Multi-View Data Analysis and Concept Extraction Methods for Text

Jean-Charles Lamirel

Synalp team – LORIA – Nancy – France, LORIA, INRIA-TALARIS Project,
615 rue du Jardin Botanique, 54600 Villers-lès-Nancy, France, <lamirel@loria.fr>

Jean-Charles Lamirel teaches information science and computer science at Strasbourg University, conducting research for the INRIA-TALARIS project in the French LORIA laboratory. His main domain of research is textual data mining based on neural networks. He has interests both in theoretical models for data mining and data mining applications. He is the creator of the concept of data analysis based on the multiple viewpoints paradigm (MVDA) with feature maximization and related metrics. Models based on these concepts have proven to outperform state-of-the-art models in the context of many challenging data mining applications.

Lamirel, Jean-Charles. **Multi-View Data Analysis and Concept Extraction Methods for Text.** *Knowledge Organization. 40(5)*, 305-319. 34 references.

ABSTRACT: In the process of textual information analysis, like in the domain of technological survey through patents analysis, or in the domain of emerging research tracking through research papers analysis, the complexity of the studied concepts and the accuracy of the questions to be answered may often lead the analyst to partition his reasoning into viewpoints. Most of the classical information analysis tools can only manage an analysis of the studied domain in a global way. The information analysis paradigm considered in this paper is an alternative paradigm called multi-view data analysis. This paradigm introduces the dimensions of viewpoints and dynamics into information analysis with its multi-view displays, its online generalization capabilities, and its inter-view communication process. The dynamic information exchange between views can be exploited, either by an analyst or in an unsupervised way, in order to perform cooperative deduction between several different analyzes that have been performed on the same data or on related data. This paper demonstrates the efficiency of a viewpoint-oriented analysis as compared to a global analysis in the domain of technological survey and research evaluation. Both objective and subjective quality criteria are taken into account for quality evaluation.

Received 17 June 2013; Revised 9 July 2013; Accepted 9 July 2013

1.0 Introduction

In complex data analysis tasks, like in those relying on textual data, the intrinsic complexity and multidimensionality of studied data have raised the development of tools that help the analysts to grasp data organization at a higher level. The huge volume of data to be managed forbids the single use of expert-based approaches leading to focus on the help of unsupervised methods. The first realistic attempt at such a technique was proposed by Kohonen (1982) through the self-organizing map (SOM) model in the 1980s. The basic principle of the SOM is that our knowledge organization at higher levels is created during learning by algorithms that promote self-organization in a spatial order (Hinton 1989; Kaski et al. 1998). Thus, the architecture form of the SOM network is based on the understanding that the representation of data features might assume the form of a self-organizing feature map that is geometrically organized as a grid or lattice. The SOM algorithm thus takes a set of N-dimensional data as input and maps it onto nodes of a two dimensional grid, resulting in an orderly feature map (Kohonen 1982). In the quantitative studies of science, the Kohonen self-organizing maps have been successfully used for mapping scientific journal networks (Campaniaro 1995) and also author co-citation data (White et al. 1998). Maps have also been successfully used for several other applications in the general area of data analysis such as clustering meeting output, for clustering socio-economic data (Varsis and Versino 1992), and for documentary database contents mapping and browsing (Lamirel 1995; Lamirel et al. 2000). Kaski et al. (1998) have implemented a specific adaptation of SOM, named WEBSOM, for the analysis of important document collections. WEBSOM's main characteristic is to include strategies for reducing the dimension of the entry document descriptions by using random projection

techniques applied on word histograms extracted from their contents. The WEBSOM method has been tested for patents abstract analysis (Kaski et al. 1998). Nevertheless, as this method only manages such an analysis in a global way, it can only provide the analyst with general overview of the concepts covered by the patents along with their interactions.

In a complex data analysis process, the fact that the usual models are only able to deal with one view of the data organization at a time might be considered a serious bottleneck to exploiting them for accurate mining tasks. The MVDA paradigm, which is presented in this paper, introduces thus the dimensions of viewpoints and dynamics into information analysis with its multi-view displays and its inter-view communication mechanism. The dynamic information exchange between views can be exploited by an analyst or in an unsupervised way in order to perform cooperative deduction between several different analyses, combining concept extraction and mapping that have been performed on the same data or on related data.

This paper will mainly focus on the study of the contribution of the viewpoint's oriented data analysis proposed by the MVDA model as compared to the global analysis proposed by the other models. Section 2 of the article deals with MVDA model presentation. Section 3 describes the associated protocols of resulting model quality measurement and resulting model feature extraction. The section describes a first experiment highlighting the added-value of a multi-view model in the domain of technological survey. Section 4 presents another experiment with high outcomes in the domain of diachronic research analysis. Finally, the conclusions and the perspectives are exposed.

2.0 The MVDA approach

Communication between self-organizing maps was first introduced in the information retrieval context for analysing the relevance of users' queries regarding the documentary database contents, through the MultiSOM model (Lamirel 1995). It represents a major amelioration of the basic Kohonen SOM model. From a practical point of view, the MultiSOM model introduces the use of viewpoints in the information analysis. The viewpoint building principle consists in separating the description space of the data into different subspaces corresponding to different criteria of analysis. Lamirel and Créhange (1994) first introduced the dynamic and unsupervised cooperation between clustering models in the context of information retrieval. This new approach was originally used for analyzing the relevance of user's queries regarding multiple semantic domains inherent in a documentary database contents. It represents a major improvement of the basic clustering approach. To the extent of both its operational and theoretical scope, the viewpoint-based approach has been extended recently by the same authors to other kinds of clustering models. Thus, the authors named the new paradigm multi-view data analysis paradigm (MVDA), enlarging the use of viewpoints associated with unsupervised Bayesian reasoning to data analysis and the data mining process, in general. Its main advantage is to be a generic paradigm that can be applied to any data analysis method and allowed to enhance the quality and the granularity of data analysis while suppressing the noise that is inherent in a global approach.

The principle of the MVDA paradigm is thus to be constituted by several data analysis models, mostly (but not exclusively) issued from clustering or unsupervised learning processes 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 data analysis 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).

