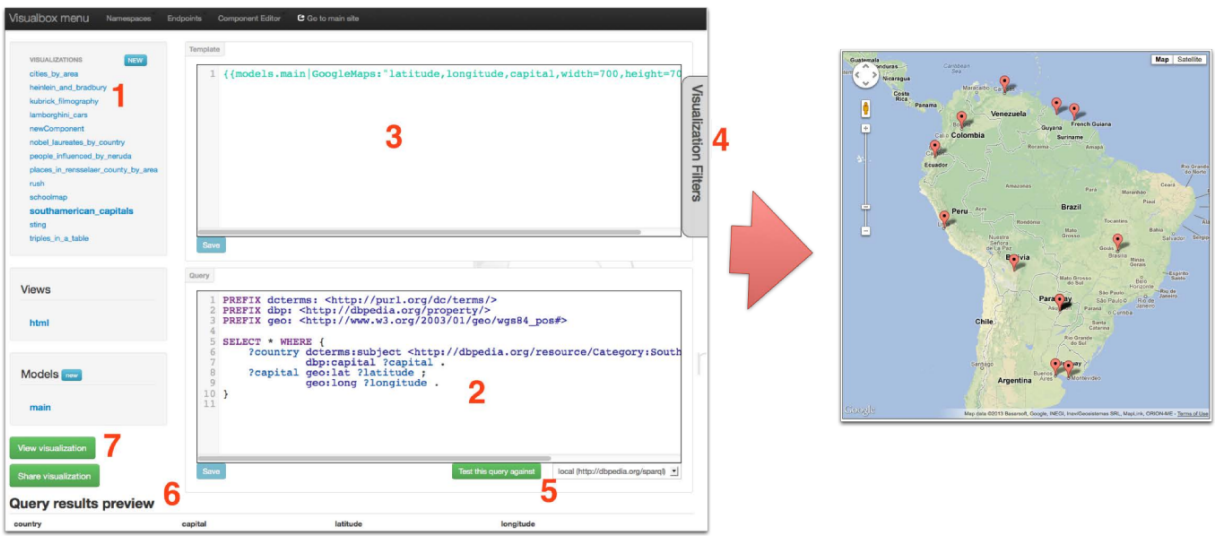


Figure 11 – Visual interface of the *VisualBox* system: region (1) shows previously created visualizations; regions (2) and (3) provide text areas for editing the SPARQL queries and the templates, respectively; region (4) shows a list of visualization filters; regions (5) and (6) show respectively a button for executing the current query against a specified SPARQL endpoint and the corresponding results, and region (7) shows a button for opening the final visualization and for sharing it on the Web. Figure extracted from GRAVES (2013).

Matrix Factorization) CHOO *et al.* (2013) is a system that applies a semi-supervised (SS-NMF) version of the non-negative matrix factorization method (NMF). To understanding the process of updating the supervised model, we must first know the original formulation of NMF. Given an array $X \in \mathbb{R}_+^{m \times n}$, and an integer $k \ll min(m,n)$, NMF finds an approximation given by

$$X \approx WH, \tag{2.1}$$

where $W \in \mathbb{R}_+^{m \times k}$ and $H \in \mathbb{R}_+^{k \times n}$ are non-negative factors. NMF is typically formulated in terms of the Frobenius norm as

$$\min_{W,H \geq 0} \|X - WH\|_F^2. \tag{2.2}$$

In a topic modeling context, X is a term-document matrix which represents the documents by a Bag-of-Words (BoW) model SALTON *et al.* (1986). k is the number of topics desired. Columns of W represent the topics as a combination of terms (or keyword) and columns of H represent the documents as a combination of topics. In addition to the X matrix and the integer k, the SS-NMF receives as input the matrices $V \in \mathbb{R}_+^{m \times k}$ and $G \in \mathbb{R}_+^{k \times n}$, which are used as references for the matrices W and H, respectively, and two diagonal matrices $M_W \in \mathbb{R}_+^{k \times k}$ and $M_H \in \mathbb{R}_+^{n \times n}$, which assign weights on the columns of V and G, respectively. Then, SS-NMF attempts to approximate W to V and H to G as

$$\min_{W,H,D_H \geq 0} \{\|X - WH\|_F^2 + \|(W - V)M_W\|_F^2 + \|(H - GD_H)M_H\|_F^2\}. \tag{2.3}$$

Due to the non-negativity constraint of its formulation, SS-NMF can provide an intuitive approach for embedding the user's knowledge to the topic modeling process. In this way, UTOPIAN can offer the following interactions: refinement of the keywords forming an existing topic (panel with label (Topic 4) in Figure 12), topic splitting/merging (shown in Figure 12 with label (3) and (1), respectively) for managing the number of topics and keyword-induced/document-induced (shown in Figure 12 with labels (4) and (2), respectively) for creating topics. The topic refinement, for example, allows the user to directly modify the V matrix in order to change keyword weights, thereby semantically changing the meaning of the topics. In order to visualize the results, UTOPIAN employs the t-distributed stochastic neighborhood embedding (t-SNE) technique MAATEN; HINTON (2008) to generate a node-link diagram that reflects the similarities between the documents in terms of their topics, as illustrated in Figure 12. Moreover, the system is designed to project the intermediate results of NMF and t-SNE, in an attempt to reflect changes in real-time, and to facilitate their tracking, via smooth animations.

ALEXANDER *et al.* (2014) proposed a system for analyzing large document collections at three levels: corpus level, document level and word level. Although each level is associated with a different visualization, they are all based on the topic modeling algorithm Latent Dirichlet

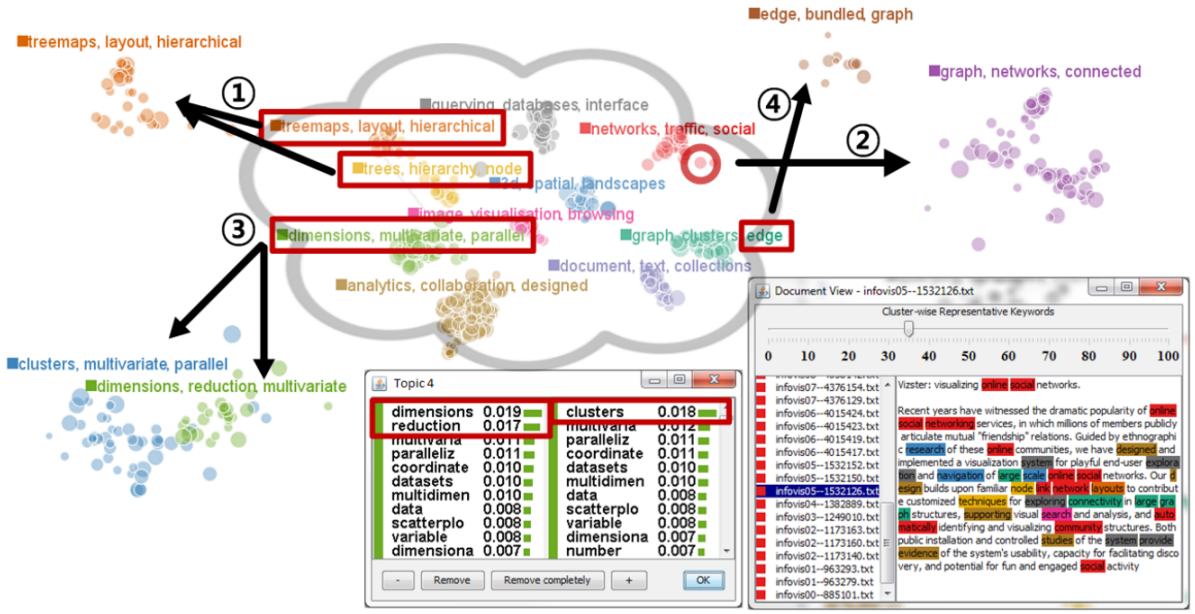


Figure 12 – Interface of the *UTOPIAN* system: regions (1) and (3) show examples of the topic splitting/merging interactions; regions (2) and (4) show examples of topic creation by document-induced/keyword-induced interactions; the panel with title "Topic 4" allows a user to refine topic keyword weights; and the panel with title (Document View) shows the textual content of the documents highlighting the representative keywords from each topic. Figure extracted from CHOO *et al.* (2013).

Allocation (LDA) BLEI; NG; JORDAN (2003). Figure 13 shows an overview of the system interface. In order to give an overview of the document collection, the corpus level has a reorderable matrix, shown on the top left side of Figure 13, where the lines represent the documents and the columns the topics. The matrix elements (weights) are represented by circles, with circle size indicating the topic proportion. The system displays only those documents and topics that are of user interest. The user can also interact with the matrix by coloring, moving and even removing rows or columns.

The document level displays the content of a given document, shown on the top right side of Figure 13, where words are labeled (as indicated by different background colors) with information from the topic model. Finally, the word level shows where a word, or group of words, fall in the ranking of individual topics, shown on the bottom of Figure 13. The visualization employs a bar graph to show the ranking, with color coded lines within these bars depicting the word's ranking within the topics and bar length indicating the relative size of the topic.

2.4 Visualization of Textual Data Based on Multidimensional Projections

The scientific literature on Visualization, which has been growing at an quick pace, includes many contributions focused on textual data. A survey of these techniques is presented

Figure 13 – Overview of the interface of the *Śerendip* system, showing the three main views: (top left) CorpusViewer; (top right) TextViewer; and (bottom) RankViewer. Figure extracted from ALEXANDER *et al.* (2014).

by ALENCAR; OLIVEIRA; PAULOVICH (2012), in which text visualization techniques are categorized into four groups, based on the input type: there are techniques whose purpose is to visually represent the content of a single textual document VIéGAS; WATTENBERG; FEINBERG (2009); WATTENBERG; VIÉGAS (2008); others are oriented to the visualization of collections of documents, generally seeking to emphasize content similarity relations PAULOVICH *et al.* (2008); PAULOVICH; MINGHIM (2008); PAULOVICH *et al.* (2012), and/or the temporal evolution of the topics ALENCAR *et al.* (2012); HAVRE; HETZLER; NOWELL (2000); CUI *et al.* (2011); yet others seek to visually represent relationships between documents, such as citation and co-citation relations, co-authorship relations, etc., usually displayed as graphs CHEN (2006a); as well as techniques for visualizing textual search results GOMEZ-NIETO *et al.* (2014). A summary table in the survey cites no less than 29 techniques/tools applicable to textual data, considering these different types of input.

Researchers from the *Visualization, Imaging and Graphics Processing* (VICG) group have addressed the development of interactive techniques for visualizing abstract data based on the use of multidimensional projection techniques, which generate data mappings in a visual space, usually two-dimensional, seeking to preserve similarity/dissimilarity relations between the instances. These mapped instances are positioned in the 2D space in order to reflect their similarity or relations to the other instances in the multidimensional space so that greater distances reflect more dissimilarity, and small distances indicate similar instances. The output of these algorithms can be presented in a variety of ways, from simple point clouds, or by using icons expressive of the data content or optimized layouts ELER *et al.* (2009); GOMEZ-NIETO *et al.* (2014), as illustrated in Figure 14. The resulting views are called “similarity maps”. The map provides an overview of the data set from which an analyst can visually identify groups of

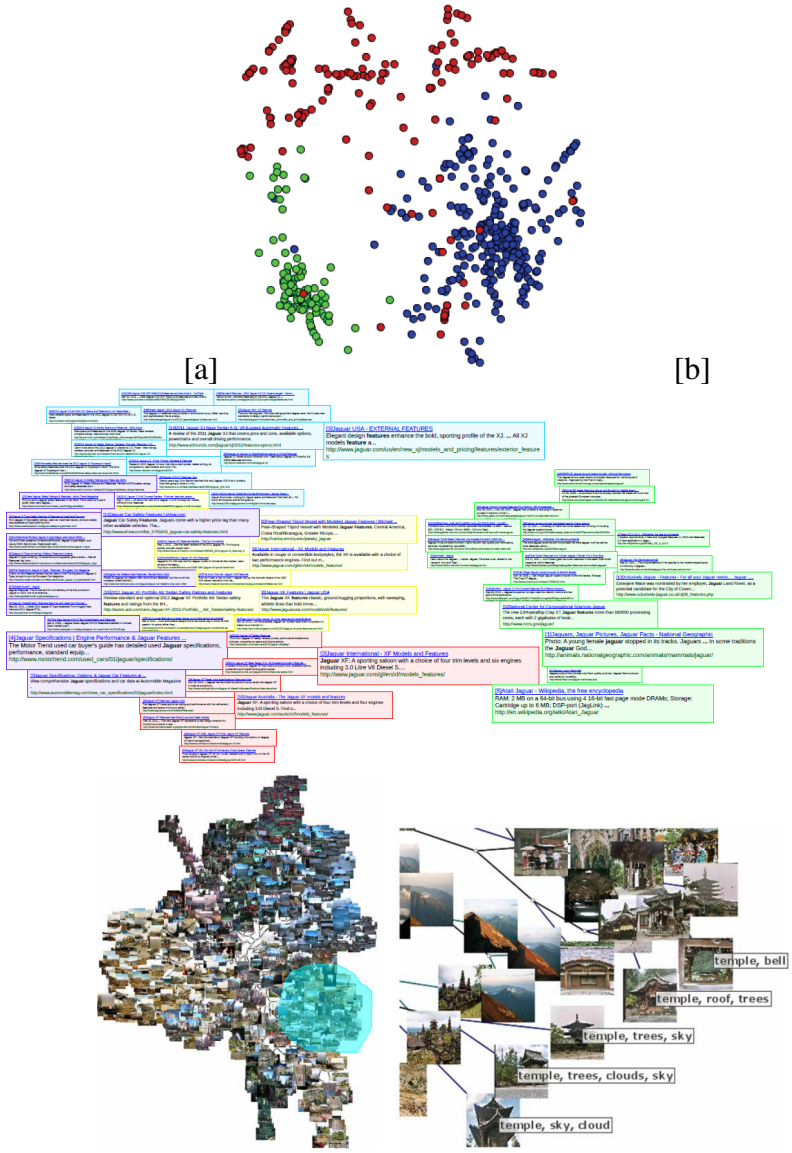


Figure 14 – Examples of: (a) document map (a collection of RSS news feeds) generated with the LSP (Least-Square Projection) multidimensional projection technique. Source: SALAZAR *et al.* (2013); (b) visualization of the result of a query using the Google search engine by the ProjSnippet technique. Source: GOMEZ-NIETO *et al.* (2014); (c) visualization of a set of photos by the NJ-Tree technique and zoom on a selected branch. Source: ELER *et al.* (2009).

similar/dissimilar elements, and focus on those groups for additional information – for example, a group of news can address related topics content-wise.

Accordingly, PAULOVICH *et al.* (2008) developed a multidimensional projection technique, named Least-Square Projection (LSP), based on least squares approximations. The projection process comprises two main steps. The first step is to select c control points that represent the best possible distribution of the data in the original space $I\!R^m$. This selection can be done, for example, by dividing the data into c clusters using the k-medoids method and selecting the medoids (the closest points of the centroid) as control points. Then, the control points must be projected in the reduced space $I\!R^d$, being $d \leq m$, by means of some Multidimensional Scaling

(MDS) technique, such as Force Scheme or Sammon's Mapping. In the second step the technique projects the points in $\mathbb{R}^m$ onto the new space $\mathbb{R}^d$. This projection is done by means of solving a linear system that is created to capture the neighborhood relations between the points in $\mathbb{R}^m$ and the Cartesian coordinates of the control points already mapped in $\mathbb{R}^d$. The linear system is defined as follows

$$Ax = b, \tag{2.4}$$

where A is a rectangular matrix $(n+c) \times n$ given by:

$$A = \begin{pmatrix} L \\ C \end{pmatrix}, c_{ij} = \begin{cases} 1 & \text{, if } p_j \text{ is a control point} \\ 0 & \text{, otherwise} \end{cases}, l_{ij} = \begin{cases} 1 & , i = j \\ -\alpha_{ij} & , p_j \in V_i \\ 0 & \text{, otherwise} \end{cases} \tag{2.5}$$

where, in case $\alpha_{ij} = \frac{1}{k_i}$, where k_i is the number of points in the neighborhood of central point p_i, the matrix L is typically called the Laplacian matrix and b is a vector

$$b_i = \begin{cases} 0 & , i \leq n \\ x_{p_{c_i}} & , n < i \leq n+c \end{cases} \tag{2.6}$$

where $x_{p_{c_i}}$ is one of the Cartesian coordinates of the control point p_{c_i}. Examples of results obtained by the LSP technique are illustrated in Figure 14 (a) and (b).

