

Visual Analysis of Topics in Twitter Based on Co-evolution of Terms

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

Abstract The analysis of Twitter short messages has become a key issue for companies seeking to understand consumer behaviour and expectations. However, automatic algorithms for topic tracking often extract general tendencies at a high granularity level and do not provide added value to experts who are looking for more subtle information. In this paper, we focus on the visualization of the co-evolution of terms in tweets in order to facilitate the analysis of the evolution of topics by a decision-maker. We take advantage of the perceptual quality of heatmaps to display our 3D data (term \times time \times score) in a 2D space. Furthermore, by computing an appropriate order to display the main terms on the heatmap, our methodology ensures an intuitive visualization of their co-evolution. An experiment was conducted on real-life datasets in collaboration with an expert in customer relationship management working at the French energy company EDF. The first results show three different kinds of co-evolution of terms: bursty features, reoccurring terms and long periods of activity.

1 Introduction

The sharp increase in the use of social web technology has led to an explosion of user-generated time-labelled texts such as news, on-line discussions and Twitter short messages. The analysis of this data has become a key issue for companies

L. Pépin (✉)
EDF R&D, Clamart, France

Equipe COD-LINA (UMR CNRS 6241), Université de Nantes, Nantes, France
e-mail: lambert.pepin@edf.fr; lambert.pepin@univ-nantes.fr

J. Blanchard • F. Guillet • P. Kuntz
Equipe COD-LINA (UMR CNRS 6241), Université de Nantes, Nantes, France
e-mail: julien.blanchard@univ-nantes.fr; fabrice.guillet@univ-nantes.fr;
pascale.kuntz@univ-nantes.fr

P. Suignard
EDF R&D, Clamart, France
e-mail: philippe.suignard@edf.fr

seeking to understand consumer behaviour and expectations. In particular, Twitter, as a social networking and microblogging service, has become one of the most visited websites (<http://www.alexa.com/topsites>).

As a consequence, topic modelling (Blei and Lafferty 2006) and information tracking (Leskovec et al. 2009) in a time-labelled corpus are of renewed interest. Roughly speaking, most recent efforts have been focused on scalability challenges and the results of the automatic algorithms often provide general tendencies at a high granularity level. However, because of the volatile nature of trending topics on Twitter and the restricted format of the published messages, automatically modelling the evolution of their “semantic content” still remains an open problem. More specifically, observations of twitter messages show that: (1) most topics have a short life span “73% of the topics have a single active period and 31% of the periods are 1 day long” (Kwak et al. 2010) and (2) a high number of new words occur daily (see Fig. 1). The topic relevance is especially sensitive in practice where experts are well informed about the main trends of their customers’ behaviour and are looking for more subtle information. An alternative to full-automation is to embed the user in the discovery process via an interactive visual support. Visual analytics which “is more than visualization and can rather be seen as an integrated approach combining visualization, human factors and data analysis” (Keim et al. 2010) has been proven efficient in guiding end-users to discover useful knowledge.

In this paper, we focus on the visualization of the co-evolution of Twitter terms in order to facilitate the analysis of the topic evolution by a decision maker. The visualization of terms with common behaviour might, indeed, favour the emergence of concepts. We propose a process in three main steps for co-evolution extraction: (1) scoring terms, (2) grouping the terms that evolve in the same way across a corpus and (3) visualizing the evolution of term scores on a heatmap. More precisely, we take advantage of the perceptual quality of heatmaps to display our 3D data (term \times time \times score) in a 2D space. Further, by computing an appropriate order to display the main terms on the heatmap, our methodology ensures an intuitive visualization

