CEPIS UPGRADE is the European Journal for the Informatics Professional, published bi-monthly at <http://cepis.org/upgrade>

Publisher
CEPIS UPGRADE is published by CEPIS (Council of European Professional Informatics Societies, <http://www.cepis.org/>), in cooperation with the Spanish CEPIS society ATI (*Asociación de Técnicos de Informática,* <http://www.ati.es/>) and its journal Novática

CEPIS UPGRADE monographs are published jointly with Novática, that publishes them in Spanish (full version printed; summary, abstracts and some articles online)

CEPIS UPGRADE was created in October 2000 by CEPIS and was first published by Novática and INFORMATIK/INFORMATIQUE, bimonthly journal of SVI/FSI (Swiss Federation of Professional Informatics Societies)

CEPIS UPGRADE is the anchor point for UPENET (UPGRADE European NETwork), the network of CEPIS member societies' publications, that currently includes the following ones:
- inforewiew, magazine from the Serbian CEPIS society JISA
- Informatica, journal from the Slovenian CEPIS society SDI
- Informatik-Spektrum, journal published by Springer Verlag on behalf of the CEPIS societies GI, Germany, and SI, Switzerland
- ITNOW, magazine published by Oxford University Press on behalf of the British CEPIS society BCS
- Mondo Digitale, digital journal from the Italian CEPIS society AICA
- Novática, journal from the Spanish CEPIS society ATI
- OCG Journal, journal from the Austrian CEPIS society OCG
- Pliroforiki, journal from the Cyprus CEPIS society CCS
- Tölvumál, journal from the Icelandic CEPIS society ISIP

Cover page designed by Concha Arias-Pérez
"Upcoming Resolution" / © ATI 2011
Layout Design: François Louis Nicolet
Composition: Jorge Llácer-Gil de Ramales

Editorial correspondence: Llorenç Pagés-Casas <pages@ati.es>
Advertising correspondence: <info@cepis.org>

Subscriptions
If you wish to subscribe to CEPIS UPGRADE please send an email to info@cepis.org with 'Subscribe to UPGRADE' as the subject of the email or follow the link 'Subscribe to UPGRADE' at <http://www.cepis.org/upgrade>

ISSN 1684-5285

Monograph of next issue (October 2011)
"**Green ICT**"

(The full schedule of **CEPIS UP**GRADE is available at our website)

CEPIS UPGRADE
The European Journal for the Informatics Professional
http://cepis.org/upgrade
Vol. XII, issue No. 3, July 2011

Monograph
Business Intelligence
(published jointly with Novática*)
Guest Editors: *Jorge Fernández-González and Mouhib Alnoukari*

* This monograph will be also published in Spanish (full version printed; summary, abstracts, and some articles online) by **Novática**, journal of the Spanish CEPIS society ATI (*Asociación de Técnicos de Informática*) at <http://www.ati.es/novatica/>.

Social Networks for Business Intelligence

Marie-Aude Aufaure and Etienne Cuvelier

Online social networks have been closely studied from sociology at the middle of the 20th century to today's computer science but enterprise social networks are still in infancy. Social networks can improve enterprise organization as well as business applications. This article introduces enterprise social networks and associated use cases, graphs used to model these complex networks and how to analyze the content and structure. We also present a project we are working on to integrate internal and external social networks in a public administration.

Keywords: Business Applications, Complex Networks, Graphs, Mining and Aggregating Graphs, Social Networks.

1 Introduction

We have now entered the era of knowledge. Ubiquitous computing as well as the constant growth of data and information lead to new ways of interaction. Users manipulate unstructured data – documents, emails, social networks, contacts – as well as structured data. They also want more and more interactivity, flexibility and dynamicity. Users expect immediate feedback, and want to find rather than search for. All these evolutions induce challenging research topics for Business Intelligence, such as providing efficient mechanisms for a unified access and model to both structured and unstructured data.

