Performing and Visualizing Temporal Analysis of Large Text Data Issued for Open Sources : Past and Future Methods

Jean-Charles Lamirel, Nicolas Dugué, Pascal Cuxac

Performing and visualizing temporal analysis of large text data issued for open sources: past and future methods

Jean-Charles Lamirel[1], Nicolas Dugué[1], and Pascal Cuxac[2]

[1] Equipe Synalp, Bâtiment B, F-54506 Vandoeuvre-lès-Nancy, France
lamirel@loria.fr, Nicolas.Dugue@loria.fr,
[2] CNRS-Inist, Vandoeuvre-lès-Nancy, France
Pascal.Cuxac@inist.fr

Abstract. In this paper we first propose a state of the art on the methods for the visualization and the interpretation of textual data, in particular of scientific data. We then shortly present our contributions to this field in the form of original methods for the automatic classification of documents and easy interpretation of their content through characteristic keywords and classes created by our algorithms. In a second step, we focus our analysis on the data evolving over time. We detail our diachronic approach, especially suitable for the detection and visualization of topic changes. This allows us to conclude with Diachronic'Explorer, our upcoming tool for visual exploration of evolutionary data.

Keywords: visualization, diachrony, clustering, feature selection, open data

1 Introduction

Databases of scientific literature and patents provide volumes of significant data for the study of scientific production. These data are also very rich and so complex. Indeed, the textual content of the publications, keywords used for archiving in these databases, the citations they contain and affiliations of the authors are as much information that it is possible to exploit for studying corpora of publications. These corpora are therefore a boon for the analysis of scientific and technical information.

In this article, we focus in particular on a major concern, which is to identify major changes related to developments in science and to describe them in a textual and visual way. Indeed, monitoring the development of transversal themes as well as detection of emerging themes or bridges between themes allows researchers to ensure of the innovative character of their area of research.

Furthermore, in managing the financing of research by the European Commission (EC), the detection of emerging issues is fundamental, as shown in the following examples. The NEST (New and Emerging Science and Technology) program was a specific EC program in FP6. Its objective was to encourage a

visionary and unconventional research at the frontiers of knowledge and at the interface of disciplines. To organize this program, the EC launched a call for support actions to follow and evaluate the projects but also to identify future research opportunities. Similarly, alongside the thematic in ICT (Information & Communication Technologies) program, the European Commission has set up, in the 7th Framework Program (FP7), the FET program (Future and emerging technologies) to promote research in the long term, or high risk, but with potentially strong impact from a societal or industrial point of view [3].

The detection of emerging technologies remains a complex process, and is therefore subject to studies in a broad spectrum of areas ranging from marketing to bibliometrics.

The selection tree proposed by [1] gives a good image of all forecasting methods that can be applied, in particular for the detection of these emerging technologies. It illustrates very well the dichotomy between quantitative methods and those based on expertise and shows the great diversity of existing approaches: Delphi and Nominal Group technique, methods based on the confrontation of the opinion of experts, scenario methods designed to scan different possible futures [20], until the methods combining the knowledge of the experts of a field and statistical techniques, allowing the identification of trends affecting causal factors.

The size and the complexity of the data that can be exploited to study the resulting databases of scientific publications require the development of quantitative methods for the detection of emerging topics by bibliometric methods, applying relatively simple statistical techniques as growth curves, or more sophisticated ones, such as automatic classification and analysis of networks [21] [4] [19] [13]. Another concern is also to provide tools capable of producing outputs exploitable by the end user. These outputs should be descriptive, intelligible and viewable.

Therefore, we separate our analysis into two parts. In the first part, we describe the quantitative and automatic methods that allow the extraction of relevant information from a corpus. In particular, these techniques offer to detect characteristic keywords from documents, or underlying topics - and keywords that describe them - referred in the documents. We also discuss of the visual exploitation which may be made of these methods. In a second part, we detail the methods for studying topic changes within a corpus whose data are anchored in time. We insist in particular on diachronic analysis methods, particularly effective to track these changes in a step by step and in a synthetic way.

Finally, in a last part, we detail Diachronic'Explorer, our open source tool for the production and viewing of diachronic analysis results. We show the effectiveness of the extraction methods that we offer through complementary and dynamic visualizations using recent technologies.

[3] See URL: http://cordis.europa.eu/fp7/ict/programme/fet_en.html

2 Exploitation of textual data and visualization

Topic identification is a technique which consists in understanding the meaning of the content of the documents of a corpus in an unsupervised way (without prior knowledge on the corpus and without human intervention) by isolating the topics underlying this content. These topics are usually represented by coordinated phrases and are often ranked in order of importance in the documents. As shown in [5], many techniques can be applied for topic identification and they might exploit research issued from several different communities, such as data mining, computational linguistics and information retrieval. We present hereafter two different types of identification techniques and their related visualization tools: the first is a widely used state of the art technique, the second is an alternative technique that we propose.

2.1 LDA

The method The LDA method is a probabilistic method for topic extraction who considers that the underlying topics of a corpus of documents can be characterized by multinomial distributions of words present in the documents [2]. According to this principle, each document is then considered as a composition of the topics extracted of the studied corpus. Figure 1 presents a list of topics produced by LDA, and their manifestation in a document. LDA uses a Dirichlet law to allow a careful choice of the parameters of the multinomial distributions. In practice, the extraction of these parameters is however complex and costly regarding computation time. It requires to exploit expectation maximization algorithms [8], which are prone to produce sub-optimal solutions, and in particular trivial or general results that are not usable in many cases, as in the context of the diachronic analysis (Section 3). This last type of analysis, which aims at comparing topics evolving over time, indeed requires working with accurate topic descriptions to isolate changes. Finally, the importance of the words in the documents can only be estimated in an indirect way by LDA and the method does not work on isolated documents. We present later a method based on feature maximization metric we have developed that does not have these drawbacks.

Vizualization using LDA In LDAExplore [6], the authors use a Treemap (Fig. 2) to represent the distribution of the importance of keywords in topics learned by the LDA. The representation of the topics in each document is also displayed, but in the form of curves where each point x-coordinate is a topic, and its weight is on the ordinate. Guille and Morales offer as a complete library for topic modelling, including LDA, which can also produce visualizations in the form of word clouds or histograms [11].

2.2 Feature maximization for feature selection

The method To introduce the feature maximisation metric and process [16], we first use an example. We then explain its use in our application framework.

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Fig. 1. Results of LDA from [2].

In Table 3, we present sample data collected from a panel of Men (M) and $Women$ (W) described by three features ($Nose\_Size$ (N), $Hair\_Length$ (C), $Shoes\_Size$ (S)). The problem of supervised classification in computer science is to learn to discriminate the class of Men of the class of $Women$ automatically by using these features. To achieve this, it is worthwhile for the algorithms to exploit the features that best separate the Men from the $Women$.