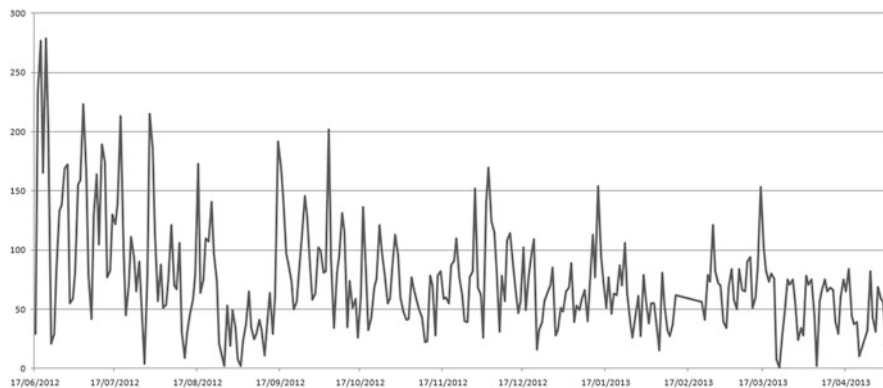


Fig. 1 Number of new words each day

of their co-evolution. Our approach was tested on a real-life dataset from the French energy company EDF.

The rest of the paper is organized as follows: Sect. 2 briefly reviews related work on topic detection and tracking (TDT) applied to Twitter and presents some visualizations adapted to time-oriented data. In Sect. 3, we put forward our methodology. In Sect. 4, we detail our experimental protocol and the results obtained in collaboration with an expert in customer relationship management.

2 Related Work

TDT (Allan et al. 1998) is concerned with three main problems: first, modelling the meanings of documents (the topics), then detecting events in a time-labelled corpus (giving boundaries to the addressed stories) and finally, tracking information over time.

Topic modelling which aims at understanding the semantic meanings of documents has been addressed in numerous papers (Kasiviswanathan et al. 2011; Caballero et al. 2012; Jo et al. 2011), from statistical approaches with Tf-Idf (Jones 1972) and Okapi-BM25 (Robertson et al. 1999) to probabilistic modelling with probabilistic-Latent Semantic Allocation (pLSA) (Hoffman 1999) and Latent Dirichlet Allocation (LDA) (Blei et al. 2003) through linear algebra with Latent semantic analysis (LSA) (Deerwester et al. 1990).

Event Detection and especially Peak Detection (Kleinberg 2002) also emphasizes the interest of time in understanding a corpus. Peak detection approaches take advantage of a time-labelled corpus to apply techniques from time series analysis and provide boundaries to the addressed stories (Marcus et al. 2011).

News Tracking addresses the problem of the evolution of topics over time. Some dynamic extensions of LDA (Hoffman et al. 2010) aim at tracking topic transitions. Graph clustering algorithms (Leskovec et al. 2009) may be used to construct groups of terms and see their evolution over time. Dynamic Clustering and Mapping algorithms (Gansner et al. 2012) may be used to construct a dynamic map which preserves the user's mental map.

Visualization. The increasing accessibility of time-oriented data has led to the development of various visual representations [see Aigner et al. (2011) for a recent overview]. The pioneering ones were dedicated to visualizing time series and data semantics were not taken into account. Over the last few years, different visualizations have been put forward to track topic evolution, but they do not explicitly take into account topic co-evolution. For instance, stacked areas charts (Leskovec et al. 2009) allow one to draw many distributions on the same chart (see Fig. 2), but are more efficient at displaying the evolution of one component with respect to the total flow than at highlighting the co-evolution of different components. Evolution graphs (Mei and Zhai 2005; Jo et al. 2011) display the evolution of topics over time, but the non-linearity of the evolution of topics (some may split in two, or two may merge into one) makes it difficult to visualize their

