A buzz and e-reputation monitoring tool for Twitter based on Galois Lattices

Etienne Cuvelier, Marie-Aude Aufaure

A buzz and e-reputation monitoring tool for Twitter based on Galois Lattices

Etienne Cuvelier[1] and Marie-Aude Aufaure[1]

Business Intelligence Team
Applied Mathematics and Systems Department (MAS)
École Centrale Paris,
Etienne.Cuvelier@ecp.fr, Marie-Aude.Aufaure@ecp.fr

Abstract. In the actual interconnected world, the speed of broadcasting of information leads the formation of opinions towards more and more immediacy. Big social networks, by allowing distribution, and therefore broadcasting of information in a almost instantaneous way, also speed up the formation of opinions concerning actuality. Then, these networks are great observatories of opinions and e-reputation. In this e-reputation monitoring task, it is easy to get a set of information (web pages, blog pages, tweets,...) containing a chosen word or a set of words (a company name, a domain of interest,...), and then we can easily search for the most used words. But a harder, but more interesting task, is to track the set of jointly used words in this dataset, because this latter contains the more shared advice about the initial searched set of words. Precisely, the exhaustive discovering of the shared properties of a collection of objects is the main task of the Galois lattices used in the Formal Concept Analysis. In this article we state clearly the characteristics, advantages and constraints of one of the more successful online social networks: Twitter. Then we detail the difficult task of tracking, on Twitter, the most forwarded information about a chosen subject. We also explain how the characteristics of Galois lattices permit to solve elegantly and efficiently this problem. But, retrieving the most used corpus of words is not enough, we have to show the results in an informative and readable manner, which is not easy when the result is a Galois Lattice. Then we propose a visualisation called topigraphic network of tags, which represent a tag cloud in a network of concepts with a topographic allegory, which permits to visualise the more important concepts found about a given search on Twitter.

1 Introduction

Since their appearance, blogs and social networks create a growing interest for observation and modelling of opinions, as illustrated by the special session on this subject of the TREC conferences since their edition of 2006 [OdRMS06]. Identifying Hot Topics in the Blogosphere was one of the tasks of the 2009 edition of this blog session [MOS09]. Social networks, like Facebook and Twitter,

with their sharing and forward features, should also permit to observe the appearance of opinions practically in real-time and then allow to detect tendencies. For instance [Kra10] uses words expressing emotions in Facebook's status of the American users to synthesize a new index modelling the concept of "Gross National Happiness".

Social networks are therefore ideal places for the observation of opinions, notably regarding a chosen subject, that can be a person (personal branding), an official institution or an industrial operator. In the case of the e-reputation, the observation of the buzz and more particularly of the negative buzz (bad buzz) is important. But monitoring a buzz and/or an ereputation is not only collect the set of information about a subject, it is also to structure this latter in a understandable way. We proposed a method for this on the most reactive of these networks: Twitter.

This article is organized in the following way: in the section 2 we detail succinctly how the Twitter network works, what are its constraints and its conventions, and what are the implied difficulties of analysis. In the section 3, we recall the basics of the Galois lattices which allow us to solve these difficulties. Finally in the section 4 we display the principles of our tool, as well as results acquired on the dataset of information relating to key word "e-reputation". We end with conclusions and perspectives of improvement.

2 Twitter and Micro-Blogging

Twitter was created in 2006 to allow its users to share easily short textual messages called *Tweets*. The system was initially conceived to share tweets via SMS, and then a limit of 140 characters was fixed to these short messages. And even if nowadays the system is mainly used via web applications and mobile phones softwares, this constraint of 140 characters is still true. The basic principles of Twitter are the following:

- a user can , with its Twitter account, generates or forwards an information using a specific field (field "Whats happening?" in figure 1);
- a user A can follows the tweets of a user B without this latter has to follow the tweets of A in return.

We see immediately that one characteristic of this social network is its asymmetric aspect.

The users which follow a Twitter account A are called his *followers*, while the users which A follows are called its *following*. The set of tweets of the following set from a given account is called his *timeline*. An illustration of each of these elements can be seen in the figure 1.

With its principle of "micro-blogging", Twitter allows to share information very quickly and then allows the diffusion of these information, but also of the opinions related to this latter. The growth of this service is nowadays important and, in April 2010, Twitter counted almost 6 million of recorded users, 300 000 new accounts a day and, on average, 55 million of tweets generated a day [Bos10].