The process of feature maximization is comparable to a feature selection process. This process is based on the feature F-measure. The feature F-measure $FF_c(f)$ of a feature f associated with a cluster c (M or W in our example) is defined as the harmonic mean of the feature recall $FR_c(f)$ and of the feature predominance $FP_c(f)$, themselves defined as follows:

$$FR_c(f) = \frac{\Sigma_{d \in c'} W_d^f}{\Sigma_{c' \in C} \Sigma_{d \in c'} W_d^f} \quad FP_c(f) = \frac{\Sigma_{d \in c} W_d^f}{\Sigma_{f' \in F_c, d \in c} W_d^{f'}} \tag{1}$$

with

$$FF_c(f) = 2 \left(\frac{FR_c(f) \times FP_c(f)}{FR_c(f) + FP_c(f)} \right) \tag{2}$$

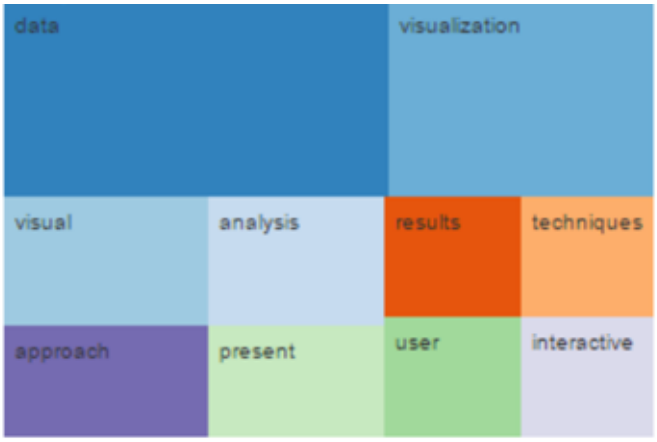


Fig. 2. Treemap based on LDA from [6].

where W_d^f represents the weight [4] of the feature f for the data d and F_c represents all the features present in the dataset associated with the class c.

The feature selection process based on feature maximization can thus be defined as a parameter-free process in which a class feature is characterized by using both his ability to discriminate the class to which it relates ($FR_c(f)$ index) and its ability to faithfully represent the data of this class ($FP_c(f)$ index). Table 4 presents how does operate the calculation of the feature F-measure of the $Shoes\_Size$ feature to the Men class.

Once the capacity to discriminate a class ($FR_c(f)$ index) and to faithfully represent the data of a class ($FP_c(f)$ index) calculated for each feature (Tab. 5), the further step consists in automatically selecting the most relevant features for distinguishing classes. The set S_c of features that are characteristic of a given class c belonging to the group of classes C is represented as:

$$S_c = \left\{ f \in F_c \mid FF_c(f) > \overline{FF}(f) \text{ and } FF_c(f) > \overline{FF}_D \right\} \text{ where} \tag{3}$$

$$\overline{FF}(f) = \Sigma_{c' \in C} \frac{FF_{c'(f)}}{|C_{/f}|} \text{ and } \overline{FF}_D = \Sigma_{f \in F} \frac{\overline{FF}(f)}{|F|} \tag{4}$$

[4] The choice of the weighting scheme is not really constrained by the approach instead of producing positive values. Such scheme is supposed to figure out the significance (i.e. semantic and importance) of the feature for the data.

Feature Recall is a scale independent measure but feature Predominance is not. We have however shown experimentally in (Lamirel et al., 2014a) that the F-measure which is a combination of these two measures is only weakly influenced by feature scaling. Nevertheless, to guaranty full scale independent behavior for this measure, data must be standardized.

Shoes_ Size	Hair_ Length	Nose_ Size	Class
9	5	5	M
9	10	5	M
9	20	6	M
5	15	5	W
6	25	6	W
5	25	5	W

Fig. 3. Sample data for supervised classification in *Men/Women* classes.

Shoes_ Size	Hair_ Length	Nose_ Size	Class
9	5	5	M
9	10	5	M
9	20	6	M
5	15	5	W
6	25	6	W
5	25	5	W

$$FR(S,M) = 27/43 = 0.62$$

$$FP(S,M) = 27/78 = 0.35$$

$$FF(S,M) = \frac{2(FR(S,M)\times FP(S,M))}{FR(S,M)+FP(S,M)} = 0.48$$

Fig. 4. Sample data and computation of feature F-measure for *the Shoes_Size* feature.

where $C_{/f}$ represents the subset of C in which the f feature is represented.

Finally, the set of all selected features S_C is the subset of F defined by:

$$S_C = \cup_{c \in C} S_C. \tag{5}$$

The features that are considered relevant for a given class are the features whose representations are better in that class than their average representation in all classes, and also better than the average representation of all features, as regards to feature F-measure. Thus, features whose feature F-measure is always lower than the overall average are eliminated, and the variable *Nose_Size* is therefore suppressed in our example ($0.3 < 0.38$ and $0.24 < 0.38$).

A complementary step to estimate the contrast may be operated in addition to the first stage of selection. The role of this step is to estimate the information gain produced by a feature on a class. It is proportional to the ratio between

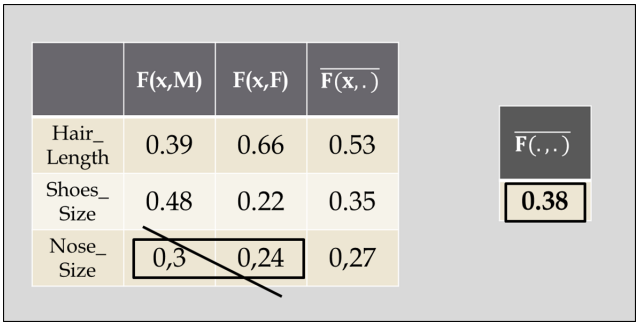


Fig. 5. Feature F-measure of the feature and related marginal average.

the value of the F-measure of a feature in the class and the average value of the F-measure of that feature in all classes. For a feature f belonging to the group of selected features S_c from a class c, the gain $G_c(f)$ is expressed as:

$$G_c(f) = FF_c(f)/\overline{FF}(f) \tag{6}$$

Finally, active, or descriptive, features of a class are those for which the contrast is greater than 1 in those above. Thus, the selected features are considered active in the classes in which the feature F-measure is higher than the marginal average:

- $Shoes\_Size$ is active in the Men class ($0.48 > 0.35$),
- $Hair\_Length$ is active in the $Women$ class ($0.66 > 0.53$).

Contrast ratio highlights the degree of activity/passivity of the features selected compared to their average F-measure. Table 6 shows how the contrast is calculated on the presented example. In this context, the contrast may thus be considered as a function that will virtually have the following effects:

- Increase the $length$ of $women'shair$,
- Increase the $size$ of the $men'sshoes$,
- Reduce the $length$ of the $men'shair$,
- Decrease the $size$ of $women'sshoes$.

