Using 'core documents' for detecting new emerging topics

Wolfgang Glänzel[1,2] and Bart Thijs[3]

[1]*Wolfgang.Glanzel@econ.kuleuven.be*
Centre for R&D Monitoring (ECOOM) and Dept. MSI, K.U. Leuven, Leuven (Belgium)

[2]*glanzw@iif.hu*
Institute for Research Policy Studies (IRPS), Hungarian Academy of Sciences, Budapest (Hungary)

[3]*Bart.Thijs@econ.kuleuven.be*
Centre for R&D Monitoring (ECOOM) and Dept. MSI, K.U. Leuven, Leuven (Belgium)

Abstract
The notion of 'core documents', first introduced in the context of co-citation analysis and later re-introduced for bibliographic coupling and extended to hybrid approaches, refers to the representation of the core of a document set according to given criteria. In the present study, core documents are used for the identification of new emerging topics. The proposed method proceeds from independent clustering of disciplines in different time windows. Cross-citations between core documents and clusters in different periods are used to detect new, exceptionally growing clusters or clusters with changing topics. Three paradigmatic types of new, emerging topics are distinguished. Methodology is illustrated using the example of three ISI Subject Categories selected from the life sciences, applied sciences and the social sciences.

1. Introduction
The detection of new emerging research topics or sub-disciplines is one of the big challenges of contemporary scientometrics. The reason for the theoretical, methodological and practical difficulties, scientists are faced with when trying to identify such topics, are complex and sometimes even superposing. The most obvious problem is that scientists themselves might not always be aware that their research field is an emerging one. So the theoretical and practical question arises of when can we speak about a *new emerging topic*. New ideas underlying the research, a rapidly growing number of publications and scientists dealing with this topic is a necessary, however not yet sufficient condition. This new topic must also be characterised by coherence, a certain independence of its "mother topic" and other disciplines and must be largely self-sustaining, i.e., it should not simply be a satellite structure fostered by or supplying input for other research areas. The question of when such a field is considered "new emerging" is not only related to its "age" but also to the time when it's literature has reached a critical mass, which is necessary to exist as and to be recognised as a independent and self-sustaining structure.
The second, rather methodological question refers to finding techniques for detecting these new emerging yet coherent structures. The idea of searching these topics in the mirror of their scholarly literature is quite obvious. Also the use of text mining for this purpose is not far to seek. The easiest way of monitoring the emergence of research topics is certainly considering the growing frequency of specific terms within a given research area. However, textual similarity based on shared terms is also related to strong citation links. Thus Jo et al. (2007) have, for instance, analysed the correlation between the term distribution and the link distribution in the citation graph, and have found that citation connectivity is correlated with textual coupling of a term representing a topic. Shibata et al. (2008) have monitored the evolution of clusters by selecting characteristic terms for each cluster, and have shown that similar topics are strongly connected through cross-citation links while papers dealing with different topics are weakly connected. They concluded that the division of a given subject into strongly connected clusters is necessary for the detection of emergence. They also introduced two topological measures to determine the role of each paper in the citation network such that

nodes with the same role should be at similar topological positions. These measures were used to decide whether there are emerging clusters.
A different dynamic approach uses sub-classification of subject domains in the sciences, social sciences and humanities, which can be done based on text mining and textual similarities between documents; extracted terms can, in addition, be used in sophisticated ways for labelling and describing the obtained clusters (cf. Lamirel et al., 2008). One possibility to monitor structural changes and the evolution of the number of clusters lies certainly in the application of incremental methods; Lamirel et al. (2010) have recently developed a diachronic multi-source approach for mining research topics.
At the large scale, the *lexical* approach has several limitations which have been discussed in earlier papers by the authors (see Glänzel and Thijs, 2011). The relatively low discriminative power, resulting in "overestimating" relationships among documents, the dimensionality problem at the large scale as well as problems arising from the underlying vocabularies at different levels of aggregation and in different time periods form some of these limitations. The advantage of the combination of text-based methods with citation-based techniques have soon been recognised (cf. Braam et al, 1991a,b, Zitt and Bassecoulard, 1994). The superiority of such hybrid methods over text-based and link-based approaches have been shown, among others, by Glenisson et al. (2005), Boyack and Klavans (2010) and Liu et al. (2010). The question arises of how such combinations of text- and link-based components could be applied to diachronic analysis if each of the components causes specific problems in the application to long-term analysis. The inappropriateness of the application of both *bibliographic coupling* and *co-citation analysis* over periods, say, longer than ten years are caused by literature ageing and the genealogy of citations is self-evident. Furthermore, the use of co-citation analysis in the context of new emerging topics is questionable since it takes time before a 'critical mass' of papers on a new research topic is reached, which is needed to produce the highly cited publications needed for co-citation based clustering (see Hicks, 1987). The use of bibliographic coupling for long-term analysis, in turn, is mainly limited by the effect of *citation genealogy*. Citations form kind of *branching process* such that the probability that the reference lists of two related papers would still strongly overlap is rather low if ten, twenty years or more have elapsed between their publication. This experience has lead us to an alternative approach which is described in the following section.