Business Intelligence (BI) has historically been based on a combination of data warehousing, the process of storing historical data in a structure designed for efficient processing, and data mining, the mathematical and statistical methods necessary to transform this raw data into valuable information for making business decisions. The increasing flow of information, called Big Data, implies that BI can no longer afford to focus solely on historical records stored in tabular form. BI is moving to Business Intelligence 2.0, which combines BI with elements from both Web 2.0 (a focus on user empowerment, social networks, and community collaboration), and the Semantic Web, sometimes called "Web 3.0" (semantic integration through shared ontologies to enable easier exchange of data).

Social Networks are a part of this evolution and can be defined as a set of social entities, such as individuals or social organizations connected by links created during social interactions. They correspond to a new form of organization, called Enterprise 2.0, decentralized and more flexible, and viewed as more efficient than traditional hierarchical organizations. Historically, social networks have been first studied from a sociological point of view. Georg Simmel states that the foundation of sociology is defined by the relations and interactions between individuals, and not the individuals themselves. Networks are produced by these interactions. Jacob Moreno used surveys to build a set of social data, and searched for configurations appearing regularly in relations between individuals (analytical usage). Mark Granovetter [1] defined the theory of "power of weak links", these links being the most efficient ones in a profes-

Authors

Marie-Aude Aufaure obtained her PhD in Computer Science from the University of Paris 6, France, in 1992, and her HDR from the University Claude Bernard Lyon 1, France, in 2002. From 1993 to 2001, she was associated-professor at the University of Lyon; then, she has integrated a French Research Centre in Computer Science (INRIA) during two years. She was professor at Supelec from 2003 to 2008. Now, she is full professor at *Ecole Centrale Paris* (MAS Laboratory) and head of the SAP Business Objects Chair in Business Intelligence. She is also scientific partner at INRIA since 2003. Her research interests deals with the analysis, retrieval and querying of unstructured data, and the combination of structured and unstructured data from a Business Intelligence perspective. The scientific topics developed in her team are related to semantic technologies, graphs, conceptual classification with a user-centric point of view and are applied to semantic information retrieval, question and answering over data warehouses, social networks and recommender systems (special focus on user modeling and personalization). She is reviewer for many journals and conferences and has deeply published in the fields of semantic technologies, data mining and databases in international journals, books and conferences. <marie-aude.aufaure@ecp.fr>

Etienne Cuvelier holds his PhD in Computer Science from the Facultés *Universitaires Notre-Dame de la Paix of Namur*, Belgium, since 2009. Prior to this he also obtained a MSc in Mathematics and a MSc in Computer Science and taught mathematics and computer science during several years before integrating the research world. He is now postdoctoral researcher at *Ecole Centrale Paris*, MAS Laboratory, in the Business Intelligence team. His work is related to the data mining and the social network analysis, and more specifically in functional data analysis, symbolic data analysis, conceptual analysis and graph mining. <etienne.cuvelier@ecp.fr>

" Enterprise social networks are still in infancy but they can improve enterprise organization as well as business applications "

> "Public Administrations also need social networks as an interface so that themselves and citizens can easily understand who does what and who says what"

sional context. Milgram [2] established in 1967 the theory of six degrees of separation related to the small-world phenomenon, which is the hypothesis that the chain of social acquaintances required to connect one arbitrary person to another arbitrary person anywhere in the world is generally short.

As with other networks, social networks can be represented as a graph: a set of nodes linked by edges representing any kind of relationship between nodes. Social networks have been deeply studied for the web (online social networks), but they are also a key element in enterprises. These enterprise social networks are more complex than online social networks like Facebook, LinkedIn or Twitter, because they do not only connect people but also all the objects manipulated in an enterprise like projects, products, etc. In the enterprise context, two main reasons lead to a strong interest for social networks: (1) a technology for internal communication, collaboration and information sharing, and (2) a technology for communicating to clients, citizens, etc. Recent studies (for example those by Singh and by Nieto et al.) have shown a growing interest in transversal collaboration leading to increase the value of collaborators' competencies, and to facilitate innovation and productivity. This form of organization is generally not easy to introduce because of the organizational structure which favors intra-team and hierarchical exchanges and also the fact that sites can be geographically distributed. Social networks are adapted to build virtual communities based on common interests for example, and can be seen as a potential solution to the above mentioned problems. According to a recent study done by Coleman Parkes Research in 2008 among 500 companies around the world, only 7% of these companies use a social network and 4% are integrating a social network, while 59% declare having no strategy for having an internal social network. This study also outlined that 60% of these companies admit that social networks will be a fundamental tool in the future for collaboration in the enterprise. The main obstacles to adopting this technology are related to the security issues (76%), the inaction of the direction (57%) and the hesitation to exploit new technologies (58%).

