

A new approach for automatizing the analysis of research topics dynamics: application to optoelectronics research

Jean-Charles Lamirel

Received: 27 January 2012 / Published online: 8 May 2012
© Akadémiai Kiadó, Budapest, Hungary 2012

Abstract The objective of this paper is to propose a new unsupervised incremental approach in order to follow the evolution of research themes for a given scientific discipline in terms of emergence or decline. Such behaviors are detectable by various methods of filtering. However, our choice is made on the exploitation of neural clustering methods in a multi-view context. This new approach makes it possible to take into account the incremental and chronological aspects of information by opening the way to the detection of convergences and divergences of research themes at a large scale.

Keywords Diachronic analysis · Clustering · Multiple viewpoint analysis · Unsupervised learning · Bayesian reasoning · Neural networks

Introduction

The literature taking into account the chronological aspect in information flows is focused on “DataStream” whose main idea is the “on the fly” management of incoming (i.e., not stored) data. In this context, the data that have been considered up to now are primarily physical measurements or Web usage data (connection, browsing, etc.). Applications on textual data (bibliographical databases, online news, etc.) are still stammering. Research on “DataStream” has been initiated, amongst other things, in 1996 by the DARPA through the TDT project (Allan et al. 1998). But the algorithms resulting from this work are intended to treat very large volumes of data (i.e., DataStream) and are thus not optimal for detecting emergent topics and for precisely following-up the evolution of a research field.

The numerous methods of clustering within this framework are organized as various families among which one can quote:

- Hierarchical methods like AHC (Voorhees 1986);
- Methods of the K -means family (MacQueen 1967);
- Density-based and graph partitioning methods;

J.-C. Lamirel (✉)

LORIA, INRIA-TALARIS Project, 615 r. du Jardin Botanique, 54600 Villers-lès-Nancy, France
e-mail: lamirel@loria.fr

- Neural clustering methods.

Most of these methods were initially defined in a non incremental way. However, in each of these families incremental versions were initiated, making it possible to take into account the temporal component of a data flow (Gaber et al. 2005). Among these methods, those which seem the most promising are the methods based on the data density and the neural methods.

One of our former studies (Lamirel and Al Shehabi 2004b) highlighted that most of the clustering methods, and especially the neural clustering methods, show high performance in the usual context of the analysis of homogeneous textual datasets. However, one of our more recent studies (Lamirel et al. 2010) has also clearly highlighted the drastic decrease of performance of all clustering methods, including classical methods, like K -means, as well as new incremental methods, when a heterogeneous or polythematic textual dataset, which can be considered as a static simulation of a time-evolving dataset, is taken as an input. Even if new incremental methods whose goal is to cope with the problems of actual methods by means of similarity measures which differ from classical Euclidean distance are promising, they are still under development (Lamirel et al. 2010).

To cope with the current defects of existing incremental clustering methods, an alternative approach for analyzing information evolving over time consists in performing diachronic analysis. This type of analysis is based on the application of a clustering method on data associated with two, or more, successive periods of time, and on the study of the evolution of the clusters contents and of their mappings between the different periods. For analyzing the evolution of the vocabulary describing the clusters of different periods, Schiebel et al. (2010) propose to construct a matrix of keywords comparison which is based on the percentage of keywords of one period which pre-exist in the clusters of another period. Thanks to this matrix, it is then possible for an expert of the domain to highlight different cluster behaviors: stability, but also merging or splitting. Even if it partly avoids exploiting the clustering methods in their critical area, an important limitation of this approach is that the process of comparison between clustering models must be achieved in a supervised way.

An alternative unsupervised solution has been proposed by Thijs and Glänzel (2010). It makes use of core documents to bridge clustering results issued from different time periods. The core documents are defined as the documents that combine high bibliographic coupling and high index terms similarities with other documents (Glänzel and Thijs 2010). In such a way, clusters of two time periods are considered as similar if they share a sufficient amount of references to the same core documents. Clusters are themselves built up using a co-clustering methodology mixing reference and contents information. This approach presents the advantage to be relatively independent of vocabulary changes between periods, but it necessitates exploiting referencing data.

Lamirel and Créhange (1994) firstly introduced the dynamic and unsupervised cooperation between clustering models in the context of information retrieval. This new approach has been originally used for analyzing the relevance of user's queries regarding the documentary database contents. It represents a major improvement of the basic clustering approach. From a practical point of view, the *MultiView Data Analysis paradigm* (MVDA), introduces the use of viewpoints associated with unsupervised Bayesian reasoning in the clustering process. Its main advantage is to be a generic paradigm that can be applied to any clustering method and that allows to enhance the quality and the granularity of data analysis while suppressing the noise that is inherent to a global approach.

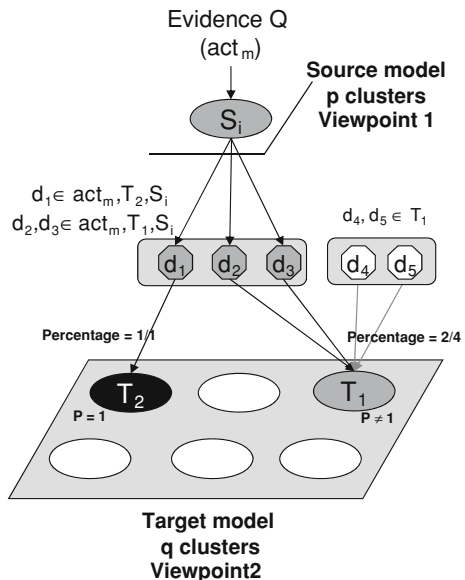
The principle of the MVDA paradigm is thus to be constituted by several clustering models that have been generated from the same data or even from data that share the same overall description space. Each model is issued from a specific viewpoint and can be generated by any clustering method. The relation between the models is established through the use of an inter-model communication mechanism based itself on unsupervised Bayesian reasoning (see Fig. 1).