Preliminary cluster labelling experiments showed that feature maximization metric has discrimination capabilities similar to Chi-squared metric, while with generalization capabilities very appreciably higher [14]. Moreover, this technique proved to have very low computation time, unlike LDA. It often has a dual function in learning and visualization, as shown by experiments in [7] or in [16]. In the classification context, it can thus optimize performance of the classifiers, while producing class profiles exploitable for the visualization of the content of the classes.

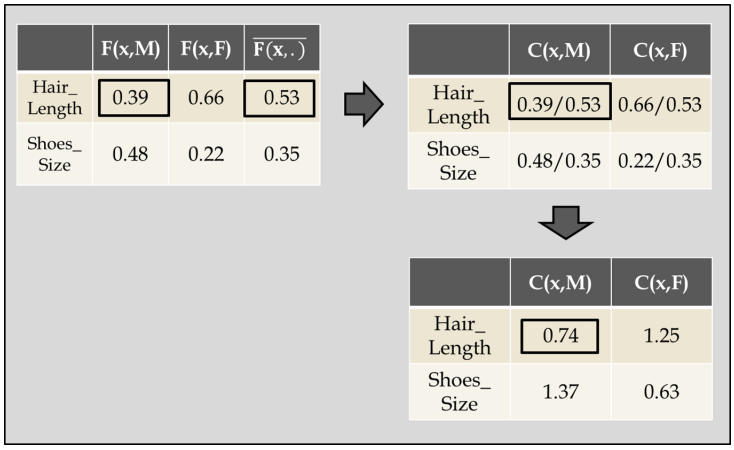


Fig. 6. Principle of computation of contrast on selected features and obtained results.

Table 7 shows an application of the method on textual data with the goal of establishing discriminative profiles of Chirac and Mitterrand presidents using data of the DEFT 2006 corpus containing around 80000 extracts of their speeches [17]. It shows in particular that contrast allows to quantify the influence of features in classes (typical terms of each speaker). Extracted features and their contrast can thus act as profiles of classes for a classifier, as well as indicators of content or meaning for an analyst.

This first example shows how to use feature maximization in the general framework of text datasets. We show more specifically in Section 3 how our feature maximization method can be applied for the more precise and difficult task of topics changes detection in corpora of scientific papers and describe in Section 4 advanced visualizations of that changes. But, to clearly illustrate the potential and the scope of the method, we firstly give hereafter a more specific example, related to synthetic visualization and automatic summary of the individual content of the documents from this method.

Visualization using feature maximization For the synthetic representation of the content in a single document, we propose an original method based on competition between blocks of text, coupled with the feature maximization metric. This approach allows to overcome the lack of metadata to describe the texts. Furthermore, it has the advantage of being independent of the language, to function without external knowledge source and parameters, and is likely to have multiple applications: generation of metadata or input data for the clustering, generation of automatic summaries and explanations of several levels of generality. It consists here in performing indexing of full text papers taking benefit of their structure. As regards to this approach, each part of the paper can be seen as a class, the paper itself being a classification: for example, the

Mitterrand			Chirac	
Contraste	Terme		Contraste	Terme
1.88	douze		1.93	partenariat
1.85	est-ce		1.86	dynamisme
1.80	eh		1.81	exigence
1.79	quoi		1.78	compatriotes
1.78	-		1.77	vision
1.76	gens		1.77	honneur
1.75	assez		1.76	asie
1.74	capables		1.76	efficacité
1.72	penser		1.75	saluer
1.70	bref		1.74	soutien
1.69	puisque		1.74	renforcer
1.67	on		1.72	concitoyens
1.66	étais		1.71	réforme
1.62	parle		1.70	devons
1.62	fallait		1.70	engagement
1.60	simplement		1.69	estime

Fig. 7. Most contrasted features (terms) in *Mitterrand* and *Chirac* speeches (extract).

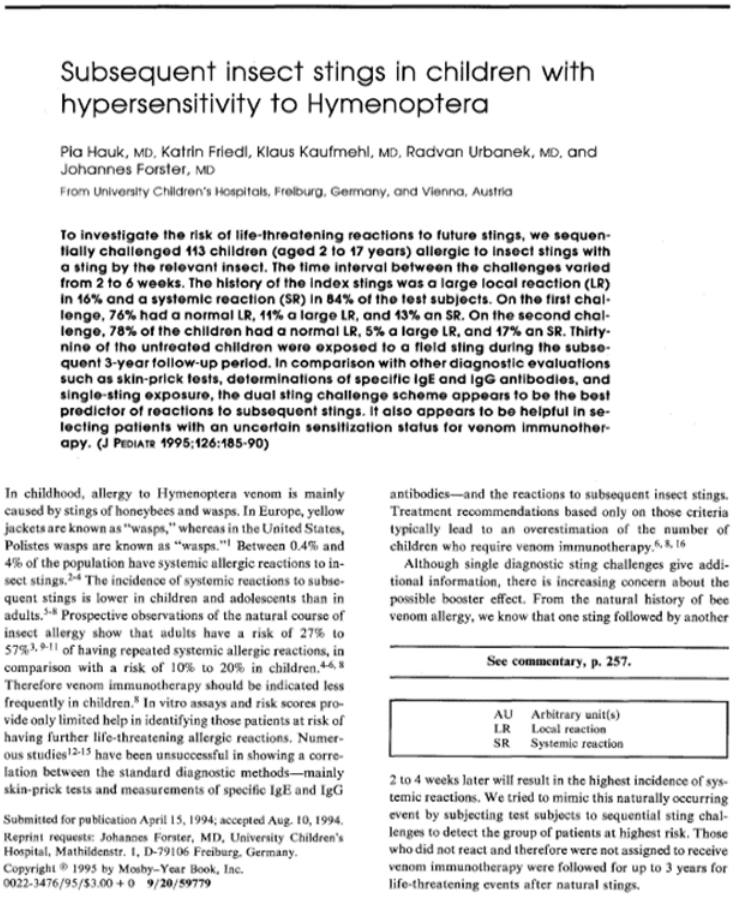
exploited classes might be:"introduction, methodology, state of the art, results, conclusion... ".

We will illustrate our point with this scientific paper (randomly selected in the ISTEX [5] reservoir): Hauk P, Friedl K, Kaufmehl K, Urbanek R, Forster J.: Subsequent insect stings in children with hypersensitivity to Hymenoptera. J Pediatr. 1995 Feb; 126 (2): 185 - 90.

This paper includes the following major parts: Introduction / Methods / Results / Discussion (Fig. 8).

[5] The ISTEX project (Initiative d'Excellence pour l'Information Scientifique et Technique) fits in the "Investment for the future" program, initiated by the French Ministry of Higher Education and Research (MESR), whose ambition is to strengthen research and French higher education on the world level. The ISTEX project main objective is to offer to the whole of the community of higher education and research, online access to the retrospective collections of scientific literature in all disciplines by engaging a national policy of massive acquisition of documentation: archives of journals, databases, corpus of texts.