Why are social networks of interest for companies? This is a technology for (1) internal communication, information sharing and collaboration, (2) information communication towards clients, (3) watching the gossip about the company (e-reputation, opinion mining) and (4) creating collective intelligence.

Social networks can be internal (model of the organization, social interactions between employees, etc.) and external (like Twitter than can be used for e-reputation).

In the next sections, we will define use cases social networks in an enterprise context and also for public administrations, the variety of graphs that can be used to model these networks, and how is it possible to analyze social networks.

2 Social Networks Use Cases

This section shows how social networks can be used in public administrations and in enterprises. Public administrations and enterprises have similar problems that can be partly addressed by the use of social networks. They need to manage the organization and to deliver services to citizens. Such services cover various domains such as highways maintenance, urban planning, assistance to persons, etc. Public administrations need social networks as an enterprise to analyze their internal networks (projects, organization etc.) and to analyze their external networks (suppliers, clients, partners). They also need social networks as an interface so that citizens and the Administration can easily understand who does what and who says what. We are implementing such scenarios in a project we have with a public administration (ARSA project with the city of Antibes in France). We are working in this project on internal and external social networks. The internal social network can be seen as an extension of the intranet and the main objective is to enhance transversal collaboration. The need expressed is to visualize and to share information and data, to model and visualize collaborative project and to navigate through the hierarchical structure of the organization. The objective of the external social network is to monitor what is said by citizens about the city through Twitter. The idea is to be able to visualize in real time citizens opinions and to give immediate feedback about the actions done by the administration. We will show in the social network analysis section how we monitor Twitter for defining e-reputation in real time.

Scenarios for enterprises can be organized around the centrality [3], importance and influence of actors, the identifications of groups and the identification of key actors, from a human resource, management and individual's perspective. Being central is: be the source or destination of numerous relations (degree centrality), be close to many actors (closeness), be central for many connections (betweeness). Many *scenarios* for using social networks are useful for human resources. Finding persons having the biggest influence can help to transmit good practices and improve social aspects. Identifying the most central groups helps in finding groups with a good communication, which are important elements of cohesion. The similarity between groups can be computed in order to analyze groups with "good" properties, apply observation on a group

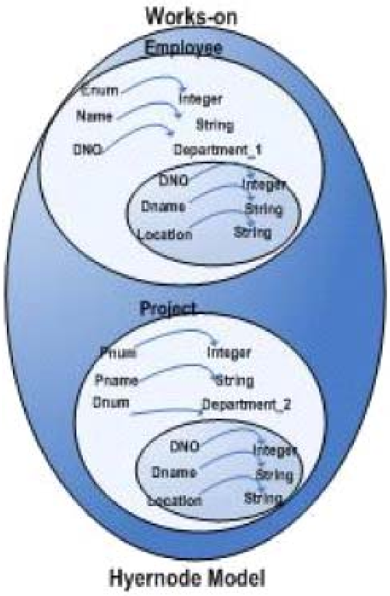

Figure 1: Example of an Hypernode Schema.

to a similar one or explain a behavior within a structure. We can also search for atypical elements, like persons who are not well integrated, and find the best group for such persons. One other application is to be able to constitute efficient teams, in order to enable innovation and to reach group cohesion. From an individual perspective, such tools can help for networking, to identify for example important relations by studying relations of other persons. Individuals can create a collaborative social network, learn from the experience of others, find an expert in the company or build a profile for finding a new job.

