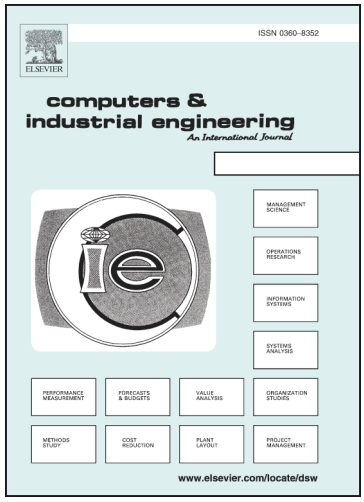

Visual analytics for exploring topic long-term evolution and detecting weak signals in company targeted tweets

Lambert Pépin, Pascale Kuntz, Julien Blanchard, Fabrice Guillet, Philippe Suignard

PII: S0360-8352(17)30047-5
DOI: http://dx.doi.org/10.1016/j.cie.2017.01.025
Reference: CAIE 4625

To appear in: *Computers & Industrial Engineering*

Accepted Date: 31 January 2017

Please cite this article as: Pépin, L., Kuntz, P., Blanchard, J., Guillet, F., Suignard, P., Visual analytics for exploring topic long-term evolution and detecting weak signals in company targeted tweets, *Computers & Industrial Engineering* (2017), doi: http://dx.doi.org/10.1016/j.cie.2017.01.025

Visual analytics for exploring topic long-term evolution and detecting weak signals in company targeted tweets

Lambert Pépin[a], Pascale Kuntz[b], Julien Blanchard[b,*], Fabrice Guillet[b] and Philippe Suignard[a]

[a] EDF R&D, 1 Av. du Général de Gaulle, Clamart, France

[b] DUKe Research Team, LINA Computer Science Laboratory of Nantes, UMR CNRS 6241, Nantes, France

[*] Corresponding author: Julien Blanchard

Tel: +33 240683066

E-mail address: julien.blanchard@univ-nantes.fr

Postal address: Polytech Nantes, rue Christian Pauc, BP 50609, 44306 Nantes, France

- We propose a visual analytics approach to track topics relative to a company from Twitter.
- The approach combines topic modeling and topic temporal evolution visualization.
- We perform an experimental analysis of dissimilarity measures to assess topic proximities.
- The approach has been used by the EDF company to detect previously unknown patterns in Twitter.

Visual analytics for exploring topic long-term evolution and detecting weak signals in company targeted tweets

Abstract

Business decision support tools, including social media data analysis, are required to help managers better understand trends and customer opinions. This paper presents a visual analytics-based approach to assist an expert user in tracking topics relative to his/her company from Twitter. Developed for visualizing topic long-term evolution and detecting weak signals, our process is composed of three complementary steps: (i) a time-dependent topic extraction based on a Latent Dirichlet Allocation, (ii) a topic relationship detection based on a dissimilarity which evaluates the topic proximities between consecutive time slots, and (iii) a topic evolution visualization inspired by a Sankey diagram popular in industrial environments to show dynamic relationships in a system. To test our approach, we have used a real-life dataset from the French energy company EDF from which we have analyzed the evolution of a corpus of more than 70 000 tweets related to this company published over one year, and detected different types of evolving patterns hidden by the data volume and commonly masked by fully automatic mining algorithms.

Keywords: social media, visual analytics, topic mining, weak signals, Twitter

1. Introduction

The massive growth in the use of social media increasingly affects company decision-making. A recent analysis of their impact [1] reports that traditional decision-making processes are being disrupted by social media, and that business decision support tools including social media data analysis are required to help managers better understand trends and customers' opinions. For instance,

for the French energy company EDF –at the origin of the use case presented in this paper– customer retention is becoming a major challenge with the low prices in the energy market associated with an increasing competition and an enhanced regulatory pressure. It relies on detailed knowledge of the customers. Besides classical surveys, customer complaints and interactive conversational tools, general public social media are today an unavoidable source of information. An innovative generation of Customer Management Systems is emerging [2] and this evolution stimulates numerous researches in social media mining. In particular, opinion mining [3] aims at improving knowledge on brand and product perception and at guiding reputation management. As recently stated [4], "monitoring these opinions related to a particular company [...] is a new challenge": social media data are vast, noisy, distributed, unstructured and dynamic. A lot of effort has been put into their linguistic specifities and scalability issues, and a range of efficient algorithms are available today for detecting major events in large tweet sets (e.g. [5]) and following the main trendy topics online [6].

Nonetheless, beyond the saillant information, companies also need to analyze the evolution of topics which concern them over time, in order to detect "weak signals" which appear at irregular intervals. From the pioneering paper of Ansof in the 70's [7], different attempts to define this concept have been proposed [8, 9, 10, 11]. Roughly speaking, in our context, weak signals represent snippets of information, hidden in social media streams, characterized by "a temporal discontinuity in their discovery" [10] and their potential interest and impact in the future. Promoted as early as 2008 in the "Social Media Roadmaps" by Ahlqvist *et al.* [12], search for weak signals in social media is today a fast-growing area of interest [11, 13] the importance of which is underlined by experts in strategic intelligence [14].

In this paper, we focus on Twitter which is becoming one of the favored social media by companies because of its popularity and the easy access to its data. We here propose an original interactive process to detect evolving relationships in topics extracted from tweets targeting a company. More precisely, we combine

topic mining with visual analytics for an interactive exploration of the temporal evolution of the topic relationships. In previous works, we tested automatic approaches with limited results. User integration in the discovery loop quickly appeared to be necessary to detect non-apparent relationships relevant for the company. As recently mentioned by Eckhoff *et al.* [11] , customers of weak signal detection are experts and "their wish to detect weak signals seems to be driven by their interest in getting more factual knowledge". Visual mining is consequently well-adapted to let the expert play the role of a heuristic which guides the algorithms to extract "knowledge nuggets" potentially interesting for him/her.

Our process is composed of three complementary steps: (i) time-dependent topic extraction, (ii) topic relationship detection, and (iii) interactive topic evolution visualization. A topic extraction based on a Latent Dirichlet Allocation (LDA) [15] is applied at each time interval; each topic is described by the probability distribution of its most representative terms. The relationships between the topics computed at consecutive time intervals are detected via a dissimilarity measure. As the results interpretation closely depends on the choice of the topic proximity measurement, we have here experimentally compared the empirical statistical distributions of seven dissimilarities to select the best adapted one. The strongest proximities and their evolution over time are visualized with a graph layout inspired by a Sankey diagram [16] which is a familiar representation in industrial environments.

The rest of the paper is structured as follows. Section 2 recalls relevant related works from the literature. Section 3 summarizes the main principles of topic extraction with LDA and presents our dynamic modeling based on a graph representation of the topic relationships. The process implementation depends on three parameters : the selected topic number, the concentration parameter of the Dirichlet distribution and the dissimilarity which measures the topic closeness. Section 4 describes numerical experimentations to set the parameter values. Section 5 presents the interactive visualization on a real life application. Our approach has been applied to a dataset from the French energy

company EDF: we have analyzed the evolution of a corpus of more than 70 000 tweets relating to this company published over one year.