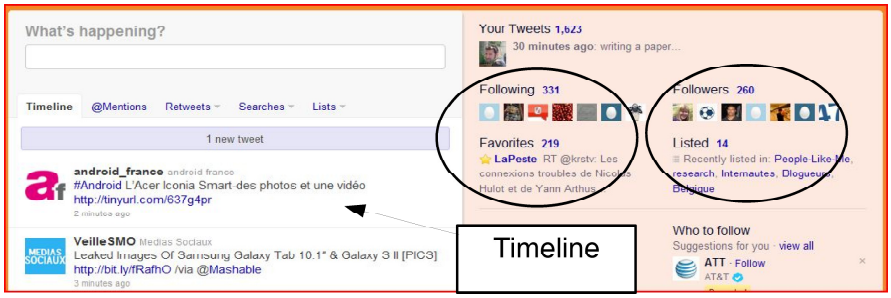

Fig. 1. Web interface for Twitter.com.

The result of this intense activity is a very big reactivity about the actuality facts, which can be illustrated by the the wikileaks case. In figure 2 we can see a peak of 2% published tweets containing the word "wikileaks" less than 24 hours after the first release of the diplomatic documents on wikileaks[Raz10]. This reactivity

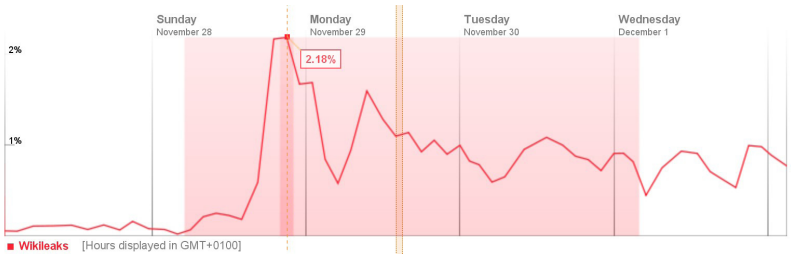

Fig. 2. Evolution of the number of tweets containing the word wikileaks since the publication on 28th of November of the diplomatics files (source: http://trendistic.com).

is obviously very interesting for the analysis of the observation of opinions.For instance [OBRS10] showed that there was a very important correlation between three existent indicators, calculated via daily inquiries and opinions formulated on Twitter regarding these subjects: the first index concerns the trust level of the American consumers, the second one is an opinion polls Obama vs. Mc Cain during the American presidential campaign and the last one was an evaluation of

the job of Obama as president. In a more predictive way, [ROK09] showed that information circulating through Twitter concerning avian flu, linked to a model of prediction of market, allows to predict more efficiently the opinion concerning the transformation of influenza into pandemic.

The use of Twitter is ruled by some conventions, and we are going to specify the most important for a minimal understanding of the platform. The first one of these conventions is the use of the arobase to name or contact a chosen user. In example below, the user Jules contacts the user Jim naming the user Catherine:

```
Jules: @Jim see you at 22h00 at @Catherine's home ?
```

Both users Jim and Catherine will see this tweet in their own time-lines. A second convention is use of the *retweet or RT*. When an user notices in his time-line an information which he wants to share with his followers, he will use the retweet function of the service (web or application) which he uses, as illustrated in the example below:

```
Jules: Inception is an awesome movie.
Jim: Not for me RT @Jules: Inception  is an awesome movie.
Catherine: LOL RT @Jim: Not for me RT @Jules:  is an awesome movie.
```

Even if in most of the cases retweets are preceded by *"RT @"*, other variants and practices coexist, as it was very well analysed in [BGL10]. For instance some users edit the retweet by adding, a *(via)* at the end of this latter, as in following example:

```
Jules: http://www.google.com is awesome!
Catherine: RT @Jules: http://www.google.com is awesome!
Jim: http://www.google.com is awesome! (via @Jules)
```

This possibility of edition of a retweet, with the constraint of the 140 characters limit creates something that we call *polymorphism of forwarded information* on Twitter. This polymorphism is illustrated in the figure 3. In this example we see that the initial tweet (tweet N. 0) was retweeted in many ways. We see a first group of retweets (N. 1.1 and N. 1.2) in which the initial tweet is unchanged, and a second group of retweets (N. 2.1 and N. 2.2), where the users changed the retweet slightly. Furthermore one of these retweets of this second group is itself a retweet of retweet (tweet N. 2.2.1). This polymorphism is a real problem when we want to measure the popularity of an information, the number of unchanged retweets is then not a sufficient, because if we limit our count to this, then we will forget the whole set of modified retweets which, in spite of their modifications, carry the same information.