Applications related to this technology become more and more important, as well as the size of the resulting social networks. We need methods to be able to model (using graphs) such complex networks, to analyze and query these networks and to be able to visualize results in an efficient and intuitive way.

3 Variety of Graphs

A wide variety of graphs can be used to model social networks, from the basic mathematical definition to more complex variations. A graph $G = (V, E)$ consists of a set V of vertices (also called nodes), a set E of edges where $E \subseteq V \times V$. This definition refers to simple and undirected graphs. The term *multigraph* is generally understood to mean that multiple edges and loops are allowed. A *graph labeling* is the assignment of labels, traditionally represented by integers, to the edges or vertices, or both, of a graph. Labeling is applied to finite undirected simple graphs. A graph can also contain more information by adding attributes to nodes or edges. Such a graph is termed an *attributed graph*. Simple graphs are not sufficient to model heterogeneous graphs based on complex objects, having multiple attributes and relations. Then, the basic structure of a graph (nodes and edges) is complemented with the use of hypernodes and hypergraphs extensions that provide support for nested structures. A *hypergraph G* is a tuple (V, E, μ), where V is a finite set of nodes, E is a finite set of edges, $\mu : E \rightarrow V^*$ is a connection function where V^* means multiple nodes (an edge can connect any number of nodes). An Hypernode can encapsulate nested nodes (see Figure 1).

A variety of graph database models [4] has also been defined. All these models have their formal foundation as variations of the basic mathematical definition of a graph. The main characteristics of these models is that the schema and instance levels are distinct, like in RDF graphs, and can easily be used to model business applications. Another characteristic is that these models encapsulate the semantic in the nodes and edges.

When graphs of extracted social networks are large, effective graph aggregation [5][6] and visualization methods are helpful for the user to understand the underlying information and structure. Graph aggregations produce small and understandable summaries and can highlight communities in the network, which greatly facilitates the interpretation. Graph aggregation differs from other methods such as graph mining, based on the graph structure (edges) and graph clustering which groups similar nodes. The aggregation is a summary based on nodes content and neighborhood.

4 Social Networks Analysis

If Social Network Analysis [7] was firstly developed by sociologists, the rise of computer science and one of its fields, data analysis, enables analysis of graphs of big sizes, and permits the development of several interesting but computationally intensive tasks.

One of these is the community detection task [8]. If the notion of community seems very natural for most people, it demands a formal, and most of the time measurable, definition to be then computed. And even if there exist many definitions of what is a community in a social network, we can say, without lack of generality, that a community is a group of individuals with high concentrations of relationships within the group and low concentrations of relation-

“ As with other networks, social networks can be represented as a graph: a set of nodes linked by edges representing any kind of relationship between nodes ” with individuals outside the group. The detection of such groups in a social network can be performed using several kinds of algorithms. The choice of the algorithm depends on many criteria, but a decisive one is often the "size" of the network (the number of nodes and the number of links), because the larger the network, the more expensive in computing time, is its execution. For the sake of illustration we give here a sketch of the ideas of some of these algorithms. The "cheapest" algorithm family is the partitional ones which try to find a partition of the individuals in *a priori* given number of clusters equal to k. The "best" partition is searched using jointly, most of the time, a distance measure and a quality criterion of the found partition based on the sum of the distances of individuals to the centre of the cluster. The most popular partitional algorithm (with several variants) is the k-means clustering. Another interesting type of algorithm is the family of hierarchical clustering algorithms which are divided into two types, depending on whether the partition is refined or coarsened during each iteration: agglomerative algorithms start with a set of small initial clusters and iteratively merge these clusters into larger ones, while divisive algorithms start with all the network as one big group, and then split the dataset iteratively or recursively into smaller and smaller clusters. At each step, the clustering algorithm musts select the clusters to merge or split by optimizing a quality criterion. Several other algorithms exist.

