An advanced diffusion model to identify emergent research issues : the case of optoelectronic devices

Edgar L. Schiebel, Marianne Hörlesberger, Ivana Roche, Claire François, Dominique Besagni

An advanced diffusion model to identify emergent research issues: the case of optoelectronic devices

Edgar Schiebel[1], Marianne Hörlesberger[1],

Ivana Roche[2], Claire François[2], Dominique Besagni[2]

[1] marianne.hoerlesberger@arcs.ac.at, edgar.schiebel@arcs.ac.at

[2] ivana.roche@inist.fr, dominique.besagni@inist.fr, claire.francois@inist.fr

[1]Austrian Research Centers GmbH, Tech Gate Vienna, Donau-City-Straße 1, 1220 Wien, Austria
Tel.: +43 (0) 50 550-4520, Fax: +43 (0) 50 550-4599

[2] INIST-CNRS, 2 Allée du Parc de Brabois, CS 10310, 54519 Vandoeuvre-lès-Nancy, France
Tel.: +33 (0) 3 83 50 46 00, Fax: +33 (0) 3 83 50 47 33

Abstract

Scientific progress in technology oriented research fields is made by incremental or fundamental inventions concerning natural science effects, materials, methods, tools and applications. Therefore our approach focuses on research activities of such technological elements on the basis of keywords in published articles.

In this paper we show how emerging topics in the field of optoelectronic devices based on scientific literature data from the PASCAL-database can be identified. We use Results from PROMTECH project, whose principal objective was to produce a methodology allowing the identification of promising emerging technologies. In this project, the study of the intersection of Applied Sciences as well as Life (Biological & Medical) Sciences domains and Physics with bibliometric methods produced 45 candidate technological fields and the validation by expert panels led to a final selection of ten most promising ones. These 45 technologies were used as reference fields.

In order to detect the emerging research, we combine two methodological approaches. The first one introduces a new modelling of field terminology evolution based on bibliometric indicators: the diffusion model and the second one is a diachronic cluster analysis.

With the diffusion model we identified single keywords that represent a high dynamic of the mentioned technology elements. The cluster analysis was used to recombine articles, where the identified keywords were used to technological topics in the field of optoelectronic devices.

This methodology allows us to answer the following questions: Which technological aspects within our considered field can be detected? Which of them are already established and which of them are new? How are the topics linked to each other?

Introduction

Early recognition of new and alternative products and production processes is a strategic necessity contributing to timely assessment and decision-making. Emerging technologies are essential to advances in research, industry and society. But the detection of emerging technologies remains quite a problem. Studies in a large range of domains in society, econmomy, ecology, politics, etc. were performed (Kajikawa Y., Yoshikawa J., Takeda Y., Matushima K. (2008), Noyons E. (2004), Roche I., Besagni D., François C., Hörlesberger M., Schiebel E. (2008), Schiebel E., Hörlesberger M. (2007)).

An interesting introduction to use bibliometrics for the identification of upcoming issues in online databases has been published by Lancaster F.W., Lee J.L. (1985). Lancaster et al

analysed the spread or migration of an issue from pure science over applied science to public awareness. They counted the occurrence of a keyword over time in different online databases that cover publications with the mentioned different categories. Additionally they suggested a procedure to identify growing issues in terms of time gradients of the number of published articles where single keywords occur.
The selection tree introduced by Armstrong & Green (2007) gives an additional picture of the landscape of the forecasting methods that could be employed to detect these emerging technologies. It illustrates the dichotomy between judgmental and quantitative forecasting methods and shows the great diversity of existing approach like the Delphi or the Nominal Group Technique ones. Forecasts are obtained in a structured way from two or more experts. Other methods combine expert domain knowledge and statistical techniques and allow the identification of causal forces acting on trends.
When the available collected data is enough to apply quantitative methods, an important question remains. It concerns the type of the data employed to forecasting: scientific literature databases and patent databases are the most often used sources to detect new themes by bibliometric analysis. They apply statistical techniques comparatively simple as growth curves analysis or more sophisticated as clustering or network analysis (Noyons, 2004; Daim et al., 2006; Mogoutov & Kahane, 2007; Kajikawa et al., 2008). The addition of a third type of data coming from state-funded research grants is very interesting and allows, if the three sources of information are analyzed as a whole and not as separate entities, to gain an understanding of the triple-helix interfaces between university, industry and government (Salerno et al., 2006; Mogoutov et al., 2008).
As a technology consists of a bunch of physical effects, materials, methods and application we used a different approach. We analysed the growing number of publications in classes of the PASCAL database. These classes were considered to be a proxy for technologies and then after having identified the most growing classes, we examined the occurrence of all keywords of all articles in each class.
The framework for this study was the PROMTECH project. It was financed by the European Commission. The principal objective in the project was to produce a methodology for the identification of promising emerging technologies. The basic hypothesis was derived on the observation that many new emerging technologies of the last decades have drawn on the technical application of physical knowledge. The growing relevance of Physics for technology can be explained by the increased complexity of today's technology. Indeed, there are many examples showing that advanced technology in applied sciences as well as in life sciences is increasingly linked to recent outcomes of research in Physics.
In this project, the study of the intersection of applied sciences as well as life (biological & medical) sciences domains and physics with bibliometric methods, validated by expert panels, led to a final selection of ten most promising technological fields.
The core idea of this contribution was to develop a methodology for detecting of new topics inside of broader technological fields. We choose the field "optoelectronic devices", because it is one of the most promising technology fields of the last decade. Light emitting diodes gain more applications in cars and housing lighting, OLED displays are introduced in electronic devices and consumer electronics. Optimisation of the luminous power and the tuning of the light spectrum are wwll known important research topics.
In order to identify promising emerging research topics and their diffusion stage, we used two approaches and two datasets in two successive periods:

- the "diffusion model", that is an approach that uses a bibliometric filter that distributes keywords in different diffusion stages in order to model the field terminology evolution (Schiebel & Hörlesberger, 2007) and

- the “diachronic cluster analysis”, using the software-tool Stanalyst® (Polanco et al., 2001), to identify research topics by recombining keywords by clustering papers.