Reference: http://www.istex.fr



Subsequent insect stings in children with hypersensitivity to Hymenoptera

Pia Hauk, MD, Katrin Friedl, Klaus Kaufmehl, MD, Radvan Urbanek, MD, and Johannes Forster, MD

From University Children's Hospitals, Freiburg, Germany, and Vienna, Austria

To investigate the risk of life-threatening reactions to future stings, we sequentially challenged 113 children (aged 2 to 17 years) allergic to insect stings with a sting by the relevant insect. The time interval between the challenges varied from 2 to 6 weeks. The history of the index stings was a large local reaction (LR) in 16% and a systemic reaction (SR) in 84% of the test subjects. On the first challenge, 76% had a normal LR, 11% a large LR, and 13% an SR. On the second challenge, 78% of the children had a normal LR, 5% a large LR, and 17% an SR. Thirty-nine of the untreated children were exposed to a field sting during the subsequent 3-year follow-up period. In comparison with other diagnostic evaluations such as skin-prick tests, determinations of specific IgE and IgG antibodies, and single-sting exposure, the dual sting challenge scheme appears to be the best predictor of reactions to subsequent stings. It also appears to be helpful in selecting patients with an uncertain sensitization status for venom immunotherapy. (J Pediatr 1995;126:185-90)

In childhood, allergy to Hymenoptera venom is mainly caused by stings of honeybees and wasps. In Europe, yellow jackets are known as "wasps," whereas in the United States, Polistes wasps are known as "wasps."[1] Between 0.4% and 4% of the population have systemic allergic reactions to insect stings.[2-4] The incidence of systemic reactions to subsequent stings is lower in children and adolescents than in adults.[5-8] Prospective observations of the natural course of insect allergy show that adults have a risk of 27% to 57%[3, 9-11] of having repeated systemic allergic reactions, in comparison with a risk of 10% to 20% in children.[4-6, 8] Therefore venom immunotherapy should be indicated less frequently in children.[8] In vitro assays and risk scores provide only limited help in identifying those patients at risk of having further life-threatening allergic reactions. Numerous studies[12-15] have been unsuccessful in showing a correlation between the standard diagnostic methods—mainly skin-prick tests and measurements of specific IgE and IgG antibodies—and the reactions to subsequent insect stings. Treatment recommendations based only on those criteria typically lead to an overestimation of the number of children who require venom immunotherapy.[6, 8, 16]

Although single diagnostic sting challenges give additional information, there is increasing concern about the possible booster effect. From the natural history of bee venom allergy, we know that one sting followed by another 2 to 4 weeks later will result in the highest incidence of systemic reactions. We tried to mimic this naturally occurring event by subjecting test subjects to sequential sting challenges to detect the group of patients at highest risk. Those who did not react and therefore were not assigned to receive venom immunotherapy were followed for up to 3 years for life-threatening events after natural stings.

See commentary, p. 257.

AU	Arbitrary unit(s)
LR	Local reaction
SR	Systemic reaction

Submitted for publication April 15, 1994; accepted Aug. 10, 1994.
Reprint requests: Johannes Forster, MD, University Children's Hospital, Mathildenstr. 1, D-79106 Freiburg, Germany.

0022-3476/95/$3.00 + 0 9/20/59779

Fig. 8. First page of the selected paper.

After extraction of the terms by a conventional PoS tagging method, the feature maximization method described Section 2.2 allows to obtain a list of specific terms for each part of the paper, weighted by their importance. Frrom that, it is possible to build up a vectorial (i.e. Bag-of-Words) representation of the paper, or alternatively, to build up a weighted graph (*paper parts/selected terms*) that will illustrate clearly the scientific contents of the text (Fig. 9).

If we follow an approach of automatic summarization [10], each selected term being weighted for each identified part of the paper, it is easy to balance the sentences containing these terms by adding their weight. We furthermore assign an additional weight to terms that are also part of the title of the paper. The curve of the weights of the sentences thus calculated for each of them always shows a plateau (Fig. 10) ; then, we choose to keep the sentences whose weight

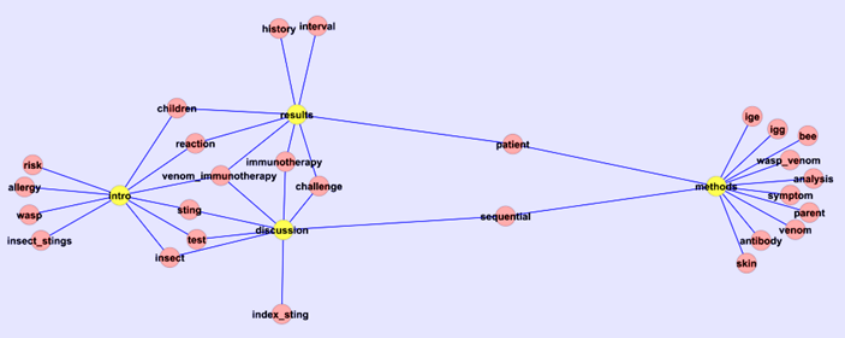


Fig. 9. Graphical representation of the content of a scientific paper with the use of its structure.

is greater or equal to this level and reorder them by rank of appearance in the text.

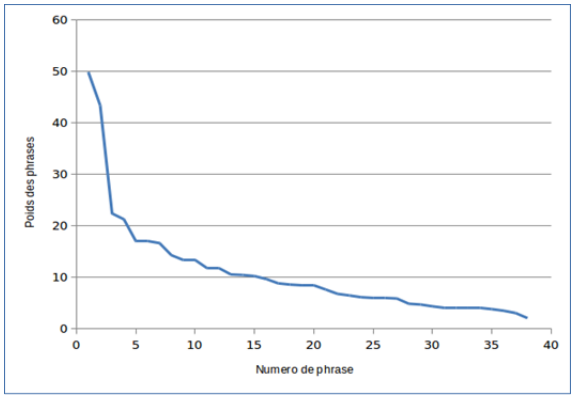


Fig. 10. Distribution of the weights of the sentences.

We then have a summary obtained by extraction of meaningful sentences of text that has generally small size (less than 12 sentences in all our experiments). For the paper used as an example, the summary is described on Figure 11.

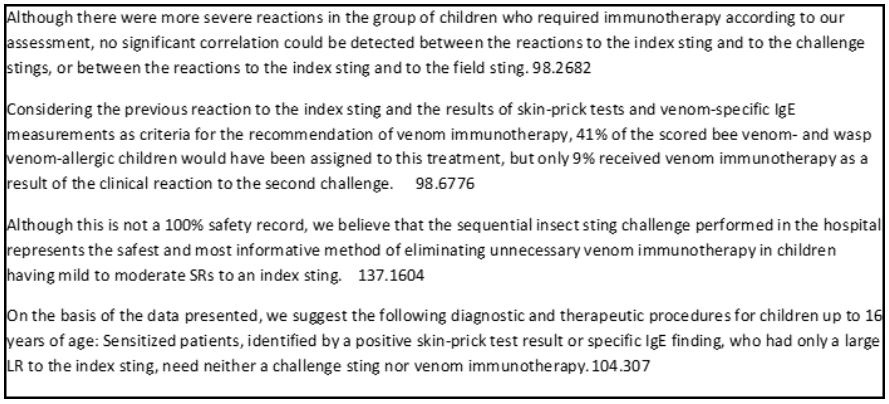


Fig. 11. Automatic summary of the paper of Figure 8 produced by sentence extraction (most relevant sentences along with their weights ranked in their order of appearance in the text).