2. Related research

Twitter is a social networking and microblogging service that enables users to send short messages with a maximum of 140 characters. Let us remark that our approach will remain consistent with the extension of the tweet size currently in testing (10 000 characters with 140 characters in the timeline and an additional text by clicking). The Twitter messages often contain creative spelling, slang and misspelling, and many specific elements differ from proper formal language. Applying classical approaches to such sparse data is not straightforward, and a key component when mining such unstructured textual data is the automatic extraction of the semantic content of documents. Therefore, most of the works done in this field rely on topic modeling. Topic modeling is the task of automatically uncovering the underlying semantic structure in document collections [17]. Popularized by the DARPA's initiative "Topic Detection & Tracking" [18], topic modeling methods were originally applied to press articles in the context of Information Retrieval, and they now know a renewed interest for the analysis of various sets of documents. More precisely, in our context, Topic Modeling was developed for four main problems: event detection, event summary, emerging trend detection and news tracking.

Event detection and event summary. Numerous definitions of an "event" have recently been proposed in the literature (e.g. [19, 5]). A general definition given by Dou *et al.* [20] describes an event by four attributes: topic, time, people and location. Event analysis in Twitter revolves around two broad steps. The first one referred to as "event detection" extracts relevant documents from tweet flows associated with specific topics within a specific time period. Event detection approaches often combine two techniques : name entity recognition which selects event-related tweets by focusing on the personalities and locations mentioned in the documents [21, 19] and peak detection which exploits the quick

response-time of Twitter users to detect bursts of interest [22]. Once an event has been identified, the second step produces a summary to answer the four traditional issues of an investigator: who ? what ? where ? and when ? Summaries come under different forms e.g. a tweet set, a dashboard with the four constitutive components of an event [20, 23], an ordered list of representative terms [24]. Most summary techniques are based on LDA modeling and several variants complete the data with external sources (e.g. documents associated with URLs mentioned in tweets [25] or press articles [26]). Event discovery methods handle time as a key feature to detect and summarize events, but they are not adapted to follow a concept life cycle.

Emerging trend detection and news tracking. Emerging trend detection –also called "new event detection" or "first story detection"– aims at identifying trends before they become hot topics. The Twitter text-stream analysis is done by tearing apart the emerging topics of discussion from the constant background. One of the most famous works on emerging trend detection in Twitter is Toretter [27]. This earthquake reporting system makes use of Twitter users as social sensors to detect natural disasters occurring in Japan. It combines an SVM classifier with a Kalman filter model to estimate the location of the event. Other approaches have been developed: they are based on incremental clustering, matrix factorization [28] and LDA [6]. When new events are detected, the analyst seeks to analyze their temporal evolution. News tracking methods, also called trend analysis, especially focus on this evolution over time. They regroup various approaches: temporal association rule mining [29], dynamic clustering [30], dynamic LDA [31] and an original correlation analysis inspired by the agenda-setting theory [32] between the evolution of topics and the evolution of opinion leaders' interests [33]. Roughly speaking, most of the recent efforts have been focused on scalability challenges and the results of the automatic algorithms still often provide general tendencies at a high granularity level.

Visual analytics. An alternative to full-automation is to embed users in the discovery process via an interactive visual support. In a recent overview on

visual analysis for social media data Schreck and Keim [34] list problems that face social data processing and conclude that "these problems can make fully automatic analysis of social media data impractical. The analysis often requires exploratory approaches, and user involvement is crucial". Visual analytics that combines data analysis, visualization and human factors [35] has been proven efficient in guiding end-users to discover useful knowledge. This approach has led to a range of proposals for text stream data. In early 2010, a state-of-the-art already inventoried thirty approaches combining various data analysis techniques and visual restitutions [36] and since then, several systems based on various metaphors have been proposed for Twitter data. Twitinfo [23] combines a timeline display of the tweet volume with additional representations (maps for geolocation, sector diagram of positive/negative sentiments, list of popular URLs) to highlight peaks of high tweet activities. VisualBackChanel [37] uses a streamgraph to visualize the main topics of persistent conversations associated with a range of events. TwitterScope [38] regroups similar tweets into clusters displayed as countries on a map whose topology evolves over time with the new data. Whisper [39] focuses on the collective responses of communities on a topic of interest and represents the social and spatio-temporal components of the diffusion process on a sunflower display. FluxFlow [40] has been designed for visualizing anomalous conversational threads in Twitter detected with a sequential anomaly detection algorithm. It combines a thread glyph for aggregating the main information and thread timeline visualizations for different perspectives (e.g. temporal trends of user volume in a retweeting thread, timestamp of each retweet event). The validation of these systems still remains a difficult open problem widely debated in the community but the user feedback is promising.

However, most of the above approaches have concentrated on visualizing main trends of temporal events and, as far as we know only FluxFlow focuses on anomalies and it can be closer to our weak signal detection goal. But, it only represents the top ranked anomalous retweeting threads and their interpretation requires a sophisticated interaction task. Here, we resort to a display familiar to

the industrial environment of our targeted users to simultaneously represent the tweet topics at different periods and the patterns of their evolution over time.

3. Dynamic topic-modeling

Our objective is to combine the efficiency of proven topic modeling for event detection with a human-centered data exploration necessary to go beyond highlights. In this section, we first recall the main steps of the Latent Dirichlet Allocation (LDA) algorithm used for topic extraction. Then, we present our extension for modeling the dynamic evolution of topics over time. The approach is based on a graph model easily adaptable to an interactive visualization.

3.1. Topic modeling with LDA

LDA is a generative probabilistic model originally used for discovering the "topics" that occur in a textual document collection D [15]. This model assumes that the occurrences of terms in documents are due to latent variables (the "topics"). Formally, a topic z is defined as a probability distribution over the terms contained in the documents. In our context, a textual document $d \in D$ is a tweet described by a sequence of words obtained from a preprocessing stage (see section 5.1 for details). The term set, called vocabulary, is denoted by V. A term occurrence at the j^{th} position in the tweet d is a word denoted $w_{d,j}$. For a pre-defined number of topics k, the model assumes that the tweets arose from the following generative process:

For each tweet d:

(1) Randomly choose a distribution over topics

$\boldsymbol{\theta_d} \sim Dir(\boldsymbol{\alpha})$

(2) For each word $w_{d,j}$ at the j^{th} position in tweet d

(a) Randomly choose a topic from the distribution over topics in step (1)

$z_{d,j} \sim Multinomial(\boldsymbol{\theta_d})$

(b) Randomly choose a term from this topic

$w_{d,j} \sim Multinomial(\boldsymbol{\phi_{z_{d,j}}})$

where $\boldsymbol{\phi_z}$ is the distribution over terms for a topic z. In step (1), the per-tweet topic distribution is drawn from a Dirichlet distribution $Dir(\boldsymbol{\alpha})$. Like the beta distribution, which is its one-dimensional restriction, the Dirichlet distribution can take various forms, which makes it a good candidate to modelize prior distributions for categorical variables.

