MAPPING PERCEPTIONS OF RISK WITH TEXT-MINING METHOD USES OF RFID CASE

Alexandre DELANOË(*), Laura DRAETTA(*)

alexandre.delanoe@telecom-paristech.fr, laura.draetta@telecom-paristech.fr

(*) TELECOM ParisTech, LTCI/CNRS, Département Sciences Economiques et Sociales

Pôle Usages-Sophia Campus Eurecom - 2229 route des Crêtes - BP 193 - 06904 Sophia Antipolis

24 mars 2009

Mots clefs :
Sociologie, perception sociale, presse grand public, statistiques lexicales, Tetralogie, RFID.
Keywords:
Sociology, social perception, news press, text mining, Tetralogie, RFID.

Résumé

Cet article traite des aspects méthodologiques d'un projet de recherche sociologique identifiant les sources potentielles de controverse liée à la technologie RFID (Radio Fréquence IDentification). Il propose de montrer les difficultés inhérentes aux statistiques lexicales lorsqu'on essaie d'expliciter les termes du débat autour de la technologie et de ses usages. Une analyse de la presse écrite francophone et anglophone a été réalisée à l'aide du logiciel de veille Tetralogie. Cette recherche permet de retracer l'évolution des représentations de la RFID dans la presse écrite francophone, de la fin des années 1990 à nos jours.

Abstract

This paper deals with the methodological aspects of a sociological project which aims at identifying potential sources of controversy related to the use of RFID (Radio Frequency IDentification). It intends to throw light on the debate, by focusing on various discussions that feed it and have generated it. We have made an analysis of

English and French speaking news media writings, using Tetralogie. This research helps to trace the evolution of representations associated to RFID in French print media from the latevvqjq 1990s till now.

1 Methodology

The issue of collective perception of risk related to the production and use of technologies for radio frequency identification is diverse and has not been institutionalized or stabilized yet. As a matter of fact, the terms of the debate around RFID are still being structured and actors are still emerging. With the aim of identifying sources of controversy related to RFID technology, the first step has been to identify the perceptions of different actors involved in the emergence of public debate and who have been contributing to it.

The initial project, RISC-Definition[1], proceeds along two identified axes. They correspond to two distinct but complementary exploratory methodological approaches : a statistical approach and a qualitative and ethnographic approach. The results were obtained through these two heuristic approaches. However, this paper presents only the first one, that is the lexical and statistical analysis we have made on a corpus of press articles.

This statistical study was conducted using Tétralogie, which is a software which is not (yet) widely used by sociologists (even if it enables to study innovation and network dynamics in industry [6]). It would therefore be useful to justify our choice to resort to this tool by specifying our methodology.

1.1 Pragmatic and reflexive sociology as first step

Presenting the discipline of reflexive and pragmatic sociology, F. Chateauraynaud and D. Torny [3] describe an approach examining the way in which the context around a controversy arises, i.e the area of mobilization in which a discourse or a text is taking shape. Following this approach, some research questions are raised: what are the qualities attributed to the enunciator of a specific view on a given issue? Does the same argument rest on different supports that cross each other? Does the author jump in the arena and argue with the actors? What ends the dispute? What contributes to its revival? What kind of discourse or texts represent more an institutional act? Inspiring from pragmatic sociology, we apply the same type of dynamic questioning to the case of RFID, which (first) appears to be a complex one.

A complex affair is usually characterized by the presence of various actors, a large number of arguments, spreading out over a long period of time. In effect, for D. Torny, achieving completeness is difficult, representativeness is unthinkable, and the case does not necessarily end at a given time. Regarding our research on RFID, the characteristics of complexity appear, even if this case is a new one and the debate is just starting. Since this topic has been treated both by the press and on the Internet, the actors appear to be varied and numerous, making completeness hardly attainable. Indeed, the varied nature of the different sources makes it difficult for the statistical analysis to be representative of the actual debate. In addition, as we have said earlier, the case is still open and very likely far from ending. The long term characteristic appears to be the most probable feature concerning the case of the ‘RFID debate’.