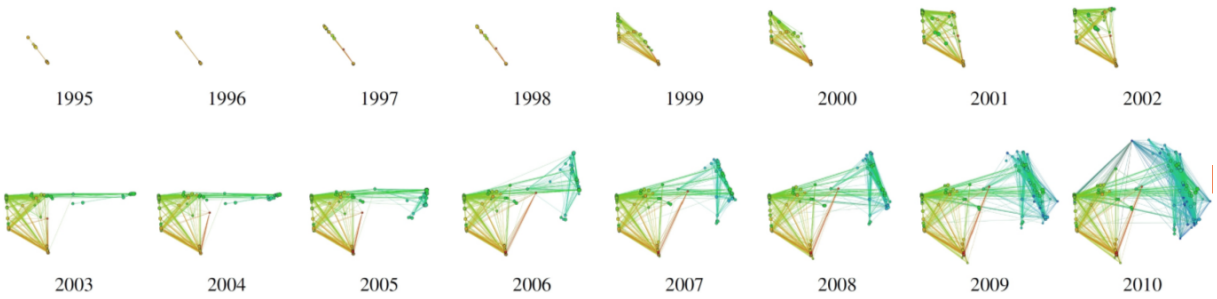


Figure 15 – Illustration of the process performed by the *Time-Aware* system for creating the visualizations. The figure shows a sequence of time-stamped document maps of a collection of articles authored by scholar Alessandro Vespignani. Each circle in the maps represents a scientific article, circle color indicates publication year and circle size maps the global citation count. Edges represent the bibliographic coupling between articles. Figure extracted from ALENCAR *et al.* (2012).

Due to its efficiency and accuracy, LSP has been extended to handle different aspects of the data sources. ALENCAR *et al.* (2012), for instance, introduced an enhanced technique, named Time-based Least Square Projection, which extends LSP to explicitly contemplate the temporal component of the data. Specifically, their solution addresses the temporal evolution of collections of scientific articles by showing a series of visualizations that emphasize changes in

their similarity patterns over time. The process consists of creating a sequence of content-based similarity maps from time-stamped sub-sets of the data. The map list created from the dataset can then be visualized sequentially, moving one step forward or backward in time, or as an animation, where the point positions are interpolated between two consecutive maps. The outcome of this process for a particular collection of papers is illustrated in Figure 15.

Another variation of LSP is introduced in the work by PAULOVICH; MINGHIM (2008), who propose a new technique, named Hierarchical Point Placement (HiPP), for handling the visual scalability limitation when dealing with large volumes of data. The process consists of two main steps: (1) the data is hierarchically organized as a tree; and (2) the tree nodes are projected, creating a multilevel visualization. The tree is created by a recursive partitioning process where the internal nodes represent groups of content-wise similar documents and the leafs represent individual documents. The tree nodes are then projected by the LSP technique in an attempt to preserve similarity relations between the instances, at the different levels of detail. Figure 16 shows a document map of a collection of scientific papers obtained with HiPP.

2.5 Final Remarks

This Chapter presented a brief survey of contributions from the literature related to this work. An interesting point to note is that most contributions in this Chapter employ multiple visualization techniques for the analysis of text corpora. Some apply different techniques simultaneously by merging two or more visualizations into one and provide multiple secondary views for

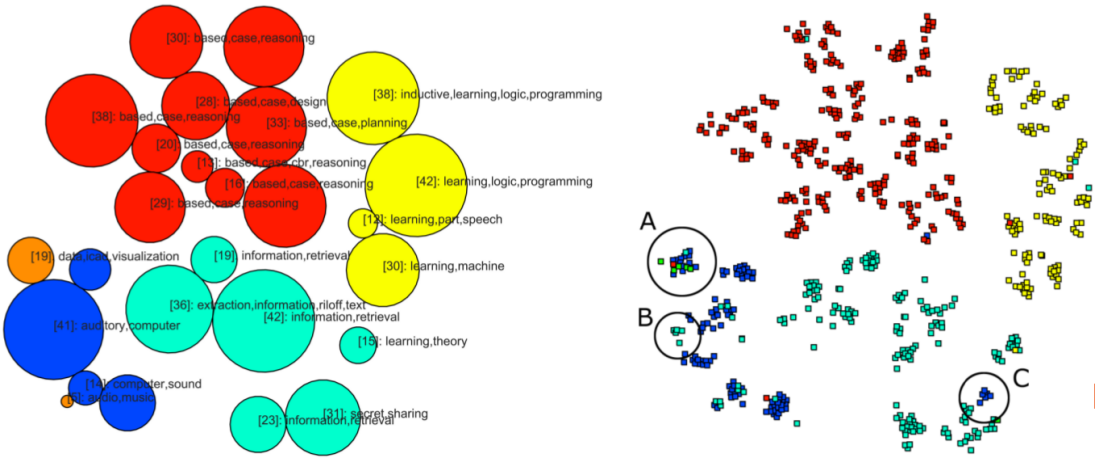


Figure 16 – A document map of a collection of scientific papers from four distinct academic subjects. The view on the left shows the top level map view, where the circles represent high-level clusters of documents. Each cluster is annotated with a set of topical words extracted from them and the colors reflect an existing classification of the documents. The view on the right shows the circles representing individual papers, accessed after expanding the clusters. The region delimited as (A) indicates highly correlated documents, whereas region (B) indicates documents apparently misplaced and region (C) shows documents that deal with the topic "parallel programming". Figure extracted from PAULOVICH; MINGHIM (2008).

context, while others leverage traditional charts or introduce new ways to graphically represent the data. Contributions on visualization for social media data, for instance, often employ some variation of the river flow metaphor, which is useful to convey the temporal evolution of some data property (e.g. topic) and can be extended to convey, for example, the temporal evolution of multiple properties simultaneously, relying on scatter plots and word clouds visualizations to provide additional information.

Another common strategy is to extend existing techniques and models in order to adapt them for specific problems. The data under analysis may include additional information enabling the development of customized solutions to the problem domain. Posts published on Twitter, for example, have a considerable amount of metadata that goes far beyond of just working as a complement to the text written by the user and may contain valuable information such as the geographic location of the tweet author, the subject being debated, and the people involved in the conversation.

Although the visualizations are central to the tools, the steps of data collection and processing cannot be neglected, being essential in the whole process to capture and model the relevant information from the data. In fact, the quality and effectiveness of the visualizations are highly dependent on the choices made in data preparation, such as the removal (cleaning) of unnecessary or problematic features, and the selection of an appropriate representation. To reduce visual clutter when handling large document collections, for example, a common approach is to organize the data in a hierarchical structure in order to create a multilevel visualization.

Some of the contributions described address challenges similar to those faced in this work. TopicPanorama, for example, needs to find correspondences among topics that appear in distinct data sources so that a user can obtain an overview of what is being discussed. They handle this in an additional preprocessing step that generates the topic graphs so as to increase the performance of the graph matching algorithm. In our own current work, groups of similar or related topics are interpreted as debates going on in two distinct data sources and a correspondence, or association between them also needs to be established. Table 1 presents a summarized comparison between the systems described in Section 2.2 and the one proposed in this work.

Table 1 – Summary of the main visualizations presented and the tasks supported by the systems described in Section 2.2. *Topics here is a generalization for the type of the attribute under analysis.

Name/Reference	Main visualizations/ layout types	Support visualization of			Year
		the hierarchical structure of *topics	the temporal evolution of *topics	the correspondences of *topics between multiple sources	
The system proposed in this work	River layout	No	Yes	Yes	2017
HierarchicalTopics DOU *et al.* (2013)	Tree and river layouts	Yes	Yes	No	2013
TopicPanorama LIU *et al.* (2014)	A Radial stacked tree combined with a density based graph displayed in a circular layout	Yes	Partially	Yes	2014
XU *et al.* (2013)	River layout combined with overlapping threads	No	Yes	No	2013
FluxFlow XU *et al.* (2013)	Multiple river layouts, a tree layout and a multidimensional projection	Yes	Yes	No	2013

CHAPTER

3

VISUAL INTERFACES FOR ACTIVE LEARNING

3.1 Initial Remarks

Active learning in machine learning algorithms, or active retrieval for the task of information retrieval, aims at improving the performance of the corresponding method by requesting labels of key instances from an oracle or information source. The key hypothesis SETTLES (2009) is that the learning algorithm will perform better with less training if it is allowed to choose the data from which it learns, which is a desirable property in scenarios where labeled instances are difficult, or expensive, to obtain. In this context, the quality of the whole process relies on the oracle's capacity of providing reliable answers and on the formulation of good *query strategies*, i.e., the process by which the algorithm evaluates the unlabeled instances in order to select the most informative one. In the particular case of the oracle being a human, it is possible to employ visualization techniques to leverage the domain knowledge of the annotator and help in making decisions. The work by HEIMERL *et al.* (2012), for instance, introduces a good example of how the performance of a classifier can be improved by employing active learning and visualization techniques for the task of document retrieval. Their system encompasses interactive visualizations in order to depict the current state of an *Support Vector Machine* (SVM) classifier and to allow a user to select instances for labeling. The classifier's state is shown using a scatterplot in which the middle of the view is used to depict the boundary of the SVM model. The horizontal axis represents the confidence value for each document and the vertical axis the diversity of the documents closest to the decision boundary. Additionally, a cluster view, which employs the LSP technique to project the instances, depicts the 100 most uncertain documents, which have the potential to speed up the classifier training if labeled.

In this chapter, we present two solutions in which visualizations were embedded into interfaces in order to present the active learning (and active retrieval) strategies and allow a better

understanding of the data. The visual interfaces were developed during a research internship at Dalhousie University, in Canada[1], motivated by an ongoing collaboration with the Research Group on "Machine Learning and Networked Information Spaces (MALNIS)"[2].

3.2 ATR-Vis: a System for Visual and Interactive Information Retrieval for Parliamentary Discussions in Twitter

The following description is based on the content of the paper by MAKKI *et al.* (2017).

Twitter is one of the most popular micro-blogs nowadays and more and more people have been using the platform as a way of opining on a variety of subjects. For politicians, political analysts and journalists, in particular, being able to know what is being said about a subject of interest, such as political decisions and bills under debate, is of the utmost importance since much of their work depends on it. However, due to the brevity and noisy nature of Twitter content, it is difficult to formulate queries that match relevant posts without retrieving unwanted content. In this context, MAKKI *et al.* (2015) proposed a framework with the specific goal of retrieving Twitter posts related to political debates and associate them to the specific debate they refer to. The main contribution of the framework is the proposal of a set of active retrieval strategies that make use of the Twitter's structural information and increase the retrieval accuracy while minimizing user effort by keeping the number of labeling requests to a minimum. However, the user involvement is restricted to the task of providing labels; any other interaction, such as choosing a different instance to be labeled, is not supported.

In order to involve the user in the retrieval process so that s/he can benefit from the active retrieval strategies while exploring the data and gaining a better understanding of the results, an interactive and exploratory tool, named ATR-Vis (Active Tweet Retrieval Visualization), was introduced in the work by MAKKI *et al.* (2017). Particularly, the visual interface (e.g. the front-end) of the ATR-Vis system was design and developed as part of this work. ATR-Vis consists of three major components for handling the problem, as illustrated in Figure 17 – the first two were introduced by MAKKI *et al.* (2015) and the third one employs a visual interface developed within the scope of this M.Sc dissertation project. The first component automatically formulates the initial query by extracting discriminative features (the most discriminative words for topical categorization) from a document representative of the user's information need (e.g. transcripts of debates or news stories), and then performs an initial unsupervised retrieval. Next, relying on the idea of query expansion XU; CROFT (1996), the initial query is iteratively refined/expanded based on the retrieved tweets. Additionally, in order to diminish problems related to the brevity and noisy nature of Twitter content, structural information of the tweets

[1] <https://www.dal.ca/> [Accessed 1 June 2017]
[2] <https://projects.cs.dal.ca/malnis/> [Accessed 1 June 2017]

such as hashtags, user mentions and URLs are also added to the list of discriminative features. Moreover, as the analysts may be interested in understanding what people think about multiple topics/debates, an additional step is performed to assign each of the retrieved tweets to a single debate – including a "non-relevant" class when the similarity score between a tweet and each of the debates is below a given threshold.

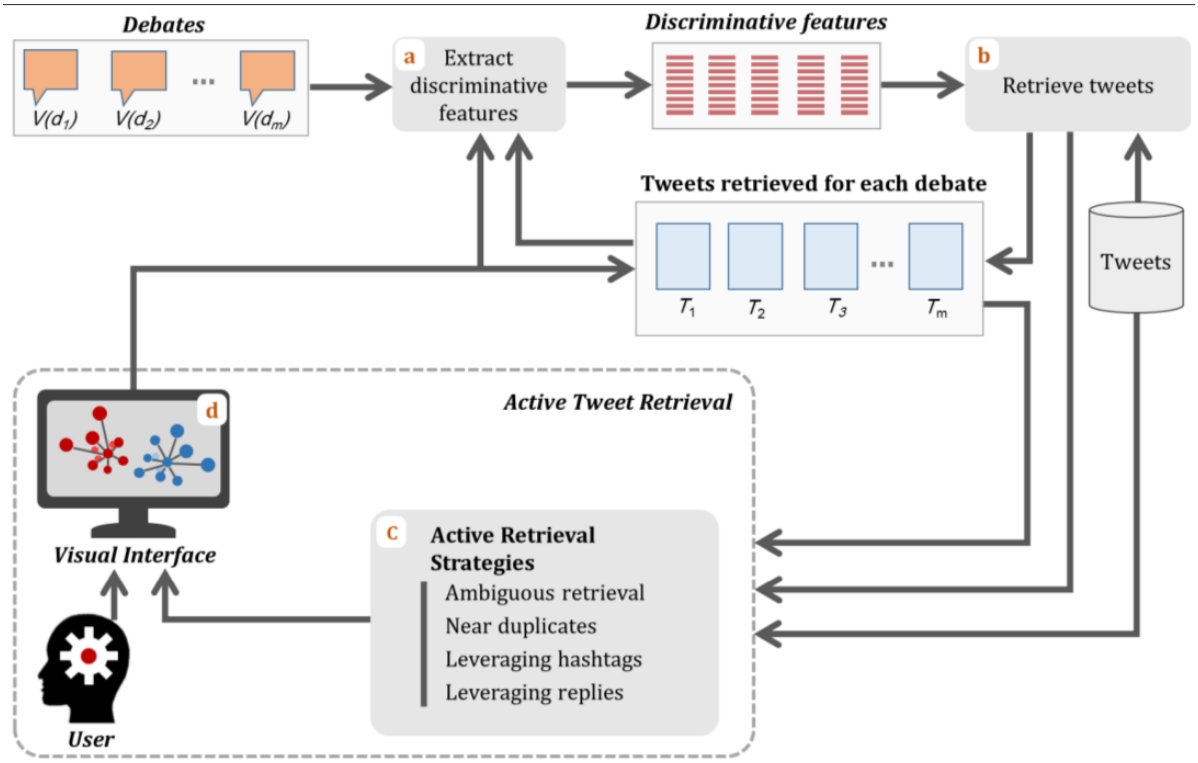


Figure 17 – The proposed framework for retrieving tweets relevant to a set of political debates. The unsupervised retrieval consists of extracting discriminative features (a) and retrieving tweets (b), and the active retrieval component selects the labeling requests (c) and updates the retrieval model based on the obtained labels (d). Figure and description extracted from MAKKI *et al.* (2017).

The second component proposes four active retrieval strategies with user involvement for improving retrieval accuracy. The strategies are referred to as: *ambiguous retrieval*, *near-duplicates*, *leveraging hashtags* and *leveraging replies*. In the first strategy (*ambiguous retrieval*), tweets that have very similar scores for more than one debate are selected. Since the method assumes that each tweet is related to at most one debate, providing the correct label for the tweet will cause the method to change the weights of the discriminative features found in that tweet for those debates. The second strategy assumes that very similar tweets (near-duplicates) should be assigned to the same debate. Therefore, the method creates clusters of near-duplicates tweets and ranks them based on their likelihood to belong to a debate and their cardinality. Then, based on this rank tweets from these clusters are selected and presented to the user to be labeled, which will cause all the tweets from that cluster to be assigned to the same debate. The third strategy introduces the concept of stop and specific hashtags. A stop hashtag would be the analog of a stop word in natural language processing, meaning that they are not discriminative for topical

categorization, while specific hashtags are just the opposite. In order to determine when a hashtag is specific, the method first ranks each hashtag in ascending order based on its debate frequency, i.e. the number of debates in which the hashtag appears in the content of the retrieved tweets for that debate, then the top ranked ones are more likely to be considered specific. These hashtags are then picked to select tweets to be labeled by the user, which will cause all the tweets containing that hashtag to be assigned to the same debate. According to Twitter's help center, a hashtag is used on Twitter to index a keyword or topic on Twitter, which therefore allows people to easily follow topics they are interested in[3]. This means that hashtags should be the most useful of the Twitter's features when trying to discover the topics addressed by a tweet. Experiments in the work by MAKKI *et al.* (2017) have shown that this assumption is partially true, since a mistake in the identification and assignment of specific hashtags could lead to a huge decrease in the accuracy. The last strategy leverages the conversational aspect of Twitter by back tracking reply tweets to their source and then grouping together all reply tweets that share the same source in order to obtain a reply chain. It is expected that a reply tweet will be addressing the same topic as the replied tweet, therefore a reply chain in which the tweets are split uniformly among all debates will be more likely to be selected to be presented to the user for labeling.