Firstly we refer to the data acquisition. Secondly the applied methodologies are illustrated. Thirdly the results of the applied methodologies to the field “optoelectronic devices” are presented. Finally we discuss the two approaches, summarize and conclude our discussion.

Data acquisition

The data framework is a sample recorded from the PASCAL database that was specifically adapted to the purpose of our approach. The PASCAL database provides a broad multidisciplinary coverage of scientific publications and contains, nowadays, about 17 million bibliographic records that are derived from the analysis of the scientific and technical international literature published predominantly in journals and conference proceedings.
In this work, we limited our study to a corpus of publications coming from journals only. The queries operated in this work were exclusively based on the classification categories given in the database and assigned to the individual publications, either manually by scientific experts or automatically based on a content analysis. These classification categories belong to the PASCAL classification scheme that is a taxonomy of every field and subfield of all the disciplines covered in the database. Each individual publication also benefits from a manual or automatic indexing by keywords. After a verification step, done by a scientific expert, that terminology can be employed in our analysis.

Methodology

In this section, we introduce the two methods we employed to follow the evolution of the technological field. In previous work, we presented their detailed description (Roche et al., 2008).
Firstly, the diffusion model is used to evaluate the term status in the considered technological field, comparing them with theirs status in a reference set the other 45 technological fields identified in PROMTECH.
Secondly, the diachronic cluster analysis recombines the dynamic elements of the technological field by a clustering approach allowing to organize articles in sub-topics and to analyse links between them.

Diffusion model

The diffusion model is based on the assumption that new findings are published in articles in a research field. Keywords that describe the research discoveries occur in the first stage in an unusual manner. In the second stage the research intensifies and more established keywords are used in the home field and in a third stage the research results like natural science effects, methodologies, test arrangements, materials and applications diffuse to other research fields or from basic research to applied research like Lancaster F.W. , Lee JL. (1985) mentioned in their publication.
In an earlier approach a bibliometric filter was used to assign the keywords of articles in three diffusion stages for the period from 1996 to 2003, Schiebel E., Hörlesberger M. (2007) to get an idea about diffusing research topics during the whole period. In this paper a time dependency is introduced. The span of time is divided into two periods. Table 1 shows this new perspective.
Following the diffusion approach it can be assumed that keywords migrate from earlier stages in the first period to later stages in the second period like the following:

1. terms from stage 1 in the first period stay in this stage or migrate in the second period to stage 2 or stage 3;

2. terms from stage 2 in the first period stay in stage 2 or migrate to stage 3;
3. terms from stage 3 stay in stage 3.

Table 1. Taxonomy of the migration pathways of research topics (descriptor terms) through different stages of the diffusion model in two time periods

Periods	***Diffusion Model*** Stage 1 Unusual Terms	Stage 2 Established Terms	Stage 3 Cross Section Terms
1996 to 1999 (period I)	1 2	3 4 5	6
2000 to 2003 (period II)			

We defined pathways as follows:

Pathway 1: the terms remain in stage 1 in both periods
Pathway 2: terms move from stage 1 in first period to stage 2 in second period
Pathway 3: terms move from stage 1 in first period to stage 3 in second period
Pathway 4: the terms remain in stage 2 in both periods
Pathway 5: terms move from stage 2 in first period to stage 3 in second period
Pathway 6: the terms remain in stage three in both periods

The terms in the different stages were identified and the migration of terms between different stages over the two time periods where traced.
This procedure has been performed for articles in the field "optoelectronic devices". Terms were extracted from the descriptor field. We constructed a diffusion matrix for each period by assigning the terms to the three stages. We used the bibliometric filter that was published in Schiebel E., Hörlesberger M. (2007). Table 2 shows the filter and the values for optoelectronic devices for the first period. As the technology filed is a part of the 45 reference technology fields the field name "optoelectronic devices" is the first selection criterion. Each condition reduced the number of keywords or classified them to be assigned to a diffusion stage.

Diachronic cluster analysis

The diachronic cluster analysis is based on the study of the evolution of cluster maps and contents obtained by a clustering method and considering data sets related to two, or more, successive time periods. In this work, after splitting the corpus in the two time periods I and II we applied a clustering algorithm on the corpus of each period, in which documents are represented by the keywords existing in the bibliographic references. The clustering tool we used (Lelu & François, 1992; Lelu, 1993) applies a non-hierarchical, non-supervised clustering algorithm, the axial K-means method that allows producing clusters presenting particular characteristics:

- they can overlap because the clustering method allows a document or a key-word to belong to more than one cluster,
- documents and key-words as the constituting elements of a cluster, are ranked by decreasing similarity with the cluster ideal type.

Each cluster is homogeneous subset of the information contained in the analyzed corpus, corresponding to a topic. After the clustering, we represent graphically its results employing two methods: a principal component analysis and a connected component analysis. The first one produces a global map of the topics. The topics that are very far apart represent clusters very different with respect to the key-words defining them.
The connected component analysis is then employed to show the relationships that exist between the clusters. It is a method based on the graph theory and defining the connected

components representing the relative proximity between clusters. The distance between two clusters is defined as the cosine of the angle between the two axes representing the clusters. A lower and an upper threshold are defined as respectively the minimum and the maximum value of the cosine between the clusters. Nine intermediate levels are defined in that interval. We consider that, at a given level, two clusters are connected if the value of the cosine associated to them is within the range of values of this level.
Finally, we analysed the evolution between the two cluster sets and the two maps by examining the vocabulary related to the clusters of each period using a comparison matrix (Roche et al., 2008).