3 Visualization of evolving data

3.1 Visualization methods

In Neviewer, Wang et al. use alluvial diagrams (Fig. 12), sometimes called Sankey visualization [28]. These visualizations were also used in [27] to view the changes in the citations between scientific disciplines.

Ratinaud uses dendogram to visualize the various topics (and their vicinities) mentioned on Twitter with the hashtag #mariagepourtous [26]. It is nonetheless Treemaps (Section 2, Fig. 2) that enable him to show the progression or regression of topics in time. On their own side, Osborne and Motta use graphics with stacked areas to follow the evolution of the amount of publications grouped into topics across time [22].

These methods have all of the interesting benefits and we show their complementary exploitation use in Section 4.

3.2 Diachronic analysis

Diachronic analysis, which consists in comparing data or results by time step, is extensively used. In linguistics, Perea uses this technique to follow the evolution of the Catalan language through time [25]. Cardon and al. study the evolution of blogs and their importance on the web in a diachronic way [3]. Similarly, the activity of bloggers and the evolution of their interests are studied in a diachronic way by [12]. The work of Wang et al, more connected to our field of applications, analyzes the thematic evolution of the research in such a way [28].

In our case, we develop a parameter-free method, directly exploitable by the user and based on feature maximization [16] to identify and describe the topics

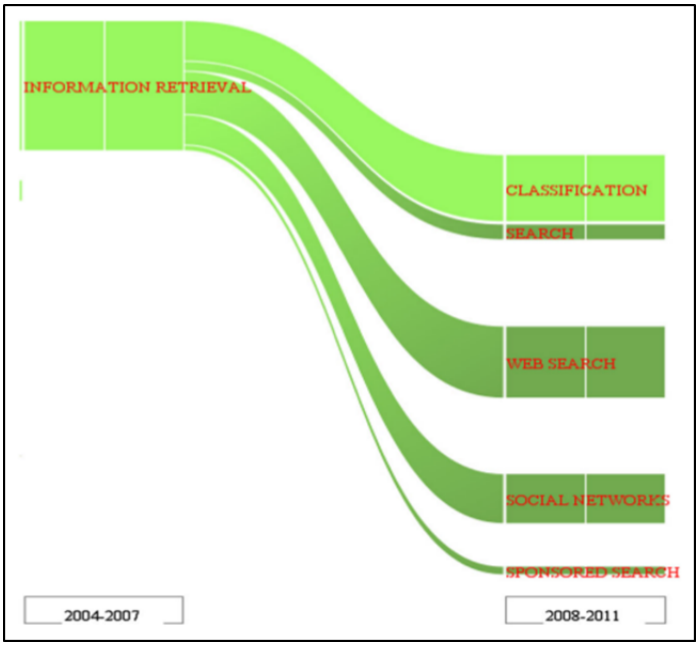


Fig. 12. Alluvial diagram from Neviewer [28].

of a corpus. This approach allows identification of keywords issued from the full text content of documents which are characteristic of each topics, conversely to methods based on keywords indexed by the publication databases [28] [23]. Furthermore, the absence of parameters in the process allows to completely automate the task of indexing the documents passing through the detection of topics and up to the visualization of their evolution. Figure 13 shows the progress of the complete method up to visualization. The whole process is also detailled hereafter:

1. We query a bibliographic database in order to build up a corpus covering several years of publication on a given topic.
2. The full text of each obtained document is treated with a conventional PoS tagging tool to extract index terms (keywords).
3. The documents and their extracted keywords are then grouped into classes corresponding to their year of publication and a feature selection based on feature maximization is applied on the keywords of each of the document group. Furthermore, a graph figuring out the interactions between keywords and document groups, using weighted links setting the strength of the relationships, is constructed. Thanks to a random walk algorithm (here Walktrap [24]), it is then possible to automatically detect groups of years (time periods) who will then serve as time steps for the diachronic clustering algorithm.

Figure 14 gives an example of contrast graph as well as resulting cutting in time periods.

4. A neural clustering algorithm [9], more stable and more efficient than the usual clustering algorithms on the textual data, is applied multiple times, with standard parameters, on the data of each time period by varying the desired number of clusters. Clustering quality criteria that are reliable for textual and multidimensional data are exploited in a further step [18] to isolate an optimal model (ideal number of clusters) for each of the periods.
5. The optimal clustering models of each period are post processed separately using feature maximization so as to extract the salient features of each cluster in each time period.
6. The feature maximization results, as well as the overall clustering results, are transmitted to the diachronic module of the Diachronic'Explorer tool presented in Section 4. This tool implements both diachronic analysis functions, based on unsupervised Bayesian reasoning [15], in order to detect thematic connections and differences between the time periods, as well as many functions of visualization of the results.

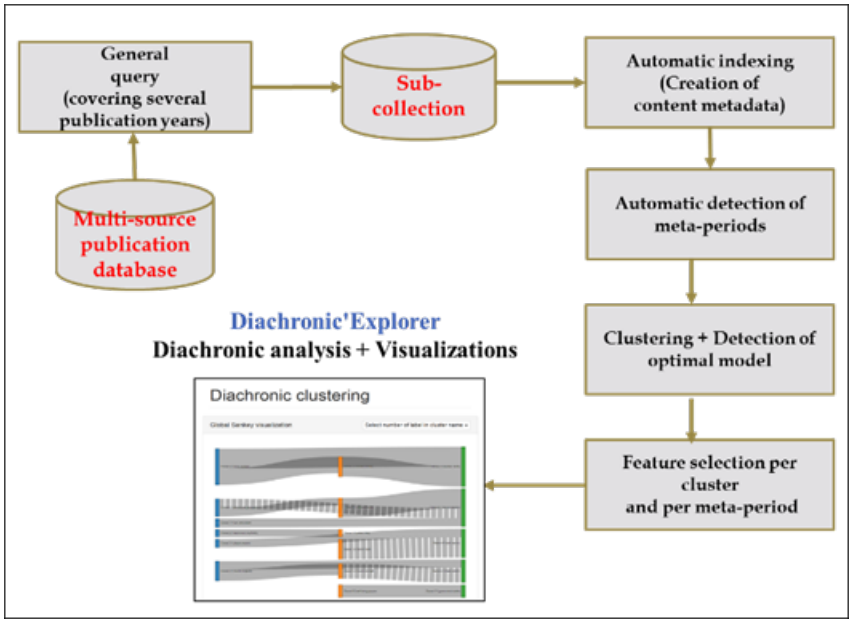


Fig. 13. Detailled diachronic analysis process (incl. DiachronicExplorer tool).