From a business intelligence point of view, as in the case of tabular data, finding "homogeneous" groups in networks allows the study of each of the found groups in order to know their characteristics, which can be valuable knowledge in a customer relationship framework for example. From a media diffusion point of view, finding such groups can be also very interesting, not only to study the characteristics of these groups, but also to find two types of majors actors in the diffusion process: actors with the most connections with other people and actors which link two or more communities. The first type of actor, those with the highest degree of centrality, can be seen as the most popular persons in their groups - thus they became preferential and economical targets for marketing actions dedicated to adoption of new products, for propagation of information and advices, and also for monitoring opinions and mood of customers. The second type of actor, those linking communities, and which have the highest betweeness, are key actors for the spreading of information through a network because they permit the transfer of information from one community to another and they are switches, which can be used in positive or negative ways, if we wish to favor diffusion or not. This latter type of actor can be easily retrieved using, for the sake of illustration, the algorithm hierarchical divisive algorithm of Girvan and Newman [7] which attempts, successfully, to detect communities in finding the bridges between the communities. Indeed, this algorithm also permits the detection of these path between communities. Figure 2 gives a picture of three communities in the set of the Facebook friends of one of the authors of this paper. These communities, detected using the algorithm of Girvan and Newman, are completely meaningful, according to the prior knowledge of the author, because we retrieve the community of its family (nodes colored in blue), the communities of the ex-students of two universities where he taught (in red and green). And we see clearly the "bridge actors" in this figure.

> "A wide variety of graphs can be used to model social networks, from the basic mathematical definition to more complex variations"

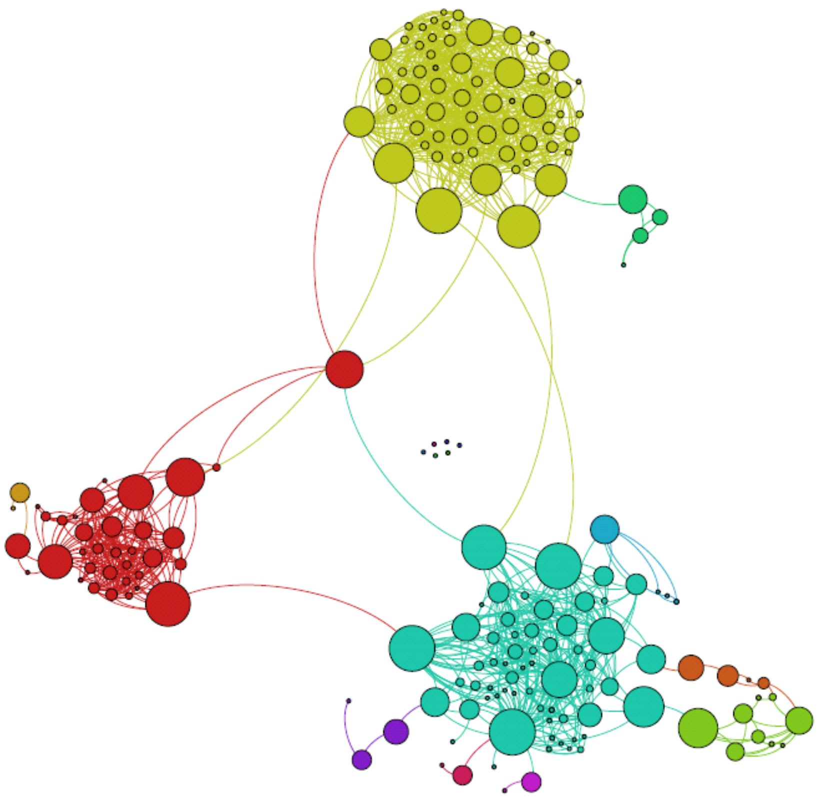

Figure 2: Communities in a Real Facebook Profile.