As is commonly done, we add a prior step (0) at the beginning of the generative process[1]: for each topic z the term distribution $\boldsymbol{\phi_z}$ is chosen from another Dirichlet distribution, which is denoted $\{\boldsymbol{\phi_z}\}_{z=1}^{k} \sim Dir(\boldsymbol{\beta})$. Hence, the two Multinomial distributions in the process have a Dirichlet prior. This allows to simplify the calculations for estimating the model parameters since the Dirichlet distribution is the conjugate prior of the Multinomial distribution. Finally, the joint probability of the model is given by:

$$p(W, Z, \theta, \phi \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{z=1}^{k} p(\boldsymbol{\phi_z} \mid \boldsymbol{\beta}) \prod_{d \in \mathcal{D}} p(\boldsymbol{\theta_d} \mid \boldsymbol{\alpha}) \prod_{j=1}^{|d|} p(w_{d,j} \mid \boldsymbol{\phi_{z_{d,j}}}) p(z_{d,j} \mid \boldsymbol{\theta_d})$$

where $W = \{w_{d,j}\}_{d,j}$ is a word set, $Z = \{z_{d,j}\}_{d,j}$ is a topic assignment for each word, $\theta = \{\boldsymbol{\theta_d}\}_d$ is a set of per-tweet topic mixtures, $\phi = \{\boldsymbol{\phi_z}\}_z$ is a set of per-topic term mixtures. $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the parameters of the Dirichlet priors.

One of the main advantages of this model relies on the reduced number of the parameters to be estimated which greatly mitigates the problem of overfitting. Different approaches have been proposed to learn the distributions. For computational efficiency, we here use a Gibbs sampling [41] where the topic proportions are alternatively estimated by fixing the others.

3.2. Temporal topic relationship modeling

In the following, we consider that the tweet set D is recorded during a time interval T. For modeling the dynamic evolution of topics over time, T is divided in n consecutive periods $T_1, \ldots, T_n$ and the corresponding tweet sets $D_1, \ldots, D_n$ are built from the tweet publication dates. The corresponding vocabularies

[1] This is the "smooth" version of LDA, which allows generalizing the model to new documents

are denoted $V_1, \ldots, V_n$. As the basic scheme of LDA does not exploit time information, we here train an LDA model for each period T_i on the tweet set D_i for extracting timely focused topics. For each T_i, we obtain a set $\widehat{\boldsymbol{Z}}_{\boldsymbol{i}}$ of k topics described by the probability distributions of their most representative terms; for a topic $z \in \widehat{\boldsymbol{Z}}_{\boldsymbol{i}}$ observed during T_i, the term distribution is denoted $\boldsymbol{\phi_{z,i}}$. However, listing the topic sets $\widehat{\boldsymbol{Z}}_{\boldsymbol{i}}$ for all periods T_i is not sufficient to follow their evolution over a long period since the detection of hidden relationships requires a lot of cognitive effort. Our analysis of the tweet content evolution is based on a sequential evaluation of the relationship degree between the topics extracted from consecutive tweet sets D_i and D_{i+1}. The proximity between two topics $z \in \widehat{\boldsymbol{Z}}_{\boldsymbol{i}}$ and $z' \in \widehat{\boldsymbol{Z}}_{\boldsymbol{i+1}}$ is measured by a dissimilarity $\delta(z, z')$: small δ values correspond to strong relationships between topics which persist over time. The choice of δ is discussed in detail in section 4.3.

The temporal topic relationships are then modeled by a layered weighted digraph G. Vertices represent topics such that each topic $z \in \widehat{\boldsymbol{Z}}_{\boldsymbol{i}}$ is represented on the layer L_i. A directed edge is added between two topics z and z' on two adjacent layers L_i and L_{i+1} if $\delta(z, z')$ is smaller than a threshold τ fixed by the user. The weight $w(z)$ of a vertex is equal to $\sum_{d \in D_i} p(z \mid d)$, the sum of the topic probabilities during T_i. Here, unlike previous models where each tweet d just contributes to its most representative topic, each tweet contributes to the weight of several topics depending on the probability $p(z \mid d)$. The weight $w(z, z')$ of a directed edge depends both on the respective weights $w(z)$ and $w(z')$ of the linked topics and on their dissimilarity $\delta(z, z')$:

$$w(z, z') = (w(z) + w(z')) \times (1 - \delta(z, z'))$$

Let us note that this model requires a filtering of the LDA results to limit the dissimilarity computation complexity. As terms with low probabilities can be considered as poor descriptors of a topic, we only retain the r most representative terms of each topic. We denote by V_i^* the set of all r-most-representative terms of $\widehat{\boldsymbol{Z}}_{\boldsymbol{i}}$:

$$V_i^* = \left\{ t \in V_i \quad | \quad \exists z \in \widehat{\boldsymbol{Z}}_{\boldsymbol{i}} \;\; \boldsymbol{\phi_{z,i}}(t) \geq \max_r(\boldsymbol{\phi_{z,i}}) \right\}$$

where $\max_r(\boldsymbol{\phi_{z,i}})$ is the r^{th} value of $\boldsymbol{\phi_{z,i}}$ sorted by decreasing order. Moreover, for the dissimilarity computation, it is useful to define the distributions $\boldsymbol{\phi_{z,i}}$ and $\boldsymbol{\phi_{z,i+1}}$ on a same set. Let us consider the common vocabulary set $U_{i,i+1} = V_i^* \cup V_{i+1}^*$. For a given topic z, the term distributions $\boldsymbol{\phi^*_{z,i}}$ and $\boldsymbol{\phi^*_{z,i+1}}$ are both described by a vector of length $\text{card}(U_{i,i+1})$ s.t. for each term $t \in U_{i,i+1}$, $\boldsymbol{\phi^*_{z,i}}(t) = \boldsymbol{\phi_{z,i}}(t)$ (resp. $\boldsymbol{\phi^*_{z,i+1}}(t) = \boldsymbol{\phi_{z,i+1}}(t)$) if $t \in V_i^*$ (resp. $t \in V_{i+1}^*$) or 0 otherwise. In this new coding, probabilities of terms associated to the topics of $\widehat{\boldsymbol{Z}}_{\boldsymbol{i}}$ (resp. $\widehat{\boldsymbol{Z}}_{\boldsymbol{i+1}}$) do not change but a 0 has been added for the terms belonging exclusively to V_{i+1}^* (resp. V_i^*).

4. Parameter settings

Results of the process described above depend on three elements: the adopted topic number r at each period, the hyper-parameter $\boldsymbol{\alpha}$ of the LDA model which influences the allocation of the tweet probabilities on each topic, and the dissimilarity δ which measures the topic link intensity between two consecutive periods. We have compared different values of $\boldsymbol{\alpha}$ and r and different definitions of δ on a real-life corpus of 70 000 tweets covering a one-year period of publication (see section 5 for details).