The third component of the system employs a visual interface. The visual interface was developed as part of this work for the active tweet retrieval strategies described in the second component and support user involvement in the process. The application is web-based built using the D3.js library BOSTOCK; OGIEVETSKY; HEER (2011) and Bootstrap[4].

For the purpose of presenting the user with the instances selected by the active retrieval strategies and to enable the user to explore the results of the processes and to provide feedback beyond what was initially requested by the system, there are two essential panels: the *Assignment* panel (panel *a* in Figure 18) and the *Context* panel (panel *b* in Figure 18). The first panel is used to present the selected instances; it shows the textual content of the posts and allows a user to drag and drop them into one of the available classes. Since ATR-Vis does not have a fixed number of classes, its debates are presented as a vertical list in a *Debates* panel (panel *c* in Figure 18). Each debate (or class) is consistently associated with a color that is preserved throughout the interface. Additionally, by clicking on the debates in the *Debates* panel, the list of system-extracted discriminative features is shown as a sequence of terms (panel *d* in Figure 18). If the user believes that a feature has been mistakenly associated with a debate, s/he can change it by dragging and dropping the feature to the appropriate debate. The second essential panel (panel *b*) provides contextual information and visual clues to assist the user in making decisions. The Force-Layout and the Ring Visualization (or chord diagram) aim at depicting the relationships between the instances. Both rely on a graph metaphor, where the nodes represent the instances and an edge connects a pair of nodes if the pairwise similarity

[3] <https://support.twitter.com/articles/49309> [Accessed 1 June 2017]

[4] <http://getbootstrap.com>: a web framework for designing websites and web applications [Accessed 1 June 2017]

score of the corresponding instances is above a user-provided threshold. Additionally, a bar chart is used to inform the debate frequency distribution of any non-stop words of the vocabulary. Visualizations and panels are coordinated so that certain interactions, such as the node or element selections performed in one of them will immediately have an effect in the others (e.g. selected elements are highlighted).

ATR-Vis presents an additional view, named *More*, that takes advantage of the tweets' structural information regarding to the leveraging of replies and hashtags, as illustrated in Figure 19. The *Reply Tree* panel (panel *f*) shows the reply-based conversation of a source tweet displayed as a tree layout. Aided by the node colors, the user can provide the labels of specific tweets or of the whole chain at once. The *Similarity Hashtag-Debate* panel (panel *g*) shows the system selected hashtags due to its potential to improve retrieval accuracy in case supervision is provided. The information is also displayed as a tree layout where the left node (root of the tree) depicts a selected hashtag and the nodes on the right depict the debates. The *Retrieved Tweets* panel (panel h) shows a sample of the retrieved tweets that include the hashtag on focus (from panel *g*) in their content.

3.3 Datasets

The following description can also be found in MAKKI *et al.* (2017).

Experiments with ATR-Vis have been conducted on three datasets. The first two are parliamentary datasets while the third one is related to news stories. The first parliamentary dataset refers to the Canadian House of Commons during the period 12-16th May 2014. The 11 debates that received most attention in the parliament during that week (measured in terms of their overall length of discussion), have been selected since these were more likely to generate an expressive number of opinions in social media. The second dataset refers to 5 mainstream debates being held in the Brazilian Federal Senate from 25th to 29th May 2015. Transcripts of the selected debates were extracted from the respective parliament websites[5,6].

Twitter's streaming API[7] has been employed to collect tweets during the weeks of interest. Since it returns a minor fraction of the total volume of tweets at any given moment (roughly 1%), it is necessary, to the maximum extent, to restrict the search to Canadian (or Brazilian) political tweets in order to gather as many relevant tweets as possible without introducing many spurious ones. Furthermore, the collection procedure should not be biased towards any specific keywords, or the resulting datasets would depend on the choice of keywords and will likely report these keywords as being relevant. Bearing this in mind were used as initial information the parliament members' Twitter accounts. Twitter user names for the Canadian Parliament members were

[5] <http://www.parl.gc.ca/Default.aspx?Language=E> [Accessed 1 June 2017]
[6] <http://dadosabertos.senado.gov.br/> [Accessed 1 June 2017]
[7] <https://dev.twitter.com/streaming/overview> [Accessed 1 June 2017]

obtained from the Politwitter website[8], while for the Brazilian case 80 Twitter user names, out of 81 senators, were identified manually. Tweets were collected that were either posted by a parliament member, replied to a post by any of them, or included one of their Twitter user names in its text. This resulted in datasets containing 16,297 and 9,625 original tweets (no retweets) for the Canadian and the Brazilian cases, respectively.

The third dataset refers to news stories that received great attention from the media from 15^{th} to 27^{th} July 2016, namely: “Terrorist attack in Nice”, “Brexit”, “Colombia’s government and FARC”, “Dallas shooting”, “Israeli-Palestinian conflict”, “Killing of Afro-Americans”, “Orlando nightclub shooting”, “Refugee crisis”, “Rio 2016 Olympics”, “Turkey attempted coup” and “US Presidential campaign”. News articles from CNN and Fox News related to each story have been considered to extract keywords and set the initial queries. The tweets collected during this period accounted for 9,277,751 after retweets and non-English posts were discarded. Twitter user names were obtained from a list created by the StatSocial[9], which presents a rank of the “1,969 most politically influential journalists and bloggers". In addition, accounts of newspapers of high circulation (e.g. The Wall Street Journal) were also used in order to gather relevant tweets, totalizing 118 user names (105 journalists and 13 newspapers). The resulting datasets are public available[10].

In order to evaluate the visual interface of the system described in Section 3.5 the Amazon Review dataset shared by BLITZER; DREDZE; PEREIRA (2007) was employed. This is a multi-domain sentiment dataset consisting of 8,000 product reviews across four domains: Book, DVD, Electronic and Kitchen.

3.4 Results

This section presents a brief description of the ATR-Vis evaluation results. The evaluation includes use cases and quantitative experiments, as well as qualitative analysis conducted with three domain experts. A more in depth description of the experiments and results can be found in the paper by MAKKI *et al.* (2017). Additionally, a new case in a scenario where the task is to retrieve tweets related to news stories is introduced.

3.4.1 Retrieval Results

The performance of ATR-Vis has been evaluated considering multiple metrics, namely, *accuracy*, *macro-precision*, *macro-recall*, *Mean Average Precision* (MAP) and *R-Precision*. Furthermore, the proposed approach was compared with a modified version of ReQ-ReC by LI *et al.* (2014), a state-of-the-art active retrieval method, and with a random-based selection strategy,

[8] <http://politwitter.ca/page/canadian-politics-twitters/mp/house> [Accessed 1 June 2017]
[9] <http://www.statsocial.com/#!/social-journalists> [Accessed 16 June 2017]
[10] <https://web.cs.dal.ca/~niri/#research?code=res201501&resrc=0> [Accessed 16 June 2017]

which serves as a baseline for the proposed approach. The results of applying both approaches to the Canadian politics dataset are shown in Table 2.

Table 2 – Accuracy, macro-precision, macro-recall, R-precision and MAP for the unsupervised retrieval method, a random active retrieval strategy, ReQ-ReC, and the ATR-Vis' selection strategies using the Canadian dataset. *The number of labeling requests reported for ReQ-ReC is an average over all debates. Table and description extracted from MAKKI *et al.* (2017).

Retrieval Method	Accuracy	Macro-Pr	Macro-Re	R-precision	MAP	#Requests
Unsupervised (1^{st} iteration)	0.61	0.74	0.55	0.67	0.70	0
Unsupervised	0.80	0.75	0.68	0.70	0.71	0
Random active retrieval	0.81	0.76	0.70	0.71	0.73	100
Req-Rec	0.29	0.26	0.70	0.66	0.64	116*
Ambiguous retrieval (1)	0.83	0.80	0.75	0.75	0.75	15
(1) + Near-Duplicates (2)	0.84	0.81	0.76	0.76	0.76	24
(1) + (2) + Hashtags (3)	0.89	0.82	0.81	0.79	0.80	60
(1) + (2) + (3) + Replies	0.92	0.83	0.86	0.82	0.84	100

3.4.2 ATR-Vis Pair Analytics Evaluation

ATR-Vis has also been evaluated by three domain experts. Evaluation was performed by means of a pair analytics process ARIAS-HERNANDEZ *et al.* (2011), which is carried out with one Subject Matter Expert (SME) and one Visual Analytics Expert (VAE).

The process consisted in three stages based on two out of seven scenarios introduced in the work by LAM *et al.* (2012). First, the SMEs are asked background questions, such as whether they need to search/analyze Twitter data in their respective professions. Then, an overview of the dataset and a showcase of ATR-VIs is presented to the SMEs by the VAE. Finally, the SMEs, assisted by the VAE, are asked to interact with the system to conduct the retrieval of tweets and to provide feedback about those interactions and the retrieval results. After the pair analytics session, three main questions were posed: "What advantages and possible other uses you find for ATR-Vis?", "Would you consider using this system for your own work/research?" and "What limitations did you find and/or what suggestions can you give us to improve the tool?"

A complete description of the process and the SMEs answers can be found in MAKKI *et al.* (2017). Answers to those questions include:

What advantages and possible other uses you find for ATR-Vis?

"I can see this being useful at various points along the process for a journalist, one is looking for people... When you are assigned to a story and you are doing background information, so one way would be to find people. Because if you find people who are actively engaged on Twitter, you can track them down, you can call them up, you can do interviews"

Would you consider using this system for your own work/research?

"This has actually exceeded all of my expectations because it just makes the possibility of my

research big [sic]" and *"This is something that I would use for every single piece of research, something that students can do master theses on"*

3.4.3 Use Cases

Visual Interface of ATR-Vis

This section illustrates the application of ATR-Vis for the retrieval of tweets related to a set of news stories that received great media attention at the time of the retrieval. This use case was made in order to showcase the suitability of the tool to a domain other than parliamentary debates. Nevertheless, the application of ATR-Vis to the Canadian Parliament debates and Brazilian federal debates can be found in MAKKI *et al.* (2017).

The user is presented with tweets that are relevant to a set of news stories rather than parliamentary debates, although some of them are of a highly political nature. Furthermore, due to the recent terrorist attacks and events of a violent nature in overall, the unsupervised retrieval has to deal with debates that are very similar to each other content-wise. For example, the very first labeling request, shown in Figure 18 with label (a), has "Colombia's Government and FARC" as the debate with the highest similarity score. Although a user could easily note the expression "radical Islamist terror", and therefore infer that this tweet more likely belongs to one of the debates related to the recent terrorist attacks, s/he may find it difficult to choose between one of them due to the lack of specific references (e.g. names and dates). However, by inspecting the article the URL "https://t.co/rHzdL2WKtb" is referring to, which is actually a post published by "@jaketapper" — the Twitter account mentioned in the request — the user could find enough evidence that the tweet is related to the attack in Nice, and therefore decide to assign it to that debate. This in turn could lead to the incorporation of the URL and the mentioned Twitter's user name to the list of discriminative features of the chosen debate.

Assuming that the user is interested in all the stories, s/he could start by taking a look at the list of tweets retrieved for each debate. However, after inspecting the number of tweets retrieved for each debate, s/he would find out that, with the exception of "US Presidential Campaign", most of them are under-represented, and s/he may want to explore the non-retrieved ones in order to find an explanation. The ring visualization, as shown in Figure 18 whith label (b), is the most recommended in this case since it shows the stratified sample of non-retrieved tweets arranged together in a portion of the circle. By interacting with the link threshold (localized on the bottom of the visualization) s/he would notice that the number of connections is already very high from the beginning (the threshold is 1.0 by default), which can indicate that the dataset is formed by closely-related topics, as stated above.

Additionally, any tweet that is highly connected to more than one debate could be an indicator of cases similar to the previous one (in which the stories are about the same topic, e.g. terrorism), and therefore its manual labeling to the correct debate will contribute to add

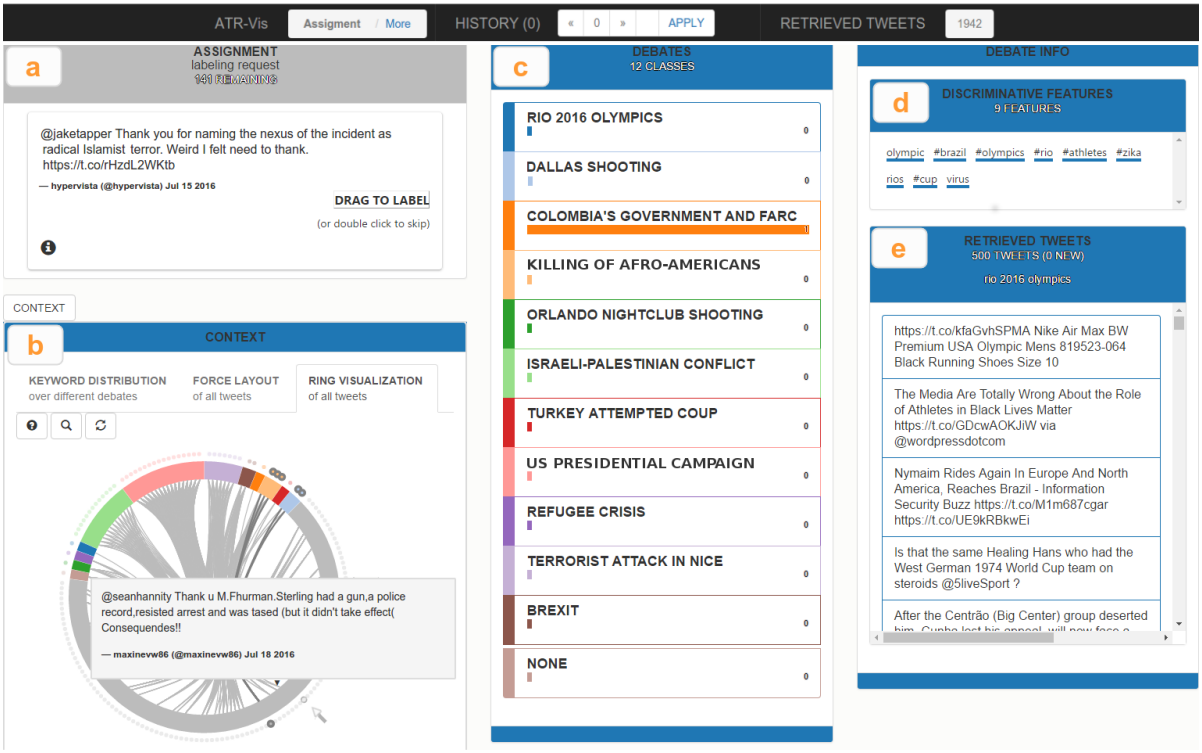


Figure 18 – Assignment View: a set of visual aids to facilitate tweet retrieval. (a) Labeling request for a Twitter post. (b) Visualization of the labeling requests in a broader context. (c) List of debates of interest. (d) Discriminative features for the selected debate. (e) Tweets retrieved by the selected debate. Description extracted from MAKKI *et al.* (2017)

features that are highly discriminative for that debate. In this case the tweet is "@seanhannity Thank u M.Fhurman.Sterling had a gun,a police record,resisted arrest and was tased (but it didn't take effect(Consequendes!!", and it has links to the debates "Killing of Afro-Americans" and "Dallas Shooting". Assuming that the user is aware of the recent events that took place in the USA related to those topics (both involving the police and black men), s/he would recognize "M.Fhurman.Sterling" as being the name of the former detective who made a statement about the killing of a black man by the police of Baton Rouge city, and therefore its assignment to the correct debate "Killing of Afro-Americans" could also contribute to increasing the retrieval accuracy.

After labeling some requests and submitting the feedback, the user may observe that the overall number of tweets retrieved is still low. In this case, s/he could take advantage of the strategy of leveraging hashtags, which is already known as likely to affect many tweets and therefore contribute to improving the retrieval recall. With this goal in mind, the user could start by selecting the hashtags appearing in many tweets, such as the hashtag "#dnc" (whichs stand for *Democratic National Committee*). It appears in 317 tweets of which 216 were mistakenly retrieved to the debate "Terrorist Attack in Nice", as indicated in Figure 19 (see area labeled (g)), whereas the correct label would be "US Presidential Campaign",

since it refers to the formal governing body of the United States Democratic Party[11]. Another straightforward approach would be to start by labeling hashtags such as “#trump”, “#clinton”, “#dallas”, “#turkeymilitarycoup”, etc., since their meaning is quite obvious to any user. Even before applying the new changes, the interface already shows an increase to 330 tweets retrieved, and after submitting the new batch of feedback s/he can observe an increase of nearly 57% in the number of retrieved tweets.