Results

The PASCAL database query delivered 3,871 articles in the field "optoelectronic devices". In the first period we had 1,797 documents and 2,345 different keywords among which 1,527 (65%) appear less than 3 times. The second period had 2,074 articles and 2,738 terms among which 1,762 (64%) occur less than 3 times.

Diffusion model

In the first period corpus, we had 2,345 different keywords. Their categorization into diffusion stages was performed by using the bibliometric filter presented in table 2.

Table 2. Filter to identify stages in the period 1996 to 1999

Indicator	Number of Terms	Comment
Technology Field: *Optoelectronic Devices*	2,345	
Minimum Number of Articles: 3	817	
TFIDFTech > 0,0008	289	
High Diffusion: small Gini < 0,95	92	Stage 3 Terms
Low Diffusion: large Gini > 0,95	197	
Relative Frequency > 0,0035	96	Stage 2 Terms
Relative Frequency < 0,0035	101	Stage 1 Terms

The TFIDFTech is a modified TFIDF measure. We take the relative frequency of articles as the local weight TF[1]. The global weight IFD[2] is defined by the inverse number of technologies.
We used a threshold that selected 250 to 300 keywords. The first third of these keywords with a Gini[3] index smaller than 0,95 consisted of terms that occur in many of the 45 technologies therefore these were defined as keywords that were well established and broadly diffused to other technologies. The remaining two thirds of keywords were divided in two halves: keywords with a low relative frequency were seldom used and the remaining words occurred more often but mostly in the respective technology field.
It can be seen that the filter has led to three nearly equal numbers of keywords for each stage. The reason is that the thresholds have been chosen in a pragmatic way, because it can not be said where the exact borders between the stages of the diffusion model are. It is a fuzzy approach: it was assumed that from N keywords a third is in each stage. With this pragmatic

[1] The relative frequency TF is measured by nij/nj with i be the index for a key word, j be the index for the technology field, nij be the number of publications a key word i occurs in technology j and nj the total number of publications in technology fieldj. where the keyword occurs at least in one article.
[2] The global weight IDF is measured by 1/ti, ti be the number of technologies where the key word i occurs at least in one publication.
[3] The Gini Index is used as a measure for the inequality distribution of occurrence of a term in the technology fields. A high value indicates a low diffusion.

choice it is not really visible if a diffusion of keywords really takes place. We just span a space to be able to observe migration. An answer can be given if we can identify individual keywords that change the diffusion stages from period one to period two.

Table 3. Keywords by diffusion stage in period 1996 to 1999 (Selection)

Stage 1: Unusual Terms	Stage 2: Established Terms	Stage 3: Cross Section Terms
Acousto-optical devices	Aerospace instrumentation	Absorption coefficients
Aluminium Nitrides	Aluminium arsenides	Aluminium
Automatic optical inspection	Aluminium compounds	Arrays
Binary compound	Amorphous semiconductors	Avalanche breakdown
Cadmium tellurides	Astronomical instruments	Bolometers
Cerium additions	Astronomical telescopes	Cadmium compounds
Chirp modulation	Avalanche photodiodes	Carrier lifetime
Clutter	Brightness	Carrier mobility
Colour displays	Carrier density	CCD image sensors
Commerce	Chemical interdiffusion	Charge injection
Conjugated polymer	Complete computer programs	Computerized simulation
Coordinates	Conducting polymers	Doped materials
…	…	…

In the second period corpus, that contained articles of the period 2000 to 2003, we found 2,738 keywords that were categorized into diffusion stages by using the bibliometric filter of Table 4. The growth of articles from period I to period II was about 16 percent. Differences to the first period in the thresholds of the filter indicators were not really remarkable.

Table 4. Filter to identify stages in the period 2000 to 2003

Indicator	Number of Terms	Comment
Technology Field: *Optoelectronic Devices*	2,738	
Minimum Number of Articles: 3	976	
TFIDFTech > 0,0007	273	
High Diffusion: small Gini < 0,94	95	Stage 3 terms
Low Diffusion: large Gini > 0,94	178	
Relative Frequency > 0,0045	82	Stage 2 terms
Relative Frequency < 0,0045	96	Stage 1 terms

In general, it can be said that the research field, “optoelectronic devices”, consisted of several research topics in the context of sourcing, detecting and controlling light. There are some very special research topics like “Superconductor-insulator-superconductor mixers”, so called SIS-mixers, that are used in space technology for the transforming of frequencies in space observation on one side of the spectrum of specialised research and LCD-displays, CCD devices for digital cameras or light emitting diodes on the other side of the spectrum of more applied research, see table 5.

Table 5. Keywords by diffusion stage in period 2000 to 2003 (Selection)

Stage 1: Unusual Terms	Stage 2: Established Terms	Stage 3: Cross Section Terms
Aluminium antimonides	Aerospace instrumentation	Aluminium compounds
Aluminium complexes	Aluminium arsenides	Amorphous semiconductors

Aluminium nitride	Avalanche photodiodes	Antimony compounds
Aspherical optics	Binary compound	Band structure
Astronomical observation	Cadmium tellurides	Binary compounds
Attitude measurement	Carrier density	Bolometers
Auger effect	CMOS image sensors	Brightness
Blackbody radiation	Dark conductivity	Cadmium compounds
Blue light	Dislocation density	Carrier lifetime
Charge coupled device	Electrical properties	Carrier mobility
Color filter	Electrical resistivity	CCD camera
Color image	Electroabsorption	CCD image sensors
…	…	…

Pathways of keywords

The dynamics in the research could be demonstrated by the pathways of the keywords.