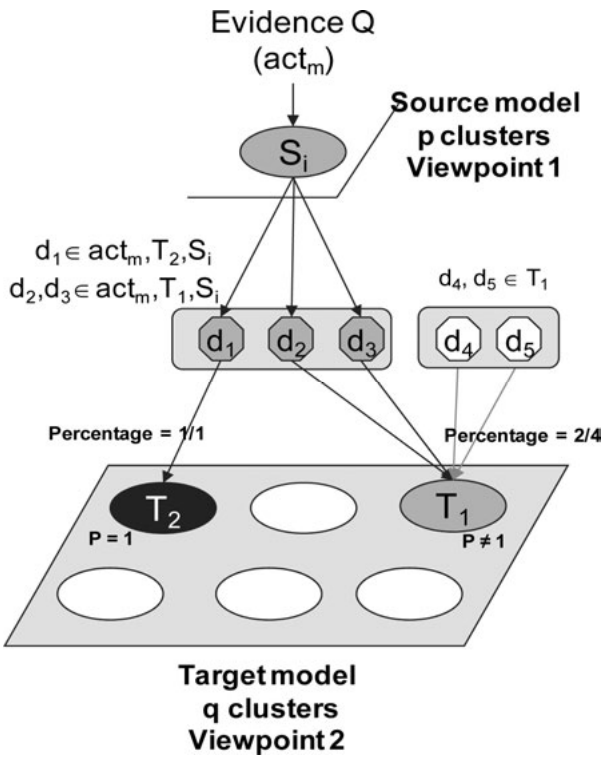

Figure 1. The MVDA inter-model (i.e., views) communication principle.

One of the assets of this paradigm is that there are various ways to define viewpoints. One possible way consists of separating the description space of the data into differ-

ent subspaces corresponding to different criteria of analysis. As an example, web pages can be simultaneously described using three different viewpoints represented by: 1) a keyword vector issued from the page full text extraction process; 2) an inlinks vector; or 3) an outlinks 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 those of the citations to and from the documents. In the Webometrics domain, such a methodology can therefore help to provide "conceptualization" to groups of links, while maintaining the opportunity to figure out specific relationships existing inside each separate domain.

The inter-view communication mechanism enables us to highlight semantic relationships between different concepts (materialized by classes or clusters) belonging to different viewpoints related either to the same data or to different data sharing common sets of features. In the MDVA paradigm, this communication is based on the use of the data or the features that have been projected onto each view as intermediary nodes or activity transmitters between views (see Fig. 1).

The view-map communication is established by a standard Bayesian inference network propagation algorithm which is used to compute the posterior probabilities of the target view's node T_k which inherited the activity (evidence Q) transmitted by its associated data or feature nodes. This computation can be carried out efficiently because of the specific Bayesian inference network topology that can be associated with the MVDA model. Hence, it is possible to compute the probability for an activity of modality act_m on the view node T_k which is inherited from activities generated on the source view. According to Lamirel and Al Shehabi (2004), this computation is achieved as follows:

$$(1)\quad P(act_m|T_k,Q) = \frac{\sum_{d\in act_m,T_k} Sim(d,S_d)}{\sum_{d\in T_k} Sim(d,S_d)}$$

such that S_d is the source node to which the data d has been associated, is the cosine correlation measure between the codebook vector of the data d and the one of its source node S_d and if it has been activated with the modality act_m from the source view.

The MVDA paradigm has thus been chosen as one of the two reference approaches of the IST-EISCTES European project (Francois 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 citation information (Al Shehabi and Lamirel 2006).

3.0 Clustering quality evaluation and feature extraction

3.1 Quality evaluation

When anyone aims at evaluating clustering (i.e., automatic concept extraction) results, or even comparing data analysis methods, he will be faced with the problem of choosing reliable quality indexes. The classical evaluation indexes for the clustering quality are based on the intra-cluster inertia and the inter-cluster inertia (Davies and Bouldin 1979). Thanks to these two indexes, a clustering result is considered good if it possesses low intra-cluster inertia as compared to its inter-cluster inertia. However, as shown in Lamirel et al. (2004), 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 with each cluster after the clustering process without prior consideration of cluster 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 () and local unsupervised Precision () indexes for a given feature f of the cluster c can be expressed as:

$$(2)\quad R_c^f = \frac{|d_c^f|}{|D^f|}, P_c^f = \frac{|d_c^f|}{|D_c|}$$

where is the set of data having the feature f in c, represents the set of data in c, the set of data with feature f.

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

$$(3)\quad \begin{aligned} MR &= \frac{1}{|\overline{C}|}\sum_{c\epsilon\overline{C}} \frac{1}{|F_c|}\sum_{f\epsilon F_c} R_c^f, \\ MP &= \frac{1}{|\overline{C}|}\sum_{c\epsilon\overline{C}} \frac{1}{|F_c|}\sum_{f\epsilon F_c} P_c^f \end{aligned}$$

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

$$F = \left\{ f \in d, d \in c \middle| \overline{W}_c^f = \mathrm{Max}_{c' \in C}\left(\overline{W}_{c'}^f\right) \right\} \quad (4)$$

where represents the set of prevalent clusters extracted from the clusters of C, which verifies:

$$\overline{C} = \{ c \in C \mid F_c \neq \emptyset \} \text{ and}$$
$$\overline{W}_c^f = \frac{\sum_{d \in c} w_c^f}{\sum_{c' \in C} \sum_{d \in c'} w_d^f} \quad (5)$$

where 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 represents a Galois sub-lattice. Therefore, the combination of this two measures enables us 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. 3) 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 us 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.

However, similar to the classical distance-based indexes, the main defect of the former indexes is that they do not permit us to detect degenerated clustering results whenever those jointly include a small number of heterogeneous or "garbage" clusters with large size and a big number of "chunk" clusters with very small size (Ghribi et al. 2010). To correct that, we have recently proposed constructing complementary feature-oriented indexes of Micro-Recall and Micro-Precision by averaging the Recall/Precision values of the peculiar properties independently of the structure of the clusters.

The averaged Micro-Recall (mR) and Micro-Precision (mP) indexes are expressed as:

$$mR = \frac{1}{|L|} \sum_{c \epsilon \overline{C}, p \epsilon S_C} \frac{|c_p|}{|P_p|} \quad ,$$
$$mP = \frac{1}{|L|} \sum_{c \epsilon \overline{C}, p \epsilon S_C} \frac{|c_p|}{|P_p|} \quad (6)$$

where L represents the size of the data description space.

A difference between the values provided by the Macro-indexes (i.e., cluster-structured averages) and those provided by the Micro-indexes (i.e., unstructured averages) indicates the presence of garbage clusters. However, it is possible to refer not only to the information provided by the indices Micro-Precision and Micro-Recall, but to the calculation of the Micro-Precision operated in a cumulative way. In the latter case, the idea is to give a major influence to large clusters which are the most likely to repatriate the heterogeneous information, and therefore, to significantly lower on their own the quality of the resulting partition. This calculation can be made as follows:

$$CMP = \sum_{i=|C_{inf}|,|C_{sup}|} \frac{1}{|\overline{C}_{i+}|^2} \sum_{c \epsilon \overline{C}_{i+}, p \epsilon S_C} \frac{|c_p|}{|P_p|} \Big/ \sum_{i=|C_{inf}|,|C_{sup}|} \frac{1}{|C_{i+}|^2} \quad (7)$$

where C_{i+} represents the subset of clusters of C for which the number of associated data is greater than i, and:

$$inf = \mathrm{argmin}_{c_i \in C} |c_i|, \quad sup = \mathrm{argmax}_{c_i \in C} |c_i| \quad (8)$$