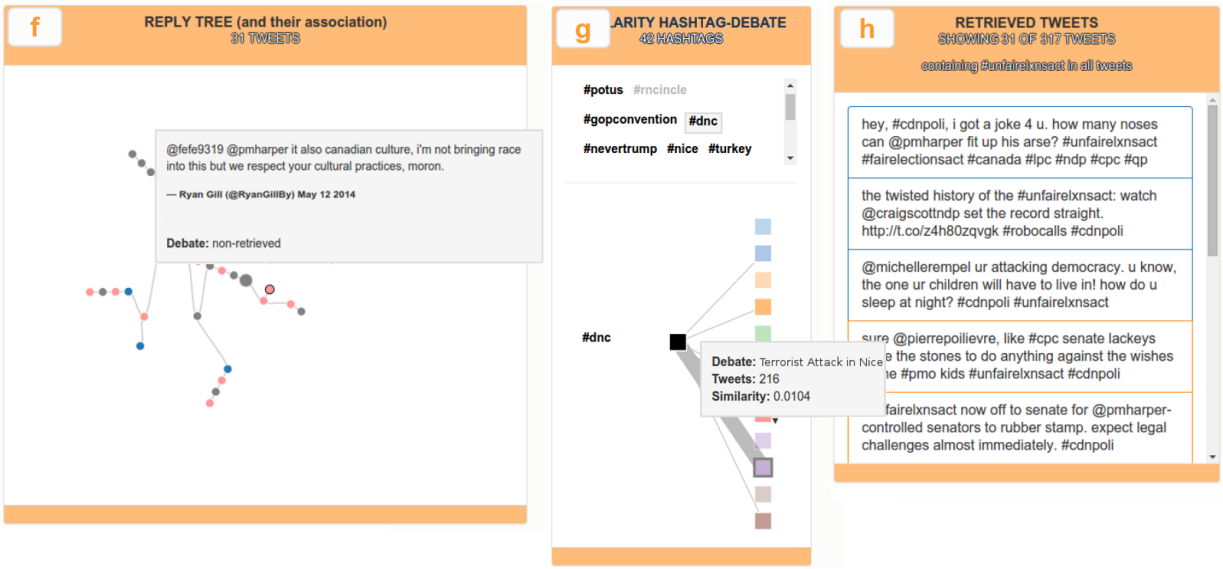


Figure 19 – More View. (f) Exploring a conversation thread on Marine Mammal Regulations. (g) Exploring how the hashtag #unfairelxnsact is associated with different debates. (h) Enumeration of tweets containing a hashtag. Description extracted from MAKKI *et al.* (2017)

3.5 Active Learning with Visualization for Text Data

The second system presents a pilot study of the usage of visualization to support active learning for text classification. The method assumes that a small group of labeled data instances is available in order to train a classifier, and that the active learner can select instances based on the outcome of this classifier and on a pool-based sample of the remaining unlabeled instances. The study employed the classifiers Support Vector Machine (SVM) and the Naive Bayes, and three query strategies for selecting key instances to be labeled by the annotator: uncertainty sampling, query-by-committee (QBC), and variance reduction. The process is illustrated in Figure 20.

As both ATR-Vis and the system described in this Section rely on active learning (or active retrieval) strategies to improve the accuracy of the process, they share a common requirement, which is the need to present the instances selected by the strategies to the user for labeling. Moreover, both aim at giving the user the opportunity to explore the results of the

[11] <https://en.wikipedia.org/wiki/Democratic_National_Committee> [Accessed 1 June 2017]

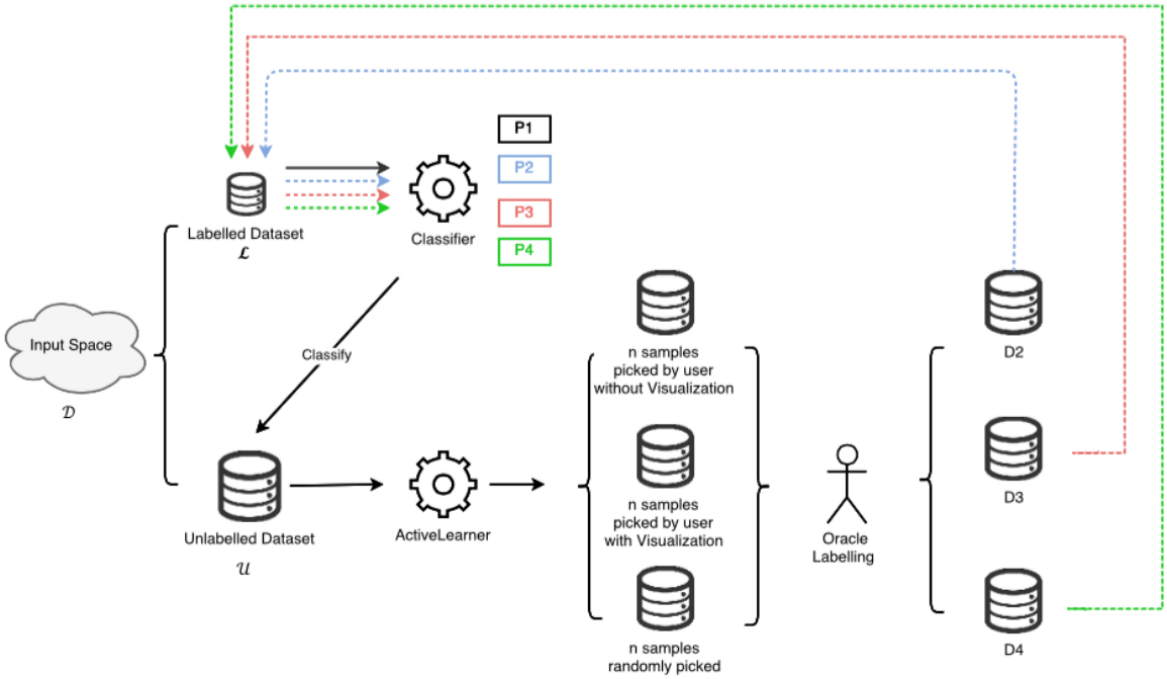


Figure 20 – Active learning workflow. Starting from the input data pool D, an initial text classifier is trained on labeled dataset L with performance $P1$. This classifier runs an active learning algorithm on the unlabeled data pool U. A human annotator labels queries from three settings. These three experiments generate different influence to the classifier: $P2$, $P3$, and $P4$, which are applied to compare the roles of visualization and active learning. Figure and description extracted from HUANG *et al.* (2017).

processes and to provide feedback beyond what was initially requested by the system. In this way, elements of the visual interface developed for the ATR-Vis system could be easily adapted and employed in other applications, such as the one by HUANG *et al.* (2017); specifically, the two essential panels: the *Assignment* panel (panel *3* in Figure 21) and the *Context* panel (panel *4* in Figure 21). The Force-Layout and the Ring Visualization (or chord diagram) are common to both interfaces, while the bar chart, is unique to the first interface, and the T-distributed Stochastic Neighbor Embedding (T-SNE), which is a dimensional reduction technique well suited to show content-based relationships between the documents, is unique to the second interface. Moreover, the two classes of the second system (i.e. negative and positive) are presented in two different panels (panels 1 and 2 in Figure 21).

3.6 Final Remarks

In this chapter two visual interfaces were introduced aimed at involving analysts in the retrieval process so that they can benefit from a set of active retrieval strategies. The visual interface comprises multiple interactive visualizations that, besides presenting labeling requests, enable the exploration of the data retrieved as well as the labeling of instances beyond those initially suggested by the system. As such, user engagement into the retrieval process by means of visualizations was shown to contribute to improved recall, as further described in the papers

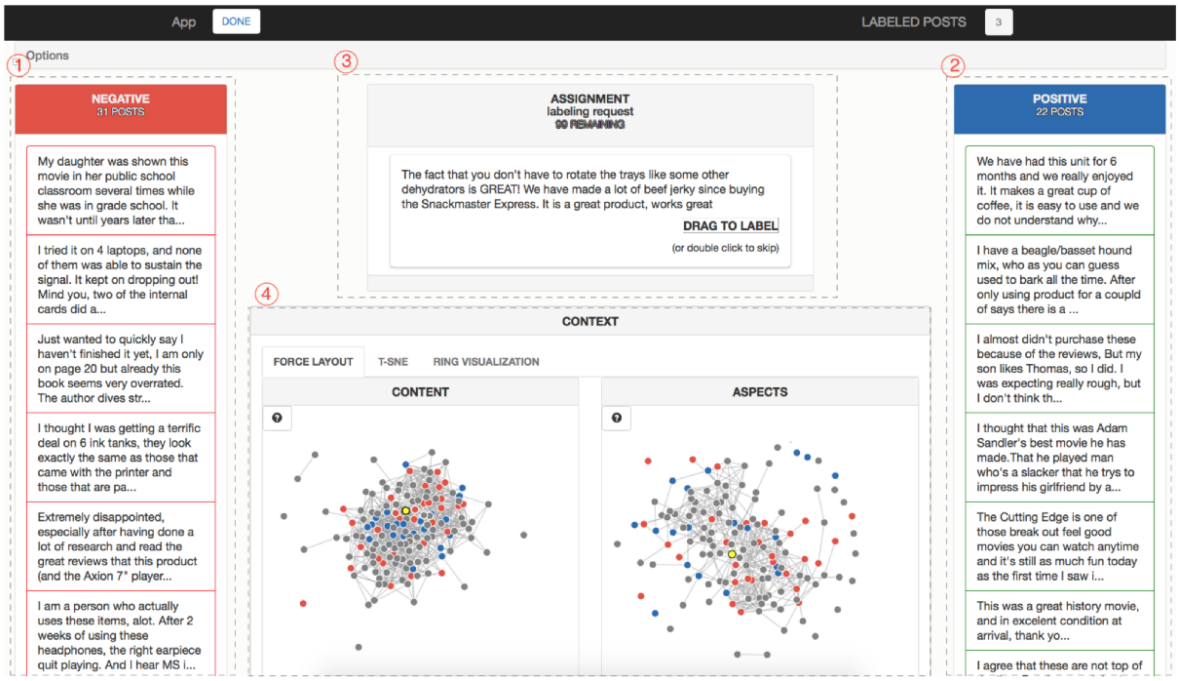


Figure 21 – Overview of HUANG *et al.* (2017) interface. Panels 1 and 2 present each of the classes and a vertical list with documents assigned to that class. Panel 3 presents the selected query. Panel 4 presents information for context through three different visualization techniques: Force Layout, T-distributed Stochastic Neighbor Embedding (T-SNE) and chord diagram. The different active learning strategies and classifiers can be changed through the navigation bar. Figure extracted from the respective work.

by MAKKI *et al.* and HUANG *et al.*.

CHAPTER

4

A TOOL FOR VISUAL ANALYTICS OF TOPICS IN TWITTER IN CONNECTION WITH POLITICAL DEBATES

4.1 Initial Remarks

In previous chapters we discussed several contributions that rely on visualization techniques and data mining methods to support the analysis of textual data obtained from one or multiple sources. Such contributions suggest that in many application domains the need of combining different techniques and methods from mining and visualization often leads to the development of a visual analytics tool. Similarly, this work proposes a method which combines text mining methods with interactive visualizations into a visual analytics tool to help analysts in correlating data from two distinct sources.

Nevertheless, there are some crucial differences between the method proposed in this chapter and the previous ones. The visualization of the temporal evolution of the data, for instance, is addressed in this work and in many contributions described earlier. However, here the visualization attempts to evidence the existing associations, or correspondences, between two data sources within a time window along with the temporal evolution of the data, instead of treating results as a single data stream. The process is described in the following sections.

Yet, the proposed method is similar in purpose to that of the ATR-Vis framework described in Chapter 3.1 as both aim at associating political tweets to parliamentary speeches. However, although in ATR-Vis tweets can be retrieved over time, the framework provides no support for visualizing the temporal evolution of the content of the retrieved tweets. Moreover, the content of the parliamentary debates remains static during the whole process, which may result in outdated debates and consequently in a decrease in the retrieval process performance.

This chapter is organized as follows. Section 4.2 discusses the challenges faced in finding

a good text representation for the textual data under analysis and details the proposed text processing approach. It also introduces a new visualization technique for conveying the temporal evolution of topics from two different sources. Section 4.3 describes the visual analytics tool designed to combine the results generated by the proposed text processing and visualization. Section 4.4 outlines the parameter settings used to perform the process. Section 4.5 presents the results of applying the proposed method to the Brazilian dataset that comprises parliamentary speeches and political tweets. Section 4.6 discusses some characteristics of the method, whereas Section 4.7 summarizes the content of this Chapter.

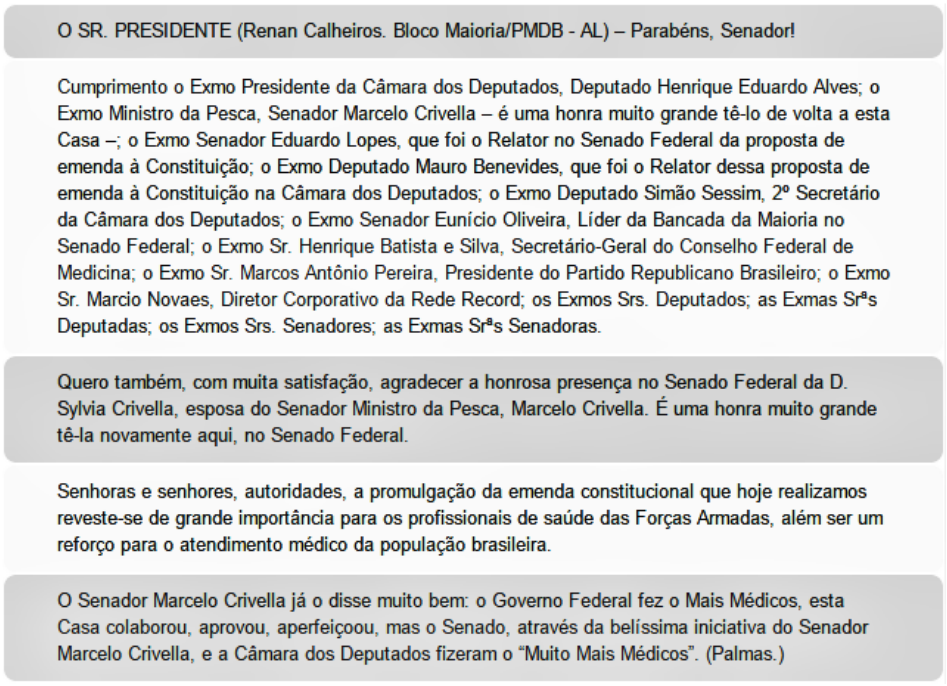
4.2 Document processing

A variety of approaches could be considered in order to achieve the goals introduced in Section 1.2. However, such an approach should not assume the existence of an initial set of debates to begin with and should account for the distinguishing characteristics of textual corpora obtained from different sources. Moreover, the process should consider the temporal evolution of the set of textual documents under analysis.

Our proposed method employs a number of techniques for handling the problem and has two major components. The first component exploits a set of techniques available in the literature in order to leverage the document representation and make it more suitable for grouping together speeches and tweets that share an underlying set of topics or talk about the same debate. In other words we start by modeling the research problem as a clustering problem. This clustering component is also responsible for tackling the challenging temporal requirements of the problem. The second component is a user interface that employs interactive visualizations to present the results prepared in the previous step and support the kind of analysis required to answering our motivating questions and goals.

4.2.1 Document Representation and Summarization

In this work we focus on two domain specific corpora, namely a collection of parliamentary speeches and a collection of tweets. The first and most remarkable difference between them is document length. While Twitter restricts the length of its posts up to 140 characters, parliamentary speeches suffer no length restrictions except those posed by delivery time. Moreover, tweets typically address one topic at a time, whereas in speeches topics can drift considerably from one paragraph to the next. For instance, a senator can start by greeting everyone in the room, then s/he makes some comments about the "hot topic" of the week, and in closing the speech s/he can revive an old debate that s/he believes still needs to be discussed. From a text mining point of view, such a text can be seen as a concatenation of segments, in which each segment consists of one or more chunks (usually sentences or paragraphs) and the boundaries are placed whenever there is a drift in the topic. Figure 22 shows an example of a parliamentary speech.


O SR. PRESIDENTE (Renan Calheiros. Bloco Maioria/PMDB - AL) – Parabéns, Senador!