4.1. Choice of the number of topics

Selecting an adequate topic number k which reflects the wealth of the corpus as faithfully as possible is a tricky issue. A basic way is to experiment the LDA model with different k values in order to select the one which leads to the more understandable results. Here, as the model is applied at each time interval, the exploration of k values is impracticable. Therefore, we have chosen to estimate the best-adapted k with an automatic approach based on the Hierarchical Dirichlet Process (HDP). This topic model is a non-parametric generalization of LDA, i.e. the number of topics is not a parameter but is learnt from data. The generative process of HDP is based on nested Dirichlet processes:

- for each tweet in D, a distribution over topics is drawn from a Dirichlet process;

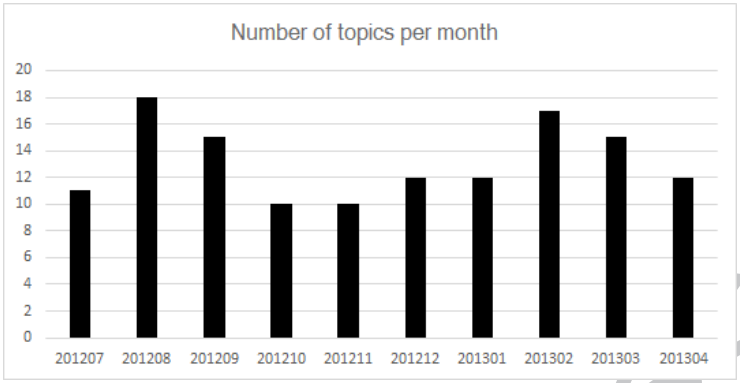

Figure 1: Number of topics discovered by a Hierarchical Dirichlet Process for each one-month period.

- for all the tweets in D, the Dirichlet processes share the same base distribution which is itself drawn from another Dirichlet process.

We refer to [42] for a detailed presentation of the HDP process. Contrary to the Dirichlet distribution in LDA, the Dirichlet process does not require to specify the problem dimensionality, i.e. the number of topics. The figure 1 shows the number of topics discovered by HDP in the tweets for each month. It shows that $k = 10$ is an adequate value on average. For the months where HDP generates more than 10 topics, some of the topics are very specific.

The strength of the Hierarchical Dirichlet Process is its independence from the parameter k. However, its main limit is still its execution time for parameter estimation. This is the reason why we use it only to empirically justify a ballpark value for k, and keep LDA at the heart of our methodology. Nevertheless, last advances on sampling complexity reduction of topic models [43] give hope for operational implementation of complex models such as HDP over the next few years.

4.2. Choice of the concentration parameter

In the LDA generative process, the parameter $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$ directly influences the topic distribution in the documents. Each component α_j can

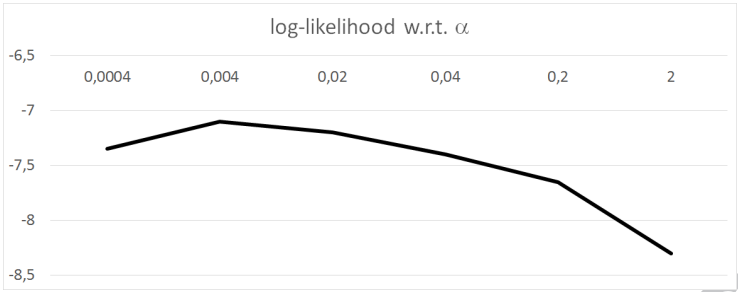

Figure 2: LDA model log-likelihood w.r.t. α.

be interpreted as an inertia for the j^{th} topic z_j (a small α_j tends to make z_j negligible on average in the documents, while a large α_j tends to make it significant). In order not to favor any particular topic, we use a symmetric Dirichlet prior, i.e. all the components α_j are equal. In this case, the α_j value directly controls the sparsity of the topic distribution:

- a high α_j value means that each document is likely to contain a mixture of most of the topics;
- a low α_j value means that each document is likely to contain a mixture of just a few topics, or even only one.

For our dataset, we have tried several α_j values between 0^+ and 2 and compared the outputted models with log-likelihood (see figure 2). The best model is obtained with $\alpha_j = 0.004$.

4.3. Choice of the dissimilarity

Numerous dissimilarities have been proposed in the literature (e.g. [44]) and the choice is tricky here as it may influence the interpretation of the topic relationships. To guide this choice we have experimentally compared the empirical distributions of seven classical measures adapted to our data: cosine, Euclidean distance, generalized Jaccard, Kullback-Leibler, Jensen-Shannon, Bhattacharyya and Hellinger (see Appendix A. for their definitions). In the experiments, ten time-interval of the same amplitude (one month) were retained and k

SPEARMAN RHO	Bhattacharyya	Cosine	Euclidean	Jaccard	Hellinger	Jensen-Shannon	Kullback-Leibler
Bhattacharyya		0.99	0.7	0.98	1.0	0.99	0.99
Cosine	0.99		0.7	0.98	0.99	0.99	0.97
Euclidean	0.7	0.7		0.68	0.68	0.68	0.68
Jaccard	0.98	0.98	0.68		0.98	0.98	0.95
Hellinger	1.0	0.99	0.68	0.98		0.99	0.99
Jensen-Shannon	0.99	0.99	0.68	0.98	0.99		0.98
Kullback-Leibler	0.99	0.97	0.68	0.95	0.99	0.98	

Figure 3: Rank correlation (Spearman Rho coefficient) between different dissimilarities.

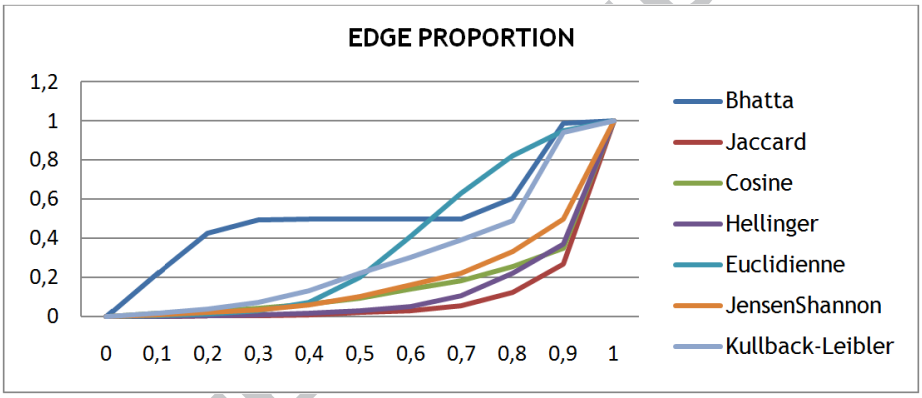

Figure 4: Number of edges in the layered digraph G w.r.t. the dissimilarity threshold τ.

was fixed at 10. For each dissimilarity, we have computed all the values between each topic pair $z \in \widehat{\boldsymbol{Z}}_{\boldsymbol{i}}$ and $z' \in \widehat{\boldsymbol{Z}}_{\boldsymbol{i+1}}$ at consecutive periods T_i and T_{i+1}, for i ranging from 1 to 9.