3.2 Feature Selection

Complementary to overall clustering model evaluation, cluster labeling is a feature selection process whose role is to highlight the prevalent features of the clusters associated with a clustering model at a given time. Labeling can be thus used both for visualizing or synthesizing clustering results (Lamirel and Ta 2008), for optimizing the learning process of a clustering method (Attik et al. 2006) and for highlighting the content of the individual clusters. It can rely on endogenous data properties or on exogenous ones. Endogenous data properties represent those being used during the clustering process. Exogenous data properties represent either complementary properties or specific validation properties. 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 (5). For a feature f of a cluster c, it is expressed as:

$$FR_c(f) = \overline{W}_c^f \quad (9)$$

The Feature Precision (FP_c) can be expressed as:

$$\frac{\sum_{d \in c} w_c^f}{\sum_{f' \in d, d \in c} w_c^{f'}}$$

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:

$$(10)\ L_c = \{f \in d, d \in c \mid FF_c = Max(FF_{c'})\}$$

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

$$(11)\ FF_c(f) = \frac{2(FR_c(f) \times FP_c(f))}{FR_c(f) + FP_c(f)}$$

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).

In the preceding section, we have introduced the MVDA model after having previously presented the SOM model. In the current section, we have presented some original approaches for data analysis result evaluation. In the next sections, we shall then use two real examples, to make some of the notions more concrete. Hence, we argue that data visualization and data analysis with the help of communicating views represents an important added-value for analysis in technology watching tasks, as well as in science watch, and in knowledge discovery in databases. In the first example, we thus propose to compare the exploitation of the MDVA paradigm in the context of the SOM model, which we have called MultiSOM, with the classical SOM model. This first example more precisely relates to the use of the MVDA paradigm in an interactive way. In the second example, we show how the MDVA paradigm can be used in a fully automated and unsupervised way for precisely analyzing changes occurring in a focused research field, though the diachronic analysis of the content of bibliographical records.

4.0 Elaborated technological survey based on MVDA paradigm

4.1 Dataset preprocessing and knowledge maps building

Our experimental dataset consists of 1,000 patent abstracts related to oil engineering technology recorded during the year 1999. One of the first tasks related to patent survey based on a multiview approach is to prepare the data by isolating different semantic domains to which potential questions to be answered could belong. This preliminary task relies on the help of a domain expert. In our experimental context which is related to oil engineering, some examples of recurrent questions that have been highlighted by the domain expert are given below:

1: "Which are the relationships between the patentees?"
2: "Which are the advantages of the different oils?",
3: "Does a patentee work on a specific oil category or oil component, for which advantage and for which use?"
4: "Which technology is used by a given patentee without being used by another one?"
5: "Which are the main advantages of a specific oil component mentioned in one patent and have these advantages been mentioned in all the patents using this component?".

A more precise analysis carried out by the domain expert on the set whole of potential questions led this latter to highlight a partition into 4 mainsemantic domains corresponding to 4 different viewpoints which are:

1: Patentees
2: Title (often contains information on the specific components used in the patents)
3: Use
4: Advantages.

One of the main aims of the expert was to be able to use each viewpoint separately in order to get precise answers to domain-closed questions (like questions 1, 2) while maintaining the possibility of a multi-viewpoint communication in order to get answers to multi-domain questions (like questions 3, 4, 5) that might also contain negation marks (like question 4).

A fifth "global viewpoint" which represents the combination of all of the specific viewpoints is also considered in order to perform our comparison between a global clustering mechanism, of the standard SOM type (i.e., WEBSOM), and a pure viewpoint-oriented clustering mechanism, of the MultiSOM type.

The role of the second phase consists of mapping the four specific viewpoints highlighted by the domain expert in the preceding phase in four different views by taking into consideration the dataset structure. A preliminary task consists of obtaining the index set (i.e., the vocabulary set) associated with each view from the full text of the patent abstracts. This task has been itself divided into three elementary steps. At step 1, the structure of the patent abstracts is parsed in order to extract the subfields corresponding to the use and to the advantages viewpoints.[1] At step 2, the rough index set of each subfield is constructed by the use of a basic computer-based indexing tool (Jouve 1999). This tool extracts terms and noun

phrases from the subfield content according to a normalized terminology and its syntactical variations. It eliminates as well usual language templates. At the step 3, the normalization of the rough index set associated with each viewpoint is performed by the domain expert in order to obtain the final index sets.

The following task consists of building maps representing the different viewpoints, using the algorithm described in section 2. Before this step, a classical IDF Normalization step (Robertson and Jones 1976) is applied to the index vectors associated with the patent abstracts in order to reduce the influence of the most widespread terms in the indexes. For each specific viewpoint, a map of 10x10 nodes (clusters) is finally generated. Two global maps representing global unsupervised patent clustering, of the WEBSOM type (Kaski et al. 1998), are also constructed. The index sets of these maps represent the union of the index sets of all of the specific viewpoints. The maps only differ one from another by the number of their nodes. The first one (GlobMin) is constrained to have the same number of nodes as each viewpoint map (i.e., 100 nodes). The second one (GlobMax) is constrained to have the sum of the number of nodes of all of the viewpoint maps (i.e., it becomes a 20x20 map comprising 400 nodes). Table 1 summarizes the results of the patent indexation and the map building. A single viewpoint map resulting from the whole map building process is presented in Fig. 2.

4.2 Evaluation

In comparison with the standard mapping methods, such as principal component analysis (PCA), multidimensional scaling (MDS) or global SOM mapping (WEBSOM), the advantage of the multi-map is the inter-map communication mechanism that the MultiSOM environment provides to the user. Each map represents a viewpoint. Each viewpoint represents a subject category. The inter-map communication mechanism assists the user to cross information be-

	PATENTTEES	TITLE	USE	ADVANTAGES	WEBSOM (GlobMin)	WEBSOM (GlobMax)
Number of indexed documents (NID)	1000	1000	745	624	1000	1000
Number of rough indexes generated (NRI)	73	605	252	231	1395	1395
Number of final indexes (NFI)	32	589	234	207	1075	1075
Numbers of map nodes with members (/100)	28	55	57	61	89	238

Table 1. Summary of the results of patents indexation and map building.