2. Methods

The basic idea of how to overcome the above-mentioned limitations is to subdivide the space "vertically" and "horizontally" into disjoint subsets. "Vertical" here means splitting the document space along the time axis, while "horizontal" stands for the classification by topics within an individual time slice. The use of different, *disjoint* time slices has the advantage that both the textual *and* the citation-based component will properly work in their own relatively short time window. For the dynamic analysis, the two time slices should not overlap since joint document sets, which are partially present in both periods through the overlaps, might cause interdependence situations and thus create links between clusters of the two periods that do, otherwise, not exist on the basis of citation-based or textual similarities of different documents.
In shorter and more recent time windows bibliographic coupling has several advantages compared to co-citation analysis (Glänzel and Czerwon, 1996), and is therefore the best candidate for the hybrid classification. The time windows do preferably not overlap in order to obtain independent structures in each time window, where, of course, the same (hybrid) clustering algorithm should be applied. After the clustering has been conducted in each time window separately, all obtained clusters or classes representing research topics in the corresponding time slice form characteristic structures defined on the basis of the underlying

(hybrid) similarity measure. These structures are visualised in Figure 1; each time window t_1 and t_2 represents an individual document space with its own cognitive structure with, e.g., six clusters in t_1 and seven clusters in t_2, respectively. These link structure among the clusters in each time slice is visualised by dotted lines. The length of the lines stand for the distance, their thickness for the strength. While the same hybrid technique can be applied in both periods (t_1 and t_2), the determination of possible correspondence between t_1- and t_2-clusters – here indicated by solid lines between the two time slices in Figure 1 – remains the core problem of the detection of new emerging topics. Both sets represent disjoint document spaces and as argued above, neither bibliographic coupling nor textual analysis should be applied to long-term classification. A second problem arises from the fact that the obtained clusters are, notably at the level of local classification, usually strongly interlinked or possibly heterogeneous. In these cases, monitoring the evolution of complete clusters over different time periods and across disjoint spaces might become rather difficult. One possible solution lies in the representation of clusters or topics by sets of specific documents characterising the topics *and* in finding an appropriate link representation over time. The solution will be outlined in the following subsection.

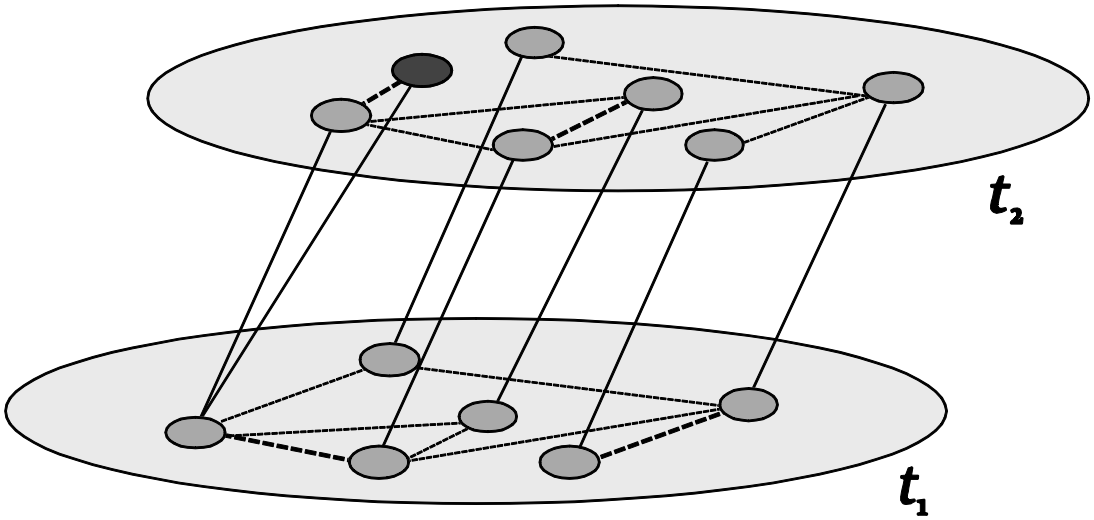

Figure 1. Sketch of a research field's changing topic structure over time (dotted lines represent internal structures, solid lines among the time slides t_1 and t_2)

2.1 Cluster representation for dynamic analysis

In an earlier study by Glänzel and Thijs (2011), the notion of 'core documents' has been extended to a hybrid approach using *bibliographic coupling* and *term frequencies*. We will recall this method in brief. The notion of a 'core' of literature has its roots in co-citation analysis (Small, 1973). The original idea by Glänzel and Czerwon (1996) was to define core documents as those publications that are strongly linked with at least a given number of other documents based on similarity measures derived from *bibliographic coupling*. For the extension, a textual component was used in accordance with the hybrid clustering algorithm. The textual component was based on stemmed terms extracted from titles and abstracts. In addition, keyword phrases were kept and stop words were removed. This hybrid approach was applied to the local level, i.e., to the aggregation level of smaller disciplines and research topics. It was stressed that simple linear combinations of the two similarity measures derived from coupling and term frequencies do often not provide satisfactory results. The reason is obvious. Citation-based and lexical similarities among individual documents are based on differently structured cosine measures as bibliographic coupling can be represented by

Boolean vector spaces (cf. Sen and Gan, 1983) and the results can be expressed by binary measures due to the uniqueness of the citation link between two document while textual similarity is based not only on the occurrence of terms but also on their frequencies. As a by-effect, hybrid similarities are usually dominated by the lexical component. In order to avoid possible biases, a *linear combination of the angles* underlying the citation- and text-based similarities has been proposed by the authors for the identification of the core documents. Consequently, core documents can be defined in the hybrid case analogously to the original definition by Glänzel and Czerwon (1996) as follows.

Definition: 'Core documents' are documents that have at least $n > 0$ links of at least a given strength $r \in (0,1)$ according to the given similarity measure. In the present case this measure r is defined as the cosine of the linear combination of the underlying angles, i.e.,

$$r = \cos\left(\lambda \cdot \arccos(\eta) + (1-\lambda) \cdot \arccos(\xi)\right), \quad \lambda \in [0,1],$$

where η is the similarity defined on bibliographic coupling and ξ the textual similarity. The λ parameter defines the *convex combination*, $\arccos(\eta)$ and $\arccos(\xi)$, respectively, denote the two underlying angles.