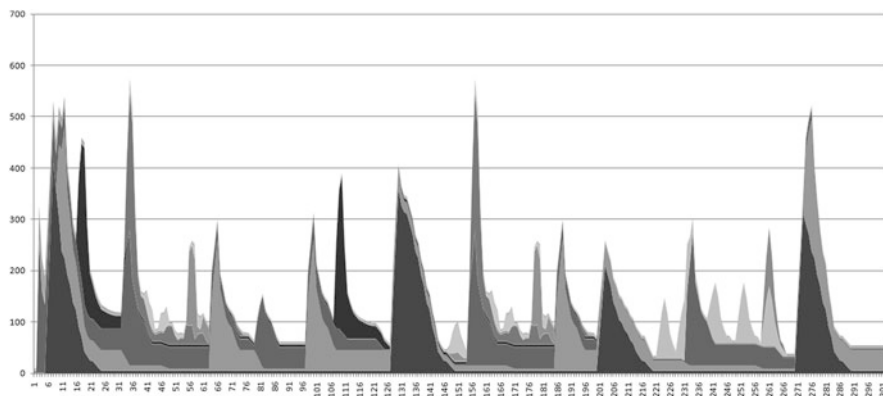


Fig. 2 Example of stacked areas chart

co-evolution. Dynamic clustering algorithms (Gansner et al. 2012) display data on a map and continuously update the visualization to provide a succinct view of evolving topics of interest. This kind of visualization is of great interest when tracking the evolution of the whole graph, but because of the dynamic representation of time, it may be difficult to visualize the co-evolution of clusters. In order to overcome this limitation, we hereafter propose a representation based on heatmaps. Clustered heatmaps have proven their efficiency in particular in bioinformatics: they are one of the most popular means of visualizing genomic data (North et al. 2005). However, as far as we know, their application to social networking analysis has not yet been explored.

3 Methodology

Heatmaps represent the values contained in a matrix as colours, and allow one to visualize 3D data in a 2D space. Because a lot of distributions can be displayed on the same map with fixed size cells, heatmaps are well suited to detecting patterns in the gradient of colours.

Central to ensuring the perceptual quality of heatmaps is the order in which data are displayed. Different orders may, indeed, lead to different perceptions. To deal with the above challenges, we propose a methodology for grouping terms that evolve in the same way and visualize their evolution.

The full process is represented in Fig. 3 and may be summarized as follows : First, we divide the total time span into fixed length time periods and select the terms to be compared. For each term, we build a raw vector, called “evolution vector”, with a length equal to the number of periods and which contains the values of the scoring function for this term and each period. Then, we compute the co-evolution matrix, which indicates, for each pair of terms, how similar the temporal evolution vectors

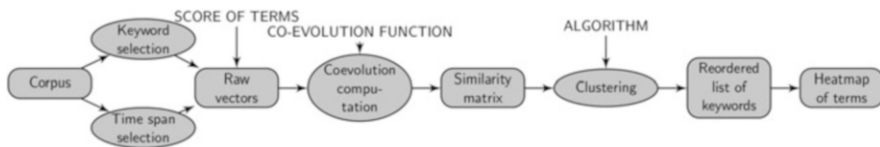


Fig. 3 Methodology summary

are. Thereafter, we apply a clustering algorithm to the co-evolution matrix in order to group together the terms that share common behaviour. Finally, we use the result of this clustering to order the terms and we visualize their evolution vectors with a heatmap.

3.1 Terms and Time Spans Selection

Given a set of tweets \mathcal{T} , recorded during a time interval t , we divide t into a set of fixed length time blocks $\{t_1, \dots, t_n\}$. Accordingly, the set of all tweets \mathcal{T} is split into n subsets $\{\mathcal{T}_1, \dots, \mathcal{T}_n\}$ where \mathcal{T}_i contains all the tweets of the period t_i . At the same time, the set of all tweets \mathcal{T} is processed to extract the keywords to be compared. First, each tweet of \mathcal{T} is tokenized into words, then any step of preprocessing may be applied to improve the quality of the clustering and hence the quality of the visualization. For instance, keywords may be unigrams or bigrams, and one may use lemme or stemme forms of words. The user may also feel it necessary to remove words from a predefined stop-list or with particularly high or low frequencies.

3.2 Scoring of Terms

The next step of our methodology is to define the measure used to represent the behaviour of each term across the corpus. Let \mathcal{W} be the set of keywords to be tracked over the corpus. We call scoring-function, a function s such as

$$s : \mathcal{W} \times \{1, \dots, n\} \rightarrow \mathbb{R} \tag{1}$$

$$w \times i \quad \mapsto s(w, i)$$

i.e. for each word $w \in \mathcal{W}$ and each period $t_i, i \in \{1, \dots, n\}$, s gives the value of the measure of interest of w for the period i . Depending on the application, s may be the number of documents containing w published during period i , or it may be a binary value corresponding to the fact that there is at least one document containing w in this period. The choice of the scoring-function is application driven and mainly depends on the interest of the user.