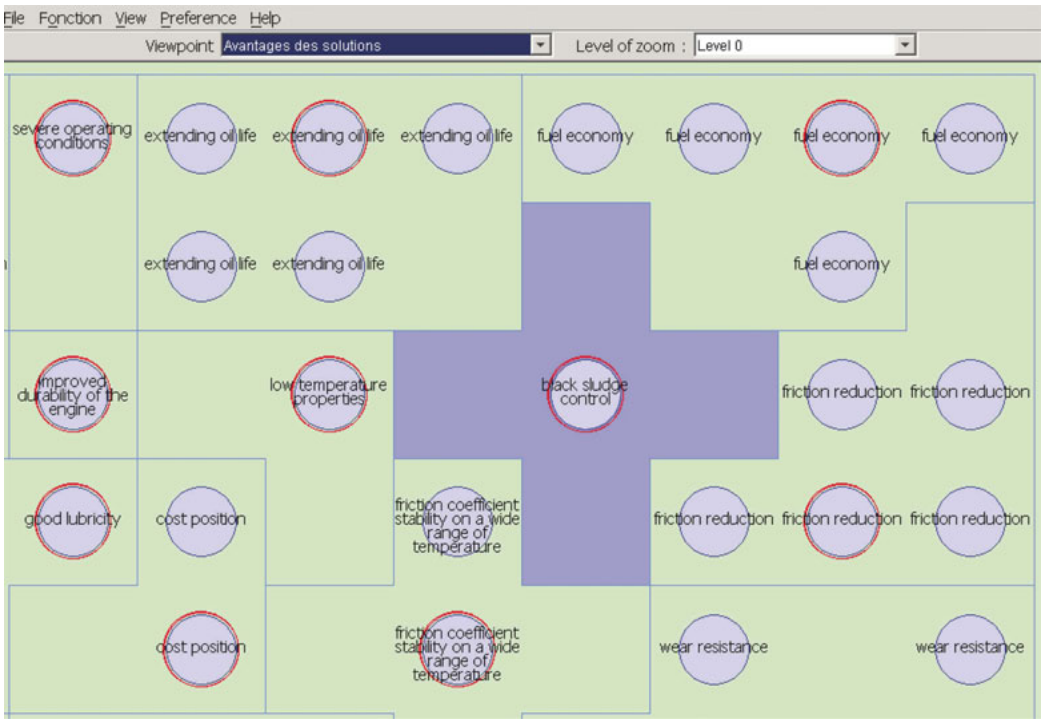

Figure 2. Example of a generated map. Partial view of a topographic map of 10 x 10 nodes.

tween the different viewpoints. In both cases, the responses of the system are given both through activity profiles on the maps and through patent examples associated with the most active node (i.e., concept) representatives of these maps. The estimation of the quality of thematic deduction can be achieved through an evaluation of the activity focalization on the target maps, as it is presented in Lamirel (1995). Fig. 4 illustrates a relevant thematic deduction between the four different viewpoints of the study.

The advantages of the MultiSOM method seem obvious to the domain expert; the original multiple viewpoints clustering approach of MultiSOM tends to reduce the noise which is inevitably generated in an overall clustering approach while increasing the flexibility and the granularity of the analyses. Moreover, with a global clustering method, like WEBSOM, important relationships between some sub-concepts are hidden in the cluster profiles and therefore are very difficult to grasp precisely. The expert found more than 35 such important relationships by the use of the MultiSOM method. A simple example is given by the comparison of Fig. 2 and Fig. 3. Other examples of more elaborated concept relationships that can be obtained only by the MultiSOM inter-map communication mechanism are provided in Results (2013a). Finally, the expert argued that the possibility of interactively activating, positively or negatively, the nodes (i.e., the concepts) on the maps represents a great help for very precisely tuning an interactive analysis process.

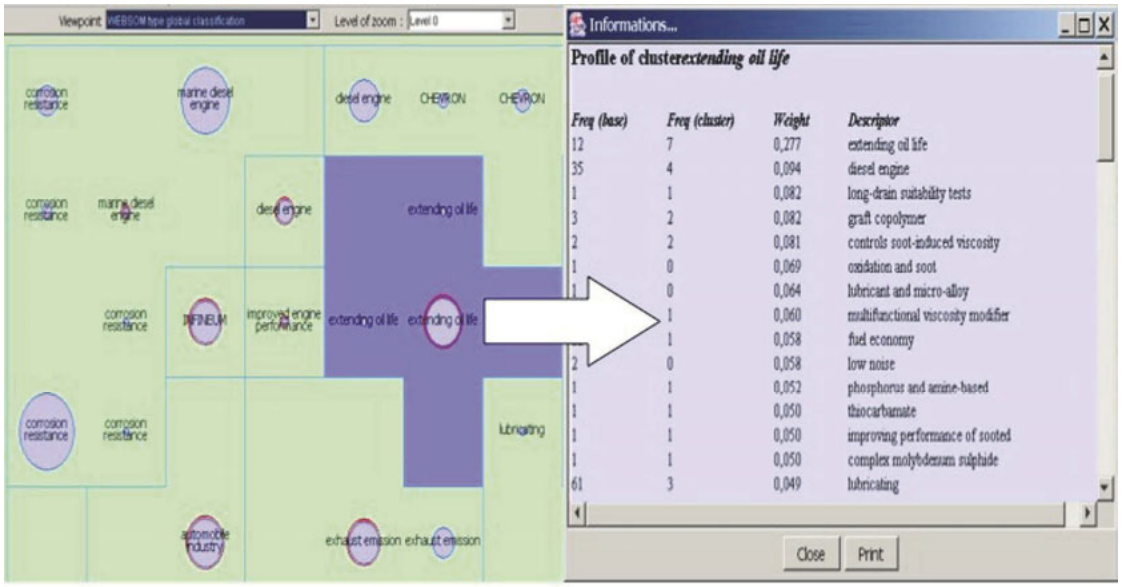

Figure 3. Results of a WEBSOM-like global mapping of 10x10 nodes (GlobMin).

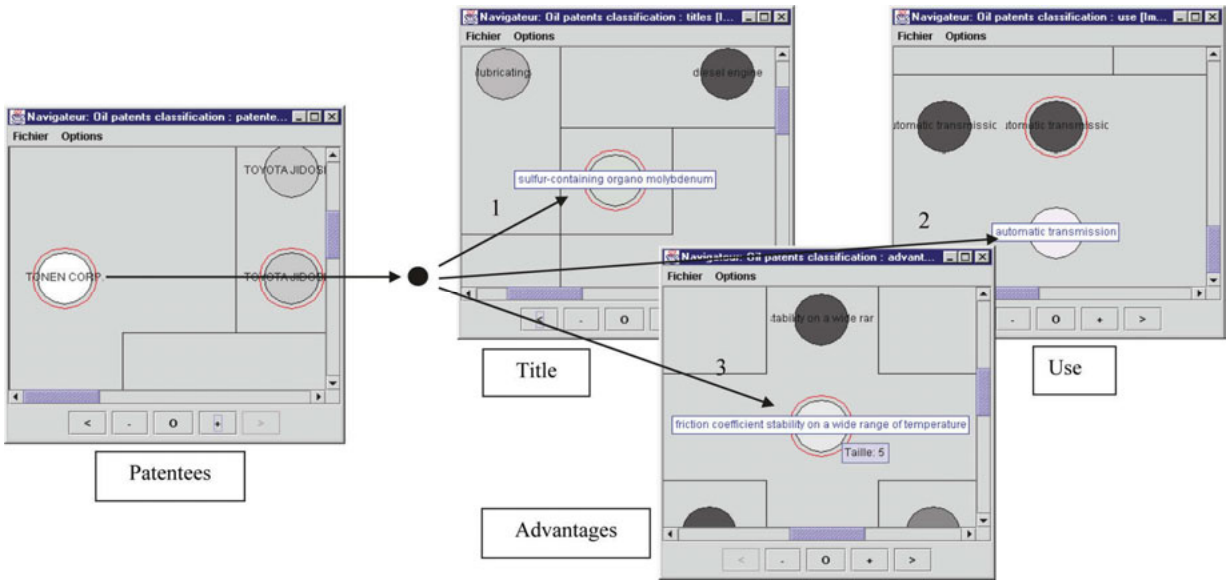

Figure 4. Example of exploitation of the inter-map communication mechanism.