Pragmatic and reflexive sociology is based on “literary technology”, which allows a sharing of knowledge among different partners, therefore aiming at the continuity of scientific research programmes. The authors we mentioned earlier use the Prospero tool, while others use software such as Alceste [7] or Tropes which has already been used in radio-frequency field [5]. But the software tool influences sociological approach [4]. Then for our research, we opted for the Tétralogie concept developed by IRIT (Institut de Recherche en Informatique de Toulouse) [2], for three reasons.. First, the nature of our research is exploratory. We did not want to *a priori* integrate a

[1] Sociological project funded by the french foundation “Santé et Radiofréquences” and by the Institut Télécom Chair “TIC & Développement Durable”.

[2] We mention this concept because the investigation is not limited to the results of the software. Our own programs (written in bash, Perl and, more generally, the tools made available by the Debian community) were used to prepare the corpus and perform successive validations.

semantic dictionary so as to avoid statistical bias. Then, the functionalities of Tétralogie allows a text search and a bibliometric statistical exploration of materials, taking into account the specific fields in the body of data, which facilitates the development of thesauri. The last but not least reason has to do with the size of our corpus, which is particularly significant (several megabytes of data): this requires the use of a stable and powerful tool, which Tétralogie definitely is. But we are very conscious of the research potentialities of the other tools. We believe that the use of the software suite - Prospero, Chelone, Tiresias - is a relevant complement to this first approach, since the integration of advanced tools would allow lemmatization, conducting collaborative ontologies and looking for updates on the web. We will most probably be using them in the future for a more detailed research work.

The tool Tétralogie is a kind of lens through which a body of text can be analyzed. The construction of the analysis is an upstream work which is both fundamental and characteristic. By "body" we mean here all the texts relating to our research. The definition of what is a "good body" of text is actually the fundamental problem for the quantitative analysis of qualitative data. Several approaches are possible. One approach is the historical method: it refers to a class of records, which have not been motivated by the analyst's choice. This is the approach we have adopted here. We use the FACTIVA database which is well known for its exhaustive coverage of the contemporary press, and whose archives go back to more than 20 years ago for English articles and more than 15 years for French publications. Then, as a second step, the Boolean equation used to search the database requires a careful choice of keywords in order to avoid targeting specific items and to gradually enrich the vocabulary for our object of study. We specify our choice of terminology in the next chapter. Finally, the body of texts is enriched using different sources. This is what we plan to achieve after this first exploration.

A good corpus of texts is also defined in terms of the type and quality of the sociological survey that it allows. In effect, it is necessary to have a good representation of the issue at hand and of its critical events. For example, if a large number of actors refer to an event, a text or an entity, these must be present in the corpus. Also, we should be able to follow the actors in their own interpretation of the events. For this purpose, we assume that the study of the press is a good starting point. Extraction from a database allows a more nuanced and fine understanding of the complex matter. Indeed, we integrate all types of discourse, ranging from warnings about the dangers of RFID to economic hope about the technological advances. We observe how the text is reflected in the analysis. We do not limit ourselves to expressions of prophecy of misfortune or warnings; we also consider the story of economic hope, problematizing about whether the two end positions are related in a dialectic relationship or not.

1.2 Research equation

The goal of this research is to capture the social perceptions of RFID and the risks related to this technology as well as its uses, which are referred to in public discourse. Thus, we analyzed the discourses from a corpus of English and French speaking press, with the aim to identify and map the terms of the debate and the actors involved.

This corpus of articles has been extracted from the FACTIVA database, which consists of daily, weekly and monthly news. We analytically built our own research equation (below) to select articles relevant for our research object relating to RFID from the database. The research request in FACTIVA allowed us to isolate two bodies, one involving English articles published between 1990 and 2008, the other gathering francophone articles published between 1998 and 2008.

For our analysis of the statistics literature, we needed to extract all the texts (selected by our equations research) from the FACTIVA database, both anglophone and francophone. Three equations have enabled search by gradual reduction, so as to obtain a satisfactory corpus for the problems we had identified and that we have presented in the introduction. Thus, the equation for synthesis of the whole process is defined as follows:

- *("radio-frequency" or "radio frequency" or "radiofrequency" or "radio fréquence*" or "radiofréquence*" or "radio-fréquence*" or RFID or IDRF or IDFR)*
- *and (risque* or risk* or santé or health or pollution* or caddie* or implantation*)*
- *and (NFC or chip* or tag* or "ID card*" or etiquette* or identification* or contactless or card* or wireless* or traceability or RFID or puce* or "sans fil")*

- *not ("International Development and Relief Foundation" or "radio frequence paris plurielle" or "radio fréquence jura" or "radio fréquence nimes" or "radio fréquence gaie" or "radio fréquence gay" or "National Finance Commission" or "National Facilitation Center" or "Not From Concentrate" or "National Factoring Company" or "New Freedom Commission" or "Al-Rubban-NFC" or "National Fiber Council")*

This equation represents our path search from interrogation on the social perception of risks associated with RFID, but also a more specific question about RFID as a possible revisiting of the health risks associated with radio frequencies.