If we do not take into account the stop words and the signs of punctuations (in our example: more, of, and, are, by, us), and restrict our work to the significant words, then we can use a set representation of the information carried, as it can be seen in figure 4. We also can see in this latter how the different inclusions allow us to show the shared part by all tweets, the core of the carried information. Inclusions also define a partial order which can be represented by a Hasse's

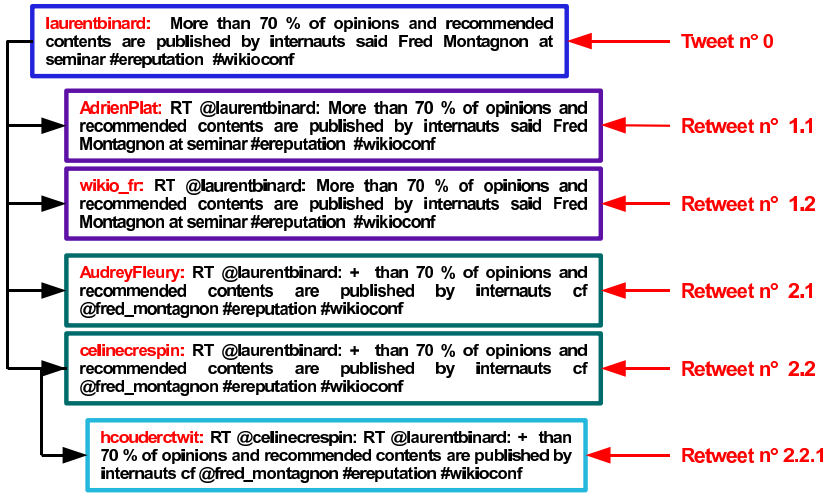

Fig. 3. Illustration of forwarded information's polymorphism.

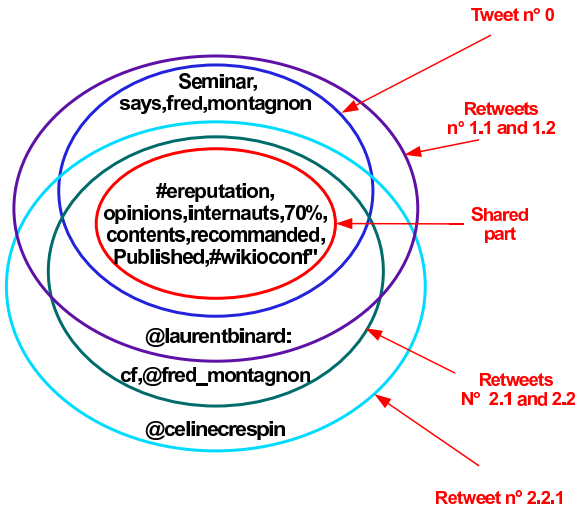

Fig. 4. Set point of view of forwarded information's polymorphism.

diagram as in the figure 5. Building such a diagram permits to establish, what are the common words in a group of tweets, as well as the different forms under which the same information was carried. Formal Concept Analysis and Galois Lattices are dedicated to the organization of information under this form.

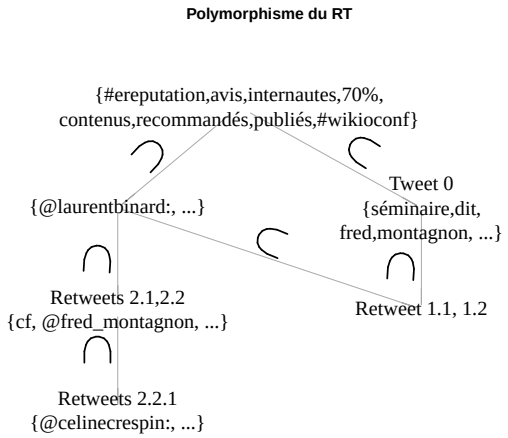

Fig. 5. Polymorphism analysis of information retweet.