The map is initially organized as a square 2D grid of nodes. The viewpoint chosen for the shown map is the "Advantages" viewpoint. The names of the clusters illustrate the concepts (considering the chosen viewpoint) that have been highlighted by the learning. After the learning, the nodes related to the same concept have been grouped into coherent areas thanks to the topographic properties of the map. The number of nodes of each area can then be considered as a good indicator of the concept weight in the database. Concepts or areas near one another are strongly related. For example, the "extending oil live" area shares some of its borders with the "black sludge control" area on the map. The proximity of these two areas illustrates the fact that oil duration strongly depends on maintaining a low level of sludge in it. The surrounding circles represent the centers of gravity of the areas.

The left part of the figure represents the WEBSOM-like mapping (i.e., without viewpoint management) of the content of the patent abstracts. The right part of the map represents the description (i.e., profile) of the "extending oil life" WEBSOM global concept. Even if a strong relationship between "extending oil life" and "black sludge control" concepts has been highlighted by the MultiSOM viewpoint-oriented clustering (see map of figure 2), this relationship has been lost by the WEBSOM-like clustering due to the inherent noise generated by the global analysis (this relationship does not appear, either in the map above, or in the "extending oil life" concept profile).

The analyst's decision to activate the area corresponding to the TONEN CORP. company on the Patentees map and to propagate the activity to the concept maps associated with the Use, Advantages, and Title viewpoints corresponds to a "viewpoints crossing query" whose explicit formulation might look like: "I want to know which are the specific areas of competence (concerning oil use, oil composition and expected advantages) of the TONEN CORP. company, if there are any." The MultiSOM application lets him interactively find that the TONEN CORP. company is a specialist in the lubrication of automatic transmissions [arrow n°2 on the map] and that it adopts for this kind of lubrication a sulfur-containing organo-molybdenum compound [arrow n°1] the main advantages of which are to provide oil with a friction coefficient that is stable on a wide range of temperatures [arrow n°3]. In this case, an inverted propagation from the target concepts also should be used to verify that these concepts relate only to TONEN CORP. areas of competence. The whiter the color of a node representing a map cluster (concept), the higher is its resulting activity.

Nevertheless, expert empirical evaluation remains insufficient to compare objectively the global approach to the viewpoint-oriented approach. For this last purpose, we propose to make use of the new objective clustering quality estimators presented in section 3 for both evaluating and optimizing the results of the clustering and of the mapping methods, especially when they are applied in the domain of textual databases.

The examination of the quality measures in table 2 gives more reliable and stable results because these measures are both independent of the clustering method and of the size of the description space. It highlights the overall superiority of the viewpoint-oriented approach as compared to a global approach with the same number of clusters (GlobMin). As soon as the number of clusters is strongly increased in the global approach (GlobMax), its quality is simultaneously increased, but the advantage of the viewpoint-oriented approach remains obvious in the average (higher Average F-value on all viewpoints than F-value of GlobMax), with a more reasonable number of nodes (i.e., clusters or concepts) per maps from a user point of view. The specific case of Title clustering should be discussed here. The bad quality of this clustering is due both to the index sparseness of this domain and to an inappropriate number of clusters, relative to the size of its associated description space. An interesting strategy would then be to make use of the proposed quality indexes in order to find the optimal number of clusters for this clustering (see Section 4 for more details). An imbalance between averaged Recall and Precision (in the favour of averaged Recall) can be observed in the case of the worse clustering (GlobMin and Titles). Such an imbalance means that documents with different feature sets are grouped in the same clusters, leading conjointly to the risk of confusion in the interpretation of the clusters' associated concepts by the user.

The former quality analysis clearly shows that the viewpoint-oriented approach enhances the quality of in-

	PATENTTEES	TITLE	USE	ADVANTAGES	MultiSOM (Global)	WEBSOM (GlobMin)	WEBSOM (GlobMax)
MACRO-R (**MR**)	0,94	0,89	0,78	0,77	0,85	0,87	0,84
MACRO-P (**MP**)	0,92	0,40	0,63	0,60	0,64	0,48	0,60
MACRO-F (**MF**)	0,93	0,55	0,70	0,67	0,73	0,61	0,69

Table 2. Summary of the results of Quality, Recall and Precision evaluation. The nearer the different values are from 1, the better are the clustering results. The F value provides a synthesis of the results of R and P.

terpretation of a clustering by both reducing the number of clusters to be consulted by the user on each viewpoint and providing him with more coherent and exhaustive clusters in terms of content.

5.0 Efficient diachronic analysis of research based on MVDA paradigm

5.1 Context of the study

The literature taking into account the chronological aspect in information flows is mainly 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 unstable. Research on "DataStream," among other things, was initiated 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 accurately detecting concept changes in specialized domains, for example, precisely following the evolution of research fields in scientific literature.

Numerous clustering methods have been used in this framework. 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 previous studies (Lamirel and Al Shehabi 2004) highlighted the fact 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 (MacQueen 1967), as well as new incremental neural and non-neural 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. 2011).

To cope with the current defects of existing incremental clustering methods, an alternative approach for sharply analyzing textual information evolving over time consists of 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 their mappings between the different periods. For analyzing the evolution of the vocabulary describing the clusters of different periods, Schiebel and al. (2010) propose constructing a matrix of keyword-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 a domain expert to highlight different cluster (i.e., concept) 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 term similarities with other documents (Glänzel and Thijs 2010). In such a way, clusters of two time periods are considered 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 content information. This approach presents the advantage of being relatively independent of vocabulary changes between periods, but it necessitates exploiting referencing data.

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.