Pathway 1

Pathway 1 reflected the research topics that stayed in the stage 1, as it is shown in table 6. There are only 3 to 9 publications per keyword in the second period and the number stayed nearly stabile over the two periods. It represents research about topics like SIS-Mixers, where only a few people work worldwide, or devices like optical fiber couplers, that were used and mentioned but were not entity of intensive research.

It can be argued that there is a high dynamic in stage one: many keywords with a low frequency of articles show a high fluctuation: they appear in a period and disappear in a second one. This could be expected because this is a stage where keywords represent new research results of which many are not interesting in later periods and only a few survive in this incubator stage. Those who survive and stay in stage 1 represent research work that persist but does not excite more researchers.

Table 6. Terms of pathway 1

Indium antimonides	Porous semiconductors
Indium nitrides	Semiconductor laser
Microcavity	Spectral response
Narrow band gap semiconductors	Submillimeter wave antennas
Nonradiative transitions	Superconducting microbridges
Optical fiber couplers	Superconductor-insulator-superconductor mixers
Photoresistors	

Pathway 2

Pathway 2 in table 7 indicates growing research dynamics because terms that occurred only seldom in the first period moved to the phase of established terms in the second period. One example were cadmium telluride, what was a semiconductor material used for thin film solar cells, so called “advanced thin film” modules. Its properties fitted perfectly to the solar spectrum. There was a high potential for high-efficiency solar cell modules with a low-cost manufacturing processes. The “quantum confined Stark effect” is used to shift the light spectrum in light emitting diodes, what is important for the optical tuning. Light emitting diodes have a very high application potential in daily life and technical applications therefore together with high potential in basic research there is a growing research interest.

Keywords of this pathway survived the incubator phase and started to grow within the same technology field: they attracted other research groups. Who found the findings usefull for their own research or started new research. Keywords with a high number of articles like

binary compound (23), p n heterojunctions (16) and Liquid Phase Epitaxy - LPE (16) have made their career.

Table 7. Terms of pathway 2 with number of publications in period 2

Binary compound, 23	p n heterojunctions, 16
Cadmium tellurides, 11	Photodiode, 14
Far infrared radiation, 12	Quantum confined Stark effect, 10
Injection laser, 11	Semiconductor-metal boundaries, 19
LPE, 16	Ultraviolet photoelectron spectra, 16
Mercury tellurides, 10	

We found only two terms that moved from stage 1 in the first period to stage 3 in the second period. Jumping ahead from the incubator stage to the use in many other fields is a big step an. Research findings with this career should have overcome bottlenecks that were relevant in several other technological fields.
The first keyword "optical tuning of light emission diodes" was very essential for LEDs to became bright enough for practical use in daily life or for special technical usage. The composition of the proper emitting spectrum was an important matter of research. The traditional semiconductor technology needed crystals but in blue LED technology the needed GaN had bad crystal properties therefore thin layers which ternary compounds were developed and are intensively investigated. This breakthrough can be tracked by the migration of "ternary compound" from stage 1 in period I to stage 3 in period II within only some years, see table 8.

Pathway 3
Terms of Pathway 3 are most important to observe breakthroughs in research and technological development.

Table 8. Terms of pathway 3 with number of publications in period 2

Optical tuning, 14	Ternary compound, 33

Pathway 4
Pathway 4 reflected established research in both time periods: Gallium arsenides and Gallium nitrides as basic materials for advanced thin film LED, MOCVD - Metal-Organic Chemical Vapour Deposition what is a technique for depositing thin layers of atoms onto a semiconductor wafer. A heavily growing field was the application of organic semiconductors in optoelectronics. Organic semiconductors are applied in optoelectronic devices such as organic light-emitting diodes (OLED), organic solar cells and organic field effect transistors (OFET), they can be easily produced and optoeletronic polymers allowed the construction of flexible displays.

Table 9. Terms of pathway 4

Aerospace instrumentation	Gallium nitrides
Aluminium arsenides	Ge-Si alloys
Avalanche photodiodes	Indium arsenides
Carrier density	Interface states
Dark conductivity	LED displays
Dislocation density	Light emitting devices
Electroabsorption	Line intensity

Electroluminescent devices	Metal-semiconductor-metal structures
Electron-hole recombination	Microelectronic fabrication
Focal planes	Minority carriers
Gallium arsenides	MOCVD
…	…

Terms in this pathway represent ongoing core research in the field of optoelectronic devices, that is stable over time and has the highest number of key terms coming from this home technology.

Pathway 5

In Pathway 5, see table 9 we found research topics that were established in the first period and migrated to applications in many other technology fields in the second period. One example is the already mentioned organic semiconductors and the thin film III-V semiconductors. Optical arrays are used in "storage and reproduction of information", "global pollution and environment", "radiotheraphy" etc. "semiconductor thin films" are important in "storage and reproduction of information" and "transistor technology".

Table 10. Terms of pathway 5 with number of publications in period 2

Aluminium compounds 193	III-V semiconductors, 449
Amorphous semiconductors, 21	II-VI semiconductors, 91
Brightness, 64	Optical arrays, 33
Chemical interdiffusion, 16	Radiation quenching, 13
Conducting polymers, 82	Schottky barriers, 22
Deep energy levels, 10	Semiconductor thin films, 54
Elemental semiconductors, 92	Tunnel effect, 39

Research topics from this pathway gain importance in other technologies. They can be crucial for the development of new applications or new properties of other technology fields.