2 Historical description

As a first step, we conducted a structural analysis of two bodies, one English, one French, to reflect their characteristics and to produce aggregate statistics on the number of articles published, their distribution by year and country of publication. Then, we focused on the contents of the French corpus to reflect the discourse conveyed and to identify the terms of (francophone) public debate.

The statistics we obtained from this database search have led us to determine and construct three categories: a first one on the social perceptions of risks related to the use of RFID, a second on the actors from whom these perceptions originate, and finally, a third on the media sources which published the articles. The categories have then been classified according to the representations that have been identified (divided by name and press sector).

This result 1 shows the distribution, by year, of English and French articles dealing with RFID published between 1991 and 2008. In the early 90's, RFID was mainly used for anti-theft applications, traceability (of animals) and identification (toll collection). Gradually, the use of RFID spread to other sectors such as the logistics industry and healthcare. During this evolution, the quantity of items has grown after 2000 and issues that were raised ranged from the economic interests of new applications to criticism of misuse. The French interest 1 for RFID is confirmed and is relative if compared to the anglophone corpus.

The graph 2 shows the different proportions of articles per media group in the francophone corpus. From a statistical point of view, this is reasonably flat. The names of these newspapers have been integrated with a thesaurus in order to group them into categories of publications. Thus, the daily press (Le Monde, Le Figaro, Liberation, but also 24 hours and many others) falls into the category “publications and general policies”. The category “publications in Finance” includes publications like l'AGEFI or La Tribune. The “publications specializing in high technology” includes 01 IT, 01DSI, Decision Science & networks, etc.. The “Publications of the European Union” refer to EurLex, Europe or Europolitique Information. As for the “global flow of information” category, it includes magazines such as Business Wire, AP French Worldstream, Factiva Press Release Service, Canada Newswire, etc. Finally, the “various publications” category includes materials other than specialized Hi-tech and finance press (chemical, automotive, media-centric). It would be interesting to note the diversity of approaches and fields producing media representations related to RFID. However, we do not achieve a sociology of media – and that is not the goal either - but try to understand the evolution in the media's treatment of RFID from perspective of the actors who contribute to the debate.

3 Actors involved

Tri-cross on the French body of texts with a filter by “actors” is presented in the figure 3. The y-axis represents the amount of hits per year in the French corpus. The (importance, size…) order of the actors on each vertical bar follows the order of the legend. The chart identifies the year 2003 as the launching of the debate surrounding the Wal-Mart company. This world leader in mass retail uses RFID for logistical purposes. The debate was first opened by the intervention of Benetton and Gillette, with their attempts to introduce RFID in their products for respectively profiling customers and preventing thefts. Even if Wal-Mart, Benetton and Gillette appear constantly