Analyzing the difference between time periods concerns different kinds of concepts' changes or similarities that could occur between the periods (appearing concepts, disappearing concepts, splitting concepts, merging concepts, stable concepts). For achieving comparison between two time periods, a label-based diachronic approach relying both on data properties (i.e., features) 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 endogenous labels can be associated with each cluster of a given time period. The identification of the concepts' relationships between two time periods is then achieved through the use of Bayesian

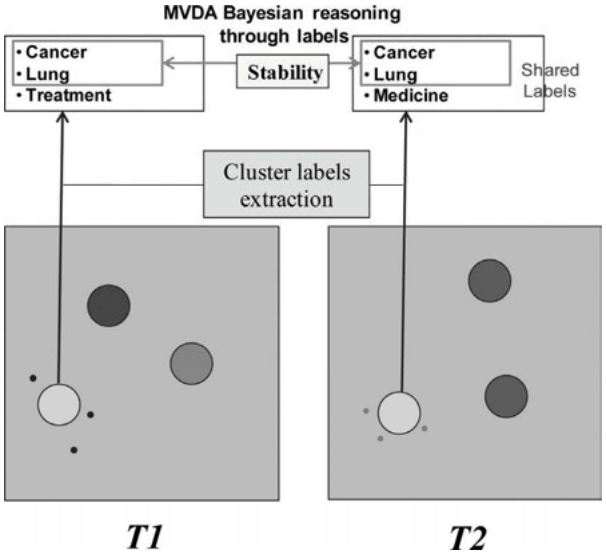

Figure 5. The label-based approach.

reasoning relying on the extracted labels that are shared by the compared periods (see Fig. 5).

5.2 Dataset preprocessing and diachronic knowledge building

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 concepts. They first employed a simple search strategy, consisting of 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 engineering and the life sciences (biology 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. 3,890 records related to these concepts were thus selected in the PASCAL database. Similarly, our approach consisted of 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 provided by the bibliographic records. In our experiment, the research concepts associated with the indexing keyword field are solely considered. For each year, a specific dataset is generated. Keywords with an overall frequency less than 3 are first removed from the record descriptions. 1,797 records indexed by 1,256 keywords are consequently kept in period 1, and 2,074 records indexed by 1,352 keywords in period 2. In a further step, the resulting vectors associated with each record are weighted using an IDF weighting scheme (Robertson and Jones 1976) in both periods in order to decrease the effect of more frequent indexes.

The clustering of the datasets associated with the two periods is achieved by the use of different clustering methods. For our experiment, we select K-means as the reference method in the category of non-neural methods, as well as various neural methods, ranging from static ones, like SOM (Kohonen 1982), NG (Martinetz and Schulten 1991) or GNG (Fritzke 1995), to incremental ones, like IGNG (Prudent and Ennaji 2005) or IGNG-F (Lamirel et al. 2011). For each method, we performed many different experiments, varying the number of clusters in the case of static methods and the vigilance parameters in the case of incremental methods. The best (i.e., optimal) clustering model for each period regarding the optimal compromise between the values of the F-average of MacroRecall/Precision indexes (Eq. 3), the F-average of the MicroRecall/Precision indexes (Eq. 4) and the F-average of the Cumulated Micro- indexes (Eq. 7) was finally kept. The values obtained highlight that the GNG neural method, which has already been proven to be especially efficient on thematically homogeneous textual data (Lamirel et al. 2011), provided the best results on our experimental dataset for both periods. Table 3 specifically presents the quality results obtained in the first period with all the methods. It highlights the fact that GNG reached high quality values with the lowest difference between the Macro- and Micro-values (most homogeneous results) and the highest CMP value (best big-sized clusters). Table 3 also highlights the inadequacy of MSE for evaluating quality in our context.

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. 10).

CLUSTERING METHOD	NBR CLUSTERS	MACRO-F	MICRO-F	CMP	MSE.
SOM	38	0,37	0,35	0,30	0,80
K-means	39	0,41	0,37	0,36	0,47
NG	40	0,43	0,39	0,38	0,70
GNG	40	0,44	0,41	0,48	0,62
IGNG	42	0,47	0,41	0,24	0,93
IGNG-F	39	0,49	0,42	0,32	0,98

Table 3. Summary of clustering results (time period 1).

The general results of the formerly described process are reported in table 4. The table also highlights some important 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 concepts (lower average number of overlapping labels), in the second period.

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 (eq. 1).

The new computation is expressed as:

$$(12)\ P(t|s) = \frac{\sum_{f \in L_s \cap L_t} FF_t(f)}{\sum_{f \in L_t} FF_t(f)}$$

where *s* represents a cluster of the source period, *t* a cluster of the target period, L_x represents the set of labels associated with the cluster x, using the cluster feature maximization approach defined by (Eq. 10), and 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:

$$(13)\ P_A(S) = \frac{1}{|Env(s)|} \sum_{t \in Env(s)} P(t|s)$$

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:

$$(14)\ A_S = \frac{1}{|S|} \sum_{s \in S} P_A(s)$$

Its standard deviation can be defined as .

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:

$$(15)\ P(t|s) > P_A(s) \text{ and } P(t|s) > A_s + \sigma_s$$

$$(16)\ P(s|t) > P_A(t) \text{ and } P(s|t) > A_t + \sigma_t$$

Cluster splitting is verified if there is more than one cluster of the target period which verifies the similarity rules (15) and (16) 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 (15) and (16) with a cluster of the target period.

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

Table 5 summarizes the results of our experiment of time period 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 clusters splitting the first period into the second period is more important than the converse number of clusters 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 built up for presentation to the analysts. Each report includes one cluster of each period, when-

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

Table 4. Overall period characteristics (datasets) and clustering optimized results (GNG).

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

Table 5. Summary of the time comparison results.

ever it is a similarity report, or one cluster of a single period, whenever it is a divergence report (i.e., an appearing or disappearing concept). 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. 10), 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 do 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 concept similarity between two periods. Said labels are displayed in decreasing order of their Feature F-measure difference with the alternative periods.

5.3 Evaluation

The results produced by our automated approach of comparison of time periods were finally compared with those of the analysis carried out by domain experts on the partitions produced over separated periods of time in the former experiment of Schiebel et al. (2010). Said analysis has mainly highlighted the two facts: 1) the general set of concepts of the studied corpus corresponded to the optoelectronic devices containing mineral or organic semiconductors; and 2) the research and applications of optoelectronics evolved from the field of the "photo-detectors" (probes, measuring instruments …) in period 1 to the field of the "electroluminescent diodes" in period 2.

The aforementioned conclusions present the disadvantage of providing only surface information on the potential concept evolutions. As is 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 an unsupervised way, while preserving the possibility of observing general orientations, such as those expressed by the experts of the PROMTECH project.

For the sake of validation, all of 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 very precisely to reveal the tremendously rich developments of the research concepts in the optoelectronic domain during the 1996-2003 period altogether, from the theoretical studies to the practical applications (from optical polymers to polymer films (figure 6), from surface emitting lasers or semiconductor lasers to vertical cavity lasers or VCSEL, etc.), from the exploitation of new chemical components to the production of new devices (from gallium arsenide to quantum well devices, etc.), or new semi-conductor types (from silicon compounds to amorphous semi-conductors, from gallium compound to wide band gap semiconductors, raise of exploitation of germanium, etc.), or the slight emergence of new semiconductor structures or organization which might become autonomous or self-assembling structures.