A first computation of the Spearman's rank correlation ρ between all the dissimilarity values for the first two periods shows an order correlation between the different measures (see figure 3). This result reassures us on some stability of the model whatever δ. However, we must also take into account the sensitivity of δ to the parameter used for the temporal topic link detection (see section 3.2): the threshold τ allows filtering the topic relationships that are retained on the layered digraph G. Choosing its default setting is a difficult

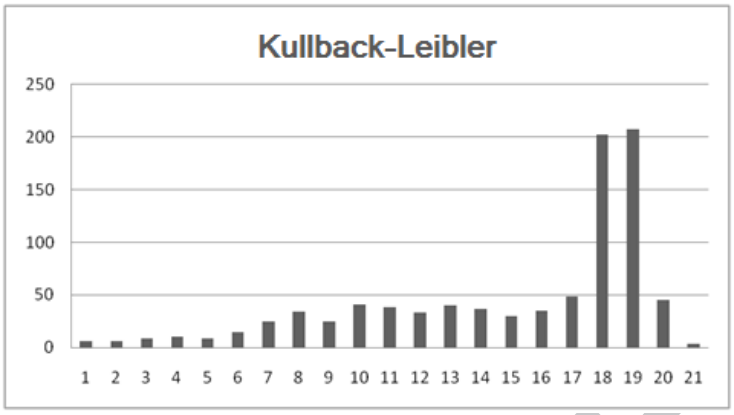

Figure 5: Observed statistical distribution of the distance values over a one-year interval for the Kullback-Leibler distance.

task when no complementary information is given *a priori*. Consequently, it can be interactively modified by the user and a certain robustness of the results is required for the intelligibility of the interpretation. More precisely, it is necessary to maintain an efficient filtering while avoiding important variations of the number of created directed edges when τ varies only slightly. Figure 4 shows the variation of the number of retained directed edges on the layered digraph for different values of τ. The dissimilarities have different behaviors. Cosine, generalized Jaccard, Jensen-Shannon and Hellinger are essentially sensitive for the large values; Bhattacharyya combines a large plateau with gross changes; Kullback-Leibler is almost linear except for the large values; and the Euclidean distance presents a sigmoid based variation. The curves associated to the Kullback-Leibler and the Euclidean distances are best suited for our interactive filtering. Moreover, their observed statistical distributions (figures 5 and 6) confirm that these measures allow a significant discrimination between their values, contrary to other measures including cosine, Jaccard, Hellinger which have very concentrated distributions.

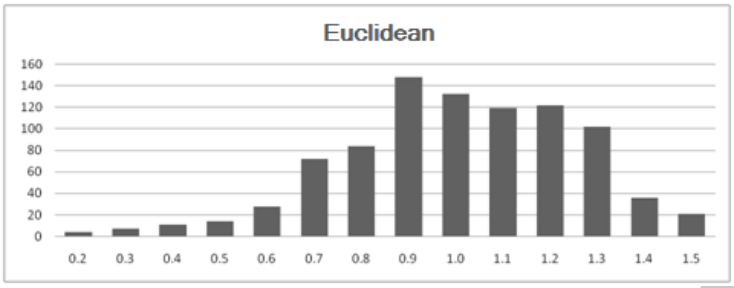

Figure 6: Observed statistical distribution of the distance values over a one-year interval for the Euclidean distance.

5. Application

In this section, we first briefly introduce the preprocessing stage for tweet transformation into sequences of words adapted to the LDA modeling process, then we present the interactive visualization of the temporal topic relationships, and finally we discuss results obtained on the EDF dataset with the support of an expert of the company.

5.1. Data acquisition and preparation

Twitter offers different applications to ease access to messages. In particular, the Streaming APIs give low latency access to representative samples of Twitter's global stream and the REST APIs allow third-party applications to collect relevant tweets in a recent period. In our work, data acquisition is based on the proprietary extraction tool ListenTwitter [45] which daily queries a REST API to retrieve french-speaking twitter messages with the keyword "edf" in reference to the French energy company "Electricité De France" (EDF). Because "edf" may also refer, in French, to "Equipe De France" (French team) a filter [46] is applied to remove all the tweets that refer to sport. Finally, our dataset contains 73023 tweets published during the period Jun. 17 2012 to May 02 2013.

Raw tweets, with their specific form of language, require a preprocessing step which transforms each tweet into a word sequence. Special characters (e.g. “, ”, \, /), accents, mentions, URL's and RT symbols are removed. A

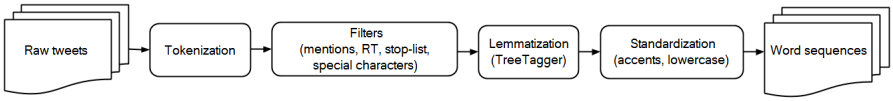

Figure 7: Preprocessing chain for the tweet transformation into word sequences adapted to the LDA modeling process.

stop-list is applied to filter out non-informative words. In order to limit term variations, we apply a lemmatization process based on the well-proven language tool independent tool TreeTagger [47]. Figure 7 recalls the main steps of the preprocessing chain.

5.2. Visual analytics

Various approaches have been proposed to visualize time-oriented data [48]. In our context, we preferred Sankey diagrams [16] which were originally developed to visualize energy flows and more generally dynamic relationships in a system. Their popularity in industrial environments makes company appropriation for this representation easy. Each weighted vertex (topic) z of the layered digraph G defined in section 3.2 is represented by a vertical rectangle of area proportional to $w(z)$ and each directed edge (z, z') is represented by a curve of width $w(z, z')$ function of the dissimilarity $\delta(z, z')$. Figure 8 presents the obtained Sankey diagram for ten one-month periods with $\tau = 0.49$ and δ is the Euclidean distance. A grey stream, width of which is proportional to the number of tweets at each period, highlights the percentage of tweets explained by the chains. The chain colors are chosen on the CIELAB hue wheel[2] in order to maximize the perceived gaps between topics without any relationship. The independent topics, which do no persist over time, are represented above the stream. By default, each topic is described by its two most representative terms but an associated term cloud can be interactively displayed for details.