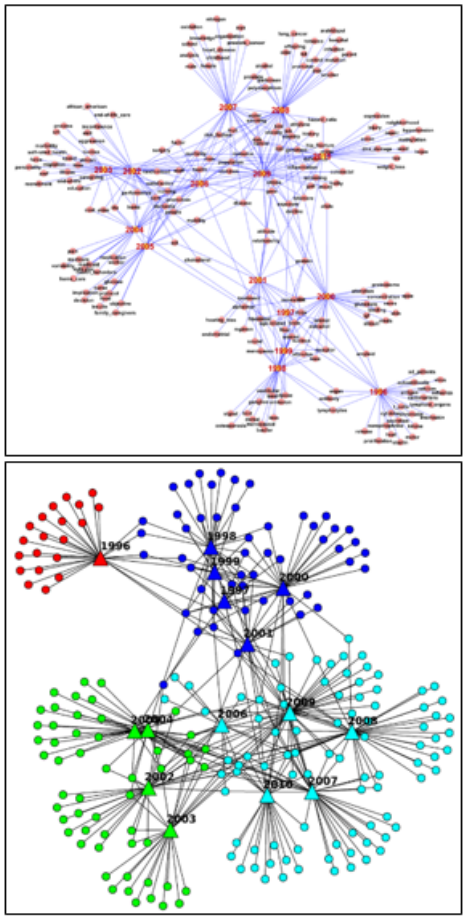


Fig. 14. Contrast graph per year (up) and cutting into time periods (down).

4 Diachronic'Explorer tool

We now introduce Diachronic'Explorer [6], our open source tool for the production and visual exploitation of diachronic analysis results.

Diachronic'Explorer is composed of two modules. The first module allows to use descriptions of the topics produced by the clustering and the feature maximization (Section 3.2) to track these topics over time and detect the changes and similarities between time periods. The second module is dedicated to the visualization of the results produced by the first. Designed in the form of web platform using modern technology, the tool offers the possibility to explore the

[6] A demo version of the tool can be found at URL: http://github.com/nicolasdugue/istex-demonstrateur

corpus through various complementary visualizations each representing a different level of granularity in the exploration of this corpus.

We will detail below how the tool can be used from the finest-grained level to more synthetic visualization of the corpus and its evolutions. For that purpose, we will take, as example corpus, the evaluation corpus operated by the French ISTEX project. This corpus comprises 9779 records related to research conducted in the general field of Gerontology/Geriatrics between 1996 and 2010. The result of the analysis steps described in Section 3 on this corpus is a division into three time periods.

First of all, the tool allows to study keywords that are particularly characteristics of the topics and to see the evolution of their importance in time. Figure 15 shows for example the main keywords and their evolution in the above-mentioned corpus. The size of a circle is proportional to the importance of the related keyword in its topic. This size is therefore conditioned by the contrast value produced by the feature maximization process (Section 3.2). Figure 16 shows the description of a topic (cluster) through keywords that are representative. The size of the rectangles is also proportional to the value of contrast.

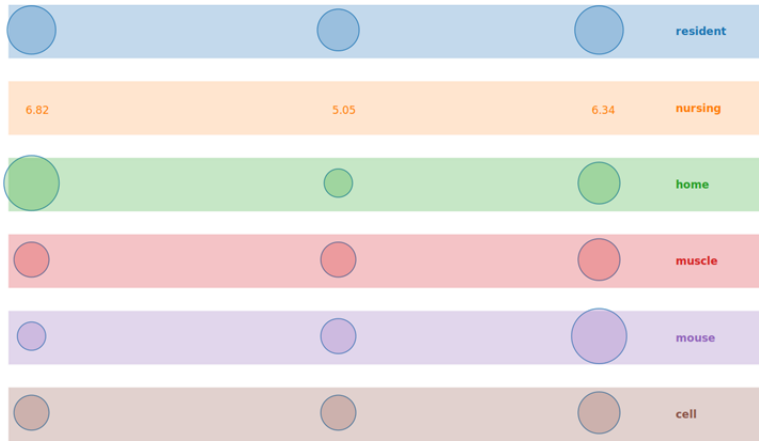


Fig. 15. Evolution of the importance of the keywords through the curse of the 3 time periods for the studied corpus. The size of the circles materialize this importance. Contrast values can also be displayed (here for the keyword nursing).

Taking some distance from the corpus, with the Figure 17, the topics of a period can be observed in a global way. In this figure, each column represents a topic, and the cells in the column are the keywords that describe this topic. The size of these cells is conditioned by the relative importance of the related keywords, in terms of contrast, in the topic. The colors represent the intensity of the contrast of the considered keywords.

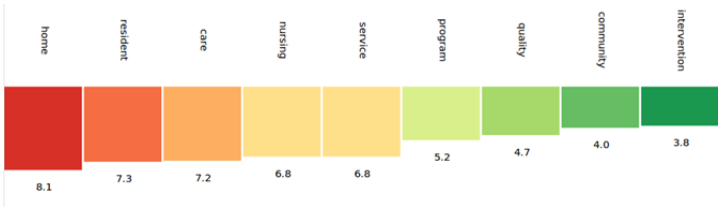


Fig. 16. List of keywords associated to a cluster and ranked by contrast values. Colors materialize the different values.

Figure 18 allows take some more distance, and to consider interactions between couples of periods. This visualization provides a detailed representation of the diachronic analysis between two periods. The blue rectangle represents the period prior to that represented by the yellow rectangle. The circles represent topics whose size is determined by the quantity of documents they contain. The label of each circle is currently the most characteristic keyword (i.e. the more contrasted one) of the period (for more details, it is possible to select several labels). In addition, the links between circles represent inter-period topic links. They are characterized by a force, which depends on both the number of keywords shared between the topics of the two periods and the contrast of those shared keywords[7]. This force provides the thickness of the link. The yellow rectangle below details similarities for each link between two topics. The dissimilarities between topics can be displayed as well, but there are not shown in the figure.

Finally, Figure 19 allows us acquire knowledge of all of the topic links between period existing within the corpus. Each color represents a period and vertical rectangles the topics of this period who have links with other periods. In grey color, we observe the inter-period link between topics. Dotted links indicate specific types of topic connections: one of the two topics has a broader descriptive vocabulary. It is possible to see the details of topics passing the mouse over the rectangles of color. Similarly, details on topics links are available on the grey areas.

To see the corpus content in its entirety, we offer also in Diachronic'Explorer an original method of visualization that shows the information in the form of a contrast graph (Fig. 20). The big circles represent topics, the small ones, the keywords. If a topic is described by a keyword then a link is present between the two circles. This visualization shows so all information in a condensed manner.

[7] The principle of computation of the strength of a topic link between periods is explained with more details in [15].

Fig. 17. Topics of a period described by their associated keywords. Each topic is a column and the height of a cell materializes the importance of the keyword related to this cell. The color represents the intensity of its contrast.