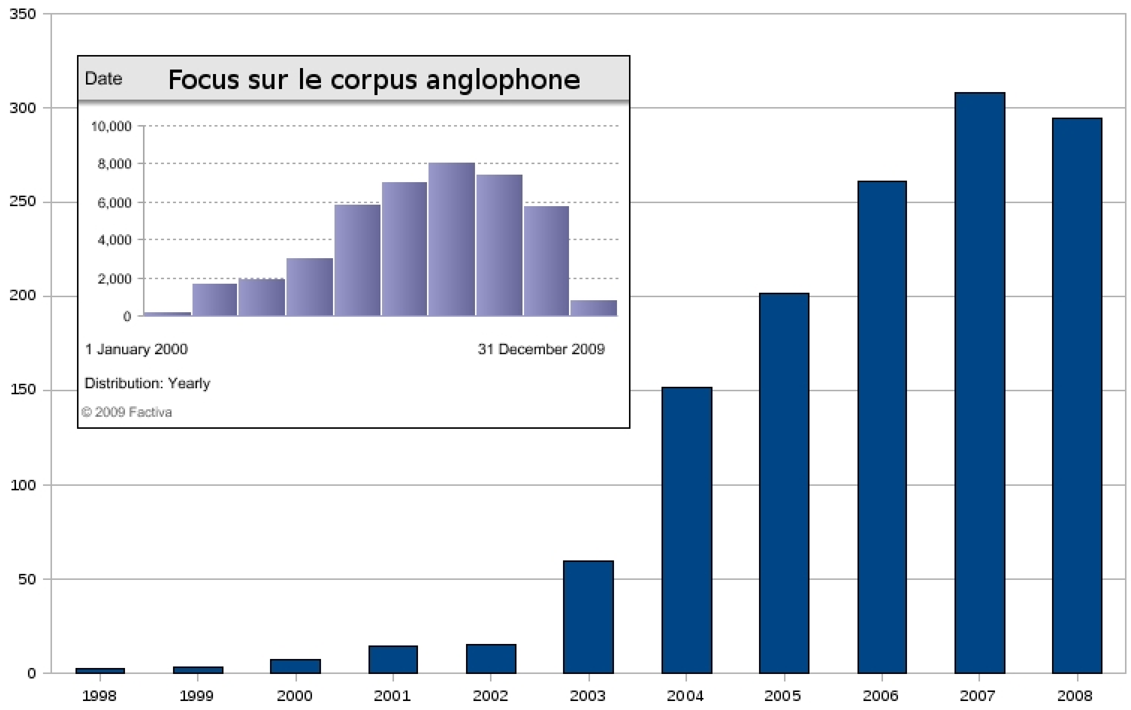

Figure 1: Historical and comparative study of anglophone and francophone corpus (n = 1362)