One of the assets of this paradigm is that there are very various ways to define viewpoints. One possible way consists in separating the description space of the data into different subspaces corresponding to different criteria of analysis. As an example, web pages can be simultaneously described using 3 different viewpoints represented by: (1) a keyword vector issued from the page full-text extraction process, (2) an in links vector, (3) an out links vector. A multi-view analysis that is performed on such data can thus highlight general relationships existing between the semantic domain of the content and the ones of the references. In the Webometrics domain, such a methodology can therefore help “thematising” groups of links, while maintaining the opportunity to figure out specific relationships existing inside each separate domain.

The MVDA paradigm has thus been chosen as one of the two reference approaches of the IST-EISCTES European project (François et al. 2003). Its most recent version has opened new perspectives for unsupervised link analysis in Webometrics by making it possible to automatically combine textual and referencing information (Al Shehabi and Lamirel 2006).

The MVDA paradigm also represents a challenging paradigm in the context of the analysis of time varying information. Hence, it allows defining efficient and precise strategies for unsupervised diachronic analyses based on the mapping into separate viewpoints of the clustering models related to the different time periods. In “A new approach for analyzing time-varying information” section, we highlight how to exploit the principles of MVDA to automatically perform such analyses. “Experimentation and

Fig. 1 The MVDA inter-models communication principle



results” section describes our first experiment and its results. “Conclusion” section draws our conclusion and perspectives.

A new approach for analyzing time-varying information

Basic principle

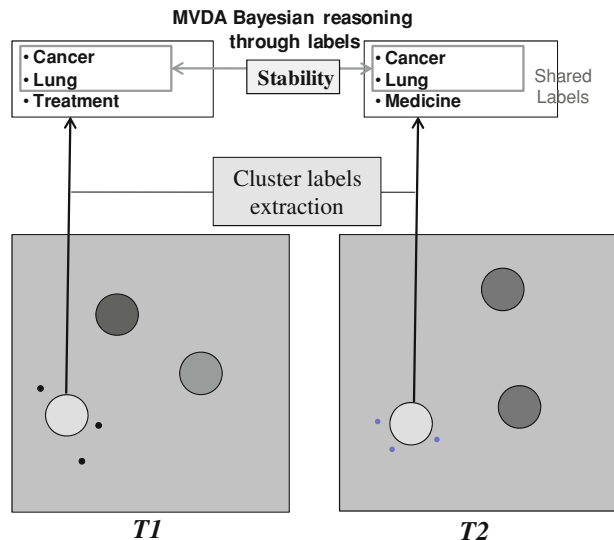
Analyzing the difference between time periods concerns different kinds of topics changes or similarities that could occur between the periods (appearing topics, disappearing topics, splitting topics, merging topics, stable topics). For achieving comparison between two time periods, a *label-based diachronic approach* relying both on data properties and on the MVDA paradigm can be thus defined. Thanks to this approach, a further step of cluster labeling is achieved after the construction of the clustering model for each time period. The purpose of the labeling step is to figure out which peculiar properties or labels can be associated to each cluster of a given time period. The identification of the topics relationships between two time periods is then achieved through the use of Bayesian reasoning relying on the extracted labels that are shared by the compared periods (see Fig. 2).

The *label-based diachronic approach* must also rely altogether on sound clustering results and on very efficient cluster labeling techniques in order to precisely identify the topics and their potential changes between time periods. The use of reliable cluster evaluation and labeling strategies becomes thus a central point in this methodology. Our set of exploited techniques is described in the next section.

Cluster validation and cluster labeling methodologies

When anyone aims at evaluating clustering results, or even comparing clustering methods, he will be faced with the problem of the choice of reliable clustering quality indexes. The classical evaluation indexes for the clustering quality are based on the intra-cluster inertia

Fig. 2 The label-based approach



and the inter-cluster inertia (Davies and Bouldin 1979). Thanks to these two indexes, a clustering result is considered as good if it possesses low intra-cluster inertia as compared to its inter-cluster inertia. However, as shown in (Lamirel et al. 2004a), the distance based indexes are often strongly biased and highly dependent on the clustering method. Thus, they cannot be easily exploited for comparing different methods, or even different clustering results issued from data whose description spaces have different sizes. Moreover, as it has been also shown in (Ghribi et al. 2010), they are often properly unable to identify an optimal clustering model whenever the dataset is constituted by complex data that must be represented in a both highly multidimensional and sparse description space, as it is often the case with textual data. To cope with such problems, our unsupervised Recall/Precision and F-measures indexes exploit the properties of the data associated to each cluster after the clustering process without prior consideration of clusters profiles. Their main advantage is thus to be independent of the clustering methods and of their operating mode.

Let us consider a set of clusters C resulting from a clustering method applied on a set of data D , the local unsupervised *Recall* (R_c^f) and local unsupervised *Precision* (P_c^f) indexes for a given feature f of the cluster c can be expressed as:

$$R_c^f = \frac{|d_c^f|}{|D^f|}, \quad P_c^f = \frac{|d_c^f|}{|D_c|} \tag{1}$$

where d_c^f is the set of data having the feature f in c , D_c represents the set of data in c , D^f the set of data with feature f .