5 Conclusion

In this paper, using a detailed state of the art on work done for analysis of textual data, and particularly that of scientific data, we have highlighted in a first time, the strategic importance of the diachronic analysis of such data, as well as the difficulties and complexities related to this type of treatment, whether it's the lack of available metadata or the parameters settings and scope problems related to usual methods of analysis, especially those of topic detection. We have also shown that there are many interesting alternatives with regard to the visualization of analysis results. This discussion has enabled us to propose, as a second step, a new methodology of analysis based on feature maximization. This methodology has many benefits to the existing, which be without parameters, to be applicable at different scale levels, from the corpus to the document, with ease of calculation, and finally, to have strong capabilities of synthesis, which

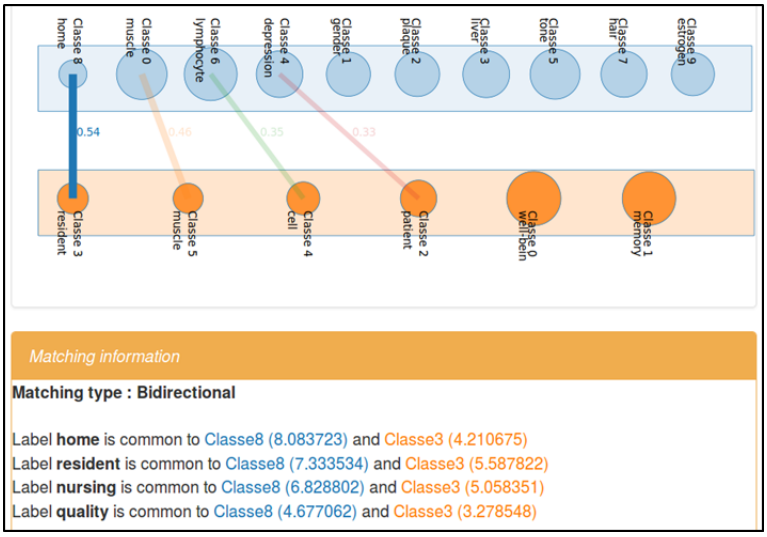


Fig. 18. Diachronic visualization of 2 periods (extract). Circles materialize topics and links between them indicates similarities or slight changes. The strength of a link is materialized by its thickness and by an indicator value (a value of 1 corresponding to a perfect match). Yellow frame (down) details the similarity between the Cluster 8 of the blue period and the Cluster 3 of the yellow period.

makes her results easily interpretable, even for complex problems and large corpora. All these decisive advantages have in particular helped us to create the DiachronicExplorer tool, which, by integrating the unsupervised Bayesian reasoning and feature maximization with many methods of visualization, provides effective solutions to deal with a problem as complex as that of the detection of changes, from the full text, within a large corpus of scientific publications whose topics evolve with time. This indirectly shows that this type of methodology, due to its great flexibility, has a field of application of much wider range than that presented in this paper. Its synthesis capabilities make it indispensable, especially upstream visualization methods, when the representation of the data itself is complex.

References

1. Amstrong, J.S. and Green, K.C. and Graefe, A.: Forescating Principles. Encyclopedia of Statistical Sciences, Lovric M. (Ed.), Springer (2011)
2. Blei, D. and Y. Ng, A. and Jordan, M.I.: Latent Dirichlet Allocation. Journal of Machine Learning Research 3, pp. 993–1022 (2003)
3. Cardon, D. and Fouetillou, G. and Roth, C.: Two Paths of Glory-Structural Positions and Trajectories of Websites within Their Topical Territory. Fith International AAAI Conference on Weblogs and Social Media -ICWSM-11 (2011)

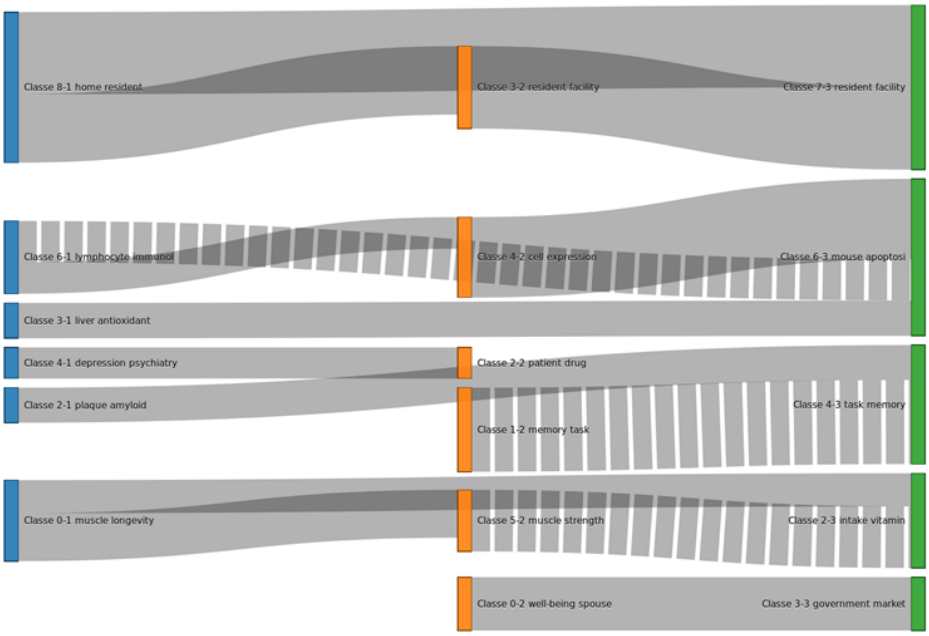


Fig. 19. Diachronic visualization of all the periods of a corpus. Each color represents one period and vertical rectangles materialize topics. In grey color, one can observe the links between the topics of the different periods.

4. Daim, T.U. and Rueda, G. and Martin, H. and Gerdsri, P.: Forecasting emerging technologies: Use of bibliometrics and patent analysis. Technological Forecasting & Social Change, 73, pp. 981–1012 (2006)
5. Dermouche, M. and Velcin, J. and Loudcher, S. and Khouas, L.: Une nouvelle mesure pour lévaluation des méthodes dextraction de thématiques : la Vraisemblance Généralisée. Actes des 13ièmes Journées Francophones dExtraction et de Gestion des Connaissances (EGC 2013), pp. 317–328, Toulouse, France (2013)
6. Ganesan, A. and Brantley, K. and Pan, S. and Chen, J.: LDAExplore: Visualizing Topic Models Generated Using Latent Dirichlet Allocation. arXiv preprint arXiv:1507.06593 (2015)
7. Falk, I. and Lamirel, J.-C. and Gardent, C.: Classifying French Verbs Using French and English Lexical Resources. International Conference on Computational Linguistic (ACL 2012), Jeju Island, Korea, July 2012 (2012)
8. Francesiaz, T. and Graille, R. and Metahri, B.: Introduction aux modèles probabilistes utilisés en fouille de données. Rapport IMAG, Université de Grenoble (2015)
9. Fritzke, B.: A growing neural gas network learns topologies. Advances in Neural Information Processing Systems 7, Tesauro, G. and Touretzky, D.S. and Leen, T. K. (Ed.), pp. 625–632 (1995)
10. Goldstein, J. and Mittal, V. and Carbonell J. and Kantrowitz, M.: Multi-document summarization by sentence extraction. Workshop on Automatic summarization, NAACL-ANLP 2000, pp. 40–48 (2000)