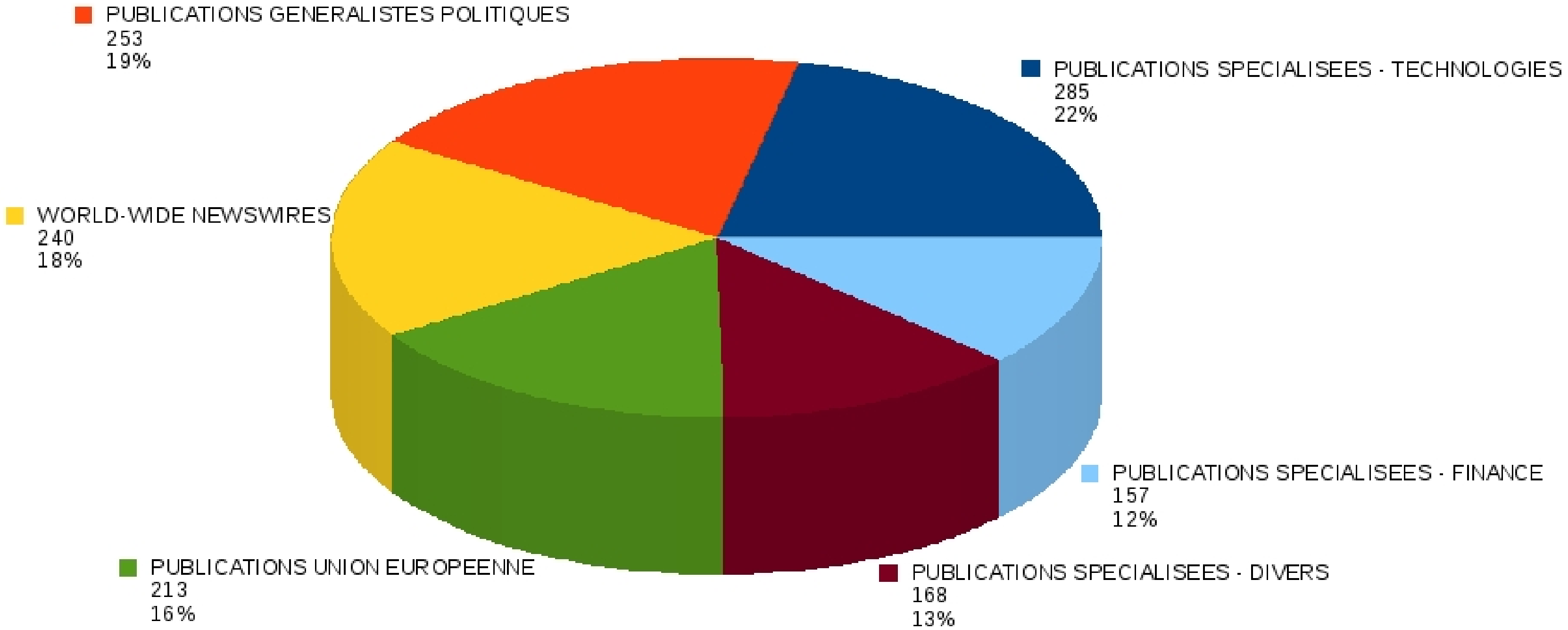

Figure 2: Quantity of articles in French corpus by media group support.

between 2003 and 2005 in the results of the statistical database, it is only from 2006 onwards that the actors - who affect regulation and critics - began to feed the discussion.

Indeed, between 2003 and 2008, the attention of the press to the RFID issue appears to move progressively from the question of the violation of consumer privacy to regulatory issues. Parliament (European) and the CNIL are cited in the texts for their regulatory actions. Concerning the stakeholders in review, on the one hand, the association CASPIAN [2] emerges, especially in 2005, with its call to boycott the "spy chip". On the other hand, the cashiers and the risk that they lose their job is one issue under constant debate.

In recent years, the category of actors on the critical side is enriched by the presence of Michel Alberganti [1] (in light purple in the graph), who published a book entitled "Under chip's eyes. RFID and Democracy (2007)". In 2008, the french collective "pièce et main d'oeuvre" made its first appearance in the debate. We note finally that the company Verichip, which markets subcutaneous RFID implants, is present throughout the body, with an increase in 2007; this increase seems to correspond to the discussion of implants in humans for personal identification and payment (in nightclubs for example).

4 Risks

The controversy appears to encourage public debate, whose terms have different levels of importance and are moving chronologically. We searched for these terms with how they weigh in the debate and their evolution, by examining the actors that appear in the public arena of RFID (promoters, regulators and stakeholders in the review) as described before. Two major subjects are emerging: the first summarizes the terminology that is shared in common by the industry and regulatory organizations, the second one is rather an issue that is specific to users, and especially those involved in the review, who are positioned as whistleblowers.

Thus, on the side of the industry and regulatory organizations, three major groups of terms appear in a regular way. The first group expresses the hope that RFID is permitted in economic terms, referring to terms such as: growth, business, inventory management. All the semantic range of "Internet of things" forms part of this category. A second group of terms - such as investment costs and return on investment - refers to the idea of economic risk. Finally, a third group reflects the discourse on the regulation via another terminology that fits the standards, their identifiers and the European Parliament.

The side of users and their spokespersons is represented by the "privacy" category which encompases all the threats relating to privacy and individual freedom. The terminology used comprises terms such as "spy", "geolocation", "automatic identification", "personal identification", "passport", "Navigo", "intrusion", "invasion", and so on. A second terminology refers to the socio professional risk with regard to cashiers and unions, and a third group of terms referring to health risk emerges. This risk is raised mainly concerning electromagnetic effects of RFID bracelets or implants under the skin.

Finally, we wanted to isolate the question of approximation of RFID with nanotechnology and clarify this point, which does not seem clear cut for the actors themselves. In the corpus (French), 85 items refer to the nano world mentioning, along with RFID, terms such as "nanoscience", "nano" "nano", "nano-object", "nanotubes" or "nanosystems". This discourse on the proximity (mixing up) of RFID and nanotechnology evokes a subsequent fear of RFID. At this stage of our research, this fear cannot be associated with a category of actors in particular. The articles discussed above and that appear in our database cannot be linked in our research and we have no means of knowing whether they actually are linked, since journalists or authors of these articles do not quote their sources. We also know that this issue is raised on the Internet by some stakeholders in the review.

The graph 4 shows the amount of items at risk category / term and year debate in the French corpus. The interpretation of the histogram should be clarified and relativized. The counting statistic is notably improved from the thesaurus which is supposed to be fed continuously. Furthermore, the concept of threshold of perception, amounting to 50 articles, is an arbitrary set used only to indicate the level of importance or seriousness at which some terms have to be taken into account in the debate.