Detect and find communities in social networks is not the only interesting task. As already explained in the "social networks use cases" section, it could also be very valuable to study what kinds

> “ The notion of e-reputation arises: ‘what is the standing of me or my society, right now on the Net?’ ”

of information cross a given network. Nowadays most online social networks are used for sharing information, mood and advice. However, very quickly following the massive adoption of such networks and practices, there arose the notion of e-reputation: "what is the standing of me or my society, right now on the Net?" Even if this e-reputation or branding is something to be built patiently, day after day, as the permanent result of an active presence on the web and social networks, any bad buzz can very quickly ruin these efforts if there is not a rapid detection and adequate reaction to such phenomenon. Efficient tools to monitor their own e-reputation become urgently needed by enterprises or official institutions in this high-speed interconnected world. A lot of such new services appears every month on the web, but mostly they use simple queries and classical statistics and then don’t give a global summary view. In [9] we have proposed a prototype of an e-reputation monitoring tool used on the Twitter network. This latter social network is one of the most used to share information, has a large audience (200 million users in April 2011,[10]), is fast growing (460,000 new accounts per day in average during February 2011, [11]) and manages a huge quantity of exchanged information, called tweets (140 millions tweets sent per day, [10]). Each piece of generated information can be forwarded, but can also be edited before the forwarding process, it then becomes a real challenge to trace the path and transformations of a "successful" tweet (i.e. a buzz).

This challenge is however crucial for an e-reputation tool. Our prototype proposes to retrieve all the groups of words that are the most forwarded on a given subject. This tool called EVARIST is developed in the framework of the ARSA project mentioned above in collaboration with the French city of Antibes. EVARIST is based on a mathematical tool called Galois lattice: briefly, if we have a set of objects and a set of attributes, with a Boolean table (called context table) such a value TRUE is given at cell (j, k) if the object j owns the property described by the attribute k, then the Galois lattice is the structure which gives all the subsets of object and attributes (called concepts) such all the objects share the same attributes. The graphical representa-