Pathway 6

Pathway 6 comprises broad applications of optoelectronic devises and research topics like CCD image sensors used in digital cameras but also natural science effects like the exciton. An exciton is a bound state of an electron in an isolator or a semiconductor and an imaginary particle called an electron hole. Of course, all optoelectronic devices like LED, photodetectors, infrared detectors, photodiodes, semiconductor lasers and used materials like cadmium compounds, Gallium and Indium Compounds are part of the cross section term in stage 3.

Diachronic cluster analysis

The diachronic cluster analysis applied on the two periods I and II, allows us to determine which topics of the second period have roots in the first one and which topics are new in the second period. The clustering algorithm produced 17 clusters involving 1,599 documents and 348 keywords in period I (P1), and 20 clusters relating 2,073 documents and 456 key-words in period II (P2). Let us notice that the very important reduction of the number of key-words associated to the clusters in both periods, around 50% with respect to the initial number of terms obtained directly from the query results, does not entail a significant loss of information at the level of the documents. Indeed we had 89% and 100% of the initial information corpora in the clusters of respectively P1 and P2.

Analysing the comparison matrix we can singularize for the second period:

- nine homonymous clusters, but only four of them seem to be stable,

- two clusters with new titles but presenting characteristics of stability,
- four clusters with low marginal values (clusters with a low inheritance from clusters of P1: are they new clusters?),
- six clusters with high marginal values (clusters with a strong inheritance from different clusters of P1).

Finally, we have detected a particular behaviour of two P1 clusters: they seem to supply at the same time and in the same way five clusters of the second period.

Having expressed these hypotheses we asked the scientific expert to validate them accordingly to both, the content of clusters, and the cluster organization in the maps. Figures 1 and 2 show the cluster maps of respectively the first and the second period. In both maps, each dot corresponds to a cluster and each line gives the connexion level between pairs of clusters. Note that we considered here only the 5 highest levels (out of 10) corresponding to the strongest connexions between clusters. In Stanalyst®, the level of connexions is code-coloured. Since these colours do not come out in print, we mentioned the level next to the connexion.

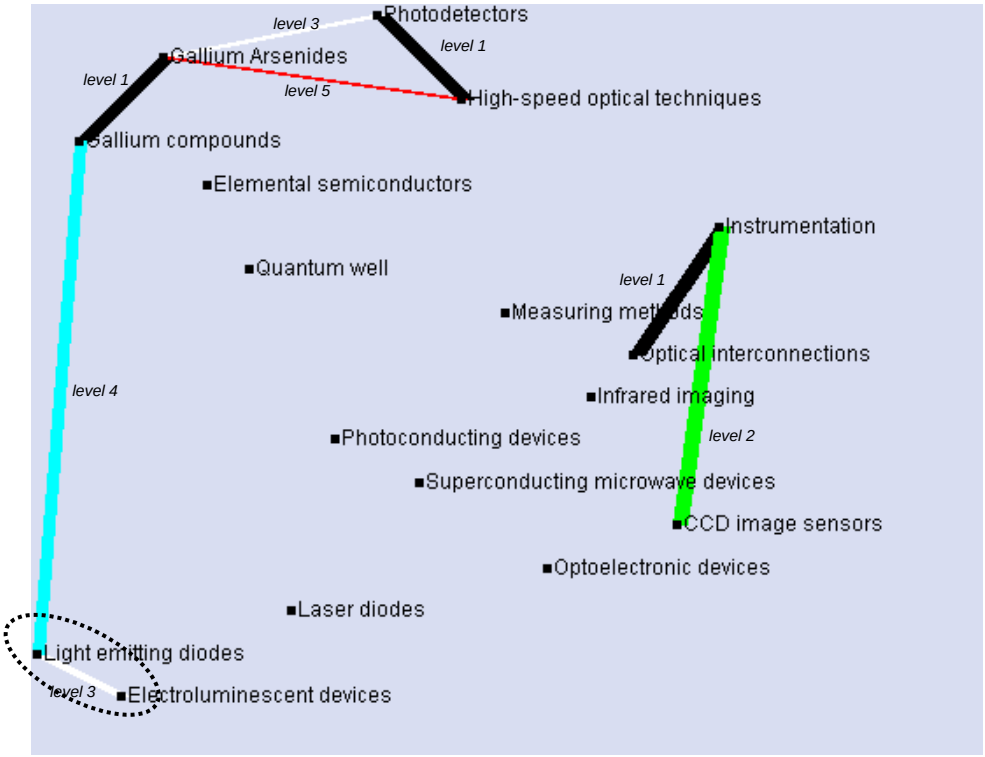

Figure 1. First period cluster map of "Optoelectronic devices" technological field.

The analysis of the expert seems to confirm the relative stability of five P2 clusters, each of them coming exclusively from an only P1 cluster that can be its homonymous or not. They are: "*Electroluminescent devices*", "*Elemental semiconductors*", "*Infrared imaging*", "*Bolometers*" and "*Photodetector*".

The keywords of P2 cluster "*Light emitting diodes*" come predominantly from its homonymous cluster in the first period, but also from others P1 clusters: "*Laser diodes*" and "*Electroluminescent devices*".

The P1 clusters "*Electroluminescent devices*" and "*Light emitting diodes*" seem to "feed" their homonymous P2 clusters but also, and with important rates, the P2 clusters "*Light emitting diodes*", "*Organic semiconductors*", "*Organic light emitting diodes*" and "*Electroluminescence*". The analysis of the cluster maps obtained for the two periods

confirms this observation. Indeed, the pair of clusters "*Electroluminescent devices*" and "*Light emitting diodes*" shows a notable evolution, coming from a marginal position in the P1 cluster map (see dotted ellipse in Figure 1) to a central role in the P2 cluster network (notice dotted ellipse in Figure 2) where they find themselves linked to the clusters which they "fed". In this network, the expert points out clusters "*Organic light emitting diodes*" and "*Organic semiconductors*" that deal with new themes coming from the Organic Electronics domain. The grouped position of these clusters on the cluster map is quite logic as well as their location in the central network that confirms the enlarging importance of these new technologies in the second period.