If RFID appears in the written press in the late 1990's, it was not until the year 2000's that it spread in the media at large. The first debate to cross this threshold of 50

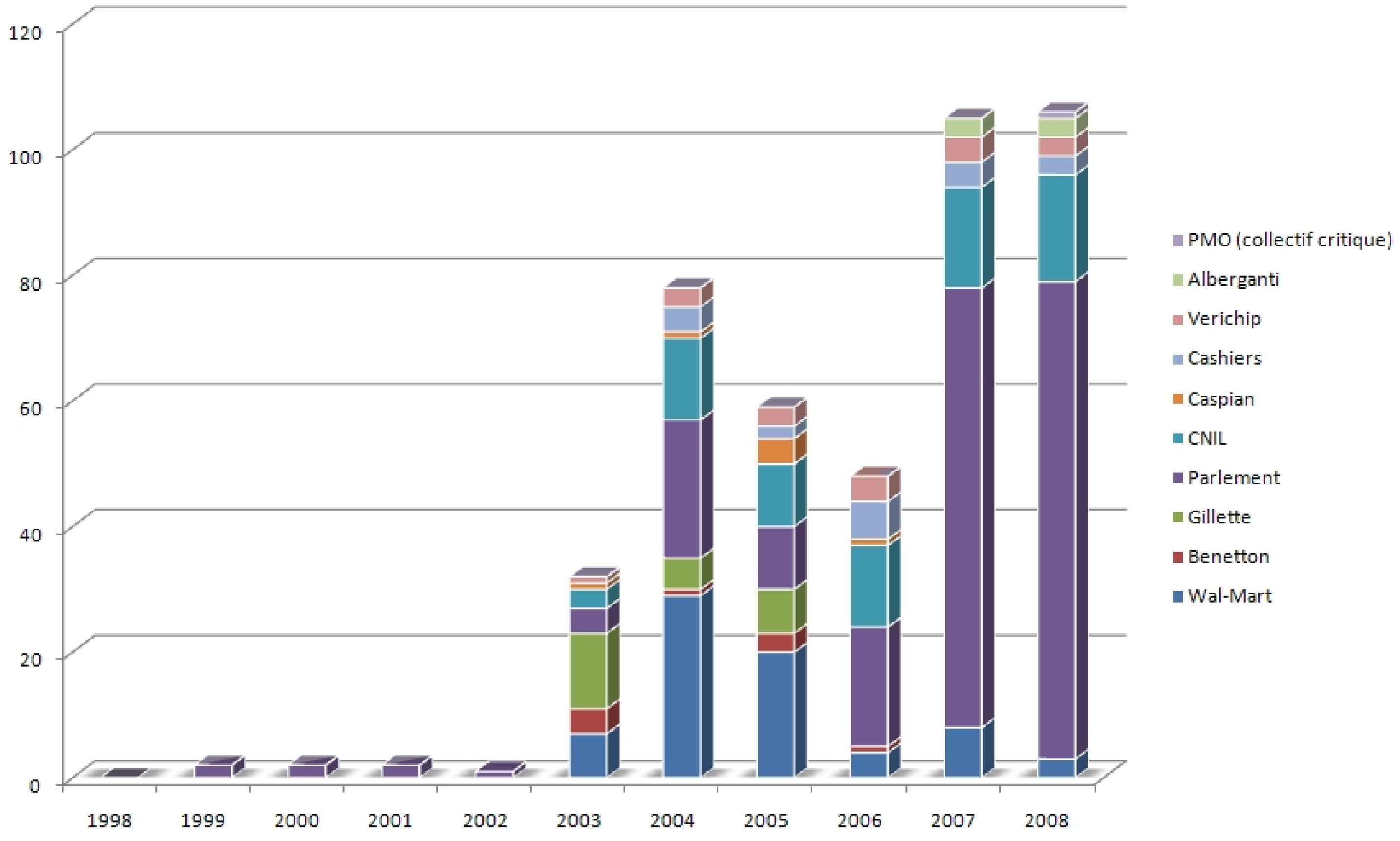

Figure 3: Emergence of actors in the debate per year.