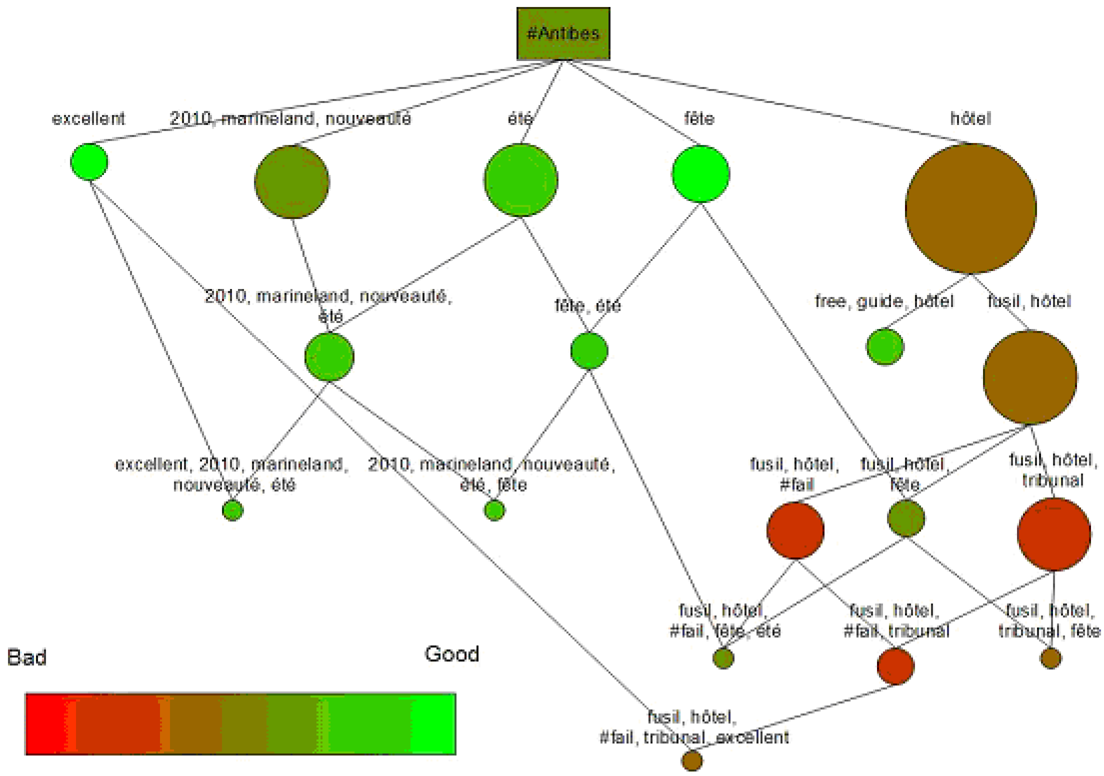

Figure 3: The EVARIST Prototype.

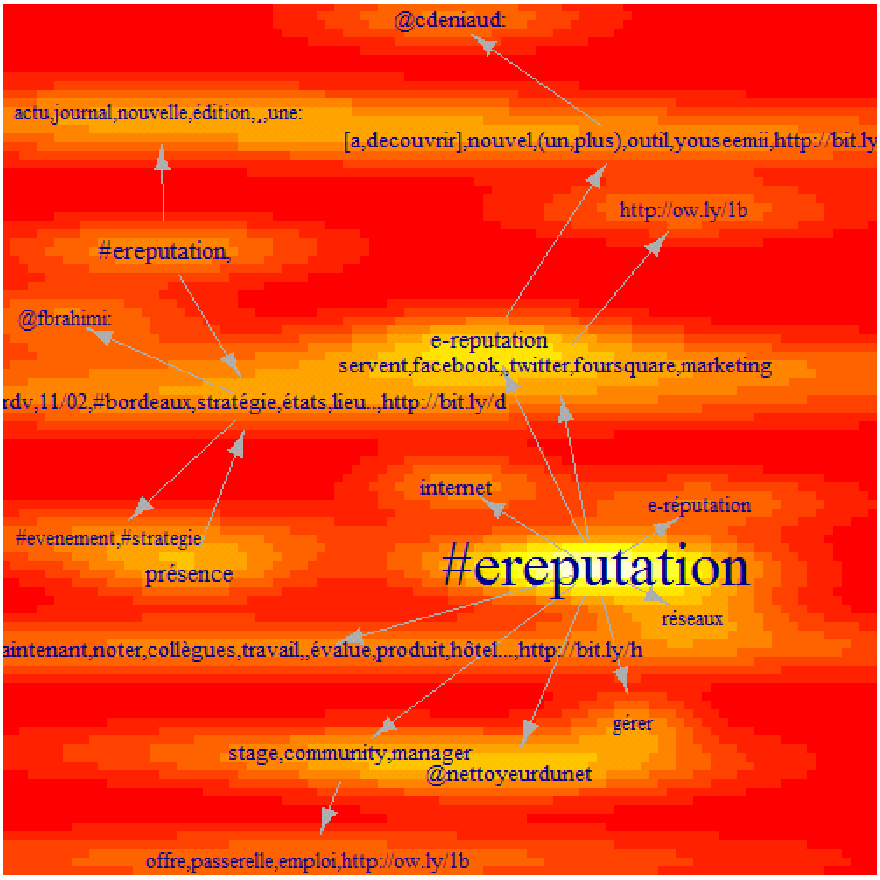

Figure 4: An EVARIST Visualization on e-Reputation.

tion of a Galois lattice is made using an Hasse diagram which show all the concepts and their inclusions.

Figure 3 shows a Galois lattice produced by EVARIST. This tool use the following steps:

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 and give to each concept a size proportional to the number of tweets owned;

5. Visualization of the results.

In Figure 3, the result of an execution of this tool for monitoring what is told on twitter about the French town is shown. In this figure, starting from the top we can see how the group of tweets can be subdivided in subgroups, and which words are shared by all the tweets of the subgroup. The size of the concept being proportional to the portion of tweets contained in the subgroup, the figure permits retrieval of the most used words in the buzzed tweets about Antibes,. A sentiment analysis process is applied to each tweet in order to determine if this latter gives a positive advice on Antibes or not. Given the color of the concept, the community manager can have also a quick look about any possible bad buzz for his town.