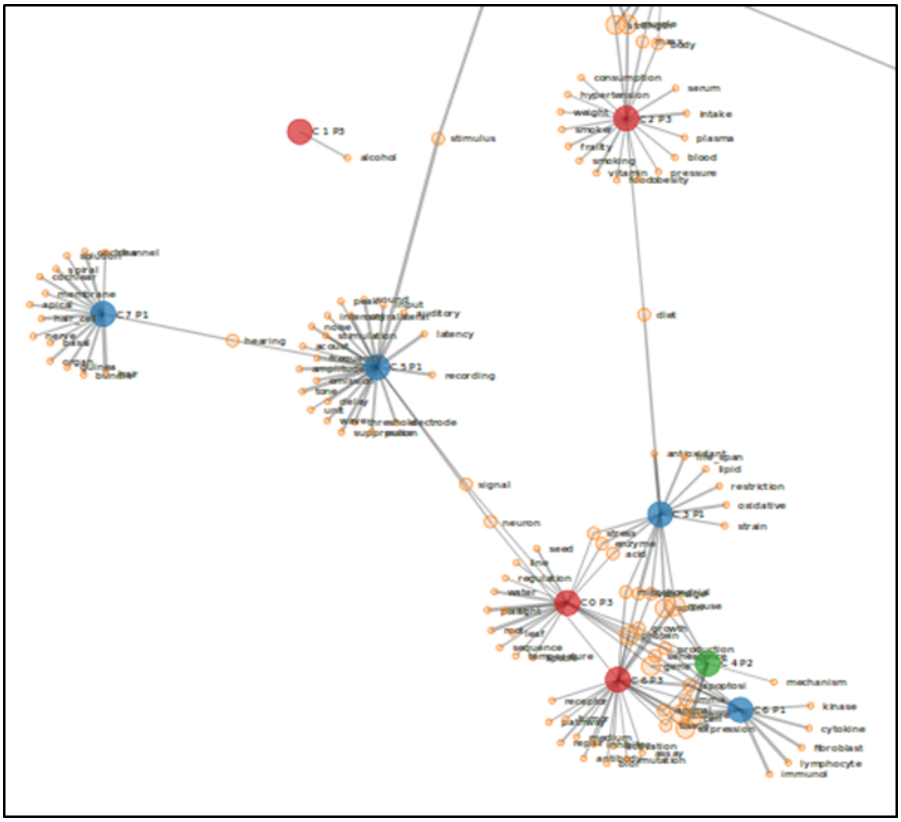


Fig. 20. Diachronic visualization of all the periods of a corpus in the form of a contrast grap (extract). Each color represents one specific period and big circle materialize the topics. In orange color one can observe the keywords related to the topics.

11. Guille, A. and Soriano-Morales, E. P.: TOM: A library for topic modeling and browsing. Actes des 16ièmes Journes Francophones dExtraction et de Gestion des Connaissances (EGC 2016), pp. 451–456, Reims, France (2016)
12. Itoh, M. and Yoshinaga, N. and Toyoda, M. and Kitsuregawa, M.: Analysis and visualization of temporal changes in bloggers' activities and interests. Visualization Symposium (PacificVis), 2012 IEEE Pacific, pp. 57–64 (2012)
13. Kajikawa, Y. and Yoshikawa, J. and Takeda, Y. and Matushima, K.: Tracking emerging technologies in energy research: Toward a roadmap for sustainable energy. Technological Forecasting & Social Change, 75, pp. 771-782 (2008)
14. Lamirel, J.-C. and Ta, A.P. and Attik M.: Novel labeling strategies for hierarchical representation of multidimensional data analysis results. Proceedings of IASTED International Conference on Artificial Intelligence and Applications (AIA), Innsbruck, Austria, February 2008 (2008)
15. Lamirel, J.-C.: A new diachronic methodology for automatizing the analysis of research topics dynamics: an example of application on optoelectronics research. Sci-

entometrics 93(1): 151–166 (2012)
16. Lamirel, J.-C. and Cuxac, P. and Chivukula, A.S. and Hajlaoui, K.: Optimizing text classification through efficient feature selection based on quality metric. Journal of Intelligent Information Systems, Special issue on PAKDD-QIMIE 2013, pp. 1–18 (2014)
17. Lamirel, J.-C. and Cuxac, P.: Une nouvelle méthode statistique pour la classification robuste des données textuelles : le cas Mitterand-Chirac. JADT, Paris, France, April 2014 (2014)
18. Lamirel, J.-C. and Cuxac, P.: New quality indexes for optimal clustering model identification with high dimensional data. Proceedings of ICDM-HDM15 - International Workshop on High Dimensional Data Mining, Atlantic City, USA, November 2015 (2015)
19. Mogoutov, A. and Kahane, N.: Data search strategy for science and technology emergence: A scalable and evolutionary query for nanotechnology tracking. Research Policy, 36, pp. 893–903 (2007)
20. Monti, R. and Roubelat, F.: La boîte outils de prospective stratégique et la prospective de dfense: rétrospective et perspectives. Actes des Entretiens Science & Défense, Paris (1998)
21. Noyons E.: Science maps within a science policy context. Handbook of Quantitative Science and Technology Research. Eds. Moed H.F., Glänzel W., Schmoch U., London, Kluwer Academic Publishers, pp. 237–255 (2004)
22. Osborne, F. and Motta, E.: Understanding research dynamics. Semantic Web Evaluation Challenge, pp. 101–107 (2014)
23. Osborne, F. and Scavo, G. and Motta, E.: Identifying diachronic topic-based research communities by clustering shared research trajectories. The Semantic Web: Trends and Challenges, pp. 114–129 (2014)
24. Pons, P. and Latapy, M.: Computing communities in large networks using random walks. Computer and Information Sciences-ISCIS, pp. 284–293 (2005).
25. Perea, M. P.: Dynamic cartography with diachronic data: Dialectal stratigraphy. Literary and linguistic computing, 28(1):147–156 (2013)
26. Ratinaud, P.: Visualisation chronologique des analyses ALCESTE: application Twitter avec l'exemple du hashtag # mariagepourtous. (2014)
27. Rosvall, M. and Bergstrom, C. T.: Mapping change in large networks. PloS one, 5(1), e8694 (2010).
28. Wang, X. and Cheng, Q. and Lu, W. Analyzing evolution of research topics with NEViewer: a new method based on dynamic co-word networks. Scientometrics, 101(2): 1253–1271 (2014)