The determination of the two parameters n and r is mainly based on experience. We would, however, like to mention that r depends on the subject whereas n depends on both r and the time window. It is clear that we can expect more links of the same strength if we increase the time window. In small and rather homogeneous disciplines, or fields, where, for instance, citations are rather frequent, r could be increased. Nevertheless, the choice of the two parameters should remain within reasonable limits. Just to mention three examples, the choice $n = 3$ and $r = 0.71$ ($\sim 45°$) would probably results in identifying follow-ups of the same author while $n = 30$ and $r = 0.71$ would in most cases not result in any satisfactory output but $n = 50$ and $r = 0.09$ ($\sim 85°$) would most likely produce a heterogeneous, weakly interlinked set of papers. As mentioned in a previous paper (Glänzel and Thijs, 2011), core documents should ideally represent about 0.1% – 1.0% of the original set. In that paper, we have used $n \approx 10$, which proved an appropriate multipurpose threshold along with $r = 0.25$ or $r = 0.34$. The choice of $\lambda = 0.875$ and $\lambda = 0.833$, respectively, guaranteed a balanced combination of the two components (cf. Glänzel and Thijs, 2011). The study demonstrated that core documents can be successfully used to represent the outcomes of document clustering. In addition, the use of core documents reduces the original document space, dependent on the choice of the parameters, by about two orders of magnitude, and can thus be considered an efficient tool for "dimensionality reduction" as well. At the same time, their use might be completely independent of clustering exercises and they might serve as representatives of the core of any given document set.

Figure 2 visualises the link environment of a core document. The core document (UT code ISI:000236168700042) in the centre of the network is entitled "A three-dimensional, multicomponent, two-phase model for a proton exchange membrane fuel cell with straight channels" and has been publishes in the journal *Energy & Fuels* in 2006. The thickness of the edges reflects the strength of the links. Figure 2 also visualises the links among the related papers of document ISI:000236168700042. Note that core documents do not only have numerous strongly linked related papers but even more weaker connections with other documents, which are, because of its quantity, not shown in Figure 2.

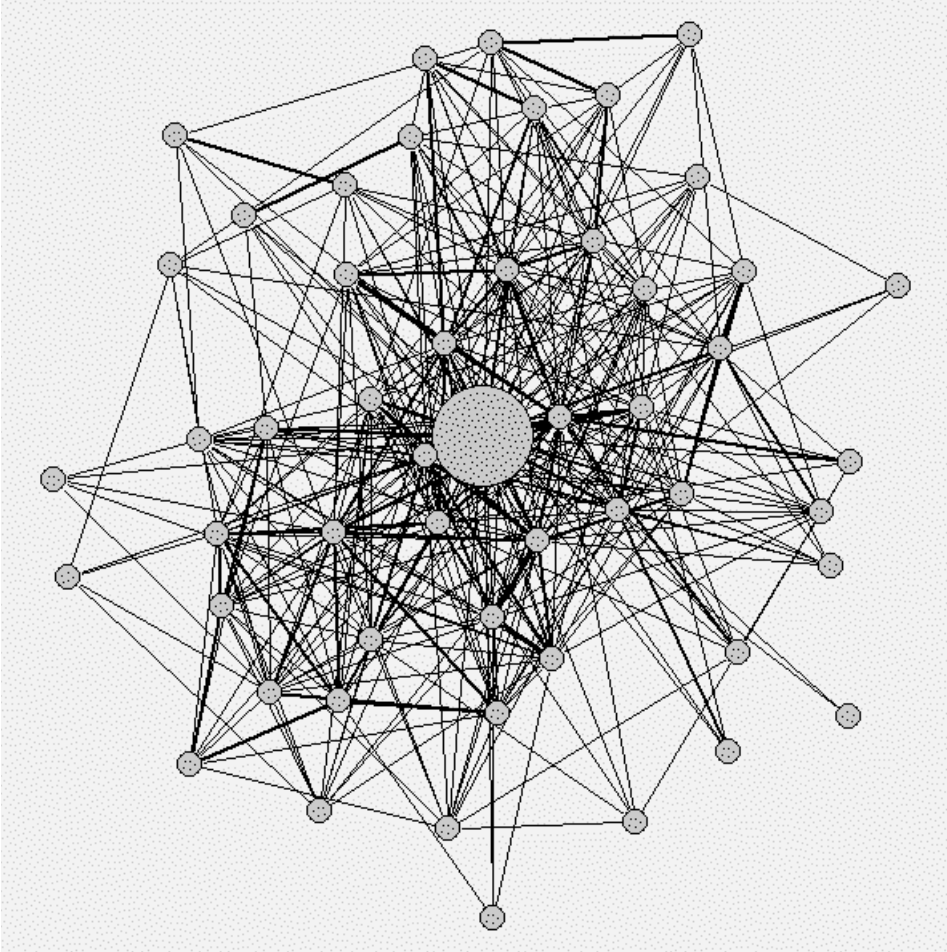

Figure 2 Visualisation of the link environment of a 'core document'
[Data sourced from Thomson Reuters Web of Knowledge (formerly referred to as ISI Web of Science]

Now the question arises of how the concordance between cluster solutions found in different time periods – as sketched in Figure 1 – can be analysed using core-document representation. Cross-citation analysis as proved means of classification of publication sets, for instance, scientific journals at a given time (cf. Leydesdorff, 2006, Zhang et al. 2009) can readily be extended to different, not necessarily overlapping time periods such as the results of independent cluster analyses in different time windows. In particular, citation links between core documents in one period to all publications in the clusters of the other period can be used to determine links between the structures of the two periods. This method is expected to reduce noise that might otherwise be caused by cross-citation links of less relevant documents far from the medoids of the clusters. On the other hand, cross-citation analysis between core documents alone usually provides an insufficient number of links necessary for creating the concordance. Therefore, whether the cited or the citing side should comprise the complete cluster and the other side should be restricted to the core-document representation only. The outcomes from such direct citation analysis are sketched in Figure 1 (solid lines). The following step concerns the detection of new emerging topics, provided that such topics exist in the discipline under study. Before we attempt to identify candidate topics in selected fields, we have a closer look at some main characteristics of emerging topics.