Then, for estimating the overall clustering quality, the averaged *Macro-Recall* (MR) and *Macro-Precision* (MP) indexes can be expressed as:

$$MR = \frac{1}{|\bar{C}|} \sum_{c \in \bar{C}} \frac{1}{|F_c|} \sum_{f \in F_c} R_c^f, \quad MP = \frac{1}{|\bar{C}|} \sum_{c \in \bar{C}} \frac{1}{|F_c|} \sum_{f \in F_c} P_c^f \tag{2}$$

Where F_c is the set of prevalent features of the cluster c that are described as:

$$F = \left\{ f \in d, d \in c \mid \bar{W}_c^f = \text{Max}_{c' \in C} \left(\bar{W}_{c'}^f \right) \right\} \tag{3}$$

where \bar{C} represents the set of prevalent clusters extracted from the clusters of C , which verifies:

$$\bar{C} = \{c \in C \mid F_c \neq \emptyset\} \text{ and } \bar{W}_c^f = \frac{\sum_{d \in c} W_d^f}{\sum_{c' \in C} \sum_{d \in c'} W_{d'}^f} \tag{4}$$

where W_x^f represents the weight of the feature f for element x .

Similarly to IR, the *F-measure* could be used to combine averaged *Macro-Recall* and *Macro-Precision* results. Moreover, we have demonstrated in (Lamirel et al. 2004) that if both values of averaged *Macro-Recall* and *Macro-Precision* reach the unity value, the prevalent set of clusters \bar{C} represents a Galois sub-lattice. Therefore, the combination of this two measures enables to evaluate to what extent a numerical clustering model can be assimilated to a Galois lattice-based natural classifier.

Macro-Recall and *Macro-Precision* indexes (Eq. 2) can also be considered as cluster-oriented measures because they provide average values of *Recall* and *Precision* for each cluster. They have opposite behaviors according to the number of clusters. Thus, these indexes permit to estimate in a global way an optimal number of clusters for a given

method and a given dataset. The best data partition, or clustering result, is in this case the one which minimizes the difference between their values.

In a complementary way, the role of clusters labeling is to highlight the prevalent features of the clusters associated to a cluster model at a given time. Labeling can be thus used both for visualizing or synthesizing clustering results (Lamirel et al. 2008), for optimizing the learning process of a clustering method (Attik et al. 2006) and for highlighting the content of the individual clusters. Some efficient cluster feature relevance indexes can be derived from our former quality indexes, using a probabilistic approach (Lamirel et al. 2010). We detail hereafter their basic definition.

The *Feature Recall* (FR_c) derives directly from (Eq. 5). For a feature f of a cluster c , it is expressed as:

$$FR_c(f) = \bar{W}_c^f \tag{5}$$

The *Feature Precision* (FP_c) can be expressed as:

$$FP_c(f) = \frac{\sum_{d \in c} W_c^f}{\sum_{f' \in d, d \in c} W_c^{f'}} \tag{6}$$

Consequently, the set of labeling features, or labels, L_c that can be considered as prevalent for a cluster c can be expressed as the set of endogenous cluster data features (i.e., unsupervised labels), or even exogenous cluster data features (i.e., external labels or supervised validation labels), which verifies:

$$L_c = \{f \in d, d \in c | FF_c = \text{Max}(FF_{c'})\} \tag{7}$$

where the *Feature F-measure* (FF_c) of a feature f of a cluster c can be defined as:

$$FF_c(f) = \frac{2(FR_c(f) \times FP_c(f))}{FR_c(f) + FP_c(f)} \tag{8}$$

As soon as *Feature Recall* is equivalent to the conditional probability $P(c|p)$ and *Feature Precision* is equivalent to the conditional probability $P(p|c)$, this former labeling strategy can be classified as an expectation maximization approach with respect to the original definition given by Dempster et al. (1977).

Experimentation and results

In the context of the PROMTECH IST project, Schiebel et al. (2010) have chosen to start from the INIST PASCAL database and to rely on its classification plan to analyze the dynamics of the various identified topics. They firstly employed a simple search strategy, consisting in the selection of the bibliographic records having at the same time a code in Physics, and a code corresponding to a technological field of application. The two selected applicative fields are the Engineering and the Life sciences (Biological Sciences and Medicine). By successive selections, combining statistical techniques and expert approaches, the authors released the 10 promising sets of themes. For their diachronic experiments, they finally selected the set of themes of the optoelectronic devices because this field is one of the most promising of the last decade. 3890 records related to these topics were thus selected in the PASCAL database.

Similarly to the former authors, our approach consisted in cutting out the resulting PROMTECH corpus in two periods, (1996–1999: period 1) and (2000–2003: period 2), to carry out for each one an automatic classification by using the content present in the bibliographic records. The structure of the records makes it possible to distinguish the titles, the summaries, the indexing keywords and the authors as representative of the contents of the information published in the corresponding article. In our experiment, the research topics associated with the indexing keywords field are solely considered. For each year period, a specific dataset is generated. For that purpose, a set of pre-processing steps is applied to the keywords field of the corresponding records in order to obtain a weighted vector representation of the information it contains. Keywords of overall frequency less than 3 are firstly removed from the record descriptions. 1797 records indexed by 1256 keywords are consequently kept in period 1, and 2074 records indexed by 1352 keywords in period 2. In a further step, the resulting vectors associated to each record are weighted using an IDF weighting scheme (Robertson and Sparck Jones 1976) in both periods in order to decrease the effect of more frequent indexes.

The clustering of the datasets associated to the two periods is achieved by the use of the “Growing Neural Gas” (GNG) neural clustering method (Frizke 1995), which has been proven to be especially efficient on thematically homogeneous textual data (Lamirel et al. 2011). Nonetheless, many different experiments are done for each period, letting vary the number of clusters. The best (i.e., optimal) clustering model of each period regarding the optimal compromise between the values of averaged *Macro-Recall/Precision* indexes defined by (Eq. 2) is finally kept. In the end, the labels of the clusters of the best models are identified in an unsupervised way by the method of cluster feature maximization described by (Eq. 7).

To enhance the quality of the comparison, three different kinds of post-processing steps are applied on the resulting optimal clustering models.