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 with different contexts, or even failing to detect them. As an example, re-

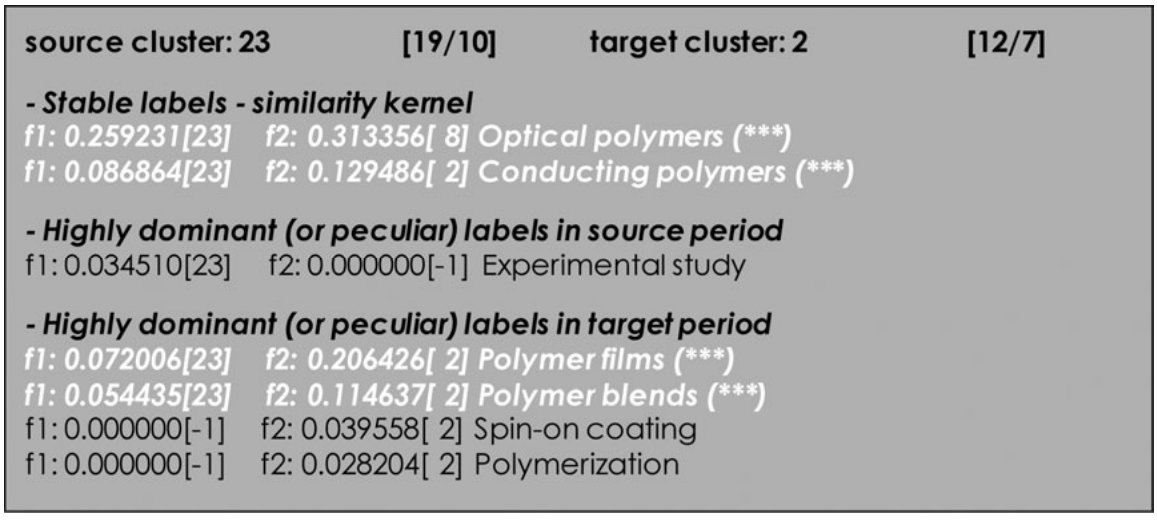

Figure 6. Similarity report related to the strong development of polymer blends and films.

ports provided confirm the progressive evolution of the optoelectronics domain from punctual developments to high scale industrial processes.

Thanks to the domain experts, automatic reports of divergence between periods, materializing disappearances or emergences of subjects (concepts), highlight more important changes in the domain than those that could be highlighted by the similarity reports. The complete disappearance of research on optical fibers during the second period is thus clearly highlighted (figure 7). 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. 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 the digital camera market in the second period (figure 8), as well as the emergence of promising research on a new generation of high efficiency optical nano-transistors (quantum dots) are also accurately figured out by the divergence reports.

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 concepts 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 concepts and search for the related papers in the exploited dataset. Table 6 synthesizes the resulting count of such papers in each period. It clearly demonstrates the efficiency of the method to detect main changes. More precise analysis would also highlight the efficiency of the related Feature F-measure to quantify the amount of change between the periods.

The complete results provided by the method cannot be presented here. They have thus been made available at a specific address (Results 2013b). However, one might already remark that such a concept-change mining process using single keyword 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.

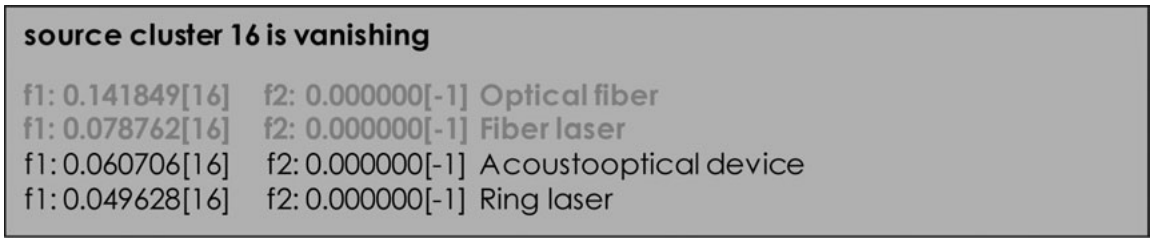

Figure 7. Divergence report related to vanishing of research on optical fibers.

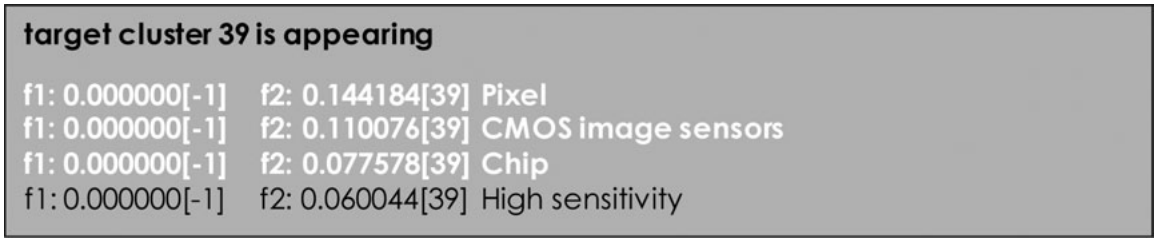

Figure 8. Divergence report related to the strong emergence of the development and integration of high sensitivity image sensors.

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

Table 6. Evolution of the paper count related to the emerging and disappearing concepts

6.0 Conclusion

In the process of textual information analysis, as in the domain of technological survey through patent analysis, or in the domain of emergent research tracking through research paper analysis, the complexity of the studied concepts and the accuracy of the question to be answered may often lead the analyst to partition his reasoning into viewpoints. Most of the classical information analysis tools can only manage an analysis of the studied domain in a global way. The information analysis paradigm which is considered in this paper is an alternative paradigm called multi-view data analysis. We have illustrated the generality of this paradigm through two different experiments.

We first presented a specific implementation of this paradigm in the form of a self-organizing multi-map model called MultiSOM. We proposed it as a visualization-based system for scientific and technical information analysis, like patent analysis. The model that this multi-map environment provides is certainly not the map but is in its original extended version an environment of intercommunication between multiple maps. We have exposed both the map generation and their intercommunication mechanism. Finally, we have shown how one can evaluate such a viewpoint-oriented approach by comparing it to a global approach using both expert judgment and method independent quality measures.

Second, we show in this paper the feasibility of an unsupervised incremental approach based on a time-step analysis of bibliographical data thanks to an alternative exploitation of the MVDA model in which viewpoints are represented by time periods. Our approach was also based on the exploitation of original and stable measures for evaluating the quality and the coherence of the data analysis results, and even for precisely synthesizing the said results. 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. Moreover, we also showed that it was not possible to achieve such results with former semi-supervised methods even with the intensive help of human experts.

Note

1 The **Patentees** and **Title** subfields are directly represented in the original patent structure and therefore do not necessitate any extraction.