2.2 Characteristics of emerging topics

The structure in the time window t_2 in Figure 1 shows a new cluster (marked in black), which is strongly linked with the leftmost cluster. The question arises of whether this cluster might

be a candidate new topic branching off the leftmost cluster which is present in both periods. The detection of emerging topics is not merely based on the number and 'size' of occurring topics or clusters but also related to the cognitive-epistemological structure of the research field under study. In principle, one can distinguish three paradigmatic cases of cluster evolution. These three cases could indicate new, emerging topics.

(1) Existing cluster with an exceptional growth,

(2) Completely new cluster with its root in other clusters and

(3) Existing cluster with a topic shift.

We just mention in passing that evolution can also occur in the opposite direction in case (1) and (2), say, as declining or vanishing topics. In the following we will show the occurrence of these cases using the example of three selected disciplines related to life sciences.

3. Results

All data for the present study have been retrieved from Thomson Reuter's *Web of Science* (WoS). Only articles, reviews and proceedings papers published in journals have been taking into consideration. The document type "Letter" has been omitted for two reasons. Reference lists of letters are often extremely short and this document type tends to lack abstracts, keywords and sometimes even an appropriate title.
Three ISI Subject Categories have been selected, particularly,

- *public, environmental & occupational health,*
- *biomedical engineering,*
- *obstetrics & gynaecology.*

In a first step a hybrid cluster analysis of three selected ISI Subject Categories has been conducted in different time periods. In a second step, core documents in the obtained clusters has been determined. In order to obtain consistent results, the *same* similarity measures are used for both clustering and core-document representation.

3.1 Public, environmental & occupational health

The first Subject Category under study is "Public, environmental & occupational health". The first publication period (t_1) is 1999-2003, the second time window is 2004-2008. The hybrid cluster analysis provided six clusters for time windows t_1 which have been suggested based on corresponding silhouette values (cf. Rousseeuw, 1987). The number of clusters increased in the second period t_2 by one. Figure 2 presents both structures in the same diagram using Pajek (Batagelj and Mrvar, 2003). Clusters of the first period are marked in red, those of the second time window in green. The size of the circles is proportional to the number of papers in the clusters. One specific keyword each has been chosen to label the clusters. The selection was arbitrary and only designed to facilitate the identification on the map and the discussion of the results. Six clusters persisted over the period of ten years, among others the huge cluster on quality of life, the big clusters on environmental risks and gender-related issues, the somewhat declining work-related cluster, the community cluster (including health surveys) and the small one related to tobacco consumption. A new cluster emerged from HIV infection. Although HIV is also present in other clusters in both periods, where HIV infection or HIV risk behaviour is related to the main topics, for instance, to quality of live, health care, tobacco consumption, to violence, substance use and disorder, to prophylaxis, treatment or to cancer risk, the new one is mainly devoted to HIV-related stigma, to social, socio-political and regional aspects of HIV, with one focus on AIDS in Africa. The 25 most important keywords are as follows.

sex; aids; hiv; south africa; sub saharan africa; infection; human immunodeficiency virus; sexually transmitted diseases; transmission; hepatitis c; risk behaviors; epidemic; users; tanzania; adherence; new york city; individuals; africa; sexual behaviour; uganda; hiv infection; hiv/aids; malawi; hiv prevention; seroprevalence

The new cluster is of type (2). The relatively strong links with the clusters on gender, life quality and community issues are obvious. The new cluster comprises 1541 documents, and is thus still relatively small.