- (1) A threshold of 3 regarding to the clusters size is applied to discard non significant clusters (chunk clusters).
- (2) To suppress the noise related to the cluster contents, we have applied a complementary strategy of identification of clusters connected label groups. This strategy operates on the labels formerly associated to the clusters. The labels that are co-occurring in the data associated to a cluster are gathered in the same group of connected labels. Noisy labels will thus tend to form isolated groups in a cluster. In a correlative way, noisy data associated to a cluster will tend to only include labels belonging to isolated groups. Moreover, this strategy permits to split clusters which include more than one important group of connected labels into homogeneous independent sub-clusters. The example of Fig. 3 shows that such a process permits to highlight both cluster noise (small-sized label groups) and independent sub-topics (large label groups) in one cluster.
- (3) The labels of the cluster connected label groups of a period whose *Feature F-measure* (Eq. 8) is under the average *Feature F-measure* of the overall clustering model of the period are considered as non significant and thus discarded from the concerned groups. The groups with no remaining labels are discarded as well.

The general results of the formerly described process are reported in Table 1. The table also highlights some important surface changes occurring on the datasets characteristics between the periods, like the increase of publication volume, the enrichment of the paper descriptions (higher average number of labels per documents) and the specialization of the topics (lower average number of overlapping labels), in the second period.

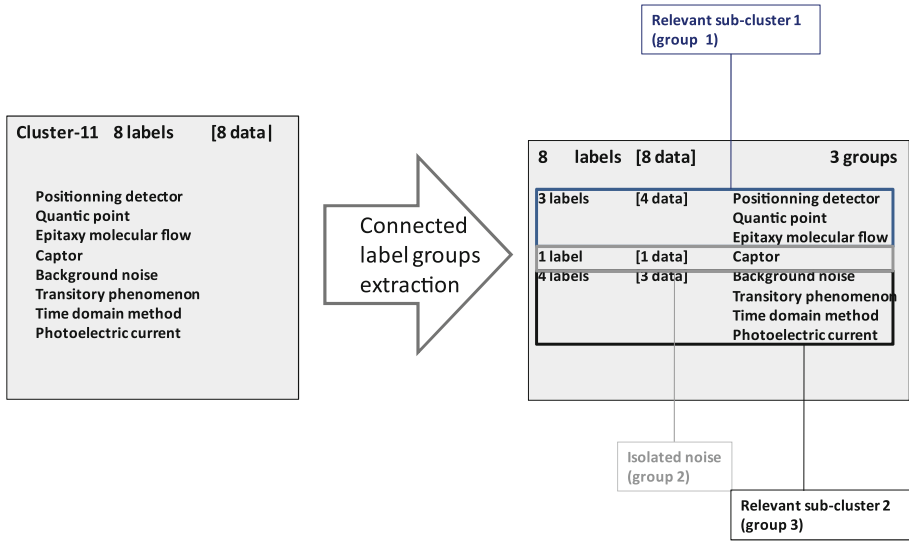


Fig. 3 Example of cluster connected label groups (i.e., connected feature groups) extraction

Table 1 Overall period characteristics (datasets) and clustering optimized results (GNG)

Time period	NBR docs	NBR labels	NBR labels (Freq. > 3)	Av. labels/doc.	Total overlap labels	Av. overlap labels/doc.	NBR clusters (optimal)	NBR clusters (size > 3)	NBR labels groups (valid)
1996–1999	1797	1256	903	8.12	903	0.503	42	40	43
2000–2003	2074	1352	947	8.43	947	0.466	49	48	50

To compute the probability of matching between clusters belonging to two time periods, we slightly modify the standard computation of the Bayesian inference provided by the original MVDA model (Al Shehabi and Lamirel 2004). The new computation is expressed as:

$$P(t|s) = \frac{\sum_{f \in L_s \cap L_t} FF_t(f)}{\sum_{f \in L_t} FF_t(f)} \tag{9}$$

where s represents a cluster of the source period, t a cluster of the target period, L_x represents the set of labels associated to the cluster x , using the cluster feature maximization approach defined by (Eq. 7), and $L_x \cap L_y$ represents the common labels, which can be called the label matching kernel between the cluster x and the cluster y .

The average matching probability $P_A(S)$ of a source period cluster can be defined as the average probability of activity generated on all the clusters of the target period clusters by its associated labels:

$$P_A(S) = \frac{1}{|Env(s)|} \sum_{t \in Env(s)} P(t|s) \tag{10}$$

where $Env(s)$ represents the set of target period clusters activated by the labels of the source period cluster s .

The global average activity A_s generated by a source period model S on a target period model T can be defined as:

$$A_s = \frac{1}{|S|} \sum_{s \in S} P_A(s) \tag{11}$$

Its standard deviation can be defined as σ_s .

The similarity between a cluster s of the source period and a cluster t of the target period is established if the 2 following similarity rules are verified:

$$P(t|s) > P_A(s) \text{ and } P(t|s) > A_s + \sigma_s \tag{12}$$

$$P(s|t) > P_A(t) \text{ and } P(s|t) > A_t + \sigma_t \tag{13}$$

Cluster splitting is verified if there is more than one cluster of the target period which verifies the similarity rules (13) and (14) with a cluster of the source period. Conversely, cluster merging is verified if there is more than one cluster of the source period which verifies the similarity rules (13) and (14) with a cluster of the target period.

Clusters of the source period that do not have similar cluster on the target period are considered as vanishing clusters. Conversely, clusters of the target period that do not have similar cluster on the source period are considered as appearing clusters.

Table 2 summarizes the results of our experiment of time periods comparison, in terms of identification of correspondences and differences. For a given period, the number of clusters implied in the comparison corresponds to its optimal number of clusters. It should be noted that the number of cluster splitting of the first period into the second period is more important than the converse number of merging into this latter period, which indicates a diversification of the research in the field of optoelectronics during the second period.