Figure 3 demonstrates that it can quickly become difficult to display all the concepts proportionally to their sizes, and at the same time display clearly all their attributes if the number of tweets and words increases. To reduce the number of attributes and concepts to be displayed, we can select only the concepts with a relative size greater than chosen threshold. In other words, in respect to the notion of buzz we can select the concepts with the more tweeted words. Finally, to reinforce a reading going from the most

“ Efficient tools to monitor their own e-reputation become urgently needed by enterprises or official institutions. We have developed the EVARIST tool to this end ” general to the most particular, we have proposed to add a topographic allegory called "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. Such a "topigraphic" map is shown in Figure 4, but without the sentiment analysis result.

5 Conclusion

Integrating social networks in enterprises and public administrations is of real interest, but somehow difficult to implement. Two implementation solutions can be considered. The first is to use an existing social network, like LinkedIn for example, and create a new group. This solution is very easy to implement but has several drawbacks: you have no access to the metrics (i.e. evolution of the number of participants), you do not own data (privacy issue), you cannot link the social network to knowledge management tools and you are dependant on a business model that is not yours!. The other solution is to use a commercial or an open source tool. In that case, you have the total control of metrics, data, content published and you can also link the tool to existing social networks. The drawback is that the implementation is not always easy. In our project with the city of Antibes, we worked with SNA[1] (Social Network Analyzer), a tool developed by SAP. The idea was to allow a user to interact from an external social network to the internal one with respect to the access rights, and conversely, to allow the public administration to be aware of the information spread over the external social networks (Twitter in this project), information that can be useful for the city.

References

[1] M. Granovetter. "Introduction for the French Reader," *Sociologica* 2: 1–8, 2007; B. Wellman. "Structural Analysis: From Method and Metaphor to Theory and Substance", Pp. 19-61 in *Social Structures: A Network Approach*, edited by Barry Wellman and S.D. Berkowitz. Cambridge: Cambridge University Press, 1988.

[2] S. Milgram,. «The Small World Problem». *Psychology Today*, 1(1), May 1967. pp 60 – 67.

[3] L.C. Freeman. A set of measures of centrality based upon betweenness. Sociometry, 40 35–41, 1977.

[4] R. Angles, C. Gutierrez. Survey of graph database models. ACM Comput. Surv. 40 (2008) 1-39.

[5] Y. Tian, R.A. Hankins, J. M. Patel. Efficient aggregation for graph summarization. In SIGMOD ’08 : Proceedings of the 2008 ACM SIGMOD international conference on Management of data, New York, NY, USA, (2008) 567–580.

[6] R. Soussi, E. Cuvelier, M.-A. Aufaure, A. Louati, and Y. Lechevallier. DB2SNA: an All-in-one Tool for Extraction and Aggregation of underlying Social Networks from Relational Databases, 2011. In: The influence of technology on social network analysis and Mining, Tansel Ozyer *et al.* (eds.), Springer, to appear.

[7] J. Scott. *Social Network Analysis*. Sage, 2000.

[8] M. Girvan and M.E.J.Newman. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA*, 99:7821–7826, 2002.

[9] E. Cuvelier and M.-A. Aufaure. A Buzz and E-Reputation Monitoring Tool for Twitter based on Galois Lattices, the 19th International Conference on Conceptual Structures (ICCS 2011), 25-29 july 2011, University of Derby, United Kingdom.

[10] B. Bosker. Twitter: We Now Have Over 200 Million Accounts (UPDATE), Huffpost Tech, 04/28/11, <http://www.huffingtonpost.com/2011/04/28/twitter-number-of-users_n_855177.html>.

[11] Twitter, #numbers, <http://blog.twitter.com/2011/03/numbers.html>.

[1] A demo is available at <http://sna-demo.ondemand.com/>.