If expert analysis leads to assert that, from a macroscopic point of view, the considered field seems to be relatively stable between the two periods because it is dealing with optoelectronic devices based on mineral or organic semiconductors, a deepest look at the application level allows the expert to observe:

- in the P1 period, the applications dealing with photodetectors (imagery, measuring instruments, sensors,…) are dominant and the electroluminescent diodes are completely isolated;
- in the P2 period, the electroluminescent diodes are dominant and are located in a central position in the cluster network.

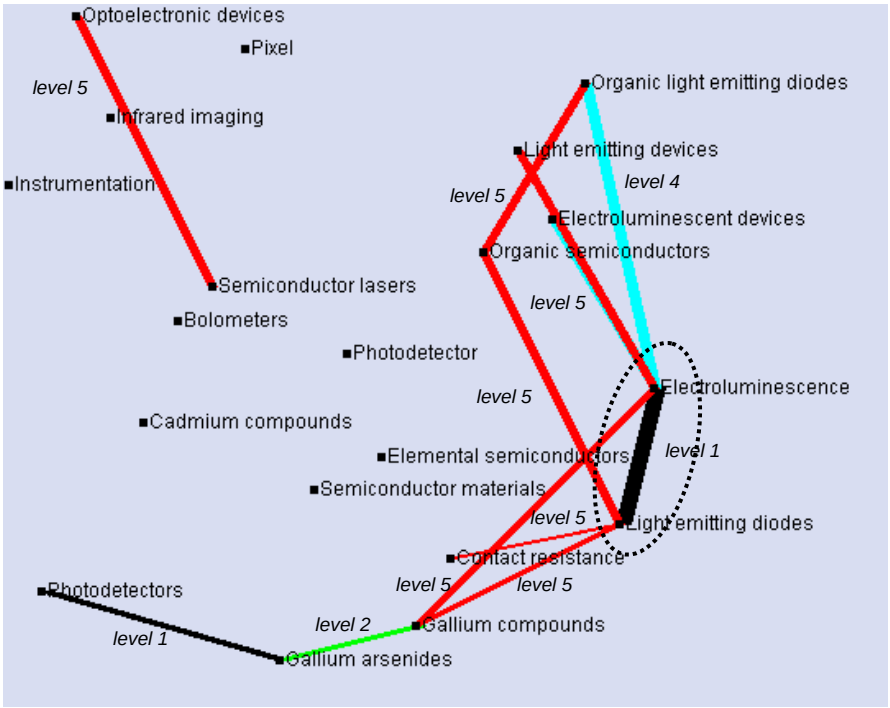

Figure 2. Second period cluster map of "Optoelectronic devices" technological field.

The analysis of the cluster contents consent the expert to detect, at the material level:

- the growing relevance of research dealing with organic semiconductors. This interest is justified by the flexibility of these materials allowing the production of flexible and miniaturized electronic devices (see clusters "*Organic semiconductors*", "*Organic light emitting diodes*" in P2);
- the rise of nanotechnologies in electronics in order to achieve a greater circuit miniaturization (internal evolution of "*Photodetectors*" and "*Gallium arsenides*" clusters);

- the renewed interest in Si-Ge semiconductors alloys and type III-V semiconductors caused by the research expansion on electroluminescent diodes already observed at the applications level (internal evolution of "*Elemental semiconductors*" cluster).

The analysis of the terminology associated to the P2 cluster "*Contact resistance*" gives an additional indication about the important role of electroluminescent diodes. Indeed, in this P2 cluster, the expert observes a massive presence of some keywords weakly represented in the first period (contact resistance, ohmic contact, solid solutions and semiconductor-metal boundaries). These notions are not new but their presence in the second period shows and confirms the growing number of optoelectronic device applications on electroluminescent diodes. Indeed, the contact quality between the diode and the electrodes is a very determinant question related to the diode performance.
Finally, the analysis of the terminology associated to the cluster "*Infrared imaging*" present in both periods, points out a very interesting trend. This cluster deals with the application of optoelectronics to the formation of images using infrared radiation, and to the processing of the acquired images for the purposes of measurements, identification or tracking. In P1, the infrared imaging technology is associated essentially to military applications and, in P2, the expert observes, with significant frequencies, the presence of keywords related to Biology and remote sensing. This singular observation allows the expert to move forward a hypothesis of either a technological transfer process from military usages to civil ones or, at least, an extension of the infrared imaging application field.

Convergences between the two analyses
The application of the diffusion model allows to select the most important terms of the field vocabulary: 288 keywords out of 2,345 in the first period and 273 keywords out of 2,738 in the second one. These selected terms are indexing the almost totality of the corpora: 1,788 documents out of 1,797 in period P1 and 2,069 documents out of 2,074 in period P2.