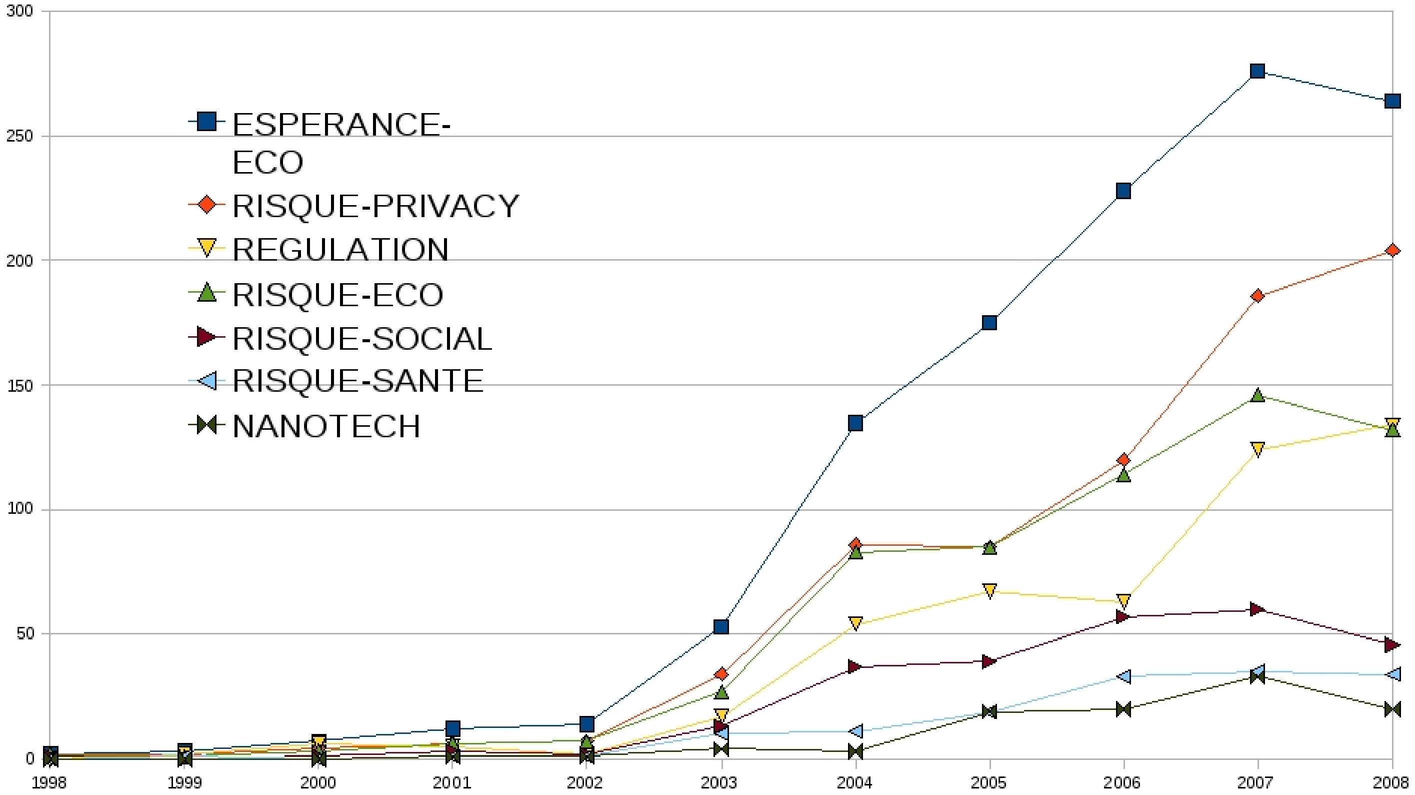

Figure 4: Historical emergence of the termes of the debate.