[2] scanline.ca/hue/cielab.html

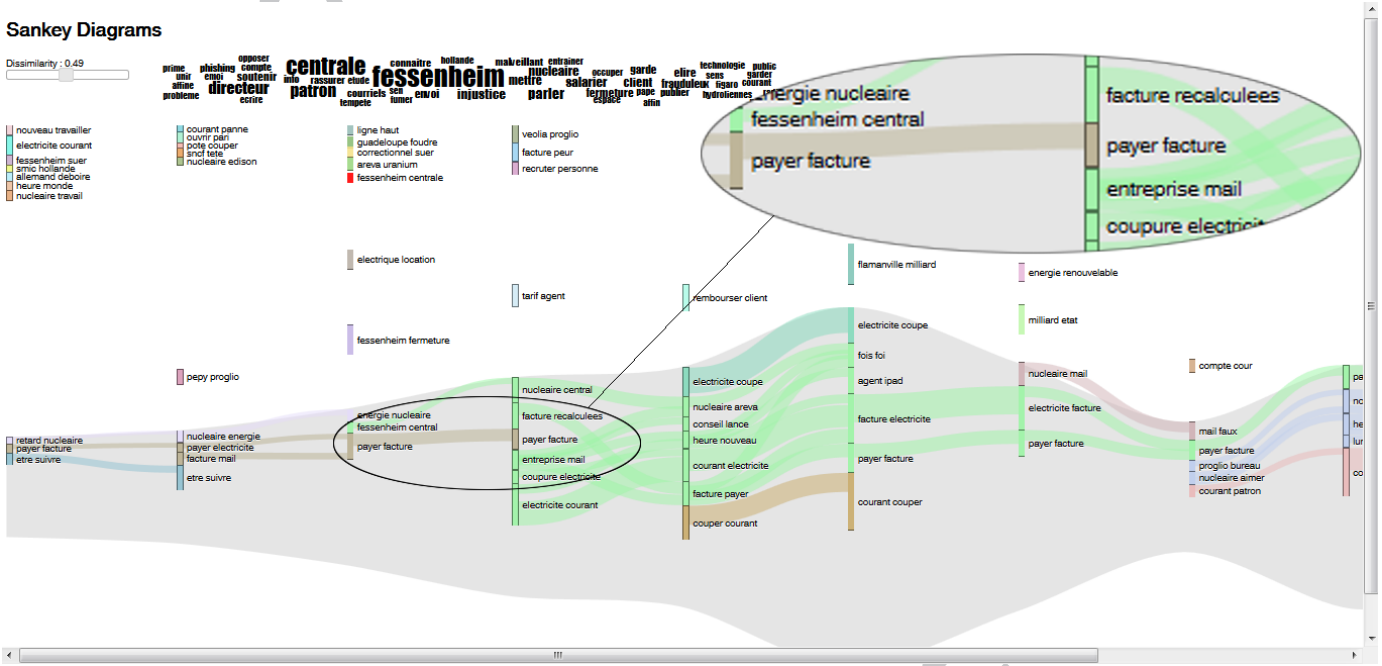

Figure 8: Visualizing topic evolution with a Sankey diagram

Each layer is associated with a one-month period. Vertices represent topics and their areas are proportional to the topic probability sums during a period. Edges represent relationships between topics at two consecutive periods. The edge width is function of the topic weights of its extremities and their dissimilarity value. The grey stream is proportional to the number of tweets at each period. Topics which do no persist over time are represented above the stream.

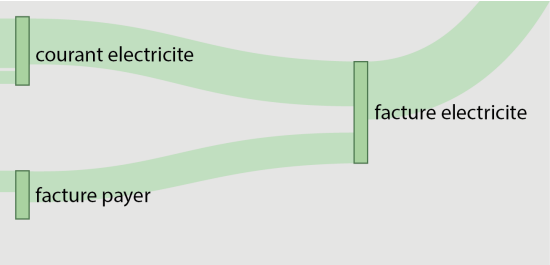

Figure 9: Two topics merging into a single one in the Sankey diagram (the topics "power, electricity" and "bill, pay" merge to form the topic "electricity bill").

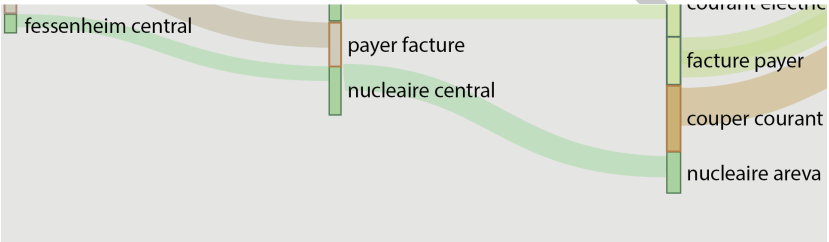

Figure 10: Zoom on the evolution of topics concerning the nuclear sector in the Sankey diagram.

The representation highlights two topic classes: precise topics representative of the company's latest news (e.g. on rumors of change at the C-level and on the departure of the company's president) and topics which persist in background throughout the period with varying importance such as nuclear energy, power bills and power outages. More subtly, the interactive Sankey diagram has allowed the detection of three different situations for the lasting subjects: (i) well-identified topics associated with the same terms throughout the period, (ii) subjects associated with close topic unions (e.g. figure 9 for the power bills) and (iii) topics which evolve over time. For instance, figure 10 shows an evolution of concerns with the nuclear sector: the topic "Fessenheim's nuclear power station" focused on a specific location is generalized to "nuclear power station" and evolves to "Areva nuclear" which is one of the key player in the nuclear industry. Detecting this evolution may be useful to adapt the way to

communicate on this sensitive topic.

On a methodological point of view, the prototype was tested at different steps of its development by an expert of the company on his own data in his everyday work environment. Moreover, the obtained results were compared on the same data with classical approaches in data mining based on itemset analysis and co-clustering [49] and the informational wealth of the discovered topic relationships with our novel approach was found to be significantly greater. In particular, as far as we know, patterns of type (ii) and (iii) were not explicitly considered by the previous approaches.

6. Conclusion

This paper discusses the problem of discovering information –and especially weak signals– relevant for a company in a large tweet set targeting this company. Our major contributions are twofold. First, we propose a human-centered mining process based on a coupling between a topic extraction algorithm and a visualization of the temporal evolution of the topics and of their relationships. This approach relies on an original adaptation of the LDA (Latent Dirichlet Allocation) model initially defined for a given data set to a temporal data set sequence. Used by an expert on a real-life tweet set collected during one year, this approach has retrieved highlights, and more interestingly it has detected different types of evolving patterns hidden by the data volume and commonly masked by the frequency-based filtering techniques. Second, we have experimentally analysed the ability of seven dissimilarities to evaluate the link robustness between pairs of topics detected at two consecutive time periods. Although a correlation analysis has confirmed a global stability of our model, a more accurate comparison of the statistical behavior of each dissimilarity has allowed us to identify the best candidates (the Kullback-Leibler and the Euclidean distances) for our user-based interactive approach.

This research direction has a great potential since the influence of social media in decision-making has become unavoidable in most business sectors; be-

sides the detection of weak signals provides some promising leads for innovation. Even if the interest in human-centered processes is now confirmed by various real-life experience feedback, research efforts still need to be pursued in two directions : (i) the evaluation of their added-value and (ii) a deeper integration of company knowledge into the discovery loop.

Citing exploratory data analysis as an example, a recent paper on the subject notes that "evaluation (...) is complex since, for a thorough understanding of a tool, it not only involves accessing the visualizations themselves, but also the complex processes that a tool is meant to support" [50]. Evaluating the added value of the human interactivity in a discovery process is indeed an open issue which is widely discussed both in the dataming and in the information visualization communities. There is still no consensual methodology and shared adapted benchmarks are very difficult to build. Here, our objective was to evaluate to what extent our approach "supports the generation of actionable and relevant knowledge" [50], and as often in visual analytics, we have considered a real-life case study with a domain expert. A process transfer is currently under discussion in a different field (the use of Twitter data in the legal field [51]). A novel application could contribute to prove its robustness.