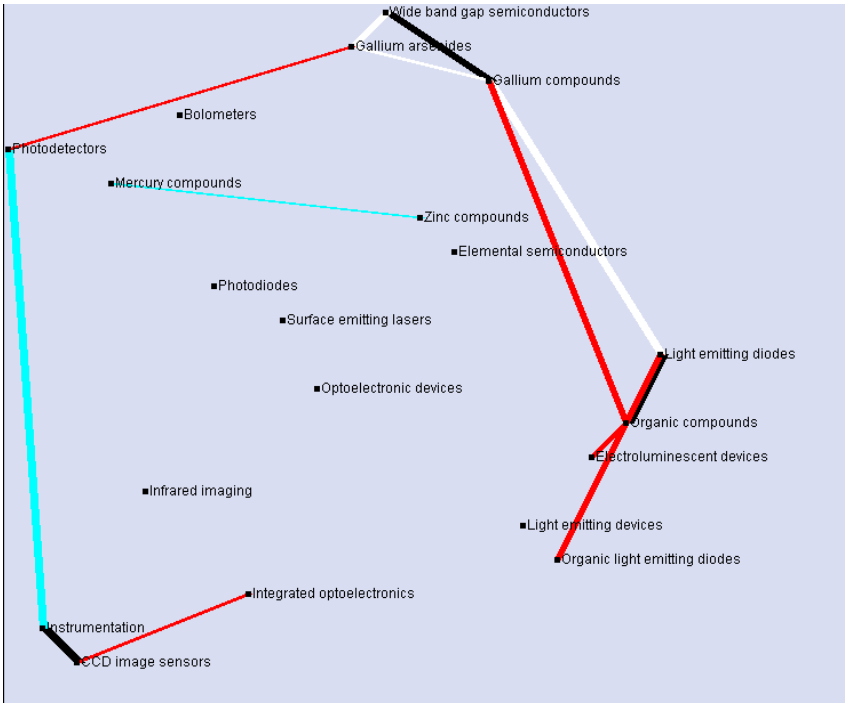

Figure 3. Second period cluster map of "Optoelectronic devices" technological field obtained from terms selected by diffusion model.

In this section, we compare, for the period P2, the cluster map obtained from the whole corpus (WC), presented in Figure 2, and the one we got by employing only the terms selected by the diffusion model (DM), presented in Figure 3.
First of all, we observed that the clustering results obtained with these selected terms are very close to the results produced from the whole corpus, although the clusters are more generic with a larger overlap between them. Then we particularly focused on the analysis, for each cluster, of the evolution of their specificity degree. The specificity degree is defined as the ratio between the number of documents present exclusively in the cluster and the total number of documents contributing to the cluster. A higher value of specificity degree means that the cluster has an isolated position very weakly linked with the rest of the cluster network. Mainly we observed three types of behaviour of the evolution of the cluster specificity degree: a) clusters maintaining it, b) clusters where it declines moderately, and c) clusters where it decreases strongly.
The first type corresponds to clusters having, in DM and WC maps, an isolated position. An interesting example is the cluster "*Infrared imaging*" that basically does not change its role in the cluster network.
In the second type, we are going to find clusters that in spite of a loss of specificity did not fundamentally change their position in the cluster network. It is the case of the clusters "*Gallium arsenides*", "*Photodetectors*" or "*Electroluminescent devices*" that indeed have relatively preserved their position in the cluster network.
The third type consists on clusters either presenting an intensification of their links to the rest of the network or completely changing their status from an isolated position to a linked one. As a notable example we have the cluster "*Instrumentation*" that presents the same number of documents in DM and WC but that shares them with a great number of other clusters in DM and wins in connectivity. But we have also the cluster "*Bolometers*" which remains isolated while receiving, in DM, a very important number of documents from other clusters more strongly connected.
Another interesting case is the cluster "*Organic light emitting diodes*" that is the only cluster with an increase of the specificity degree. This could explain its location in the cluster network of DM that shows this cluster slightly less connected.
From an other point of view, the analysis of the results obtained from the two methodological approaches operated in this work reveals some quite interesting convergence points. Let us consider the terms of the six pathways presented in table 1. According to their diffusion stage, the terms of the first period do not diffuse in a homogeneous way to the second one. Only 26% of the unusual terms (26 / 100) occur in the pathways 1, 2 and 3. 54% of the cross-section terms (50 / 92) stay in their stage (pathway 6), and 60% of the established terms (58 / 96) occur in the pathways 4 and 5.
The distribution of the terms present in the six pathways according to their presence in the clusters of both periods is given in Table 10. For example, in the first cell, the fraction 2/13 means that 2 terms among the 13 of the pathway 1 (table 6) are present in at least one cluster of each period.

Table 11. Distribution of the terms of the six pathways according to their presence in the clusters of both periods

		Period 2		
		Unusual	Established	Cross-section
Period 1	Unusual	**Pathway 1:** 2 / 13	**Pathway 2:** 3 / 11	**Pathway 3:** 1 / 2
	Established	-	**Pathway 4:** 21 / 44	**Pathway 5:** 7 / 14
	Cross-section	-	-	**Pathway 6:** 22 / 50

The inheritances between the same diffusion stage are the strongest specially for established and cross-section ones.

In order to analyse the role of these terms in the thematic structuring of the technological field we can also analyse the projection of the terms of the pathways on the cluster maps of the two periods. Concerning the terms of pathways 1 and 2 we can observe the stability of their location in isolated clusters in both periods. These unusual terms correspond to fundamental new concepts not frequent enough to participate to a cluster constitution. However we can observe that one of the two terms of pathway 1 in both periods (superconducting microbridges) seems to migrate from an isolated cluster ("*Superconducting microwave devices*") in period P1 to a cluster ("*Bolometers*") located in a quite important sub-network of clusters in period P2. Conversely, the term of pathway 3 in both periods (ternary compounds) declines, going from a very strongly positioned cluster ("*III-V semiconductors*") in period P1 to an isolated cluster ("*Ternary compounds*"). The terms of the pathways 4 and 6 appear, in the two period maps, dispersed in both isolated and linked clusters with nevertheless, in the second period, a low concentration in the isolated clusters. The spread of this great number of established and cross-section terms all over the clusters is not really surprising. Indeed, this characteristic shows that these terms are representative of all the clusters. Finally, the terms of the pathway 5 are, in both periods, concentrated in the principal cluster network. A deeper analysis is necessary to explain this observation.