Finally, clusters similarity and divergence reports are automatically build up for presentation to the analysts. Each report includes one cluster of each period, whenever it is a similarity report, or one cluster of a single period, whenever it is a divergence report (i.e., an appearing or disappearing topic). In the case of a similarity report, the similarities between the clusters of the compared periods are identified by shared groups of labels(i.e., matching kernels), extracted from the clusters maximized features (Eq. 7), which we have also named core-labels. These core-labels illustrate in a specific way the nature of the temporal correspondences. The labels of the clusters of each period which does not belong to the matching kernel of a similarity report are also considered separately. They are used to figure out small temporal changes occurring in the context of an overall topic similarity between two periods. Said labels are displayed in decreasing order of their Feature F-measure difference with the alternative periods. If a specific label of a given period does not exist in the alternative period, or if its *Feature F-measure* is under the *Average Feature F-measure* (\overline{FF}) of the overall clustering, it is marked as absent of the latter period (see Fig. 4).

Table 2 Summary of the time comparison results

Time period	NBR groups	NBR match	NBR disappear	NBR appear	NBR split	NBR merge
1996–1999	43	33	10	–	7	–
2000–2003	50	38	–	12	–	3

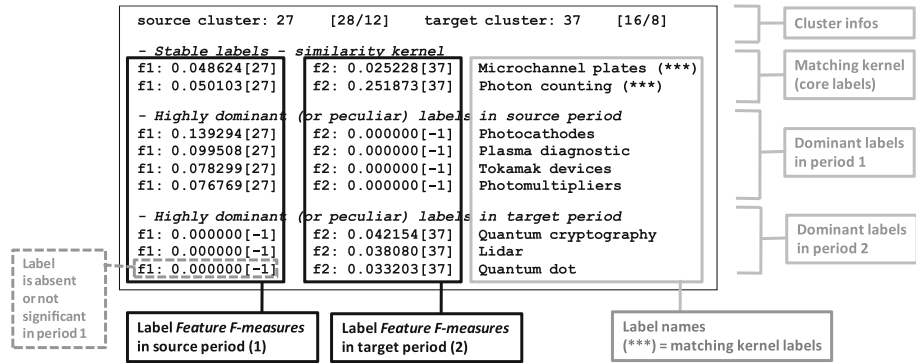


Fig. 4 Structure of a basic (i.e., rough) similarity report

In a final step, reports are slightly adapted using an automatic process in order to highlight the most important information that they provide. For similarity reports, an automatic core label migration process is used to better figure out to which period each core label is mostly related. The migration of one core label to a given period is applied if the *Feature F-measure* of this label is twice more important in this period than in the other one. Moreover, for all reports and for all labels, *Feature F-measure* difference gradations are computed based both on the *Average Feature F-measure* (\overline{FF}) and on the *Feature F-measure* standard deviation (σ_{FF}) of the clustering. For a given label l , the gradation k is thus computed as:

$$k = (|FF_s(l) - FF_t(l)|) \% (\overline{FF} + \sigma_{FF}) \tag{14}$$

where $FF_s(l)$ represents the *Feature F-measure* of the label l in the source period s , and $FF_t(l)$ its *Feature F-measure* in the target period t .

In all the reports, the *Feature F-measure* difference gradations are highlighted for $k = 2$ and $k > 3$. The complete example of the adaptation of the similarity report presented at Fig. 4 is given at Fig. 5.

The results produced by our automated approach of comparison of time periods were finally compared with those of the analysis carried out by experts of the domain on the partitions produced over separated periods of time in the former experiment of Schiebel et al. (2010).

Said analysis has mainly highlighted the following facts:

- (1) General set of topics of the studied corpus corresponded to the optoelectronic devices containing mineral or organic semi-conductors,
- (2) The research and applications of optoelectronics evolved from the field of the “photo-detectors” (probes, measuring instruments, etc.), in period 1, to the field of the “electroluminescent diodes”, in period 2.

The above-mentioned conclusions present the disadvantage to provide only surface information on the potential topics evolutions. As is it shown in the upcoming parts, the examination of the reports of similarities as well as those of divergences provided by our new diachronic method of analysis shows that it is possible to obtain both synthetic and precise conclusions, together with clear indications of tendencies (growth or decrease) in a unsupervised way, while preserving the possibility of observing general orientations, such as those expressed by the experts of the PROMTECH project.

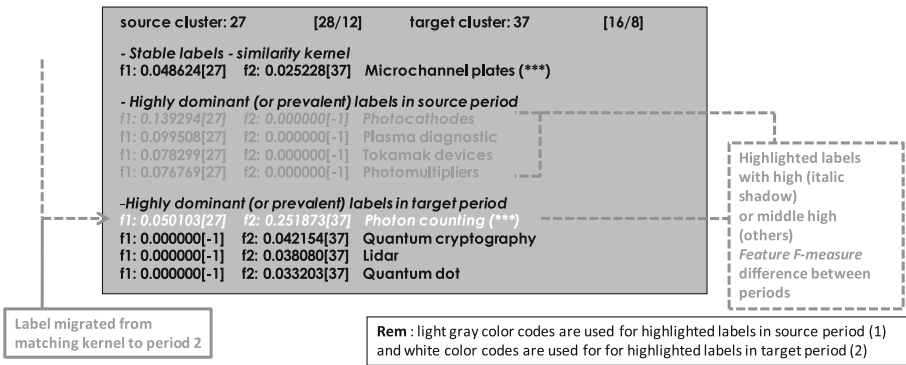


Fig. 5 Structure of an adapted (i.e., finalized) similarity report

For the sake of validation, all the adapted similarity and divergence reports have been made available to a pool of French INIST librarians specialized in the optoelectronics domain. Looking to these reports, the librarians clearly point out that the latter, whilst maintaining both a sufficiently general description level and an accurate contextual background, make it possible to very precisely reveal the tremendously rich developments of the research topics in the optoelectronic domain during the 1996–2003 period, altogether, from the theoretical studies to the practical applications (from optical polymers to polymer films (Fig. 6), from surface emitting lasers or semi-conductor lasers to vertical cavity lasers or VCSEL [(Fig. 7), etc.] from the exploitation of new chemical components to the production of new devices (from gallium arsenide to quantum well devices [(Fig. 8), etc.] or new semi-conductors types (from silicon compounds to amorphous semi-conductors (Fig. 9), from gallium compound to wide band gap semiconductors [(Fig. 10), raise of exploitation of germanium, etc.], or the slight emerging of new semiconductors structures or organization which might become autonomous or self-assembling structures (Fig. 11).