In our approach, user knowledge which contributes to weak signal detection is implicitly integrated into the discovering process via the visual interaction. A complementary approach is to improve this integration through an explicit representation of this knowledge such as ontologies. Originally defined in the early 1990s as "formal, explicit specifications of shared conceptualizations" [52], ontologies have known an increasing development with the growth of the Semantic Web and they are now the cornerstone of semantic technologies. Coupling ontologies with discovering algorithms has proven efficient in data mining, in particular for reducing the number of potentially interesting rules extracted from large databases [53]. However, this process is still in its infancy for social media data. A very recent survey on the analysis of social media through semantics confirms that, for Twitter, "the majority of approaches to event detection do not utilise ontologies or other sources of semantic information" [54]. In the

past few years various ontologies have attempted to model different aspects of social media (e.g social relationships with FOAF, user-related information with. GUMO). A deeper analysis is required to estimate their operational utility in our context where topics deal with the specific area of interest of a company. Domain ontologies (e.g. [55]) focused on a specific domain and often linked to a specific application are certainly more appropriate, and when not available, the framework UEML (Unified Enterprise ModeLing) which has been applied on a variety of enterprise activities [56] may be useful to guide their construction. Generally speaking, the combination of data mining with semantic technologies and visualization for driving knowledge discovery process is a very promising challenge for the next few years.

Acknowledgments. The authors would like to thank XXXXX for his helpful contribution to the prototype.

References

[1] D. J. Power, F. Burstein, R. Sharda, Reflections on the past and future of decision support systems: Perspective of eleven pioneers, in: Decision Support, Springer, 2011, pp. 25–48.

[2] J. Ajmera, H. Ahn, M. Nagarajan, A. Verma, D. Contractor, S. Dill, M. Denesuk, A CRM System for Social Media: Challenges and Experiences, in: Proceedings of the 22nd International Conference on World Wide Web, WWW'13, 2013, pp. 49–58.

[3] B. Liu, Sentiment Analysis and Opinion Mining, CA:Claypool Publishers, 2012.

[4] P. Gundecha, H. Liu, Mining social media: a brief introduction, INFORMS 9 (4) (2012) 1–17.

[5] A. McMinn, Y. Moshfeghi, J. Jose, Building a large-scale corpus for evaluating event detection on twitter, in: Proceedings of the 22nd International

Conference on Information and Knowledge Management, WWW'13, 2013, pp. 409–418.

[6] J. Lau, N. Collier, T. Baldwin, On-line Trend Analysis with Topic Models:# twitter Trends Detection Topic Model Online., in: COLING, 2012, pp. 1519–1534.

[7] H. Hansoff, Managing strategic surprise by response to weak signals, California Management Review 18 (2) (1975) 21–33.

[8] B. Coffman, Weak Signal Research, Part I: Introduction, Journal of Transition Management (1997) .

[9] I. Ilmola, O. Kuusi, Filters of weak signals hinder foresight: Monitoring weak signals efficiently in corporate decision making, Futures 32 (8) (2006) 908–924.

[10] S. Sidhom, P. Lambert, Information Design for Weak Signal detection and processing in economic intelligence: case study on health resources, in: Proceedings of the 4th International Conference on Information Systems and Economic Intelligence , 2011, pp. 315–321.

[11] R. Eckhoff, M. Markus, M. Lassnig, S. Schön, No outstanding surprises when using social media as source for weak signals ?, in: Proceedings of the 9th International Conference on Digital Society , 2015, pp. 59–63.

[12] T. Ahlqvist, A. Bck, M. Halonen, S. Heinonen, Social Media Roadmaps - Exploring the futures triggered by social media, Tech. rep., VTT Tiedotteita - Research Notes 2454 (2008).

[13] C. Charitonidis, A. Rashid, P. J. Taylor, Weak Signals as Predictors of Real-World Phenomena in Social Media, in: Proceedings of the IEEE/ACM International Conference on Advances in Social Network Analysis and Mining , 2015, pp. 864–871.

[14] M. Harrysson, E. Métayer, H. Sarrazin, The strength of weak signals., McKinsey Quarterly (February) (2014) .

[15] D. Blei, A. Ng, M. Jordan, Latent Dirichlet Allocation, The Journal of Machine Learning Research 3 (2003) 993–1022.

[16] H. Sankey, The Thermal Efficiency Of Steam-Engines , in: Minutes of the Proceedings of the Institution of Civil Engineers, Vol. 125, Thomas Telford, 1896, pp. 182–212.

[17] D. M. Blei, J. D. Lafferty, Topic models, in: Text Mining: Classification, Clustering, and Applications, Vol. 10, Chapman & Hall/CRC Press, 2009, pp. 71–89.

[18] J. Allan, J. G. Carbonell, G. Doddington, J. Yamron, Y. Yang, Topic detection and tracking pilot study final report, Tech. rep., Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop (1998).

[19] A. Ritter, O. Etzioni, S. Clark, et al., Open domain event extraction from twitter, in: Proceedings of the 18th International Conference on Knowledge Discovery and Data Mining, SIGKDD'12, 2012, pp. 1104–1112.

[20] W. Dou, X. Wang, D. Skau, W. Ribarsky, M. Zhou, Leadline: Interactive visual analysis of text data through event identification and exploration, in: Proceedings of the Conference on Visual Analytics Science and Technology, VAST'12, 2012, pp. 93–102.

[21] T. Hua, F. Chen, L. Zhao, C. Lu, N. Ramakrishnan, STED: Semi-supervised Targeted-interest Event Detection in Twitter, in: Proceedings of the 19th International Conference on Knowledge Discovery and Data Mining, KDD '13, 2013, pp. 1466–1469.

[22] R. Parikh, K. Karlapalem, ET: Events from Tweets, in: Proceedings of the 22nd international conference on World Wide Web companion, 2013, pp. 613–620.

[23] A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, R. C. Miller, Twitinfo: aggregating and visualizing microblogs for event exploration, in: Proceedings of the Annual Conference on Human Factors in Computing Systems, CHI '11, 2011, pp. 227–236.

[24] F. Chua, S. Asur, Automatic Summarization of Events From Social Media, in: Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM'13, 2013.

[25] O. Jin, N. N. Liu, K. Zhao, Y. Yu, Q. Yang, Transferring topical knowledge from auxiliary long texts for short text clustering, in: Proceedings of the 20th International Conference on Information and Knowledge Management, CIKM'11, 2011, pp. 775–784.

[26] W. Gao, P. Li, K. Darwish, Joint topic modeling for event summarization across news and social media streams, in: Proceedings of the 21st International Conference on Information and Knowledge Management, CIKM'12, 2012, pp. 1173–1182.

[27] T. Sakaki, M. Okazaki, Y. Matsuo, Earthquake shakes Twitter users: real-time event detection by social sensors, in: Proceedings of the 19th International Conference on World Wide Web, WWW'10, 2010, pp. 851–860.

[28] S. P. Kasiviswanathan, P. Melville, A. Banerjee, V. Sindhwani, Emerging topic detection using dictionary learning, in: Proceedings of the 20th International Conference on Information and Knowledge Management, CIKM'11, 2011, pp. 745–754.

[29] M. Adedoyin-Olowe, M. M. Gaber, F. Stahl, TRCM: a Methodology for Temporal Analysis of Evolving Concepts in Twitter, in: Artificial Intelligence and Soft Computing, Springer, 2013, pp. 135–145.

[30] E. Gansner, Y. Hu, S. North, Visualizing Streaming Text Data with Dynamic Graphs and Maps, in: Post-proceedings of the 20th International

Symposium of Graph Drawing, Vol. 7704 of Lecture Notes in Computer Science, Springer, 2012, pp. 439–450.