Conclusions

The approach of alternate utilisation of different bibliometric and/or informetric methods and scientific expertise helps validating and completing the results obtained at each step of the work as well as to get the experts personal input on the matter at hand.

For the diffusion model it can be said that the bibliometric filter to assign keywords to stages and the splitting of the article corpus in two periods opens a view insight of the emergence of research topics in a technology. At a first glance it is a formal taxonomy to disassemble what is called a technology in its "atoms" of research and traces the breakthroughs if they happen. However, it is not a prognosis neither a prediction but terms in the unusual phase show potentials of future developments in a technology. To discuss well-founded emergences and research potentials it is important to mirror the results with experts who are part of the science producing system in the reflected technology.

The diachronic cluster approach allows to have a global view of the field landscape at two successive time periods. The analysis of the cluster contents and their relative position on the cluster maps supplies indications about their similarity with respect to their respective associated keywords. The observation of cluster maps allows to detect exceptional topics and interesting topic sub-sets. Let us point out nevertheless that the expert validation remains, at each step, absolutely necessary to validate and position the analysis results in the field context by giving them a scientific foundation.

The results coming from diffusion model lead to select the most important vocabulary without disturbing the cluster maps organization, and produce a new interpretation of diachronic clustering results by introducing a new term categorization that can be projected on the clusters.

Acknowledgments: This work was carried out thanks to a European Union funding: Project N° 15615 (NEST) - 6th Research and Development Framework Plan. The project acronym is PROMTECH, and the project full title is "Identification and Assessment of Promising and Emerging Technological Fields in Europe". The Consortium was composed by the Austrian Research Centers GmbH – ARC (Vienna, Austria), the Fraunhofer Institut für Systemtechnik und Innovationsforschung (Karlsruhe, Germany) and the Institut de l'Information Scientifique et Technique – INIST-CNRS (Nancy, France). We would also to warmly thank our colleagues from ARC and INIST-CNRS, particularly Mrs. Nathalie Vedovotto, that took part

very actively to the different project steps by bringing us their scientific and documentary expertises.

References

Armstrong J.S. & Green K.C. (2007). *http://www.forecastingprinciples.com/selection_tree.html*

Daim T. U., Rueda G., Martin H., Gerdsri P. (2006). Forecasting emerging technologies: Use of bibliometrics and patent analysis. *Technological Forecasting & Social Change*, 73 pp. 981-1012

Ferber R. (2003). Information retrieval. *Suchmodelle und Data-Mining-Verfahren für Textsammlungen und das Web, Heidelberg: dpunkt*

Gini Index (2008). *http://en.wikipedia.org/wiki/Gini_coefficient*

Kajikawa Y., Yoshikawa J., Takeda Y., Matushima K. (2008). Tracking emerging technologies in energy research: Toward a roadmap for sustainable energy. *Technological Forecasting & Social Change*, 75, pp.771-782

Lancaster F.W. , Lee JL. (1985). Bibliometric Techniques Applied to Issues Management: A Case Study. *Journal of the American Society for Information Science*, 36, 8, pp 389-397

Lelu A. (1993). Modèles neuronaux pour l'analyse de données documentaires et textuelles. *PhD Dissertation, Université de Paris 6*

Lelu A., François C. (1992). Hypertext paradigm in the field of information retrieval: A neural approach. *4th ACM Conference on Hypertext*, Milano, November 30th–December 4th

Mogoutov A., Cambrosio A., Keating P., Mustar P. (2008). Biomedical innovation at the laboratory, clinical and commercial interface: A new method for mapping research projects, publications and patents in the field of microarrays. *Journal of Informetrics*, 2, pp.341-353

Mogoutov A., Kahane B. (2007). Data search strategy for science and technology emergence: A scalable and evolutionary query for nanotechnology tracking. *Research Policy*, 36, pp.893-903

Noyons E. (2004). Science maps within a science policy context. *In: Handbook of Quantitative Science and Technology Research, Moed H.F., Glänzel W., Schmoch U (Eds.), Kluwer Academic Publishers*, London, pp. 237-255

Polanco X., François C., Royauté J., Besagni D., Roche I. (2001). Stanalyst®: An integrated environment for clustering and mapping analysis on science and technology. *In: Proceedings of the 8th ISSI*, Sydney, July 16th -20th

Robertson S. (2004). Understanding Inverse Document Frequency: On theoretical arguments for IDF. *Journal of Documentation*, 60, n° 5, 503-520

Roche I., Besagni D., François C., Hörlesberger M., Schiebel E. (2008). Identification and characterisation of technological topics in the field of Molecular Biology. *To be published*

Salerno M., Landoni P, Verganti R. (2006) The role of funded projects content analysis in early stage disciplines exploration: The case of nanotechnology. *In paper presented at the SPRU 40th anniversary conference – The future of science, technology and innovation policy*

Schiebel E., Hörlesberger M. (2007). About the identification of technology specific keywords in emerging technologies: The case of "Magnetoelectronics". *In: Proceedings of ISSI 2007, 11th International Conference of the International Society for Scientometrics and Informetrics, Torres-Salinas D., Moed H. F. (Eds.)*, Madrid, June 25th -27th, pp. 691-69

Spärck J. K., Robertson S. (2006). Inverse Document Frequency - The IDF page. *Retrieved November 22, 2006 from: http://www.soi.city.ac.uk/~ser/idf.html*

van Rijsbergen C.J. (1979). Information Retrieval, London: Butterworths

White H. D., McCain K. W. (1989). Bibliometrics. *Annual Review of Information Science and Technology*, vol. 24, pp. 119-186