Another interesting point concerning the behavior of the proposed method is that the vocabulary changes which are related to slight or contextual thematic evolutions might well be merged in the same similarity report, without thus associating those changes to different contexts, or even missing to detect them. As an example, the similarity report of Fig. 12 helps to confirm the progressive evolution of the optoelectronics domain from

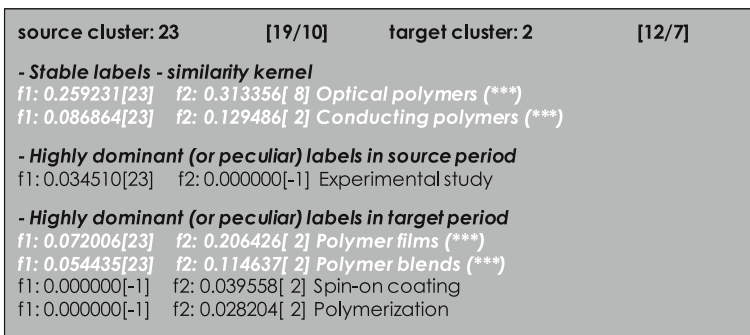


Fig. 6 Similarity report related to the strong development of polymer blends and films

source cluster: 14	[18/6]	target cluster: 14	[29/7]
- Stable labels - similarity kernel			
f1: 0.035721[14]	f2: 0.041813[14]	Surface emitting laser (***)	
- Highly dominant (or peculiar) labels in source period			
f1: 0.148633[14]	f2: 0.057783[14]	Semiconductor laser (***)	
f1: 0.078080[14]	f2: 0.033436[14]	Laser diodes (***)	
f1: 0.026498[14]	f2: 0.000000[-1]	Surface	
f1: 0.026027[14]	f2: 0.000000[-1]	Waveguide laser	
- Highly dominant (or peculiar) labels in target period			
f1: 0.000000[-1]	f2: 0.068895[14]	Light sources	
f1: 0.000000[-1]	f2: 0.039487[14]	Laser beam applications	
f1: 0.000000[-1]	f2: 0.029637[14]	Vertical cavity laser	
f1: 0.000000[-1]	f2: 0.025024[14]	VCSEL	

Fig. 7 Similarity report related to the slight emergence of vertical cavity lasers (or VCSEL)

source cluster: 41	[10/8]	target cluster: 47	[14/7]
- Stable labels - similarity kernel			
f1: 0.184306[41]	f2: 0.105946[47]	Infrared detector (***)	
f1: 0.175818[41]	f2: 0.105074[47]	Semiconduct. quantum wells (***)	
f1: 0.090714[41]	f2: 0.129986[47]	Aluminium compound (***)	
- Highly dominant (or peculiar) labels in target period			
f1: 0.048841[41]	f2: 0.141585[47]	Gallium arsenides (***)	
f1: 0.000000[-1]	f2: 0.080026[47]	Quantum well devices	

Fig. 8 Similarity report related to the emergence of quantum well devices for infrared detection

source cluster: 15	[22/9]	target cluster: 24	[20/8]
- Stable labels - similarity kernel			
f1: 0.038370[15]	f2: 0.044230[24]	Silicon compound (***)	
- Highly dominant (or peculiar) labels in source period			
f1: 0.043265[15]	f2: 0.000000[-1]	MIS structure	
f1: 0.026522[15]	f2: 0.000000[-1]	Diamond	
- Highly dominant (or peculiar) labels in target period			
f1: 0.061132[15]	f2: 0.222402[24]	Amorphous semiconductors (***)	
f1: 0.054647[15]	f2: 0.131473[24]	Hydrogen (***)	
f1: 0.000000[-1]	f2: 0.067403[24]	Selenium	
f1: 0.000000[-1]	f2: 0.039028[24]	Plasma CVD coatings	

Fig. 9 Similarity report related to the strong development of amorphous semiconductors based on hydrogen and selenium

source cluster: 28	[9/6]	target cluster: 42	[7/2]
- Stable labels - similarity kernel			
f1: 0.186473[28]	f2: 0.149314[42]	Gallium compound (***)	
- Highly dominant (or peculiar) labels in target period			
f1: 0.082561[28]	f2: 0.186924[42]	Wide band gap semiconduct. (***)	

Fig. 10 Similarity report related to the strong development of wide band gap semiconductors

source cluster: 37	[13/4]	target cluster: 15	[17/8]
- Stable labels - similarity kernel			
f1: 0.058058[37] f2: 0.071444[15] Semiconductor epitaxial layers			
- Highly dominant (or peculiar) labels in source period			
f1: 0.178256[37] f2: 0.041720[15] Semiconductor growth (***)			
f1: 0.098386[37] f2: 0.035582[15] Molecular beam epitaxy (***)			
f1: 0.089059[37] f2: 0.033129[15] VPE (***)			
- Highly dominant (or peculiar) labels in target period			
f1: 0.000000[-1] f2: 0.048815[15] Leakage current			
f1: 0.000000[-1] f2: 0.041294[15] Semiconductor superlattices			