articles is about economic hope and appears in 2003. At that time Wal-Mart, Gillette and Benetton had already initiated their experimental introduction of RFID. These experiments together with the impulsion brought by the FDA (Food and Drugs Administration), contribute to an economic hope which probably led to a proliferation of RFID tags in users' lives. But once these applications of technology were broadcasted by the media, the first perceptions of risk started to emerge. Privacy risk and economic risk crossed the threshold of the public debate at the same time in 2004. At about the same period, the economic model introduced by Wal-Mart does not meet the expectations, while actors are entering in a suspicion phase. The public dispute with Gillette and Benetton is the event trigger. The issue of economic and social costs call into question the possibility of the inevitable extending of RFID to all the objects of everyday life. After the euphoria period, disillusion becomes visible. Regulation is asked for by both the general opinion (from the criticism of actors) and institutions, especially regarding the economic expectations whose implications seem to be problematic. However, this demand / supply of regulation is not immediate: it crosses the threshold of 50 articles in 2004/2005 to finally succeed in 2007. The socio professional risk (the cashiers losing their jobs), is also emerging in 2006/2007 when the economic hope is in full swing.

Finally, health risks and their probable connection with nanotechnology are emerging in the debate. In effect, in 2005 FDA gave its approval for the use of subcutaneous implants on humans. However, the health risk issue and its supposed connection with nanotechnology are not significant in this chronological approach, since they do not cross the threshold of 50 articles. In this perspective, two questions arise. What is the future of the health risk perception as the perception is very recent ? Should the connection between nanotechnology and health risk be taken into account? If we consider the fact that assimilation of nanotechnology to RFID involves *de facto* health problems, then we should combine the two thesauri.

5 Synthesis

Text mining made possible the mapping of the different actors involved in this technology as stakeholders in the public debate (such as economic developers, regulators, stakeholders). Furthermore, we found terms of debate with an approach that is both synchronous and diachronic. RFID refers to a technological complex whose history is devoid of axiological interpretations.

Radio-frequency identification first appeared in the press during the 1990s to grow exponentially as from 2003. Initially its apperance in press articles was due to the economic hope linked to the industries producing applications based on this technology. But over time it was rather the potential negative impacts which have been discussed. Four types of risk are identified then: economic risk, risk of invasion of privacy, potential socio-professional risk and, finally, health risk. A controversy quickly appears between economic hope and critical reaction, the latter calling for social responsibility on the part of both producers and users. The debate is made possible and grows exponentially from a common horizon: the application of RFID to all objects of everyday life. The replacement of bar code by RFID on all manufactured products led to both opportunities and risks, the latter concerning the infringement of individual liberties and the threat of the loss of employment for cashiers. The perception of health risks, the last issue which appeard in the debate, is currently emerging, raising the question of the continuous monitoring of the social perception related to RFID technology, all media combined.

But this methodological approach led us to some text mining problems that are not solved yet. As a matter of fact, we could hardly link the terms of debate with the actors. For that reason, we cannot really map the perception of risk according to actors involved, but only underline problematic subjects that arise all through the debate. However, we could eventually use textual discriminant analysis to focus specifically on this aspect [8]. But then, another problem dealing with the concept of threshold of perception appears. As from how many articles and actors involved can we consider that the terms in debate are significant enough, and therefore could be catagorized as early warning signals?

References

[1] M. Alberganti. *Sous l'œil des puces. La RFID et la démocratie*. Actes Sud, 2007.

[2] K. Albrecht and L. McIntyre. *Spychips. How major coporations and government plan to track your every move with RFID*. Nelson Current, 2005.

[3] F. Chateauraynaud. *Prospero, une technologie littéraire pour les sciences sociales*. CNRS éditions, Paris, 2003.

[4] D. Demazière. Des logiciels d'analyse textuelle au service de l'imagination sociologique. *Bulletin de méthodologie sociologique*, 85, 2005. http://bms.revues.org/index978.html.

[5] N. Dillenseger-Honoré. *Le règlement des conflits dans une controverse sociotechnique - Les risques sanitaires liés à la téléphonie mobile*. PhD thesis, Université Louis Pasteur Strasbourg I, Strasbourg, 2004.

[6] B. Dousset and B. Gay. Cartographie de réseaux d'alliances et analyse stratégique. *Revue des Sciences et Technologies de l'Information, Hermès Science Publications*, 11:p. 37–51, 2 2006.

[7] N. Kalampalikis. L'apport de la méthode alceste dans l'étude des représentations sociales. In J.-C. Abric, editor, *Méthodes d'étude des représentations sociales*, pages 147–163. Erès.

[8] L. Lebart and A. Salem. *Statistique Textuelle*. Dunod, 1994. 216 p.