3 Formal Concept Analysis - Galois Lattices

Formal Concept Analysis [Wil80] is based on *Galois Lattices* [BM70], [Bir40], which can be used for conceptual classification [CR93], [Wil84]. A Galois Lattice allows to group, in a exhaustive way, objects in classes, called *concepts*, using their shared properties, and is usually based on a boolean matrix, called the *context matrix* denoted C. Rows of C represent a group of *objects* O, and the columns, a group of *attributes* A used for the description the objects. To introduce this notion of lattice, we will use the lattice shown in figure 5. The corresponding context matrix can be seen in table 1. The possession of property $a \in A$ by an object $o \in O$ materializes the existence of a relation I between them: aIo. The existence of this relation I between O and A is meant in the matrix of context C by a value "true " (and "false" otherwise) or by any mark (and anything otherwise). The triplet $K = (O, A, I)$ is called one *formal context* or simply a context.

Words Tweets	Seminar, says, fred, montagnon	#ereputation, opinions, internauts, 70%, contents, recommanded, Published, #wikioconf	@laurentbinard:	cf, @fred_montagnon	@celinecrespin
0	x	x			
1.1	x	x	x		
1.2	x	x	x		
2.1		x	x	x	
2.2		x	x	x	
2.2.1		x	x	x	x

Table 1. Table de contexte pour les tweets.

The *intention* of a group $X \subset O$ is the set of attributes owned jointly by all objects of X and, given by the function f:

$$f(X) = \{a \in A | \forall o \in X, oIa\}. \tag{1}$$

Conversely the *extension* of a group $Y \subset A$ is all objects which jointly own all attributes of Y and, given by the function g:

$$g(Y) = \{o \in O | \forall a \in Y, oIa\}. \tag{2}$$

The couple (f, g) is called a *Galois connexion*.

A *concept* is any couple $C = (X, Y) \subset O \times$A, such as *the objects of* X *are the only ones to have all attributes of* Y, in other words $X \times Y$ form, except two permutations of O and of A, a maximum rectangle in C:

$$f(X) = Y \ \& \ g(Y) = X. \tag{3}$$

To illustrate this notion of concept, we can notice in table 1 that the set $X = \{$*Tweet 1.1, Tweet 1.2*$\}$ gives a concept because $f(X) = \{$*Seminar, says, fred, montagnon, #ereputation, opinions, internauts, 70%, contents, recommanded, Published, #wikioconf @laurentbinard:* $\} = Y$ and $g(Y) = X$, and this concept is then $(\{$*Tweet 1.1, Tweet 1.2*$\}, \{$*Seminar, says, fred, montagnon, #ereputation, opinions, internauts, 70%, contents, recommanded, Published, #wikioconf @laurentbinard:*$\})$, while the set $X' = \{$*Tweet 1.2, Tweet 2.1*$\}$ does not give a concept because $f(X') = \{$*#ereputation, opinions, internauts, 70%, contents, recommanded, Published, #wikioconf @laurentbinard:*$\} = Y'$ and $g(Y') = \{$*Tweet 1.1, Tweet 1.2, Tweet 2.1, Tweet 2.2, Tweet 2.2.1*$\} \neq X'$.

The *Galois lattice* is a *poset* of concepts L with the following partial order $\leq$:

$$(X_1, Y_1) \leq (X_2, Y_2) \Leftrightarrow X_1 \subseteq X_2 (\text{ or } Y_1 \supseteq Y_2). \tag{4}$$

The Galois lattice is denoted $T = (L, \leq)$ and, its representation is done using a *Hasse diagram*, as in figure 5 for the species. Two types of display exist for the labels of concepts, the full labelling and the reduced labelling. For the full labelling, all objects and all attributes of a concept are displayed, while in the reduced labelling, attributes and objects are displayed only once in the lattice. Attributes are displayed the first time they are met when going through the lattice top-down, while it is the contrary for objects, as we can see it in figure 5.

The construction of the lattice can be made using, for instance, the Bordat's algorithm [Bor86], which compute recursively all the existing concepts starting from the concept $(\emptyset, f(\emptyset))$, computing for each found concept the set of its sub-concepts. A good review of other algorithms for Galois lattices generation can be found in [KO01] which gives also a comparison of performances.

One of the main advantages of the lattice classification is that for a given context table the resulting lattice is unique (no execution instability), and is exhaustive (all existing concepts will be found). In our case this classification is going to allow us to find all the groups of words in a set of given tweets, as seen in figure 5.