3.3 Co-evolution of Terms

For each keyword $w \in \mathcal{W}$, we define the evolution vector $\mathbf{ev}(w)$ of w by $\mathbf{ev}_i(w) = s(w, i)$, $\forall i \in \{1, \dots, n\}$. For each pair of keywords (w_1, w_2) , we compute the co-evolution of w_1 and w_2 as the co-evolution of their temporal evolution vectors $(\mathbf{ev}(w_1), \mathbf{ev}(w_2))$. The co-evolution function is symmetric. Intuitively, the more similar two vectors are, the higher their co-evolution is. Depending on the expert's objectives, the co-evolution function may be a similarity measure such as Jaccard, or a correlation measure.

By computing the co-evolution of each pair of keywords, we build a $|\mathcal{W}| \times |\mathcal{W}|$ co-evolution matrix, \mathcal{M}_S . Rows and columns of \mathcal{M}_S are labelled with keywords ordered the same way, such as : $\forall i, j \in \{1, \dots, |\mathcal{W}|\}$,

$$\mathcal{M}_S(i, j) = \text{sim}(\mathbf{ev}(w_i), \mathbf{ev}(w_j)) \quad (2)$$

3.4 Clustering

The key point here is that the order in which keywords are shown in \mathcal{M}_S does not depict any correlation between terms. The goal of this step is to reorder terms of \mathcal{W} with respect to their correlation. Hence, we apply a hierarchical clustering to \mathcal{M}_S in order to build groups of terms and display co-occurring terms close to each other. At the beginning, each term is assigned to a distinct class and at each step the algorithm merges the two most similar classes, continuing until only a single class remains.

3.5 Heatmap

Once we get a meaningful order for the terms of our vocabulary \mathcal{W} , we build a heatmap of term scores, the columns of which correspond to the n periods, displayed in chronological order, and with $|\mathcal{W}|$ rows corresponding to the reordered terms of \mathcal{W} . The colour of each cell corresponds to the score provided by the scoring function for this term and this period.

4 Experimental Settings

Our experiments use a corpus of tweets exclusively written in French, but our methodology does not make any assumptions about the language and may be applied to any kinds of “keywords”.

4.1 Dataset

The dataset used for this experiment is composed of all tweets containing the word “edf”, in reference to the French energy company “Electricité De France”, published during the period June 17, 2012 to May 02, 2013. Because “edf” may also refer, in French, to “Equipe De France” (French team) a filter is applied to remove all the tweets that refer to sport. This filtering is not straightforward and will not be detailed here. This step leads to a corpus \mathcal{T} of 73,023 tweets.

Each tweet is tokenized so that tokens are separated by a non letter-character. Mentions, URLs, “RTs” and accents are removed and the terms are converted to lowercase. The lemme of each term is obtained with TreeTagger¹ and we build our vocabulary \mathcal{W} by retaining the 512 most frequent pairs of consecutive terms.

By computing the number of tweets published each hour, we observed very low activity at night, thus we divided our corpus into periods of a day spanning from 4 a.m. to 4 a.m. This leads to 307 blocks of time $\{t_1, \dots, t_{307}\}$ and 307 subsets of our corpus $\{\mathcal{T}_1, \dots, \mathcal{T}_{307}\}$, each \mathcal{T}_i containing all tweets published during the period i .

4.2 Methods

In order to depict the evolution of the popularity of each *keyword* over time, we define the scoring function s such as

$$\begin{aligned} s : \mathcal{W} \times \{1, \dots, n\} &\rightarrow \mathbb{R} \\ w \times i &\mapsto N(w, i) \end{aligned} \quad (3)$$

where $N(w, i)$ is the number of documents containing the *keyword* w published during the period t_i . At first glance, this function may seem too simplistic. But because we compute the correlation of each pair of *keywords* with a rank correlation coefficient, any monotonic transformation of the data would not have affected the results. Nonetheless, the choice of the scoring function is important and must be done with respect to the co-evolution function. This choice is mainly data driven and depends on the application. It could be a binary value depicting the presence or absence of the *keyword* or a normalization like TF-IDF.