Fig. 11 Similarity report related to the slight emergence of semiconductors superlattices

source cluster: 24	[23/9]	target cluster: 33	[27/13]
-No stable labels			
- Highly dominant (or peculiar) labels in source period			
f1: 0.266901[24] f2: 0.068167[33] Optical fabrication (***)			
f1: 0.045998[24] f2: 0.000000[-1] Integrated circuit technology			
f1: 0.042258[24] f2: 0.000000[-1] Interference filter			
f1: 0.041773[24] f2: 0.000000[-1] Semiconductor technology			
- Highly dominant (or peculiar) labels in target period			
f1: 0.077799[24] f2: 0.213749[33] Optical design techniques (***)			
f1: 0.000000[-1] f2: 0.055834[33] Aberrations			
f1: 0.000000[-1] f2: 0.039636[33] Ray tracing			

Fig. 12 Similarity report related to changes in optoelectronics devices production scale

source cluster 16 is vanishing			
f1: 0.141849[16] f2: 0.000000[-1] Optical fiber			
f1: 0.078762[16] f2: 0.000000[-1] Fiber laser			
f1: 0.060706[16] f2: 0.000000[-1] Acoustooptical device			
f1: 0.049628[16] f2: 0.000000[-1] Ring laser			

Fig. 13 Divergence report related to vanishing of research on optical fibers

punctual developments to high scale industrial processes (evolution of the concept of optical fabrication to the one of optical design).

Thanks to the experts, automatic reports of divergence between periods, materializing disappearances or emergences of subjects (topics), play the role of highlighting more important changes in the domain than the ones that could be highlighted by the similarity reports. The complete disappearance of research on optical fibers during the second period is thus clearly highlighted (Fig. 13). Conversely, the full appearance of new research works on phosphorescence, jointly with the very significant development of those on fluorescence, is also correctly highlighted in such a way (Fig. 14). Last but not least, the emergence of research works on high-resolution optical sensors and on their integration on chips, directly related to the important development of digital camera market in the second period (Fig. 15), as well as the emergence of promising research on new generation of high efficiency optical nano-transistors (quantum dots) (Fig. 16) are also accurately figured out by the divergence reports.

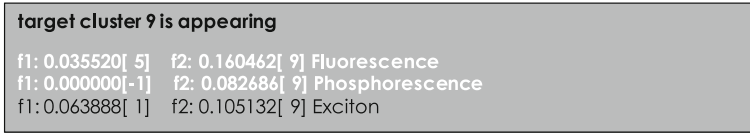


Fig. 14 Divergence report related to emergence of new research in phosphorescence and the highly significant development of the one on fluorescence

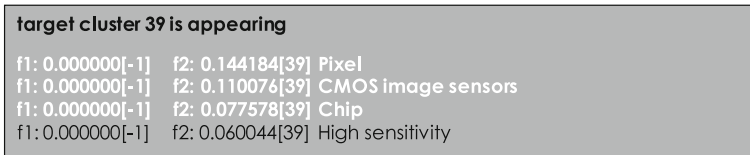


Fig. 15 Divergence report related to the strong emergence of the development and integration of high sensitivity image sensors

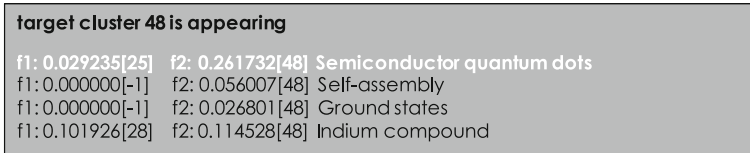


Fig. 16 Divergence report related to very highly significant development of research on new small scale transistors with high efficiency (quantum dots)

Table 3 Evolution of the paper count related to the emerging and disappearing topics between the two time periods

Cluster ref.	Topic main keywords	Feature F-measure difference between periods	Paper count in period 1 (1996–1999)	Paper count in period 2 (2000–2003)
16	Optical fiber	0.14	28	13
9	Fluorescence	0.12	18	36
39	CMOS image sensors	0.11	0	18
39	Pixel	0.14	0	26
48	Semiconductor quantum dots	0.23	16	74

An objective validation of the results of the proposed approach can also be achieved by looking up to the evolution of the count of the papers related to the main emerging or disappearing topics highlighted by the approach between the two periods. For that purpose we use the top-ranked keywords (i.e., the maximized ranked features or labels) associated with said topics and search for the related papers in the exploited dataset. Table 3 synthesizes the resulting count of such papers in each period. Figure 17 proposes year-based evolution curves of the above mentioned paper counts. Both techniques clearly demonstrate the efficiency of the method to detect main changes. They also highlight the efficiency of the related Feature F-measure to quantify the amount of change between the periods.

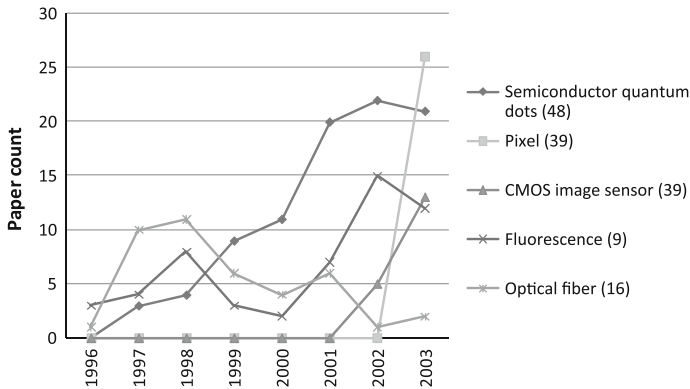


Fig. 17 Year-based evolution of the paper count related to the emerging and disappearing topics

The complete results provided by the method cannot be presented here. They have thus been made available at a specific address (Results 2011).

However, one might already remark that such a topic change mining process using single keywords information was until now impossible to reach with the existing methods, which, in addition, remained at most semi-supervised. It thus makes this new approach particularly promising.