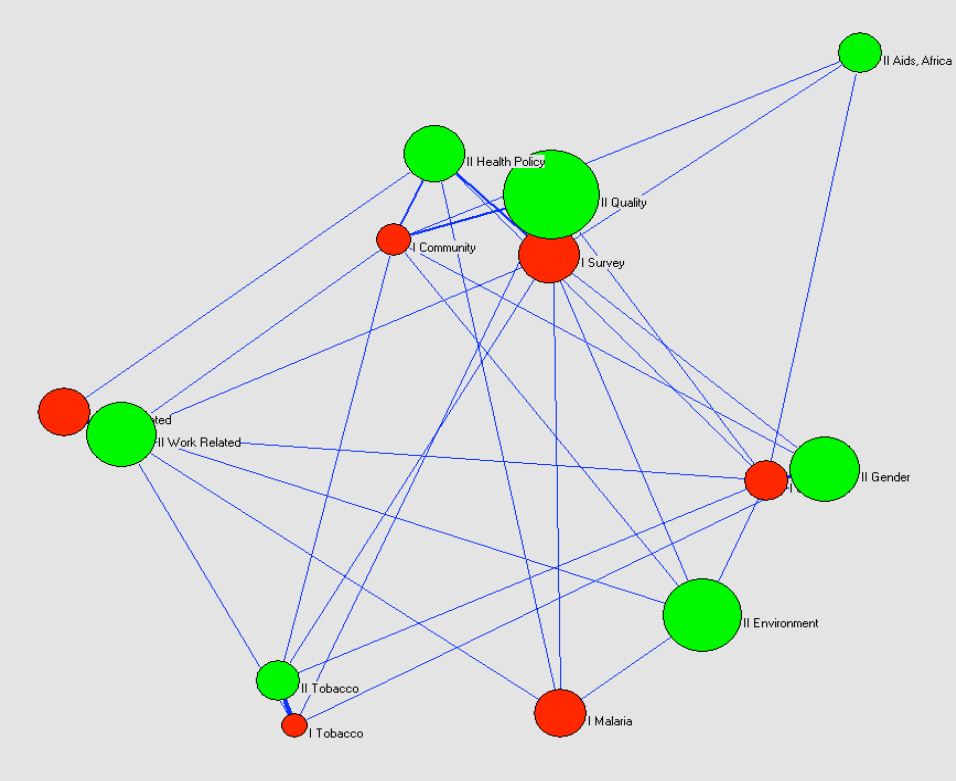

Figure 2. Cluster representation of the Subject Category 'Public health' according to Pajek; Kamada-Kawai (Red: 1999-2003, Green: 2004-2008)
[Data sourced from Thomson Reuters Web of Knowledge (formerly referred to as ISI Web of Science]

3.2 Biomedical engineering

The second subject category selected for this study is "Biomedical engineering". The underlying publication periods for this discipline are again 1999-2003 (t_1) and 2004-2008 (t_2). The hybrid cluster analysis provided eight clusters in the first and nine clusters in the second period. The structures in both periods are visualised in Figure 3 based on the Kamada-Kawai Model using Pajek. As in the previous Figure, clusters of the first period are marked in red, those of the second time window in green and the size of the circles is proportional to the number of papers in the clusters. In this Subject Category we have identified two emerging fields, the biggest cluster in the centre of the diagram has grown by more than one third as compared with period t_1. The topic, which we labelled as *Brain-Computer Interface* comprised 4160 papers in t_1 and 5638 documents in t_2. Even in the first period it was one of the biggest clusters, so that it can be considered type (1) according to the classification in subsection 2.2. The 25 most important keywords are as follows

signals; blood flow; classification; eeg; neural networks; pattern recognition; independent component analysis; hemodynamics; patterns; elasticity; elastography; models; simulation; velocity; identification; time series; carotid

artery; ultrasound; potentials; neural network; brain; localization; selection; recognition; numerical simulation

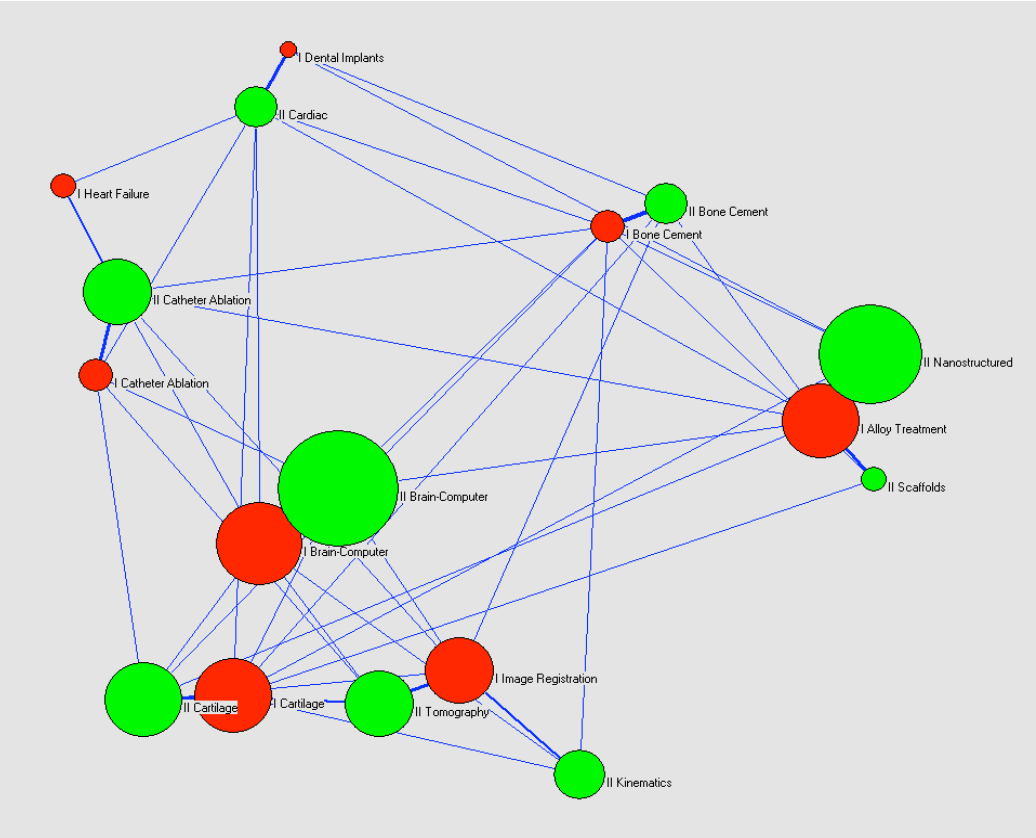

Figure 3. Cluster representation of the Subject Category 'Biomedical engineering' (Pajek; Kamada-Kawai) (Red: 1999-2003, Green: 2004-2008)
[Data sourced from Thomson Reuters Web of Knowledge (formerly referred to as ISI Web of Science]

Although the phrase 'Brain-Computer Interface' does not occur among the 25 most important keywords, acronyms (BCI) and variants of this phrase (e.g., brain machine interface, asynchronous brain interface) were among the most frequent terms in the titles of the core documents.