Cumprimento o Exmo Presidente da Câmara dos Deputados, Deputado Henrique Eduardo Alves; o Exmo Ministro da Pesca, Senador Marcelo Crivella – é uma honra muito grande tê-lo de volta a esta Casa –; o Exmo Senador Eduardo Lopes, que foi o Relator no Senado Federal da proposta de emenda à Constituição; o Exmo Deputado Mauro Benevides, que foi o Relator dessa proposta de emenda à Constituição na Câmara dos Deputados; o Exmo Deputado Simão Sessim, 2º Secretário da Câmara dos Deputados; o Exmo Senador Eunício Oliveira, Líder da Bancada da Maioria no Senado Federal; o Exmo Sr. Henrique Batista e Silva, Secretário-Geral do Conselho Federal de Medicina; o Exmo Sr. Marcos Antônio Pereira, Presidente do Partido Republicano Brasileiro; o Exmo Sr. Marcio Novaes, Diretor Corporativo da Rede Record; os Exmos Srs. Deputados; as Exmas Srªs Deputadas; os Exmos Srs. Senadores; as Exmas Srªs Senadoras.

Quero também, com muita satisfação, agradecer a honrosa presença no Senado Federal da D. Sylvia Crivella, esposa do Senador Ministro da Pesca, Marcelo Crivella. É uma honra muito grande tê-la novamente aqui, no Senado Federal.

Senhoras e senhores, autoridades, a promulgação da emenda constitucional que hoje realizamos reveste-se de grande importância para os profissionais de saúde das Forças Armadas, além ser um reforço para o atendimento médico da população brasileira.

O Senador Marcelo Crivella já o disse muito bem: o Governo Federal fez o Mais Médicos, esta Casa colaborou, aprovou, aperfeiçoou, mas o Senado, através da belíssima iniciativa do Senador Marcelo Crivella, e a Câmara dos Deputados fizeram o "Muito Mais Médicos". (Palmas.)

Figure 22 – Example of a speech made in the Brazilian Parliament. In this example, the former president of Brazil's Federal Senate Renan Calheiros devoted at least the three first paragraphs to congratulate (the Portuguese word is "parabéns") and to greet (e.g., Portuguese words like "cumprimento" and "agradecer", which mean "greeting" and "to thank", respectively) some of those attending the session. The president then makes some statements about the importance of the promulgation (the Portuguese word is "promulgação") of a particular constitutional amendment.

Since we model the problem as a clustering problem, finding precise segment boundaries is not required and for simplicity and efficiency we treat each sentence as a single document. Moreover, by assuming that a sentence typically addresses a single topic, we obtain documents that are more similar content-wise and therefore we can use a similar approach to handle texts from both sources. On the other hand, short textual segments (e.g., tweets and sentences) suffer from sparsity due to a lack of contextual information, which can result in ambiguous texts and therefore hinders the clustering of documents.

Another significant difference between the collections is the vocabulary employed. Although people can freely express their opinions in Twitter, they must do it bearing in mind that the available space is strictly limited and therefore it is often not possible to provide a complete or well formulated statement. In such a scenario, correct grammar usage and spelling are not as important as picking meaningful words and expressions that capture the user's intent and provide additional information through Twitter's features as, for example, hashtags that indicate the theme of the post. On the other hand, parliamentary speeches often adopt a very formal vocabulary and exaggerate in the amount of greetings and felicitations. Thus, several text segments add little semantic content to an ongoing debate.

As a result, the vocabulary used in each domain may be quite different even for documents

addressing the same topic, thus hindering joint analyses based uniquely on feature selection. In this scenario, no matter how well the method performs by removing noisy words and increasing the importance of keywords, the resulting set of features may not match between the two corpus.

A more advanced approach which could potentially help in the matching would be transforming the feature space in order to obtain a reduced feature space where the features are typically formed by linear combinations of those in the original space AGGARWAL; ZHAI (2012). However, even though the semantic relationship between the words has some influence in the transformation, the large difference in writing styles between the domains creates a semantic gap that is not fully captured in the process. Nevertheless, the process can still be effective if the semantic relationships between the domains are improved.

A possible approach to decrease this gap, and consequently help with the small term overlap problem of short texts, is to use a larger text corpora to identify groups of semantically similar words as topics, and then connect the documents through these topics. Specifically, rather than using the well known Bag-of-Words (BoW) representation, which uses the word frequencies as features, we use the frequency of topical clusters to obtain something similar to a Bag-of-Concepts (BoC) representation, as illustrated in Figure 23.

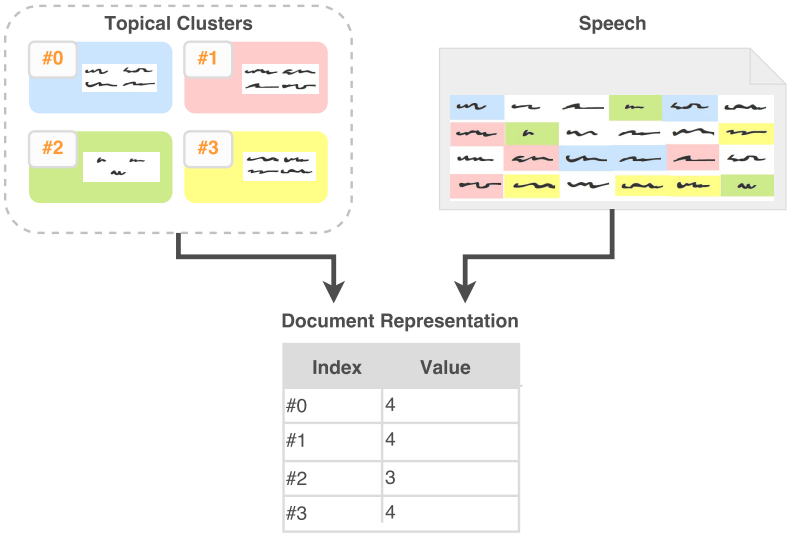


Figure 23 – Illustration of a hypothetical document representation in terms of the frequency distribution of topical clusters.

In order to obtain the topical clusters, we follow a similar approach to that introduced by SAKAHARA; OKADA; NITTA (2014) for text segmentation. Their method consists of applying affinity propagation to cluster words based on two different similarity measures. The first similarity is obtained from a *word2vec* model trained on revisions of Wikipedia articles for the purpose of grouping together semantically similar words. The second similarity is based on computing word collocation frequencies within the document under analysis SAKAHARA; OKADA; NITTA (2014). Each similarity is actually a similarity matrix of the cosine distances computed between pairs of word vectors generated by the models. Finally, the matrices are

combined and used as input to the clustering technique, which will produce the topical clusters.

Although the method by SAKAHARA; OKADA; NITTA (2014) was intended for text segmentation, we notice that the processes of segmentation and clustering have some overlap, which allows us to employ the first stage of their method to enrich the text representation. Moreover, their method is domain-independent, which is particularly suitable to our problem since we must handle different domains. However, in order to obtain more accurate topical clusters and taking advantage of the fact that political issues often become news, we train the *word2vec* model on an existing news corpus. The collocation similarity is obtained after concatenating documents from both sources. The combined similarity is input into the clustering technique to obtain the word clusters that characterize the topics.

Finally, each document (i.e., a speech sentence or a tweet) is represented in terms of their word-cluster (topic) distribution, and the final document clustering can be performed. This procedure is similar to that employed for obtaining word clusters, but instead of word vectors we use the leveraged representation of the documents to generate the document similarity matrix to be input to the document clustering technique — this is known as a two phase clustering procedure AGGARWAL; ZHAI (2012). The whole process is illustrated in Figure 24.

Note that the second phase of the clustering step, as indicated in Figure 24 with label (Phase 2), produces clusters with documents coming from both sources. This essentially means that we are grouping together speech sentences and tweets that share the same underlying set of topics, which is precisely our intend by leveraging the document representation. Moreover, by considering that a debate can comprise one or more topics, each resulting cluster can be seen as a representative of a debate.

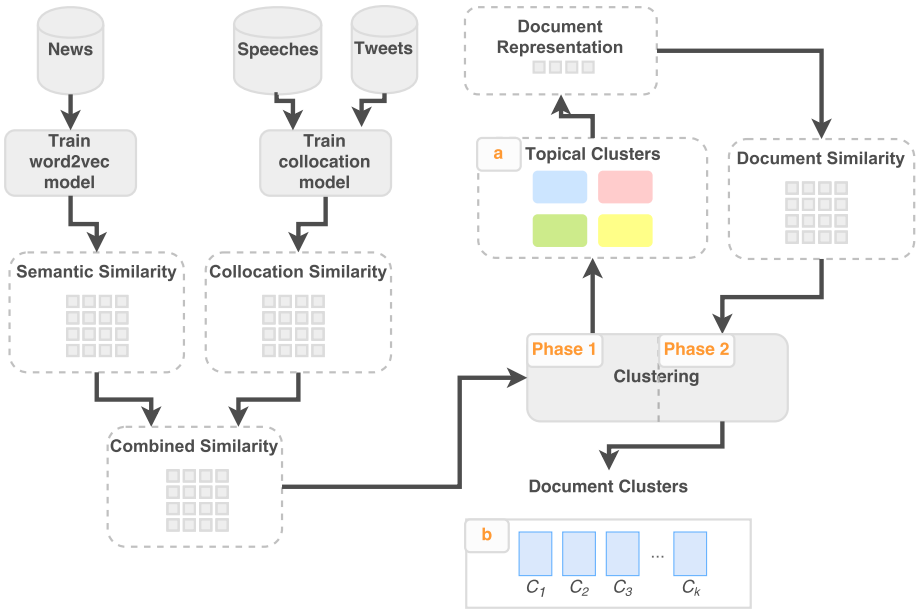


Figure 24 – The proposed method consists of a first phase (Phase 1) for obtaining the topical clusters (a), which are used to generate the document representation, and a second phase (Phase 2) for obtaining the clusters of topically-related documents (b).

As mentioned in previous chapters, hashtags provide important structural information and are often employed to retrieve tweets related to a specific target topic. Therefore, we perform an additional step that leverages hashtags in order to improve cluster quality. The procedure consists of initially identifying representative hashtags h_i from the set of all hashtags h. A hashtag is considered representative if its relative frequency $r(h_i)$, as given by Equation 4.1, is higher than a user-provided threshold.

$$r(h_i) = f(h_i) \log \frac{n}{c(h_i)} \tag{4.1}$$

where $f(h_i)$ is the overall frequency of the ith hashtag, n is the total number of clusters and $c(h_i)$ is the number of clusters in which the ith hashtag occurs. This frequency measure is somewhat analagous to the standard *tf-idf* measure used in information retrieval. Once representative hashtags h_i are identified, any document containing a representative hashtag is assigned to the cluster c_j for which h_i is most frequent. Moreover, any cluster $c_p, p \neq j$ with more than $c\%$ documents including a representative hashtag h_i will be merged with cluster c_j, where c is a user-provided theshold.

It is also worth mentioning that, although tweets may be associated with multiple hashtags, only one (the first) is considered in our processing of tweets.

4.2.2 Temporal Evolution

Finding the debates that have been discussed over a certain a time window can unveil valuable information about what is drawing people's attention on the parliament and on the social media, and also what is of interest to both audiences. However, more complete analyses of the data capable of showing, for example, where the first mention to a subject occurred, or which topics have the longest lifespans, are only possible if the temporal component is taken into account.

Since one of our goals is to track the temporal evolution of debates, including their emergence and/or fading, the initial set of clusters must be updated from time to time to include the new documents arriving. Even though tweets are published continuously in time, the speeches in plenary sessions are published at the end of each day and therefore we choose to perform the re-clustering on a daily basis.

For a good trade-off between accuracy and computational cost, we adopted a version of the framework MONIC SPILIOPOULOU *et al.* (2006) for detecting and tracking changes in clusters. Specifically, we adopt the notion of cluster "matching", which defines whether two clusters at distinct time moments are likely to be the same cluster or not. The concept is illustrated in Figure 25. An overlapping threshold value is defined (based on their common records), and the two clusters are considered to be the same if their overlapping value is above the threshold. The verification must be performed for a candidate cluster against all others, and therefore there

may be multiples cases in which the overlapping value is above the threshold. The candidate cluster is considered to be the same as the one for which the overlapping value is maximum.

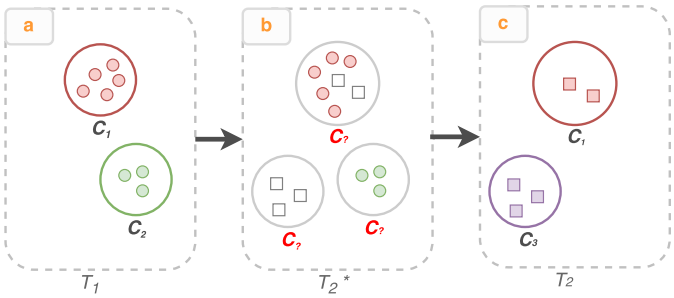


Figure 25 – Illustration of the cluster matching process. At timepoint T_1 (a) there are 8 documents distributed in two clusters C_1 and C_2. At timepoint T_2^* (b) the document set is clustered again, now including the documents added at this current time point (T_2), which results in three yet unlabeled clusters. At time point T_2 (c) one of the unlabeled clusters is identified as being the same as C_1 (as most of its documents come from C_1) and a new cluster C_3 emerges. An additional step is performed to remove documents from previous time points and any empty clusters.

Note that the overlapping value between cluster X at timepoint t_i and cluster Y at timepoint $t_j > t_i$ requires that Y includes documents that were already available at timepoint t_i or before. Eventually, the number of documents in Y can become very large to the point of causing performance problems. In order to overcome this problem, and possibly improve the quality of the clusters, we "forget" older documents as compared to the current timepoint. In the original paper describing MONIC, authors defined an aging function. Our implementation defines a constant value: documents are removed when they become older than c days, where c is a user-defined constant.

As a result, a time-stamped set of clusters is generated. Although for the cluster matching we need to analyze documents published at least on two different days, for the final result each cluster only keeps those documents that exhibit the same day as the cluster, as indicated in Figure 25. Finally, the clusters are input to the second component that will create the interactive visualizations.

4.2.3 Visualization

From Chapter 2, we note that many visualization techniques are available and several applications can benefit from them. Although there is no explicit rule that establishes that a specific technique should be chosen for solving a particular problem, the summary overview of text visualization techniques by ALENCAR; OLIVEIRA; PAULOVICH (2012) reveals that the river flow metaphor is often employed in visualizations designed to support temporal analysis of document and news collections. This is due to its ability to convey the temporal thematic changes in documents, rather than, for example, content-based relationships which are better conveyed by document similarity maps, such as those created, e.g., with multidimensional projections.

Contributions described in Chapter 2 confirm that this metaphor is still widely employed for visualizing topic evolution in documents.

We also picked the river metaphor as the underlying metaphor for our novel visualization of the results generated by the process described in Section 4.2.2. The whole idea behind the use of clustering in this work is to be able to identify the debates that are common to both sources and those that are unique to one of them. In order words, we want to establish an association between textual data coming from two distinct sources, based on their common topics. Therefore, we employ the river metaphor in a way that it simultaneously presents not only the temporal evolution of the topics in the two distinct sources but also what they have (or not) in common. Figure 26 describes the core elements of the proposed visualization.

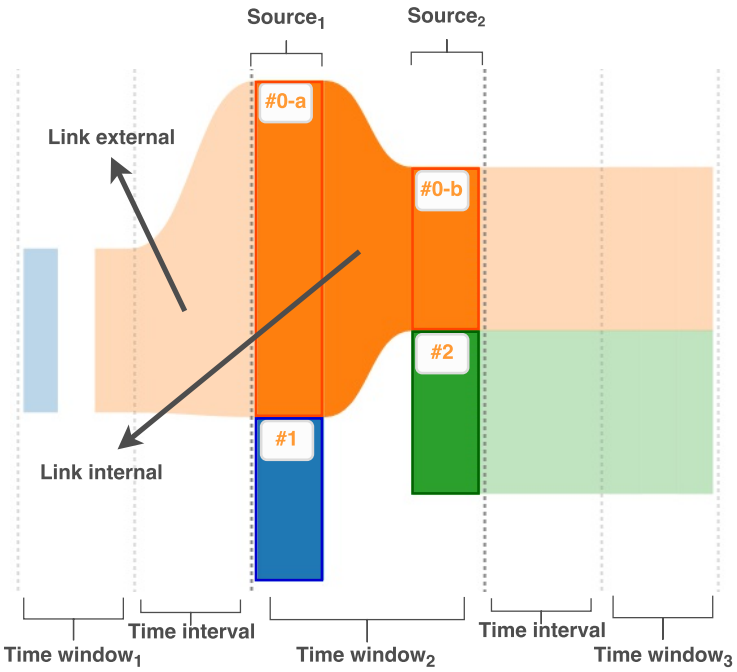


Figure 26 – Core elements of the proposed visualization. The x axis is split into time windows and time intervals. Within a time window, space is split into two opposite areas, one associated with each textual source (parliament or twitter, in this case). Within the text source area for a given time-stamp, each colored rectangle represents a cluster of topical documents relative to this specific time-stamp. Links internal to a time window and external to other time windows through a time interval, indicated by filled areas, depict the associations that exist between the two sources within the current time-stamp and over time, across different time stamps. (#0-a) and (#0-b) illustrate a hypothetical case of a cluster with documents from both sources that has been split into two areas, which are linked to depict their association. Their connection is shown by filling the area in-between with the same color as the clusters .