Conclusion

We show in this paper the feasibility of an unsupervised incremental approach based on a time-step analysis of bibliographical data. This analysis has been carried out thanks to the exploitation of a specific model of data analysis managing multiple views on the data, namely the MVDA model. It was also based on the exploitation of original and stable measures for evaluating the quality and the coherence of the clustering results, and even for precisely synthesizing clusters content. To our knowledge, our approach represents the first approach that has been proposed for fully automatizing the process of analysis of time evolving textual information using single textual content. Our experimentation proved that this approach is reliable and that it can produce precise and significant results on a complex dataset constituted of bibliographic records, like a European reference dataset related to the research domain of optoelectronic devices.

In a near future, we plan to evaluate the application of this approach within the French INIST institute for the tasks of scientific and technological survey based on large scientific databases. Within this framework, the automated detection of the evolutions of research themes and groups is essential because it gives to the information analysts the possibility of carrying out exploratory studies at a large scale.

To help to figure out the robustness of our method to high vocabulary change, we finally plan to precisely compare it with the recent diachronic approaches based on co-clustering of lexical and bibliographical information (Thijs and Glänzel 2010).

Acknowledgments The author wishes to thank Pascal Cuxac (INIST-CNRS) for his valuable help in the results validation task.

References

- Al Shehabi, S., Lamirel, J.-C. (2004). Inference Bayesian Network for Multi-topographic neural network communication: A case study in documentary data. In *Proceedings of ICTTA*, Damas, Syria, April 2004.
- Al Shehabi, S., Lamirel, J.-C. (2006). Evaluation of collaboration between European universities using dynamic interaction between multiple sources. *Journal of Information Management and Scientometrics*, 1(3).
- Allan, J., Carbonell, J., Doddington, G., Yamron, J., Yang, Y. (1998). Topic detection and tracking pilot study, final report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, Virginia.
- Attik, M., Lamirel, J.-C., Al Shehabi, S. (2006). Clustering analysis for data with multiple labels. In *Proceedings of the The IASTED International Conference on Databases and Applications (DBA)*, Innsbruck, Austria, February 2006.
- Davies, D., & Bouldin, W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1, 224–227.
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood for incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, B39, 1–38.
- François, C., Hoffmann, M., Lamirel, J.-C., Polanco, X. (2003). *Artificial Neural Network mapping experiments*. EICSTES (IST-1999-20350) Final Report (WP 9.4), September 2003.
- Frizke, B. (1995). A growing neural gas network learns topologies. In G Tesauro, D. S Touretzky, T. K Leen (Eds.), *Advances in neural information processing systems 7* (pp. 625–632). Cambridge: MIT Press.
- Gaber, M., Zaslavsky, A., Krishnaswamy, S. (2005). Mining data streams: A review. *SIGMOD Record*, 34(2).
- Ghribi, M., Cuxac, P., Lamirel, J. C., Lelu, A. (2010). Mesures de qualité de clustering de documents: Prise en compte de la distribution des mots-clés. In *EvalECD'2010 Workshop*, Hamamet, Tunisia.
- Glanzel, W., & Thijs, B. (2010). Using 'core documents' for the representation of clusters and topics. *Scientometrics*, 88(1), 297–309.
- Lamirel, J.-C., & Al Shehabi, S. (2004b). Comparison of unsupervised neural clustering methods for mining Web and textual data. In *SCI 2004*, Orlando, FL, USA, July 2004.
- Lamirel, J.-C., Créhange, M. (1994). Application of a symbolico-connectionist approach for the design of a highly interactive documentary database interrogation system with on-line learning capabilities. In *Proceedings ACM-CIKM 94*, Gaithersburg, MD, USA, November 1994.
- Lamirel, J.-C., Al-Shehabi, S., François, C., & Hoffmann, M. (2004). New classification quality estimators for analysis of documentary information: Application to patent analysis and web mapping. *Scientometrics*, 60(3), 445–462.
- Lamirel, J.-C., Ta, A. P., & Attik M. (2008). Novel labeling strategies for hierarchical representation of multidimensional data analysis results. In *IASTED International Conference on Artificial Intelligence and Applications (AIA)*, Innsbruck, Austria, February 2008.
- Lamirel, J.-C., Boulila, Z., Ghribi, M., Cuxac, P. (2010). A new incremental growing neural gas algorithm based on clusters labeling maximization: application to clustering of heterogeneous textual data. In *Proceedings of IEA-AIE 2010*, Cordoba, Spain, June 2010.
- Lamirel, J.-C., Mall, R., Cuxac, P., Safi, G. (2011). Variations to incremental growing neural gas algorithm based on label maximization. In *Proceedings of IJCNN 2011*, San José, CA, USA, August 2011.
- MacQueen, J. B. (1967). Some methods of classification and analysis of multivariate observations. In L. Le Cam & J. Neyman (Eds.), *Proceedings 5th Berkeley Symposium in Mathematics, Statistics and Probability* (Vol 1, pp. 281–297), University of California, Berkeley, USA, 1967.
- Results (2011). <https://sites.google.com/site/diacresults2012>.
- Robertson, S. E., & Sparck Jones, K. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27, 129–146.
- Schiebel, E., Hörlesberger, Roche, I., François, C., & Besagni, D. (2010). An advanced diffusion model to identify emergent research issues: The case of optoelectronic devices. *Scientometrics*, 83(3), 765–781.
- Thijs, B., Glänzel, W. (2010). A new hybrid approach for bibliometrics aided retrieval. In *Sixth International Conference on Webometrics, Informetrics & Scientometrics, and 11th COLLNET Meeting*, Mysore, India, October 2010.
- Voorhees, E. M. (1986). Implementing agglomerative hierarchical clustering algorithms for use in document retrieval. *Information Processing and Management*, 22, 465–476.