The second emerging topic is of type (2) again. We have labelled it *kinematics*. It is strongly linked to Image registration in t_1, and has medium-strong links with cartilage, bone cement, and BCI. Most relevant themes are joint and muscle kinematics during motion and the corresponding models. The 25 most important keywords are

muscle; knee; biomechanics; force; forces; joint; motion; stability; kinematics; reliability; movement; rehabilitation; injury; patterns; skeletal muscle; hip; parameters; stroke; performance; spine; lumbar spine; tendon; pain; kinetics; strategies.

The cluster contains 2639 documents.

3.3. Obstetrics & gynaecology

The third subject category is "Obstetrics & gynaecology". Again t_1 = 1999-2003 and t_2 = 2004-2008 have been chosen as disjoint time slices. The hybrid cluster analysis provided seven clusters for each period. The topic structures in both periods are visualised in Figure 4 using the same visualisation as in the previous Figures. In this Subject Category we have

identified one emerging fields, which can be characterised as type (3) according to the classification in 2.2.

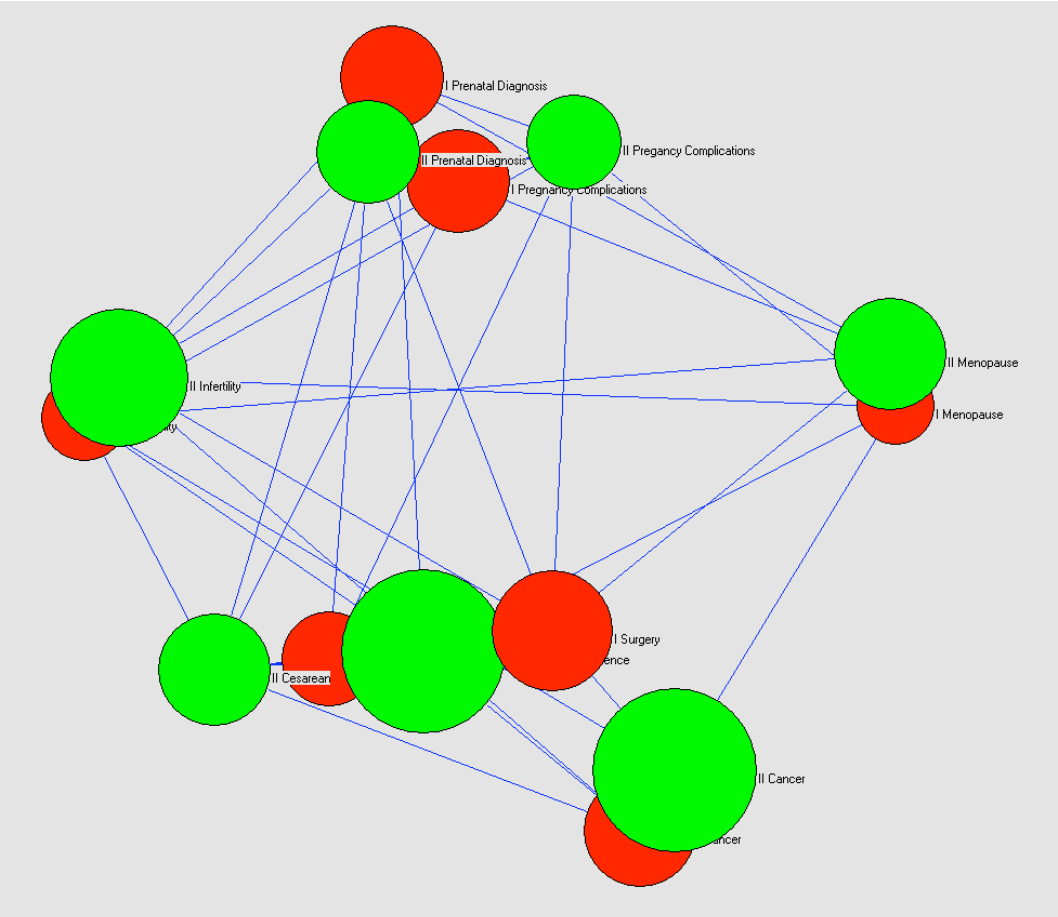

Figure 4. Cluster representation of the Subject Category 'Obstetrics & gynaecology' (Pajek; Kamada-Kawai) (Red: 1999-2003, Green: 2004-2008)
[Data sourced from Thomson Reuters Web of Knowledge (formerly referred to as ISI Web of Science]

The topic labelled *Prenatal Diagnosis* comprises 4823 documents in 2004-2008. The corresponding cluster in the first period was with 4782 papers just slightly smaller. However, we have found a certain focus shift in this topic which might justify to speak about a new emerging topic. The 25 most important keywords in t_2 are

> prenatal diagnosis; doppler; placenta; flow; ultrasound; hypoxia; anomalies; size; blood flow; fetus; congenital heart disease; blood; abnormalities; ultrasonography; pregnancies; three dimensional ultrasound; phenotype; 3 dimensional ultrasound; sheep; fetal; in utero; fetal heart rate; malformations; 3d ultrasound; trophoblast

The most important keywords used in both periods do not indicate any significant change since those comprises the same general terms and phrases, such as *prenatal diagnosis*, *doppler*, *anomalies*, *defects*, *ultrasound*, *blood flow, hypoxia, ultrasonography*, *heart, placenta*, *malformations, fetuses*, etc. The comparison of the topics of the core documents in the two periods, however, reveals some changes, which are, of course, not expected to be dramatic since the two periods are seamlessly adjoining.