[31] J. Vosecky, D. Jiang, K. W. Leung, W. Ng, Dynamic Multi-faceted Topic Discovery in Twitter, in: Proceedings of the 22nd International Conference on Information & Knowledge Management, CIKM'13, 2013, pp. 879–884.

[32] M. E. McCombs, D. L. Shaw, The agenda-setting function of mass media, Public Opinion Quarterly 36 (2) (1972) 176–187.

[33] P. Xu, Y. Wu, E. Wei, T. Peng, S. Liu, J. Zhu, H. Qu, Visual analysis of topic competition on social media, IEEE Transactions on Visualization and Computer Graphics 19 (12) (2013) 2012–2021.

[34] T. Schreck, D. Keim, Visual analysis of social media data, Computer 46 (5) (2013) 68–75.

[35] D. Keim, F. Mansmann, J. Schneidewind, H. Ziegler, Challenges in visual data analysis, in: Proceedings of the 10th International Conference on Information Visualization, 2006, pp. 9–16.

[36] A. Šilić, B. D. Bašić, Visualization of text streams: A survey, in: Knowledge-Based and Intelligent Information and Engineering Systems, Springer, 2010, pp. 31–43.

[37] M. Dork, D. Gruen, C. Williamson, S. Carpendale, A visual backchannel for large-scale event, IEEE Transaction on Visualization and Computer Graphics 16 (6) (2010) 1129–1138.

[38] E. R. Gansner, Y. Hu, S. C. North, Interactive visualization of streaming text data with dynamic maps, Journal of graph algorithms and applications 17 (4) (2013) 515–540.

[39] N. Cao, Y.-R. Lin, X. Sun, D. Lazer, S. Liu, H. Qu, Whisper: Tracing the spatiotemporal process of information visualization in real time, IEEE

Transaction on Visualization and Computer Graphics 18 (12) (2012) 2649–2658.

[40] J. Zhao, N. Cao, Z. Weng, Y. Song, Y. Lin, R. Collins, Flux Flow: Visual analysis of anomalous information spreading on social media, IEEE Transaction on Visualization and Computer Graphics 20 (12) (2014) 1773–1782.

[41] D. Newman, A. Asuncion, P. Smyth, M. Welling, Distributed algorithms for topic models, J. Mach. Learn. Res. 10 (2009) 1801–1828.

[42] Y. Teh, M. Jordan, M. Beal, D. Blei, Hierarchical dirichlet processes, Journal of the american statistical association 101 (476) (2006) .

[43] A. Li, A. Ahmed, S. Ravi, A. Smola, Reducing the sampling complexity of topic models, in: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2014, pp. 891–900.

[44] M.-J. Lesot, M. Rifqi, H. Benhadda, Similarity measures for binary and numerical data: a survey, International Journal of Knowledge Engineering and Soft Data Paradigms 1 (1) (2009) 63–84.

[45] A. Beck, X. XXX, Procédé d'identification d'une donnée comme pertinente ou hors sujet, Tech. rep., Patent Filling FR3016712(A1) (2015).

[46] A. Stoica, X. XXX, X. XXX, Extraction, clustering and visualization for strategic monitoring (in french) , in: Actes de la conférence Veille Stratégique et Technologique, 2013, pp. 50–53.

[47] H. Schmid, Probabilistic part-of-speech tagging using decision trees, in: International Conference on New Methods in Language Processing, Manchester, UK, 1994, pp. 44–49.

[48] W. Aigner, S. Miksch, H. Shuman, C. Tominsky, Visualization of time-oriented data, Springer, 2011.

[49] X. XXX, X. XXX, X. XXX, X. XXX, X. XXX, Visual analysis of topics in Twitter based on co-evolution of terms , in: Data Science, Learning by latent structures and knowledge discovery (Post-proceedings of the European Conference on Data Analysis), Springer, 2014, pp. 1217–1226.

[50] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, S. Carpendale, Empirical studies in information visualization: seven scenarios, IEEE Transactions on Visualization and Computer Graphics 18 (9) (2012) 1520–1536.

[51] P. Saurel, Y. Girard, X. XXX, Designing a tweet aggregator with semi-supervised learning of legal ontologies (in french), in: Actes du 13ème atelier sur la fouille de données complexes, 2016, pp. 63–72.

[52] T. Gruber, A translation approach to portable ontology specification, Knowledge Acquisition 5 (1993) 199–220.

[53] C. Marinica, X. XXX, Knowledge-based interactive postmining of association rules using ontologies , IEEE Transactions on Knowledge and Data Engineering 22 (6) (2010) 784–797.

[54] K. Bontcheva, D. Rout, Making sense of social media streams through semantics: a survey, Semantic Web 5 (5) (2014) 373–403.

[55] R. Mizoguchi, Tutorial on ontological engineering - Part I: Introduction to ontological engineering, New Generation Computing 21 (4) (2003) 365–384.

[56] A. Opdahl, G. Berio, M. Harzallah, R. Matulevicius, Ontology for entreprise and information systems modelling, Applied Ontology 7 (1) (2014) 49–92.

Appendix A. Dissimilarity measures.

The dissimilarity between two probability distributions ϕ_1 and ϕ_2 can be measured by:

- cosine:

$$cosineDis(\phi_1, \phi_2) = 1 - \frac{\sum \phi_{1i}.\phi_{2i}}{\sqrt{\sum \phi_{1i}^2}.\sqrt{\sum \phi_{2i}^2}}$$

- Euclidean distance:

$$eucliDis(\phi_1, \phi_2) = \sqrt{\sum (\phi_{1i} - \phi_{2i})^2}$$

- generalized Jaccard:

$$genJaccDis(\phi_1, \phi_2) = 1 - \frac{\sum \min(\phi_{1i}, \phi_{2i})}{\sum \max(\phi_{1i}, \phi_{2i})}$$

- Kullback-Leibler:

$$\begin{aligned} KLDis(\phi_1, \phi_2) &= \frac{1}{2} KL(\phi_1, \phi_2) + \frac{1}{2} KL(\phi_2, \phi_1) \\ &= \frac{1}{2} \sum \phi_{1i}. \log(\frac{\phi_{1i}}{\phi_{2i}}) + \frac{1}{2} \sum \phi_{2i}. \log(\frac{\phi_{2i}}{\phi_{1i}}) \end{aligned}$$

- Jensen-Shannon:

$$JSDis(\phi_1, \phi_2) = \frac{1}{2} \sum \phi_{1i}. \log(\frac{\phi_{1i}}{\frac{1}{2}(\phi_{1i} + \phi_{2i})}) + \frac{1}{2} \sum \phi_{2i}. \log(\frac{\phi_{2i}}{\frac{1}{2}(\phi_{1i} + \phi_{2i})})$$

- Bhattacharyya:

$$BCDis(\phi_1, \phi_2) = -\ln(\sum \sqrt{\phi_{1i}.\phi_{2i}})$$

- Hellinger:

$$HDis(\phi_1, \phi_2) = \sqrt{1 - \sum \sqrt{\phi_{1i}.\phi_{2i}}}$$