This step leads to a temporary matrix of size 512×307 which contains the number of documents in which each term appears in each period. We use it to compute the correlation of each pair of terms with *Kendall's tau* coefficient. This method, unlike, for example, *Pearson's* and *Spearman's* coefficients is not only sensitive to linear relationships between two terms, but also detects non-linear relationships.

¹<http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>.

An agglomerative clustering algorithm is applied to the co-evolution matrix and we deduce from its result an order of the list of terms. At each step of the agglomerative clustering algorithm, the two most similar classes are merged using the *complete linkage criterion* chosen, here, for its simplicity and its tendency to favour compact clusters. For each class C_1, C_2 ,

$$\text{sim}(C_1, C_2) = \min_{x \in C_1, y \in C_2} (\text{sim}(x, y)).$$

This simple criterion has led to interpretable results but a comparison with other criteria is planned in the near future.

4.3 Results

This experiment was conducted in collaboration with an expert in natural language processing and customer relationship management working at the Research & Development department of EDF. Our expert was asked to visually detect groups of terms on the heatmap constructed by our methodology. Due to space constraints, we have only provided an extract of the heatmap, with a small number of *keywords* centred on their period of activity (see Fig. 4). This extract highlights three patterns of co-evolution visually detected by the expert.

In the first pattern (A) terms are associated with a high score for a few days (around August 17, 2012) and are not employed for the rest of the time. This pattern conforms with the short-life topics described by Kwak et al. (2010) and relies on

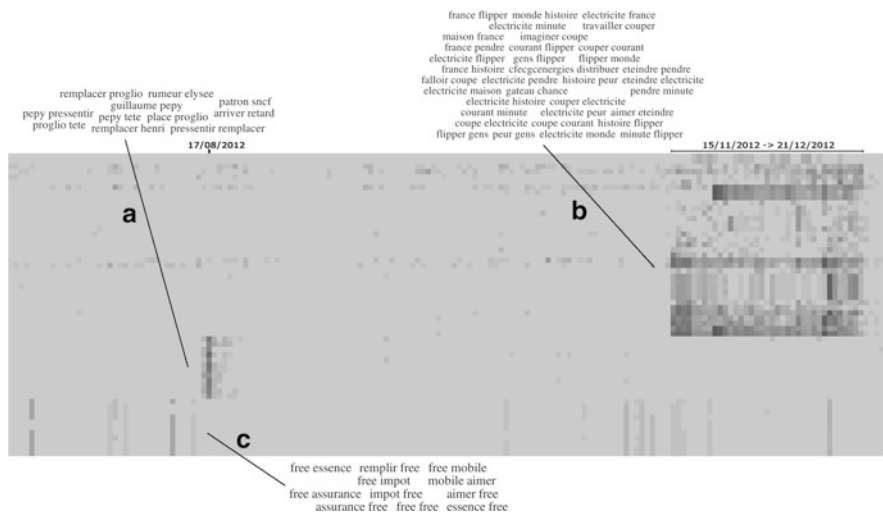


Fig. 4 Extract of the heatmap for a small number of terms centred on their period of activity

a rumour of change at the head of the executive board of EDF. This rumour was discussed in the traditional media and is not a novelty. However, detecting and quantifying the impact of this bursty feature in the social media remains valuable.

The second pattern (B) is a group of terms showing a long period of activity and may not be detected by a peak detection method. This pattern can be explained by a significant amount of documents about the end of the world announced in December 2012. In these messages Twitter users make jokes about the ancient Mayan prophecy and suggest switching off the power for 10 min to scare people. The joke stops short the day of the prophecy as it is no longer a topical issue.

Finally, in the third pattern (C), terms are employed together on many occasions and create what we call a recurring pattern. The volume of each appearance is low and may not trigger an alert, but put together these appearances provide a significant amount of information. These terms come from messages in which Twitter users humorously call for low prices in the energy sector following the lead of French telecommunications company Free. The interest of this pattern lies in the fact that, unlike the discussion about the Mayan prophecy, the joke is not circumstantial. This topic appears many times and is regularly discussed by Twitter users.