Table 1. 30 core documents representing the cluster 'Prenatal diagnosis' in 1999-2003
[Data sourced from Thomson Reuters Web of Knowledge (formerly referred to as ISI Web of Science]

ISI-UT	*Document title*

000078007500003	Screening for chromosomal abnormalities at 10-14 weeks: the role of ductus venosus blood flow
000079002900050	Can nuchal cord cause transient increased nuchal translucency thickness?
000079446500004	Pathophysiology of increased nuchal translucency in chromosomally abnormal fetuses
000079446500005	The clinical relevance of fetal nuchal translucency
000080583900028	Fetal pulse oximetry: Duration of desaturation and intrapartum outcome
000080745800001	First-trimester nuchal translucency screening for fetal aneuploidy
000083214800006	Screening for fetal aneuploidies and fetal cardiac abnormalities by nuchal translucency thickness measurement at 10-14 weeks of gestation as part of routine antenatal care in an unselected population
000083563000007	Fetal pulse oximetry and other monitoring modalities - Future directions
000084383500003	Cardiac defects in chromosomally normal fetuses with abnormal ductus venosus blood flow at 10-14 weeks
000086703800002	Comparison of fetal cell recovery from maternal blood using a high density gradient for the initial separation step: 1.090 versus 1.119 g/ml
000087920300001	One-stop clinic for assessment of risk of chromosomal defects at 12 weeks of gestation
000089181600032	Pregnancy outcome and infant follow-up of fetuses with abnormally increased first trimester nuchal translucency
000089438000011	Fetal cells in cervical mucus and maternal blood
000089517900002	First trimester umbilical artery pulsatility index in fetuses presenting enlarged nuchal translucency
000089613200016	Increased fetal nuchal translucency: possible association with esophageal atresia
000090127600004	Nuchal translucency and its relationship to congenital heart disease
000166295800003	Intrapartum fetal pulse oximetry. Part 2: Clinical application
000167136700015	Reverse flow in the umbilical vein in a case of trisomy 9
000167281400005	Ductus venosus blood flow in chromosomally abnormal fetuses at 11 to 14 weeks of gestation
000167690400005	Relationship between fetal nuchal translucency and crown-rump length in an Asian population
000168001000014	The value of minor ultrasound markers for fetal aneuploidy
000168452800003	Ductus venosus studies in fetuses at high risk for chromosomal or heart abnormalities: relationship with nuchal translucency measurement and fetal outcome
000168452800004	The role of ductus venosus blood flow assessment in screening for chromosomal abnormalities at 10-16 weeks of gestation
000168452800013	Prenatal diagnosis of mosaic trisomy 8 in a fetus with normal nuchal translucency thickness and reversed end-diastolic ductus venosus flow
000168994000002	Early screening for chromosomal abnormalities: new strategies combining biochemical, sonographic and Doppler parameters
000170088300004	Fetal nuchal translucency and normal chromosomes: a long-term follow-up study
000170088300006	Screening for Down syndrome using first-trimester ultrasound and second-trimester maternal serum markers in a low-risk population: a prospective longitudinal study
000171275400009	Prediction of fetal anemia with Doppler measurement of the middle cerebral artery peak systolic velocity in pregnancies complicated by maternal blood group alloimmunization or parvovirus B19 infection
000171288100004	Nuchal translucency thickness and outcome in chromosome translocation diagnosed in the first trimester
000172328800012	Measurement of fetal nuchal translucency thickness by three-dimensional ultrasound

Table 1 presents thirty core documents of the cluster ‘Prenatal Diagnosis’ in the first period while Table 2 shows the core documents in the second period. The titles of the documents are presented along with the ISI–UT codes serving as unique document identifiers. The results show that, above all, TTTS/FFTS (*Twin-to-Twin Transfusion Syndrome*, also known as *Feto-Fetal Transfusion Syndrome*) and examination of *fetal nasal bones* have become important research topics. Many papers in the first period deal with the screening for cardiac or chromosomal abnormalities like Down-syndrome by the application of *nuchal translucency thickness* (NTT) measurement while in the latter period only one paper deals with NTT. Due to technological advancement in 3D ultrasound the attention shifted towards the fetal nasal bone and to the detection of other syndromes. This might illustrate that emerging topics are not necessarily new ones but might evolve within existing topics indicating new research directions.

4. Conclusions

Core documents proved to be more than supplements to term representation of the outcomes of clustering exercises. Although introduced independently from classification and clustering exercises in order to identify important nodes in the network of documented scientific communication (Glänzel and Czerwon, 1996), they can seamlessly be incorporated into the framework of (hybrid) clustering techniques (cf. Glänzel and Thijs, 2011). As was shown in the present study, core documents can be used to identify important nodes in the hybrid citation/lexical network of the found clusters and their cross-citations links with corresponding clusters in other periods can provide information about possible emergence. At the same time, these papers form an ideal representation of emerging topics with interesting

properties for the information retrieval since following their links one can readily identify further relevant documents related to their research topics.
Finally, we would like to mention some validity issues as well. For a number of ISI disciplines, including , and subjects that have been delineated on the basis of special search strategies, we have asked experts who have confirmed our findings. Thus emerging topics identified by the above-described method and the danger of false positives is minor and not very likely. However, the requirement of a critical mass of documents might imply the danger of not identifying new emerging fields if the number of relevant publications does not meet this criterion. In such situations most quantitative and computerised techniques will be faced with similar problems that can only be overcome by communication with experts.