References

Allan James, Carbonell, Jaime, Doddington, George, Yamron Jonathan and Yang Yiming. 1998. Topic detection and tracking pilot study: Final report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, Lansdowne, Virginia.*

Al Shehabi, Shadi and Lamirel Jean-Charles. 2006. Evaluation of collaboration between European universities using dynamic interaction between multiple sources. *Journal of information management and scientometrics* 1 n.3.

Attik, Mohammed, Al Shehabi, Shadi and Lamirel, Jean-Charles. 2006. Clustering analysis for data with multiple labels. In *DBA'06 Proceedings of the 24th IASTED international conference on Database and applications, Innsbruck, Austria, February*, pp. 50-7.

Campanario, J. M. 1995. Using neural networks to study networks of scientific journals, *Scientometrics* 33: 23-40.

Davies, David L. and Bouldin, Donald W. 1979. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence* 1: 224-7.

Dempster A.P., Laird N.M. and Rubin D.B. 1977. Maximum likelihood for incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (statistical methodology)* 39: 1-38.

François, Claire, Hoffmann, Martial, Lamirel, Jean-Charles, Polanco, Xavier and Al Shehabi, Shadi. 2003. *Artificial neural network mapping experiments. EICSTES (IST-1999-20350) Final Report (WP 9.4).* Available http://www.eicstes.org/EICSTES_PDF/Deliverables/Artificial%20neural%20network%20mapping%20experiments.PDF

Frizke, Bernd. 1995. A growing neural gas network learns topologies. In Tesauro, Gerald, Touretzky, David and Leen, Todd, eds., *Advances in neural information processing systems: Proceedings of the 1994 Conference*, Cambridge MA, MIT Press, pp. 625-32.

Gaber, Mohamed Medhat, Zaslavsky, Arkady and Krishnaswamy, Shonali. 2005. Mining data streams: A review. *ACM SIGMOD Record* 34 n.2: 18-26.

Glänzel, Wolfgang and Thijs, Bart. 2010. Using 'core documents' for the representation of clusters and topics. *Scientometrics* 88: 297-309.

Ghribi, Maha, Cuxac, Pascal, Lamirel Jean-Charles and Lelu Alain. (2010). Mesures de qualité de clustering de documents: Prise en compte de la distribution des mots-clés. In *Évaluation des méthodes d'Extraction de Connaissances dans les Données- EvalECD'2010 Workshop, Hamamet, Tunisia.*

Hinton, G. E. 1989. Connectionist learning procedures. *Artificial intelligence* 40: 185-234.

Jouve O. 1999. *Les nouvelles technologies de la recherche d'information, Séminaire Documentation, Paris.*

Kaski, Samuel, Honkela, Jukka, Lagus, Krista and Kohonen, Teuvo. 1998. WEBSOM-self organizing maps of document collections. *Neurocomputing* 21: 101-17.

Kohonen, Teuvo. 1982. Self-organized formation of topologically rrect feature maps. *Biological cybernetics* 43: 56-59.

Lamirel, Jean-Charles. 1995. *Application d'une approche symbolico-connexionniste pour la conception d'un système documentaire hautement interactif.* Ph.D. dissertation. Nancy, France: Université Nancy-I.

Lamirel, Jean-Charles and Al Shehabi, Shadi. 2004. Comparison of unsupervised neural clustering methods for mining Web and textual data. *SCI 2004, Orlando, FL, USA*, July. If the reference is right (I did not find anything about it).

Lamirel, Jean-Charles, Boulila, Zied, Ghribi, Maha and Cuxac, Pascal. 2010. A new incremental growing neural gas algorithm based on clusters labeling maximization: Application to clustering of heterogeneous textual data. In *The 22nd International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems (IEA-AIE), Cordoba, Spain.*

Lamirel, Jean-Charles and Créhange, Marion. 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, Maryland, USA, November 94*, New York : ACM, pp.155-63.

Lamirel, Jean-Charles, Ducloy, Jacques and Oster, Gérald. 2000. Adaptative browsing for information discovery in an iconographic context, In *Conference Proceedings RIAO, Paris*, 2, pp 1657-72.

Lamirel, Jean-Charles, François, Claire, Al-Shehabi, Shadi and Hoffmann, Martial. 2004. New classification quality estimators for analysis of documentary information: application to patent analysis and web mapping. *Scientometrics* 60: 445-62.

Lamirel, Jean-Charles, Mall, Raghvendra, Cuxac, Pascal and Safi, Ghada. 2011. Variations to incremental growing neural gas algorithm based on label maximization. In *The 2011 International Joint Conference on Neural Networks (IJCNN), San José, CA, USA, August 2011.*

Lamirel, Jean-Charles, Ta, Anh Phuong and Attik, Mohammed. 2008. Novel labeling strategies for hierarchical representation of multidimensional data analysis results. In AIA '08 Proceedings of the 26th IASTED International Conference on Artificial Intelligence and Applications, Anaheim, CA: ACTA Press, pp. 169-74.

MacQueen, J. 1967. Some methods of classification and analysis of multivariate observations. In *Proceedings Fifth Berkeley Symposium on Mathematics, Statistics and Probability* 1, Berkeley: University of California, pp 281-97.

Martinetz, Thomas and Schulten, Klaus. 1991. A "neural gas" network learns topologies. In Kohonen, T., Makisara, K., Simula, O. and Kangas J., eds., *Artificial Neural Networks*, Amsterdam: Elsevier, pp 397-402.

Prudent, Y. and Ennaji, A. 2005. An incremental growing neural gas learns topology. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005. IJCNN '05* 2, pp. 1211-16

Results (2013a). https://sites.google.com/site/msomesults 2012.

Results (2013b). https://sites.google.com/site/diacresults 2012.

Robertson, Stephen E and Sparck Jones, Karen. 1976. Relevance weighting of search terms. *Journal of the American Society for Information Science* 27: 129-46.

Schiebel, Edgar, Hörlesberger, Marianne, Roche, Ivana, François, Claire and Besagni, Dominique. 2010. An advanced diffusion model to identify emergent research issues: The case of optoelectronic devices. *Scientometrics* 83: 765-81.

Thijs, Bart and Glänzel, Wolfgang. 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*.

Varsis A. and Versino C. 1992. Clustering of socio-economic data with Kohonen maps. In *Proceedings of third International Workshop on Parallel Applications in Statistics and Economics, Prague, Czechoslovakia.*

White, Howard D., Lin, Xin and McCain, Katherine W. 1998. Two modes of automated domain analysis: Multidimensional scaling vs Kohonen feature mapping of information science authors. In Edited by Mustafa el Hadi, Widad, Maniez, Jacques and Politt, Steven. A., eds, *Structures and relations in knowledge organization. Proceeding of the Fifth International ISKO Conference, Lille, 25-29 August 2000*, pp 57-63.