The visualization is built around the set of time-stamped clusters given as input. The clusters are visually encoded as colored rectangles, as shown in Figure 26 with labels (#0-a), (#0-b), (#1) and (#2). Rectangle height is proportional to the number of documents assigned to the cluster and its color is uniquely determined by the cluster index, so each color identifies a different cluster.

To accommodate the clusters the x-axis is split into d time windows, where d is the number of time-stamps, and only a user-defined number of windows are visible at each interaction. Following a focus+context approach MUNZNER (2014), the central time window is brought into focus (its elements are magnified) and the remaining windows are shown for context. Each time window is mapped to a time-stamp so that it only comprises clusters that share the same time-stamp. Within each window, the left side and the right side represent documents from each of the two sources. In this case, the left side refers to Parliament speeches and the right side refers to posts from Twitter. Specifically, if a cluster is formed by documents coming from a single source, then it is placed on its corresponding source's side, otherwise the cluster representation is split into two rectangles at opposite sides, so that each side refers only to those documents originating from the same source, as shown in Figure 26 for clusters (#0-a) and (#0-b).

The rectangles are arranged vertically following the order established by their corresponding cluster index. If rectangles on both sides have the same cluster index, then a connection area is drawn to depict associations between the two sources within a time window — the boundary shapes delimiting the connection area are defined by cubic Bézier curves. In order to smooth the task of tracking debate evolution and give it a more fluid appearance, connecting shapes (external links) are also drawn to connect clusters from different time-stamps.

4.3 Visual Interface

We propose a system that supports visual analysis of a collection of tweets and text from speeches occurring in parliament houses, aiming to afford a better understanding of the existing relationships between the two along time. The interface relies on the clustering approach and visual metaphor previously described in Section 4.2 coupled with multiple interactive visualizations.

The proposed visual interface consists of two main views: *Flow View* and *Content View*. The *Flow View* was designed to accommodate the proposed visualization and the *Content View* presents multiple secondary views for providing the user with details about the method and the data. The system is a web-based application. The frontend was build using D3.js BOSTOCK; OGIEVETSKY; HEER (2011) and Bootstrap[1], while the backend was written in Python. Sections 4.3.1 and 4.3.2 explain the *Flow View* and *Content View* view respectively.

4.3.1 Flow View

The visual interface comprises two major views, namely the *Flow View* (Figure 27) and the *Content View* (Figure 28). Similar to the visual interface for retrieval of tweets described in Chapter 3, each debate is consistently associated with a single color that is preserved in both views. However, as the text processing stage can create an arbitrary number of clusters

[1] <http://getbootstrap.com> [Accessed 1 June 2017]

it may not be possible to map a unique color to each of the cluster indexes. Moreover, the human view can distinguish only about 12 bins of color when considering noncontiguous small regions MUNZNER (2014). Nevertheless, this number can be a little higher (around 21 bins of color) if large regions are considered. Therefore, the system defines a minimum size for the regions, which are used to map the clusters to the visual space, that allows the use of a categorical colormap that can discriminate around twenty colors[2].

Actually, in order to obtain good quality clusters the method usually generates a lot more than twenty clusters for each time-stamp. Yet, many of them comprise only a few documents or include semantically meaningless text segments that would only introduce unnecessary visual cluttering. Bearing this in mind, the system shows only those clusters containing more than a user-defined number of documents and those that appear at least two times (regardless of the source and the time-stamp). Nevertheless, a user is still allowed to display all the clusters if needed and/or filtering them by informing a set of keywords that must occur with a certain frequency in the documents for the cluster to be shown.

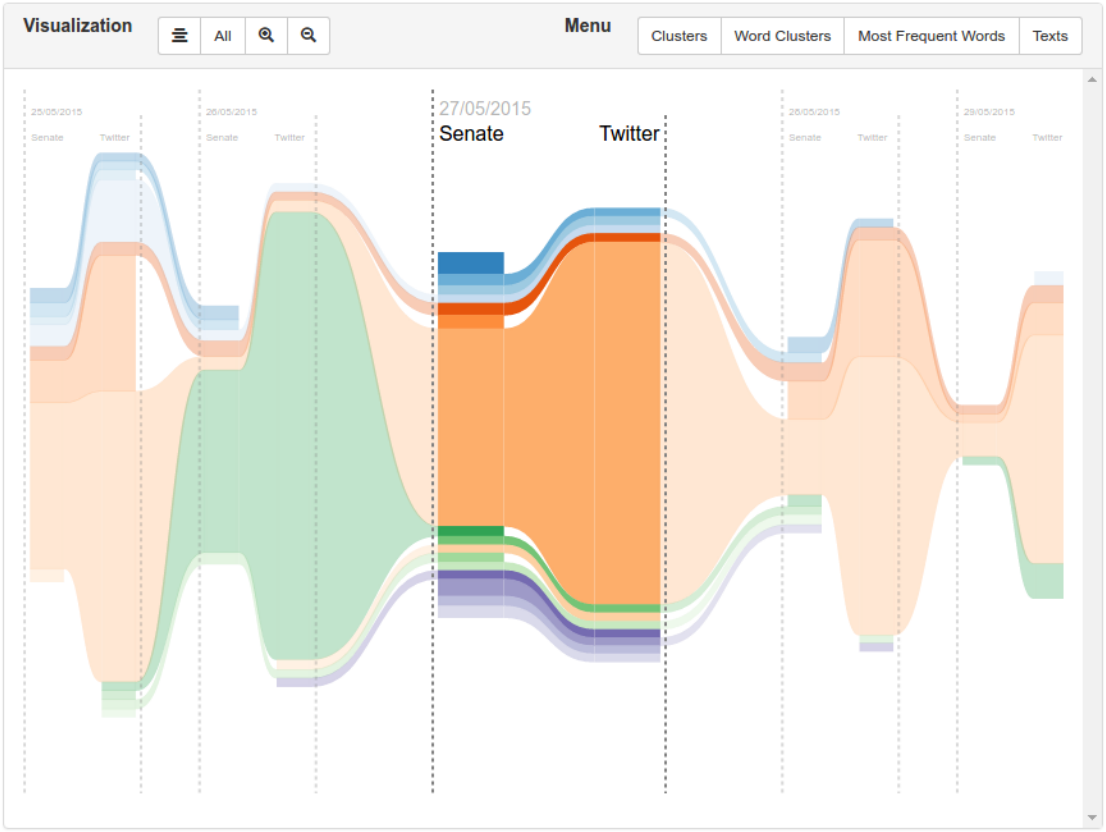


Figure 27 – *Flow View*: the visual interface designed to accommodate the proposed visualization allows a set of user interactions, such as displaying all the clusters and changing the focus time window.

Additional interactions include clicking on a region to the left/right side of the central window, which will cause a smooth animation of the whole visualization in the opposite direction

[2] d3 category20c: <http://bl.ocks.org/aaizemberg/78bd3dade9593896a59d> [Accessed 1 June 2017]

so as to bring the previous/subsequent time window into focus. Whenever the user hovers the mouse over a debate within the central window, a tooltip will be shown with additional information and the whole path of that debate will be highlighted by de-saturating the color of the other debates.

Moreover, some interactions are coordinated between the two views. For example, if a user wants to know more about a specific debate, s/he can select it by clicking on its respective stripe (or rectangle) and further explore its textual content in the *Content View*. This selection will persistently highlight the whole path relative to that debate until the user double clicks on a neutral area in the visualization.

The user can also quickly access specific panels in the *Content View* by clicking on the *Menu* buttons on the top right of the *Flow View*, which will cause the selected panel to be automatic scrolled to the center of the *Content View*. Finally, in order to obtain a more symmetrical layout, the system vertically aligns the debates to the center of the visualization. Yet, if the user prefers a layout vertically aligned to the top, s/he can modify the layout by clicking on the *Centralize* button of the *Visualization* menu on the top left of the *Flow View*.

4.3.2 Content View

This view allows a more in-depth analysis of the textual contents and clusters found by the method. The information is presented in multiple panels arranged vertically to the right side of the *Flow View* as a scrollable list, as shown in Figure 28.

The first panel (Panel *a*) shows a list of all the clusters generated by the method and allows the user to remove a cluster from the visualization if s/he believes it contains meaningless documents and/or s/he wants to see how this operation will affect the layout of the visualization in general. In order to facilitate the analysis this panel also enables the user to provide labels to the clusters.

The second panel (Panel *b*) allows the user to explore the topical clusters (or word clusters) distribution of a specific cluster so that s/he can know which are the words that contribute most for the formation of that cluster and gain a better understanding of the results. Additionally, Panel *c* shows the top 50 terms of greater importance that appear in the documents of the selected cluster based on their TF-IDF values.

Finally, the last panel (panel *d*) presents the actual data (the textual content of the speeches and tweets, in this case) as two scrollable lists displayed side by side. The documents are displayed in chronological order of their time-stamp starting from the top. If a cluster has been previously selected by the user, this panel will show only the documents belonging to that cluster, otherwise it will show a stratified sample of the document collections. There is also a thick vertical outline next to each of the texts which is colored with the same color of its cluster. In addition, hovering the mouse over a document in this panel causes the visualization to place

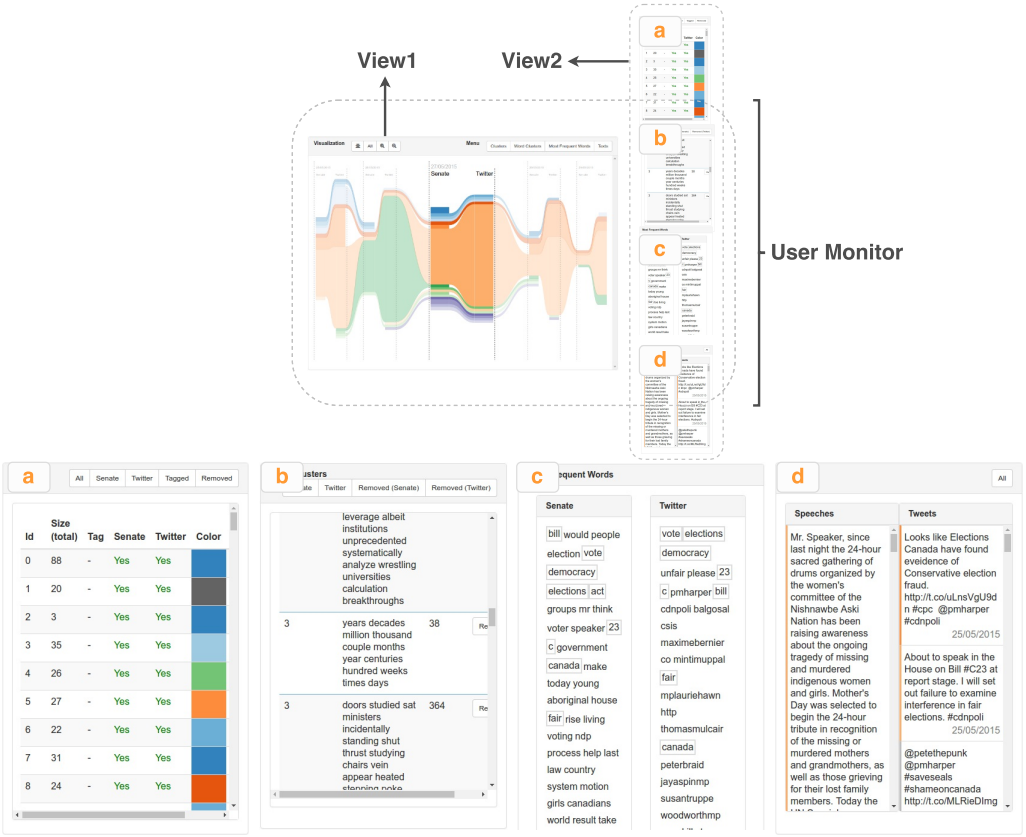


Figure 28 – *Context View*: a secondary view that allows a more in-depth analysis of the textual contents and clusters found by the method.

into focus (into the center of the *Flow View*) the time window relative to the same time-stamp of that document.

4.4 Parameter Settings

As in the case of SAKAHARA; OKADA; NITTA (2014), the clustering algorithm Affinity Propagation FREY; DUECK (2007) was employed to cluster the word vectors and, specifically in the method described in this chapter, it was also used to cluster the documents (i.e, speech sentences and tweets). The choice of Affinity Propagation for clustering is mainly due to this technique not requiring the number of clusters to be defined *a priori*, which allows us to consider a varying number of clusters over time, as new clusters may be found. It also produces good quality clusters as compared to more standard and widely employed techniques, such as the K-Means algorithm. Nevertheless, the number of clusters to be found is affected by the parameter *preference*, which has been, after some experimentation, set to 0.0 in this work. This means that the algorithm will adopt the median similarity value, from the similarity matrix given as input, in its internal computations. A drawback of this clustering technique is its computational cost, which is not a particularly critical issue in our specific problem, since it is applied to data on a daily basis, i.e., on small sized data sets.

The proposed method also requires resorting to two additional datasets: CHAVEFolha[3] and 20 Newsgroups[4]. These datasets were employed to train a *word2vec* model for the Portuguese and the English languages, respectively. Model training considered the following parameters: architecture, *skip-gram*; word vector dimensionality, 300; random downsampling threshold, 1.10^{-3}; context window size, 10; minimal word count, 5. A detailed explanation of this parameters can be found in MIKOLOV *et al.* (2013).

In the step that leverages hashtags, we consider a hashtag to be representative if its relative frequency $rf(h_i)$ is above 15, and the cluster $c_p, p \neq j$ is merged with cluster c_j when over 40% of its documents include a representative hashtag h_i. In the temporal evolution, we remove documents as they become older than 5 days. Finally, in the visualization the number of windows has been set to 5 and a cluster must contain at least 2 documents in order to be shown.

4.5 Results

The ATR-Vis system described in Chapter 3 and the system described in this Chapter have a similar purpose, as both aim at associating political tweets to parliamentary speeches. It was therefore possible to conduct the experiments on the same collection of documents.

Figure 29 shows the results of applying the proposed method to the Brazilian dataset. Although the dataset refers to only 5 debates, the text processing stage was able to find about 1,400 clusters. However, this does not necessary means that the generated clusters are useless. The highlighted cluster (or stripe) in Figure 29, for example, comprises documents in which the textual content consists of greeting phrases, such as “Thank you very much, Mr. President”, that although may be irrelevant to the analyst, represents a group of content-wise similar documents properly formed. Nevertheless, the visual clutter caused by the large number of clusters can hinder the analysis.

An alternative to this problem, as mentioned in Section 4.3.1, is to filter unwanted clusters by informing a set of keywords such as “financiamento” (funding) and “campanha” (campaign) that must appear in the documents for the cluster to be shown. The updated visualization is depicted in Figure 30, which shows that the debate labeled “Financiamento de Campanha” (the label was assigned manually) was a subject matter in parliament and on Twitter for at least three consecutive days. Moreover, a more in-depth analysis can be made by accessing the panels in the *Content View*, as illustrated in Figure 31, which shows the textual content (left) and the topical clusters distribution (right) of the documents assigned to the highlighted cluster.