5 Conclusion

In this paper, we have presented a methodology for highlighting co-evolution patterns of Twitter terms on a visual support. Experiments on real-life data have led us to identify different classes of terms associated with characteristic behaviour. This first analysis plays the role of a proof-of-concept. We have presented here feedback from one expert only, and even if the first results are promising, we are faced with a problem of visualization evaluation. Generally speaking, this question is known to be highly problematic in visual analytics (Jankun-Kelly et al. 2007). However, here, we can rely on experience gained by using heatmaps in different domains. In addition to addressing the problem of evaluating the visualization, future works will include comparing co-evolution functions and clustering algorithms.

References

- Aigner, W., Miksch, S., Schumann, H., & Tominski, C. (2011). *Visualization of time-oriented data*. London: Springer.
- Allan, J., Carbonell, J. G., Doddington, G., Yamron, J., & Yang, Y. (1998). *Topic detection and tracking*. Pilot Study Final Report.
- Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning* (pp. 113–120).
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3, 993–1022.

- Caballero, K. L., Barajas, J., & Akella, R. (2012). The generalized Dirichlet distribution in enhanced topic detection. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM)* (pp. 773–782).
- Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R. A. (1990). Indexing by latent semantic analysis. *The Journal of the American Society for Information Science*, 41(6), 391–407.
- Gansner, E., Hu, Y., & North, S. (2012). Visualizing streaming text data with dynamic maps. In *Proceedings of the 20th International Conference on Graph Drawing* (pp. 439–450).
- Hoffman, M., Bach, F. R., & Blei, D. M. (2010). Online learning for latent Dirichlet allocation. In *Advances in Neural Information Processing Systems* (pp. 856–864).
- Hoffman, T. (1999). Probabilistic latent semantic analysis. In *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence (UAI)* (pp. 289–296). San Francisco, CA: Morgan Kaufmann.
- Jankun-Kelly, T. J., Ma, K.-L., & Gertz, M. (2007). A model and framework for visualization exploration. *IEEE Transactions on Visualization and Computer Graphics*, 13(2), 357–369.
- Jo, Y., Hopcroft, J. E., & Lagoze, C. (2011). The web of topics: Discovering the topology of topic evolution in a corpus. In *Proceedings of the 20th International Conference on World Wide Web (W3C)* (pp. 257–266).
- Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *The Journal of Documentation*, 28(1), 11–21.
- Kasisviswanathan, S. P., Melville, P., Banerjee, A., & Sindhvani, V. (2011). Emerging topic detection using dictionary learning. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (ICKM)* (pp. 745–754).
- Keim, D. A., Mansmann, F., & Thomas, J. (2010). Visual analytics: How much visualization and how much analytics? *ACM SIGKDD Explorations Newsletter*, 11(2), 5–8.
- Kleinberg, J. (2002). Bursty and hierarchical structure in streams. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)* (pp. 91–101).
- Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web (W3C)* (pp. 591–600).
- Leskovec, J., Backstrom, L., & Kleinberg, J. (2009). Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)* (pp. 497–506).
- Marcus, A., Bernstein, M. S., Badar, O., Karger, D. R., Madden, S., & Miller, R. C. (2011). Twitinfo: Aggregating and visualizing microblogs for event exploration. In *Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems (CHI)* (pp. 227–236).
- Mei, Q., & Zhai, C. (2005). Discovering evolutionary theme patterns from text: An exploration of temporal text mining. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (KDD)* (pp. 198–207).
- North, C., Rhyne, T.M., & Duca, K. (2005). Bioinformatics visualization: Introduction to the special issue. *Information Visualization*, 4(3), 147–148.
- Robertson, S. E., Walker, S., Beaulieu, M., & Willet, P. (1999). *Okapi at Trec-7: Automatic ad hoc filtering, VLC and interactive track* (pp. 253–264). Gaithersburg: Nist Special Publication SP.