4 E-buzz Monitoring

In the following section we propose to analyze a set of tweets, in order to find the most tweeted words or groups of words in the original set. For this we propose the following approach:

1. Getting the tweets including a chosen word or group of words,
2. Cleaning the tweets (suppressing stop words, punctuations,...);
3. Stating the table of context with the tweets as objects and the words as attributes;
4. Building the corresponding Galois lattice;
5. Visualisation of the results.

To illustrate our methods we are going to use it on a set of 50 tweets retrieved from a search on key word "#ereputation". Here below for the sake of illustration we give the 5 first tweets of this set:

```
Tweet 1: overclub: #ereputation : your opinion on multiple  technology
watch solutions... for you which is the best tool?
Tweet 2: AudreyFleury: #eReputation Internauts watch over their
ereputation Strategies http://ow.ly/1a7fWM
Tweet 3: AudreyFleury: RT @laurentbinard: + than 70% of opinions and
recommended contents are published by  internauts cf @fred_montagnon
#ereputation #wikioconf
Tweet 4: hcouderctwit: RT @celinecrespin: RT @laurentbinard: + than
70% of opinions and recommended contents are published by  internauts
cf @fred_montagnon #ereputation #wikioconf
Tweet 5: wikio_fr: RT @laurentbinard: At seminar #wikioconf, Serge
Alleyne, founder of  #nomao, announces an presents its local
#ereputation solution with #wikiobuzz...
```

Application of step 1 to 4 is easy on such a modest lattice, while the fifth step, the visualisation of the results is less trivial. In figure 6 we display the whole lattice giving to each concept a size proportional to the number of tweets contained. We notice that in spite of its small size (59 found concepts) it is difficult to display all the concepts proportionally to their sizes, and in the same time display clearly their attributes, even when using the reduced labelling. To reduce the number of attributes to be displayed, we can select only the concepts with a relative size (number of objects of the concepts divided by the number of objects in the context table) greater than chosen threshold. In other words, in respect to the notion of buzz we can select the concepts with the more tweeted words. That is what we have done in figure 7 with a threshold of 10%, but it does not increase enough the readability because the attributes of our lattice are, most of the times, groups of words more or less long. Another type of visualization is necessary for these concepts of the more retweeted words. Of course, we can try to display the

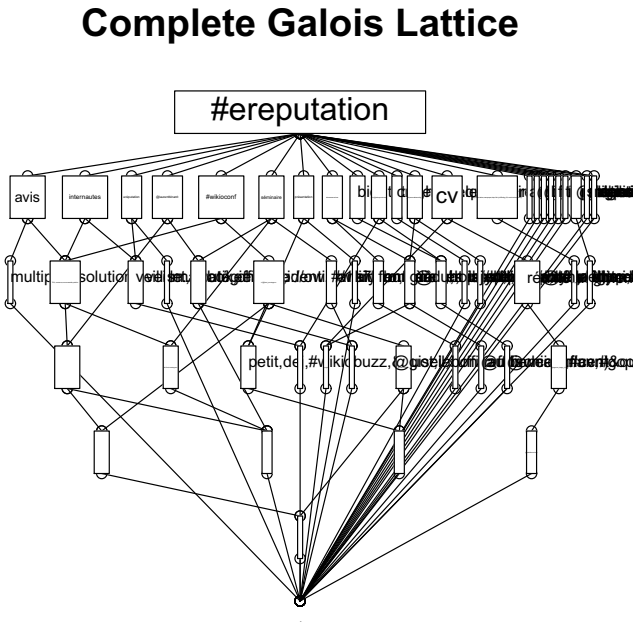

Fig. 6. The Galois lattice for the tweets.

different words of the found concepts using a classical tag cloud, giving to the tags a size proportional to the corresponding concept, but even if there exists solutions to display the associated tags near from the other one like in [KL07], we loose in this case the inclusion links between a concept and its sub-concepts. Moreover, as a sub-concept can have many super-concepts, it complicates the grouping task. That is why we propose to display the more important concepts using a network

of proportional tags (figure 8), in which the links betweens the concepts will be materialized by edges. These edges will be directed, going from the concept toward its sub-concepts. For the nodes layout we use the Fruchterman-Reingold method [FR91], because this technique optimizes the distance between the nodes and allows us to increase the readability of the tags. Finally, to reinforce a reading going from the most general to the most particular, we have decided to add a topographic allegory similarly to the topigraphic maps proposed by [FFM+08]. We call the result a *topigraphic network of tags*. To do this, for each point of the resulting graphic, we add a level, these levels being pictured using the classical level curves. To compute the level of all the points of the graphic we use a bi-dimensional gaussian probability densities mixture, using as means the centers of the tags and, as standard deviations the width and the height of these tags. Finally, to give a height proportional to the concepts' sizes exactly at the centers of these tags, we normalize the heights of the gaussians multiplying them by the standard deviations and the by the desired heights. The resulting mixture is the topigraphic function T:

$$T(x,y) = \sum_{i=1}^{k} \frac{s_i}{2\,\pi}\,\mathrm{e}^{-\frac{(x-x_i)^2\,(x-y_i)^2}{2 l_i^2 h_i^2}} \tag{5}$$

where:

- k is the number of displayed concepts,
- x_i and y_i are the coordinates of the i^{th} concept,
- l_i and h_i are the width and the height for the tag of the i^{th} concept,
- s_i is the size of the i^{th} concept.

Of course, as we have changed the volumes under the surfaces our topigraphic function T is not a probability density any more, but this property is not necessary in our case.

The final result can be seen in figure 9. In this figure, from the concept representing the starting key word {*#ereputation*}, we can see that the more important sub-concepts are {*opinions*}, {*internauts*} et {*#wikiconf*}, and these three concepts contain also the concept {*70%, contents, recommended, published*}, while only the concept {*#wikiconf*} contains the concept {*cf, @fred_montagnon*}. On the other hand we see three concepts shown independently of the first ones: {*cv*}, {*@laurentbinard:*} et {*daily, read, twitter, newspaper, http://bit.ly/d, (19, contributions, todays)*}. The main idea of this visualisation is to let the reader's look slide from the "top" (the more general concepts) toward the "valleys" (the less general concepts). The constructions and the displaying of the Galois lattice and of the topigraphic network of tags were made in the *R* statistical environment [R D10], using for the lattice part our own package *galois* (to be published on CRAN).

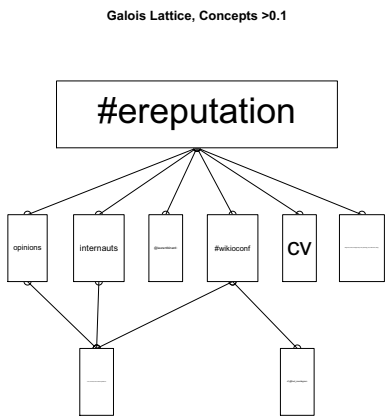

Fig. 7. The concepts containing more than 10% of tweets.

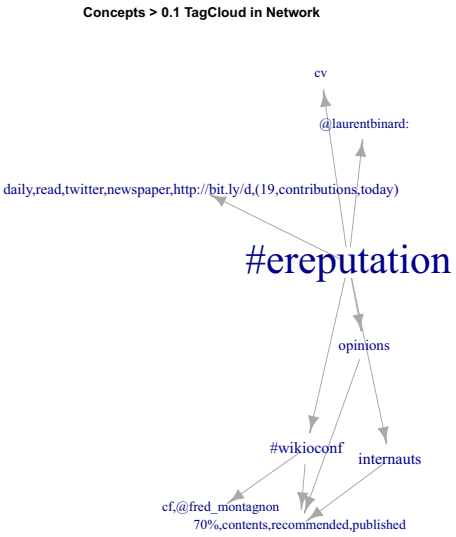

Fig. 8. A network of tags of concepts containing more than 10% of tweets.

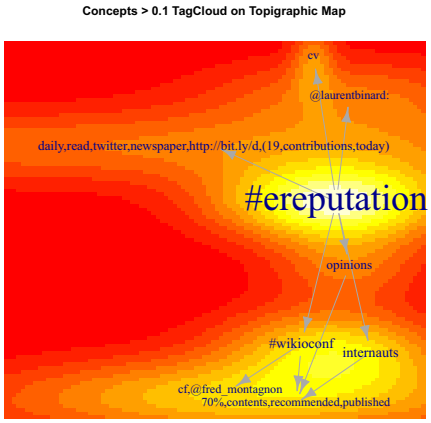

Fig. 9. The topigraphic network of tags of concepts containing more than 10% of tweets.

5 Conclusions and perspectives

In this paper we have presented a new technique for monitoring the buzz on the micro-blogging platform, Twitter. This technique is based on Galois lattices et and proposes as visualisation of the resulting concepts a topigraphic network of proportioned tags. This kind of display, limited to the more important concepts allows us to picture the tags belonging of a concept in a more readable manner than using directly the lattice. The main idea is to make "slip" the reader's look, from the more general concepts, displayed at the "tops" toward the more particular concepts placed more in the "valleys", arrows of the network being able to be seen as "lanes" to guide toward linked concepts.