[3] <http://www.linguateca.pt/CHAVE/> [Accessed 1 June 2017]
[4] <http://qwone.com/~jason/20Newsgroups/> [Accessed 1 June 2017]

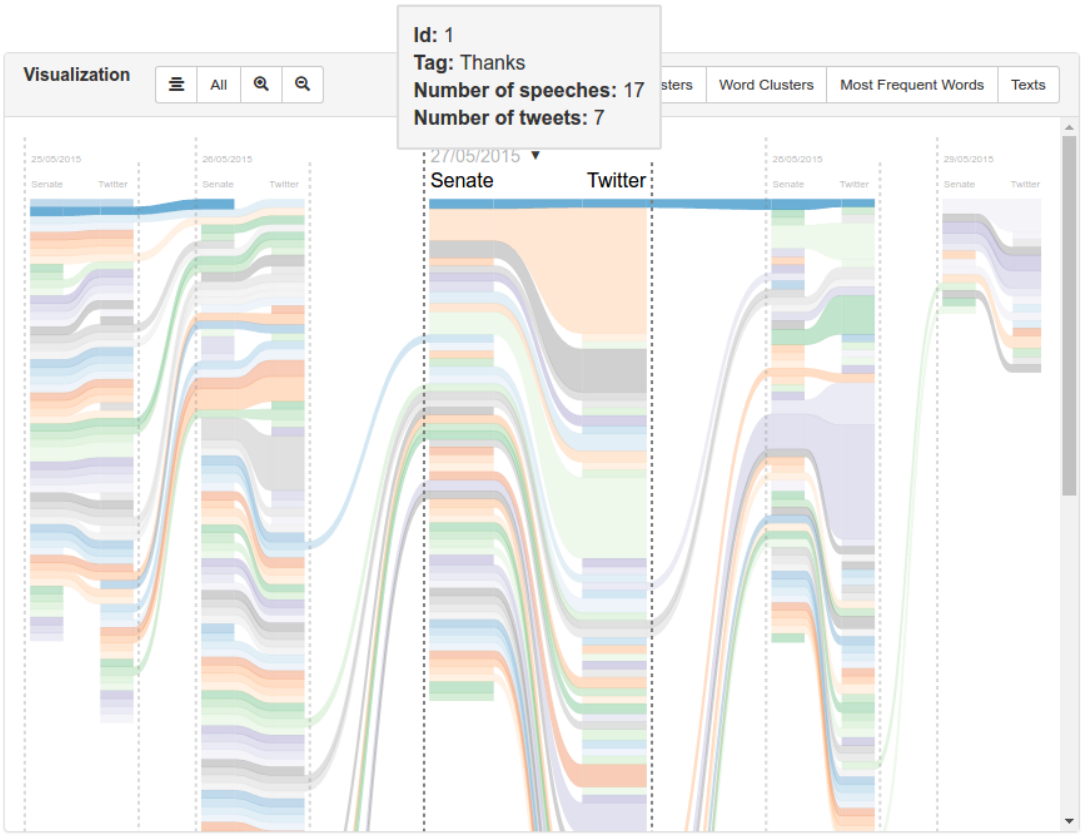


Figure 29 – Visualization of the clusters and existing relationships between the parliamentary speeches and political tweets comprised in the Brazilian dataset. The highlighted cluster, with label "Thanks", consists of greeting phrases, such as "Thank you very much, Mr. President".

4.6 Discussion

When a cluster of documents disappears for c consecutive days, the documents assigned to that cluster will be removed and the method will not be able to match that cluster in the following days, which means that a debate that appears at intervals greater than c days will be shown as two distinct debates.

The quality of the debates (or clusters) relies on the method's capability of finding meaningful topical clusters to represent the documents. For example, noisy words such as "speaker" and highly frequent words such as "vote" (very frequent in the political topics) are not discriminative enough and can interfere in the quality of the topical clusters if the unsupervised method cannot remove them before the document clustering step. One possibility is to incorporate supervision into the process by allowing the user to inspect and modify the topical clusters found. However, the whole process would have to be performed again.

Another issue is how to keep the document representation up-to-date. The *word2vec* implementation employed in this work allows updating the weights of the word vectors generated at the beginning of the process. However, we cannot yet incorporate new word vectors into the model, implying that the vocabulary will always remain the same.

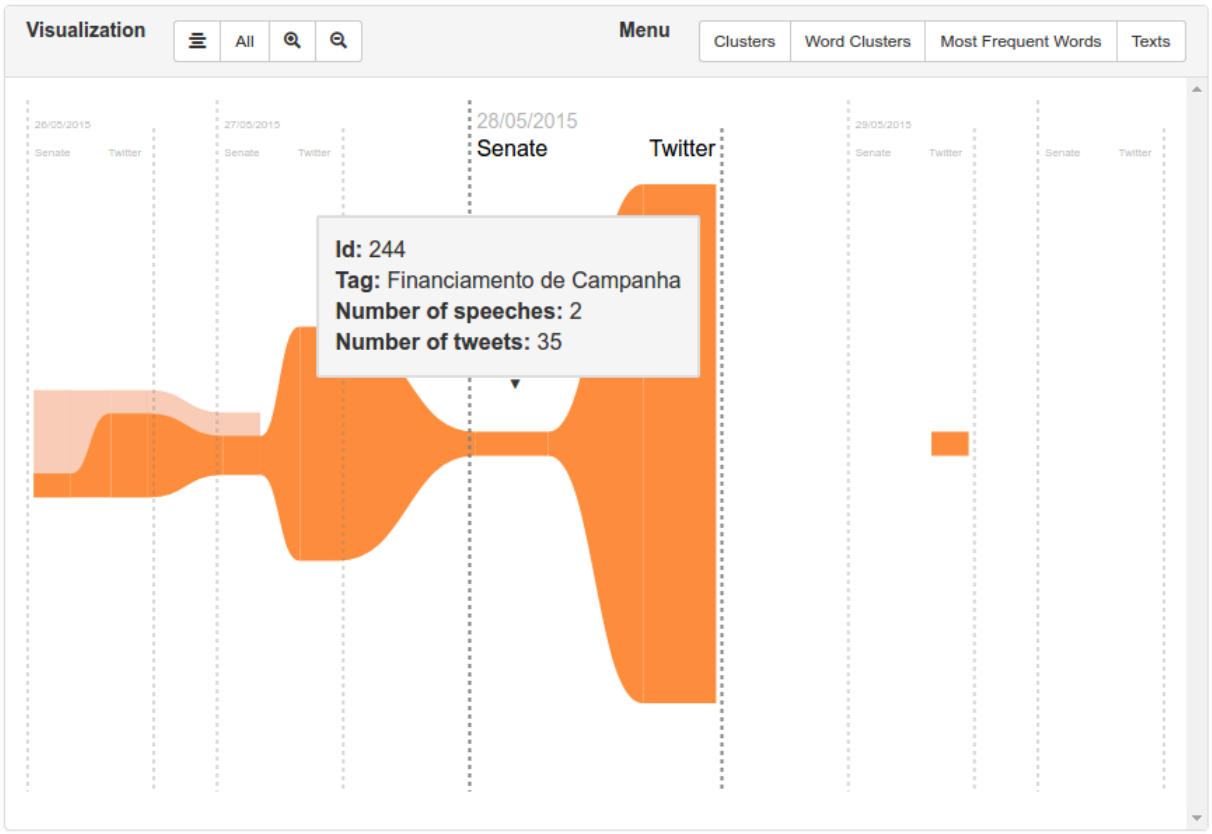


Figure 30 – Visualization applied to the same dataset displayed in Figure 29 after filtering unwanted clusters. The keywords informed were “campanha” (campaign) and “financiamento” (funding).

4.7 Final Remarks

This work introduced a novel visualization, based on the river flow metaphor, for analyzing the temporal evolution of topics in Twitter in connection with political debates. Moreover, a domain-independent text segmentation framework was adapted to associate (by clustering) Twitter content with parliamentary speeches. A simplified version of MONIC is then employed to track the temporal evolution of the debates (or clusters) and produces a set of time-stamped clusters. Finally, a visual interface was proposed and developed that incorporates the proposed visualization.

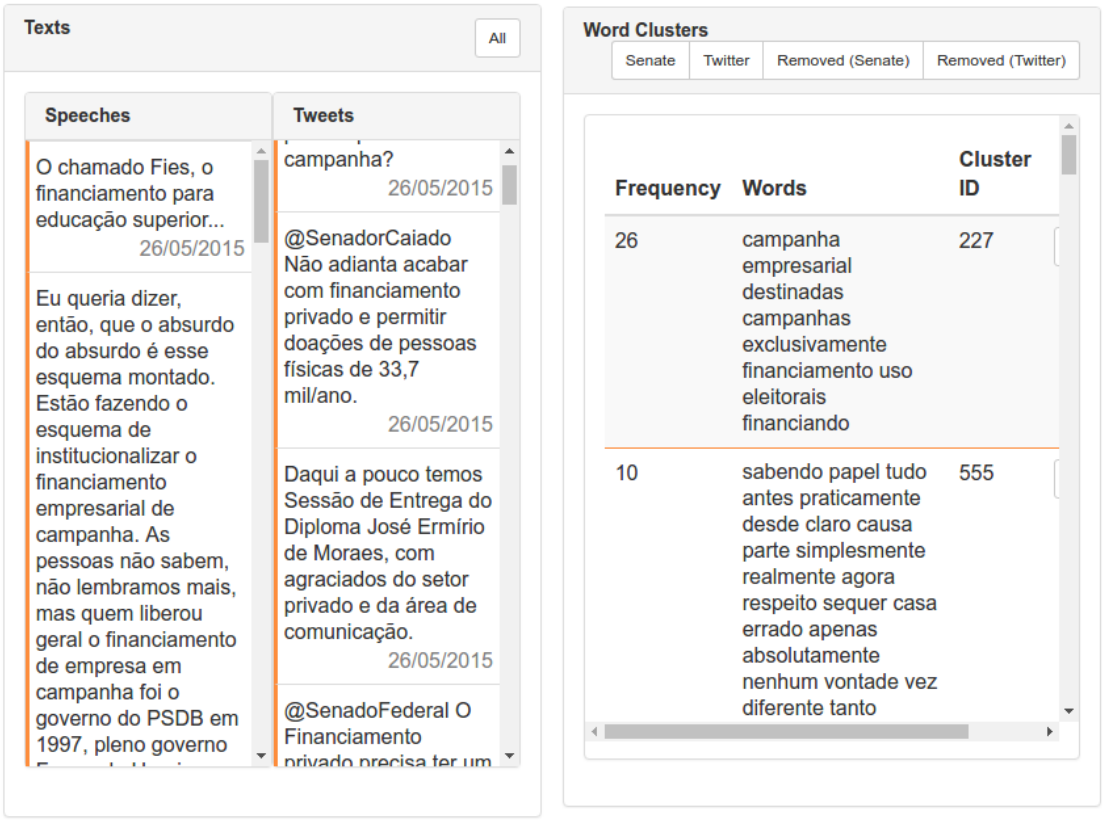


Figure 31 – A more in-depth analysis of the cluster labeled “Financiamento de Campanha”. The textual content (left) and the topical clusters distribution (right) of the documents assigned to the selected cluster.

CHAPTER

5

CONCLUDING REMARKS

This work investigated the applicability of text visualization techniques for the analysis of open government and social media data. Two visual interfaces were developed, which used in combination with active learning/retrieval strategies support user involvement in the retrieval of information and textual classification processes. The active learning/retrieval strategies are presented through visualization techniques and are intended to enable non-technical users to interact with the processes in order to improve system accuracy. Moreover, the visual interfaces developed allow the exploration and inspection of the data and the results, besides allowing the selection of instances beyond those suggested by the system.

The ATR-Vis system was evaluated through use cases and quantitative experiments, as well as a qualitative analysis conducted with three domain experts described in details in the work by MAKKI *et al.* (2017). Evaluation was conducted on two datasets corresponding to the retrieval of tweets related to parliamentary debates being held in Canada and Brazil, and one related to a set of top news stories that received great media attention at the time. The use of different datasets has shown that the framework works well with different domains/languages without any additional model refinement. Furthermore, all three experts provided positive feedback regarding the system and acknowledged the need for this type of tools for accurate retrieval of tweets. The visual interfaces were developed and evaluated within a collaboration with researchers from the MALNIS research group at the Dalhousie University.

Furthermore, a system was developed for the visual analysis of the temporal evolution of topics in Twitter in connection with political debates in the Brazilian Senate. For this purpose, a method was proposed to handle data from two distinct sources, one from social media (specifically, Twitter) and the other from open government data (specifically, speeches at the Brazilian Senate). As a consequence, an additional step was performed to take advantage of the structural information made available in Twitter.

As evidenced in the analysis of related work, the text pre-processing step is not trivial

and may require combining multiple techniques for the desired analysis to be feasible. In the method developed, the preprocessing requires identifying and associating topics on the Twitter posts and in the debates held at the Senate. In order to do so, the problem has been modeled as a clustering problem to automatically identify the topics/debates, avoiding any user involvement at this stage and allowing the identification of new topics over time.

Moreover, the association between texts from both sources was embedded in the clustering phase through a textual representation based on identifying groups of similar words. This approach was adopted in an attempt to establish semantic relationships between the documents from both sources. This was achieved by treating each paragraph of the senators' speeches as a single independent document. Although this strategy introduced processing problems related to handling short texts, the clustering of speeches presented good quality, which unfortunately was not the case of the clustering of tweets in general, even after adopting the enhanced representation.

In order to convey the temporal evolution of two different sources while simultaneously showing their associations, a new visualization technique was introduced. Once again, in agreement with findings from related work, it adopts the river metaphor, which has been widely employed to visualize the temporal evolution of documents or topics. Assuming that the topics/debates provided as input are of good quality, it is possible, for example, to answer questions such as whether the same debate was discussed in the two sources for several consecutive days, which source contributed most for a particular subject, or where a debate was first introduced.

Several relevant limitations have been mentioned in Chapter 4, such as the inability to detect clusters removed after not being detected for c days, the need to use documents analyzed in previous days to set the overlapping parameter for the MONIC algorithm (also using the c-days interval), and the fact that the textual representation becomes outdated if the vocabulary in either source undergoes major changes. Nonetheless, the major difficulty found, which has not been solved satisfactorily, is related with the limited and poor content of the tweets. This limitation has a direct impact in the quality of the clusters and consequently in their association, thus strongly affecting the quality of the resulting visualization.

Moreover, the high number of clusters identified, albeit not being a problem in itself, introduces visual clutter in the visualization. An alternative, as mentioned in Chapter 4, would be allowing users to modify the word clusters found. Specifically, assuming that word clusters that are too frequent are somewhat analogous to stop words, and therefore not discriminative enough from a clustering perspective, these clusters containing only irrelevant words could be removed before the document clustering stage. Finally, another possibility would be using a more recent text corpora to identify groups of semantically similar words. Since the representation of the tweets is the main issue, a corpus with a vocabulary more similar to that of Twitter would be preferable.

Besides incorporating these previous alternatives into the solution, other possibilities can be considered in future work. A more complete and user-oriented evaluation could be carried out

to assess whether a bag-of-concepts representation presents a significant improvement compared with a more common representation such as a bag-of-words, as well as the usefulness and acceptance of the visualization technique proposed. Specifically, a quantitative analysis could be performed by manually labeling a percentage of the instances (speeches and tweets), randomly selected, of a specific cluster. In this way, it would be possible to assess the quality of a cluster by the debate frequency of the selected instances. Moreover, a pair analytics evaluation, similar to that performed for ATR-Vis could be carried out to evaluate the visualization and the visual interface as a whole.

An interesting analysis would be the possibility of knowing whether there is a polarization of sentiments in a particular subject and how it changes over time. This could be done employing algorithms that identify and categorize the sentiment expressed in a piece of text (e.g. speeches and tweets) to obtain the prevailing opinion in the texts assigned to a particular debate (clusters, in our case) within a given time interval. Although it would be relatively easy to include an additional sentiment analysis step in the processing component, incorporating this new information into the visual interface is not trivial. Nevertheless, some contributions in the literature could serve as a starting point, as discussed in the survey by PANG; LEE (2008). One possibility would be, for instance, adding a small vertical bar within the tooltip shown when hovering the mouse over the debates. The proportion of sentiments associated to a debate could then be color encoded to reflect the proportion of positive and negative sentiment, e.g., by mapping a green color to positive sentiments and a red color to negative ones.

It is worth mentioning that there is an extensive research area concerned with open government and open data, which is defined in terms of several principles, standards and policies based on four essential characteristics: accessible, accurate, analyzable, and authentic[1]. The investigation conducted in this research contemplated just a subset of the many possible paths required to improve support to data analysis in this context.

[1] <https://opengovdata.io/2014/principles/> [Accessed 1 June 2017]

BIBLIOGRAPHY

AGGARWAL, C. C.; ZHAI, C. A survey of text clustering algorithms. In: **Mining Text Data**. Springer US, 2012. p. 77–128. ISBN 978-1-4614-3223-4. Available: <http://link.springer.com/chapter/10.1007%2F978-1-4614-3223-4_4>. Citations on pages 60 and 61.

ALENCAR, A. B.; BöRNER, K.; PAULOVICH, F. V.; OLIVEIRA, M. C. F. de. Time-aware visualization of document collections. In: **Proceedings of the 27th Annual ACM Symposium on Applied Computing**. New York, NY, USA: ACM, 2012. (SAC '12), p. 997–1004. ISBN 978-1-4503-0857-1. Available: <http://doi.acm.org/10.1145/2245276.2245469>. Citations on pages 38 and 40.

ALENCAR, A. B.; OLIVEIRA, M. C. F. de; PAULOVICH, F. V. Seeing beyond reading: a survey on visual text analytics. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, John Wiley & Sons, Inc., v. 2, n. 6, p. 476–492, 2012. ISSN 1942-4795. Available: <http://dx.doi.org/10.1002/widm.1071>. Citations on pages 38 and 63.

ALEXANDER, E.; KOHLMANN, J.; VALENZA, R.; WITMORE, M.; ; GLEICHER, M. Serendip: Topic model-driven visual exploration of text corpora. **IEEE Symposium on Visual Analytics Science and Technology 2014**, p. 173–182, 2014. Citations on pages 16, 36, and 38.