Even if our proposal is only at a prototype stage, some improvement can be considered. The first one is the introduction of some interactivity to allow the user to select the sub-concept he wish to explore. We consider also to "develop" the shortened URLs (bit.ly, is.gd, tinyURL,..) in order to do not count twice a same final URL shortened using two different services. In the stop words suppression step, before stating the table of context we plan to a particular treatment could be reserved for smileys, which are meaningful. In a next step we envisage to use the sentiment analysis to assess the positivity or negativity of the found concepts, which is an interesting notion in the buzz and e-reputation monitoring. Finally the multi-language is an important but exciting challenge for such a tool.

References

[BGL10] Danah Boyd, Scott Golder, and Gilad Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *Proceedings of the 43rd*

Hawaii International Conference on Social Systems (HICSS), 2010.

[Bir40] Garrett Birkhoff. *Lattice Theory*, volume 25. American Mathematical Society, New York, 1940.

[BM70] M Barbut and B Monjardet. *Ordre et classification, Algebre et combinatoire, Tome 2*. Hachette, 1970.

[Bor86] J. Bordat. Calcul pratique du treillis de galois dune correspondance. *Mathématique, Informatique et Sciences Humaines*, 24(94):31–47, 1986.

[Bos10] Bianca Bosker. Twitter user statistics revealed. http://www.huffingtonpost.com/, 04 2010.

[CR93] C Carpineto and G Romano. Galois: An order-theoretic approach to conceptual clustering. In *Proc. Of the 10th Conference on Machine Learning, Amherst, MA, Kaufmann*, pages pp. 33–40, 1993.

[FFM+08] Ko Fujimura, Shigeru Fujimura, Tatsushi Matsubayashi, Takeshi Yamada, and Hidenori Okuda. Topigraphy: visualization for large-scale tag clouds. In *Proceeding of the 17th international conference on World Wide Web*, WWW '08, pages 1087–1088, New York, NY, USA, 2008. ACM.

[FR91] T.M.J. Fruchterman and E.M. Reingold. Graph drawing by force-directed placement. *Software - Practice and Experience*, 21(11):1129–1164, 1991.

[KL07] Owen Kaser and Daniel Lemire. Tag-cloud drawing: Algorithms for cloud visualization. In *WWW2007 Workshop on Tagging and Metadata for Social Information Organization,*, Banff, Alberta, May 2007.

[KO01] Sergei O. Kuznetsov and Sergei A. Obedkov. Comparing performance of algorithms for generating concept lattices. In *Concept Lattices-based Theory, Methods and Tools for Knowledge Discovery in Databases (CLKDD'01)*, Stanford, July 30, 2001., 2001.

[Kra10] Adam D. I. Kramer. An unobtrusive behavioral model of gross national happiness. In *Proceedings of the 2010 conference on Human Factors and Computing Systems (CHI 2010)*, 2010.

[MOS09] C. Macdonald, I. Ounis, and I. Soboroff. Overview of the trec 2009 blog track. In *NIST Special Publication 500-278: The Eighteenth Text REtrieval Conference Proceedings (TREC 2009)*, 2009.

[OBRS10] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, Washington, DC*, 2010.

[OdRMS06] I. Ounis, C. Macdonald M. de Rijke, G. Mishne, and I. Soboroff. Overview of the trec 2006 blog track. In *NIST Special Publication 500-272: The Fifteenth Text REtrieval Conference Proceedings (TREC 2006)*, volume 272, pages 17–31, 2006.

[R D10] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010. ISBN 3-900051-07-0.

[Raz10] Massimo Razzi. 28 novembre 2010 : un jour historique. http://www.courrierinternational.com/article/2010/11/29/28-novembre-2010-un-jour-historique, November 2010.

[ROK09] Joshua Ritterman, Miles Osborne, and Ewan Klein. Using prediction markets and twitter to predict a swine flu pandemic. In *1st International Workshop on Mining Social Media*, 2009.

[Wil80] R Wille. *Restructuring lattice theory, Ordered sets I*. Rival, 1980.

[Wil84] R Wille. Line diagrams of hierarchical concept systems. *Int. Classif.*, 11:77–86, 1984.