ARIAS-HERNANDEZ, R.; KAASTRA, L. T.; GREEN, T. M.; ; FISHER, B. Pair analytics: Capturing reasoning processes in collaborative visual analytics. **Proceedings of the 2011 44th Hawaii International Conference on System Sciences (HICSS)**, IEEE Computer Society, Washington, DC, USA, p. 1–10, 2011. Available: <http://dx.doi.org/10.1109/HICSS.2011.339>. Citation on page 51.

BIZER, C.; HEATH, T.; BERNERS-LEE, T. Linked data-the story so far. **International Journal on Semantic Web and Information Systems (IJSWIS)**, p. 1–22, 2009. Citation on page 35.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. **Journal of Machine Learning Research**, v. 3, p. 993–1022, jan 2003. Citations on pages 24, 26, and 37.

BLITZER, J.; DREDZE, M.; PEREIRA, F. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. **ACL**, v. 7, p. 440–447, 2007. Citation on page 50.

BOSTOCK, M.; OGIEVETSKY, V.; HEER, J. D3 data-driven documents. **IEEE Transactions on Visualization and Computer Graphics**, IEEE Educational Activities Department, Piscataway, NJ, USA, v. 17, n. 12, p. 2301–2309, Dec. 2011. ISSN 1077-2626. Available: <http://dx.doi.org/10.1109/TVCG.2011.185>. Citations on pages 48 and 65.

BREHMER, M.; INGRAM, S.; STRAY, J.; MUNZNER, T. Overview: The design, adoption, and analysis of a visual document mining tool for investigative journalists-aware. **IEEE Transactions on Visualization and Computer Graphics**, p. 1077–2626, 2014. Citations on pages 33 and 34.

BREITMAN, K.; SALAS, P. E.; CASANOVA, M. A.; SARAIVA, D.; GAMA, V.; FILHO, J. V.; MAGALHãES, R. P.; FRANZOSI, E.; CHAVES, M. Open government data in brazil. **IEEE**

Intelligent Systems, v. 27, n. 3, p. 45–49, 2012. Available: <http://dblp.uni-trier.de/db/journals/expert/expert27.html#BreitmanFSSGCCF12>. Citation on page 21.

CARD, S. K.; MACKILAY, J. D.; SHNEIDERMAN, B. **Readings in Information Visualization: Using Vision to Think**. [S.l.]: Morgan Kaufman, 1999. Citation on page 22.

CHEN, C. Citespace ii: Detecting and visualizing emerging trends and transient patterns in scientific literature. **Journal of the American Society for Information Science and Technology**, John Wiley & Sons, Inc., New York, NY, USA, v. 57, n. 3, p. 359–377, Feb. 2006. ISSN 1532-2882. Available: <http://dx.doi.org/10.1002/asi.v57:3>. Citation on page 38.

____. **Information Visualization: Beyond the Horizon**. [S.l.]: Springer-Verlag New York, Inc., 2006. ISBN 184628340X. Citation on page 22.

CHOO, J.; LEE, C.; REDDY, C. K.; PARK, H. Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization. **IEEE Transactions on Visualization and Computer Graphics**, v. 19, n. 12, p. 1992–2001, 2013. Citations on pages 36 and 37.

CUI, W.; LIU, S.; TAN, L.; SHI, C.; SONG, Y.; GAO, Z.; QU, H.; TONG, X. Textflow: Towards better understanding of evolving topics in text. **IEEE Transactions on Visualization and Computer Graphics**, IEEE Computer Society, Los Alamitos, CA, USA, v. 17, n. 12, p. 2412–2421, 2011. ISSN 1077-2626. Citation on page 38.

DOU, W.; YU, L.; WANG, X.; MA, Z.; RIBARSKY, W. Hierarchicaltopics: Visually exploring large text collections using topic hierarchies. **IEEE Transactions on Visualization and Computer Graphics**, v. 19, n. 12, p. 2002–2011, 2013. Citations on pages 26, 27, and 43.

ELER, D. M.; NAKAZAKI, M. V.; PAULOVICH, F. V.; SANTOS, D. P.; ANDERY, G. F.; OLIVEIRA, M. C. F. de; NETO, J. B.; MINGHIM, R. Visual analysis of image collections. **The Visual Computer**, v. 25, n. 10, p. 923–937, 2009. Citations on pages 38 and 39.

FREEMAN, L. C. Visualizing social networks. **Carnegie Mellon: Journal of Social Structure: Visualizing Social Networks**, 2000. Available: <http://www.cmu.edu/joss/content/articles/volume1/freeman.pdf>. Citation on page 25.

FREY, B. J.; DUECK, D. Clustering by passin messages between data points. **Science**, American Association for the Advancement of Science, v. 315, n. 5814, p. 972–976, 2007. ISSN 0036-8075. Available: <http://science.sciencemag.org/content/315/5814/972>. Citation on page 68.

GLORIA, M. J. K.; MCGUINNESS, D. L.; LUCIANO, J. S.; ZHANG, Q. Exploration in web science: Instruments for web observatories. In: **Proceedings of the 22nd International Conference on World Wide Web Companion, WWW '13 Companion**. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2013. p. 1325–1328. ISBN 978-1-4503-2038-2. Available: <http://dl.acm.org/citation.cfm?id=2487788.2488170>. Citation on page 21.

GOMEZ-NIETO, E.; ROMAN, F. S.; PAGLIOSA, P.; CASACA, W.; HELOU, E. S.; OLIVEIRA, M. C. F.; NONATO, L. G. Similarity preserving snippet-based visualization of web search results. **IEEE Transactions on Visualization and Computer Graphics**, IEEE Educational Activities Department, Piscataway, NJ, USA, v. 20, n. 3, p. 457–470, Mar. 2014. ISSN 1077-2626. Available: <http://dx.doi.org/10.1109/TVCG.2013.242>. Citations on pages 22, 38, and 39.

GRAVES, A. Creation of visualizations based on linked data. In: **Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics**. New York, NY, USA: ACM, 2013. (WIMS '13), p. 41:1–41:12. ISBN 978-1-4503-1850-1. Available: <http://doi.acm.org/10.1145/2479787.2479828>. Citation on page 35.

GRAVES, A.; HENDLER, J. Visualization tools for open government data. In: **Proceedings of the 14th Annual International Conference on Digital Government Research**. New York, NY, USA: ACM, 2013. p. 136–145. ISBN 978-1-4503-2057-3. Available: <http://doi.acm.org/10.1145/2479724.2479746>. Citation on page 22.

HAVRE, S.; HETZLER, B.; NOWELL, L. Themeriver: Visualizing theme changes over time. In: **Proceedings of the IEEE Symposium on Information Vizualization 2000**. Washington, DC, USA: IEEE Computer Society, 2000. (INFOVIS '00), p. 115–. ISBN 0-7695-0804-9. Available: <http://dl.acm.org/citation.cfm?id=857190.857680>. Citations on pages 26 and 38.

HEIMERL, F.; KOCH, S.; BOSCH, H.; ERTL, T. Visual classifier training for text document retrieval. **IEEE Transactions on Visualization and Computer Graphics**, v. 18, p. 2839–2848, 2012. Citation on page 45.

HUANG, L.; MATWIN, S.; CARVALHO, E.; MINGHIM, R. Active learning with visualization for text data. **submitted for publication in ESIDA**, 2017. Citations on pages 16, 55, and 56.

KALAMPOKIS, E.; TAMBOURIS, E.; TARABANIS, K. On publishing linked open government data. In: **Proceedings of the 17th Panhellenic Conference on Informatics**. New York, NY, USA: ACM, 2013. (PCI '13), p. 25–32. ISBN 978-1-4503-1969-0. Available: <http://doi.acm.org/10.1145/2491845.2491869>. Citation on page 21.

KIETZMANN, J. H.; HERMKENS, K.; MCCARTHY, I. P.; SILVESTRE, B. S. Social media? get serious! understanding the functional building blocks of social media. **Segal Graduate School of Business, Simon Fraser University, 500 Granville Street, Vancouver, BC V6C 1W6, Canada**, v. 54, p. 241–251, 2011. Available: <http://beedie.sfu.ca/files/PDF/research/McCarthy_Papers/2011_Social_Media_BH.pdf>. Citation on page 25.

LAM, H.; BERTINI, E.; ISENBERG, P.; PLAISANT, C.; CARPENDALE, S. Empirical studies in information visualization: Seven scenarios. **IEEE Transactions on Visualization and Computer Graphics**, IEEE Educational Activities Department, Piscataway, NJ, USA, v. 18, n. 9, p. 1520–1536, 2012. ISSN 1077-2626. Available: <http://dx.doi.org/10.1109/TVCG.2011.279>. Citation on page 51.

LI, C.; WANG, Y.; RESNICK, P.; MEI, Q. Req-rec: High recall retrieval with query pooling and interactive classification. **Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval**, ACM, New York, NY, USA, p. 163–172, 2014. Available: <http://doi.acm.org/10.1145/2600428.2609618>. Citation on page 50.

LIU, S.; WANG, X.; CHEN, J.; ZHU, J.; GUO, B. Topicpanorama: a full picture of relevant topics. **IEEE Symposium on Visual Analytics Science and Technology 2014**, p. 183–192, 2014. Citations on pages 15, 27, 28, and 43.

LOPEZ, V.; KOTOULAS, S.; SBODIO, M. L.; LLOYD, R. Guided exploration and integration of urban data. In: **Proceedings of the 24th ACM Conference on Hypertext and Social Media**. New York, NY, USA: ACM, 2013. (HT '13), p. 242–247. ISBN 978-1-4503-1967-6. Available: <http://doi.acm.org/10.1145/2481492.2481524>. Citation on page 21.

MAATEN, L. van der; HINTON, G. E. Visualizing high-dimensional data using t-sne. **Journal of Machine Learning Research**, v. 9, p. 2579–2605, 2008. Citation on page 36.

MAKKI, R.; CARVALHO, E.; SOTO, A. J.; BROOKS, S.; OLIVEIRA, M. C. F. de; MILIOS, E.; MINGHIM, R. Atr-vis: Visual and interactive information retrieval for parliamentary discussions in twitter. **To appear in ACM Transactions on Knowledge Discovery from Data**, 2017. Citations on pages 46, 47, 48, 49, 50, 51, 52, 53, 54, 56, and 73.

MAKKI, R.; SOTO, A. J.; BROOKS, S.; MILIOS, E. Active information retrieval for linking twitter posts with political debates. **Machine Learning and Applications (ICMLA)**, 2015. Citation on page 46.

MIKOLOV, T.; SUTSKEVER, I.; CHEN, K.; CORRADO, G.; DEAN, J. Distributed representations of words and phrases and their compositionality. **Proceedings of the 26th International Conference on Neural Information Processing Systems**, Curran Associates Inc., USA, p. 3111–3119, 2013. Available: <http://dl.acm.org/citation.cfm?id=2999792.2999959>. Citation on page 69.

MUNZNER, T. **Visualization Analysis and Design**. [S.l.]: A K Peters/CRC Press, 2014. Citations on pages 65 and 66.

OLIVEIRA, M. C. F.; LEVKOWITZ, H. From visual data exploration to visual data mining: a survey. **IEEE Transactions on Visualization and Computer Graphics**, v. 9, n. 3, p. 378–394, 2003. Citation on page 22.

PANG, B.; LEE, L. Opinion mining and sentiment analysis. **Foundations and Trends in Information Retrieval**, Now Publishers Inc., Hanover, MA, USA, v. 2, n. 1-2, p. 1–135, 2008. ISSN 1554-0669. Available: <http://dx.doi.org/10.1561/1500000011>. Citation on page 75.

PAULOVICH, F.; NONATO, L.; MINGHIM, R.; LEVKOWITZ, H. Least square projection: A fast high-precision multidimensional projection technique and its application to document mapping. **IEEE Transactions on Visualization and Computer Graphics**, v. 14, n. 3, p. 564–575, May-June 2008. ISSN 1077-2626. Citations on pages 22, 38, and 39.

PAULOVICH, F. V.; MINGHIM, R. HiPP: A novel hierarchical point placement strategy and its application to the exploration of document collections. **IEEE Transactions on Visualization and Computer Graphics**, IEEE Educational Activities Department, Piscataway, NJ, EUA, v. 14, n. 6, p. 1229–1236, Nov. 2008. ISSN 1077-2626. Citations on pages 38 and 41.

PAULOVICH, F. V.; TOLEDO, F. M. B.; TELLES, G. P.; MINGHIM, R.; NONATO, L. G. Semantic wordification of document collections. **Comp. Graph. Forum**, John Wiley & Sons, Inc., New York, NY, USA, v. 31, n. 3pt3, p. 1145–1153, Jun. 2012. ISSN 0167-7055. Available: <http://dx.doi.org/10.1111/j.1467-8659.2012.03107.x>. Citation on page 38.

ROSEN-ZVI, M.; GRIFFITHS, T.; STEYVERS, M.; SMYTH, P. The author-topic model for authors and documents. In: **Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence**. Arlington, Virginia, United States: AUAI Press, 2004. (UAI '04), p. 487–494. ISBN 0-9749039-0-6. Available: <http://dl.acm.org/citation.cfm?id=1036843.1036902>. Citation on page 26.

SAKAHARA, M.; OKADA, S.; NITTA, K. Domain-independent unsupervised text segmentation for data management. **2014 IEEE International Conference on Data Mining Workshop**,

n. 14886912, p. 7, 2014. ISSN 2375-9259. Available: <http://ieeexplore.ieee.org/document/7022635/>. Citations on pages 60, 61, and 68.

SALAZAR, F. S. R.; PINHO, R. D. de; MINGHIM, R.; OLIVEIRA, M. C. F. de. A study on the role of similarity metrics in visual text analytics. In: INSTICC, PORTUGAL. **Proceedings of the 4th International Conference on Information Visualization Theory and Applications (IVAPP)**. [S.l.], 2013. p. 429–438. Citations on pages 16, 22, and 39.

SALTON; GERARD; MCGILL; J., M. **Introduction to Modern Information Retrieval**. New York, NY, USA: McGraw-Hill, Inc., 1986. ISBN 0070544840. Available: <http://dl.acm.org/citation.cfm?id=576628>. Citation on page 36.

SALTON, G.; WONG, A.; YANG, C. S. A vector space model for automatic indexing. **ACM Communications**, ACM, New York, NY, EUA, v. 18, n. 11, p. 613–620, 1975. ISSN 0001-0782. Citation on page 24.

SETTLES, B. **Active Learning Literature Survey**. [S.l.], 2009. Citation on page 45.

SPILIOPOULOU, M.; NTOUTSI, I.; THEODORIDIS, Y.; SCHULT, R. Monic: modeling and monitoring cluster transitions. **Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**, p. 706–711, 2006. ISSN 1-59593-339-5. Available: <http://doi.acm.org/10.1145/1150402.1150491>. Citation on page 62.

VIéGAS, F. B.; WATTENBERG, M.; FEINBERG, J. Participatory visualization with wordle. **IEEE Transactions on Visualization and Computer Graphics**, v. 15, n. 6, p. 1137–1144, 2009. Available: <http://dblp.uni-trier.de/db/journals/tvcg/tvcg15.html#ViegasWF09>. Citation on page 38.

WATTENBERG, M.; VIÉGAS, F. B. The word tree, an interactive visual concordance. **IEEE Transactions on Visualization and Computer Graphics**, IEEE Educational Activities Department, Piscataway, NJ, USA, v. 14, n. 6, p. 1221–1228, Nov. 2008. ISSN 1077-2626. Available: <http://dx.doi.org/10.1109/TVCG.2008.172>. Citation on page 38.

WONG, P.; THOMAS, J. Visual Analytics. **IEEE Computer Graphics and Applications**, v. 24, n. 5, p. 20–21, 2004. Citation on page 22.

XU, J.; CROFT, W. B. Query expansion using local and global document analysis. In: **Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval**. New York, NY, USA: ACM, 1996. (SIGIR '96), p. 4–11. ISBN 0-89791-792-8. Available: <http://doi.acm.org/10.1145/243199.243202>. Citation on page 46.

XU, P.; WU, Y.; WEI, E.; PENG, T.-Q.; LIU, S.; ZHU, J. J. H.; QU, H. Visual analysis of topic competition on social media. **IEEE Transactions on Visualization and Computer Graphics**, v. 19, n. 12, p. 2012–2021, 2013. Citations on pages 15, 27, 28, 29, and 43.

ZHAO, J.; CAO, N.; WEN, Z.; SONG, Y.; LIN, Y.-R.; COLLINS, C. Fluxflow: Visual analysis of anomalous information spreading on social media. **IEEE Transactions on Visualization and Computer Graphics**, v. 20, n. 12, p. 1773–1782, 2014. Citations on pages 29, 30, and 31.