Table 2. 30 core documents representing the cluster 'Prenatal diagnosis' in 2004-2008
[Data sourced from Thomson Reuters Web of Knowledge (formerly referred to as ISI Web of Science]

ISI-UT	*Document title*
000188734200003	Middle cerebral artery Doppler velocimetric assessment in two cases of hydrops fetalis without fetal anaemia
000188798500003	Maternal ethnic origin and fetal nasal bones at 11-14 weeks of gestation
000189184600003	Pilot study on the midsecond trimester examination of fetal nasal bone in the Chinese population
000220287300003	Likelihood ratio for trisomy 21 in fetuses with absent nasal bone at the 11-14-week scan
000220287300005	Assessment of the fetal nasal bone at 11-14 weeks of gestation by three-dimensional ultrasound
000220805400009	Fetal anemia: new technologies
000223418300003	How can we diagnose and manage twin-twin transfusion syndrome?
000225437500001	A case of extreme unconjugated fetal hyperbilirubinemia
000225683400004	The association between fetal nasal bone hypoplasia and aneuploidy
000227927800003	Multicenter study of first-trimester screening for trisomy 21 in 75,821 pregnancies: results and estimation of the potential impact of individual risk-orientated two-stage first-trimester screening
000227962700014	Improving the accuracy of fetal foot length to confirm gestational duration
000229361200012	First-trimester ductus venosus, nasal bones, and Down syndrome in a high-risk population
000229361200029	Contemporary treatments for twin-twin transfusion syndrome
000229812000009	Qualitative venous Doppler flow waveform analysis in preterm intrauterine growth-restricted fetuses with ARED flow in the umbilical artery - correlation with short-term outcome
000231276700010	Relationship between nuchal translucency thickness and prevalence of major cardiac defects in fetuses with normal karyotype
000232013700038	Correlation between middle cerebral artery peak systolic velocity and fetal hemoglobin after 2 previous intrauterine transfusions
000232013700056	Antenatal predictors of neonatal outcome in fetal growth restriction with absent end-diastolic flow in the umbilical artery
000232382400009	Insights into the pathophysiology of twin-twin transfusion syndrome
000232382400010	Management of fetofetal transfusion syndrome
000234160600003	Arterial and venous Doppler in the diagnosis and management of early onset fetal growth restriction
000234387300003	Twin-Twin Transfusion Syndrome: Where do we go from here?
000234813400009	Doppler and biophysical assessment in growth restricted fetuses: distribution of test results
000234813400010	Perinatal outcome in monochorionic twin pregnancies complicated by amniotic fluid discordance without severe twin-twin transfusion syndrome
000235556300003	Ultrasonographic evaluation of fetal nasal bone in a low-risk population at 11-13+6 gestational weeks
000235556300011	Mid-facial anthropometry in second-trimester fetuses with trisomy 21: a three-dimensional ultrasound study
000236945900004	The relation between fetal nasal bone length and biparietal diameter in the Korean population
000238491600005	Nasal bone length at 11-14 weeks of pregnancy in the Korean population
000238926700017	Nasal bone in first-trimester screening for trisomy 21
000239896300008	Interest of foetal nasal bone measurement at first trimester Trisomy 21 screening
000239931300012	First-trimester examination of fetal nasal bone in the Chinese population

Acknowledgement

Methodology has partially been developed in the context of the ERACEP project within the Coordination and Support Actions (CSAs) of the ERC work programme. The authors wish to acknowledge this support.

References

Batagelj, V. & Mrvar, A. (2003). *Pajek – analysis and visualization of large networks*. In: M. Jünger, P. Mutzel (Eds.), Graph drawing software. Springer, 77–103.

Boyack, K.W. & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology*, in press.

Braam, R.R., Moed, H.F. & van Raan, A.F.J. (1991a). Mapping of science by combined cocitation and word analysis, Part 1: Structural aspects. *Journal of the American Society for Information Scie*nce, 42 (4), 233–251.

Braam, R.R., Moed, H.F. & van Raan, A.F.J. (1991b). Mapping of science by combined cocitation and word analysis, Part II: Dynamical aspects. *Journal of the American Society for Information Science*, 42 (4), 252–266.

Glänzel, W. & Czerwon, H.J. (1996), A new methodological approach to bibliographic coupling and its application to the national, regional and institutional level. *Scientometrics*, 37, 195–221.

Glenisson, P., Glänzel, W., Janssens, F. & de Moor, B. (2005), Combining full text and bibliometric information in mapping scientific disciplines. *Information Processing & Management*, 41 (6), 1548–1572.

Glänzel, W. & Thijs, B. (2011), Using 'core documents' for the representation of clusters and topics. *Scientometrics*, in press, doi: 10.1007/s11192-011-0347-4.

Hicks, D. (1987), Limitations of co-citation analysis as a tool for science policy. *Social Studies of Science*, 17, 295–316.

Jo, Y., Lagoze, C. &, Giles, CL (2007), *Detecting research topics via the correlation between graphs and texts*. KDD-2007 Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 370-379.

Janssens, F., Glänzel, W. & de Moor, B. (2008), A hybrid mapping of information science. *Scientometrics*, 75 (3), 607–631.

Lamirel, J.C., Ta A.P. & Attik, M. (2008), *Novel labeling strategies for hierarchical representation of multidimensional data analysis results*. In: A. Gammerman (Ed.), Proceedings of the 26th IASTED International Conference on Artificial Intelligence and Applications, 11-13 February 2008, Innsbruck, Austria. ACTA Press, Track 595-138, p. 169–174.

Lamirel, J.C., Safi, Gh., Pryankar, N., & Cuxac, P. (2010), *Mining research topics evolving over time using a diachronic multi-source approach*. The Fourth International Workshop on Mining Multiple Information Sources - ICDM 2010.

Leydesdorff, L. (2006), Can scientific journals be classified in terms of aggregated journal-journal citation relations using the Journal Citation Reports? *Journal of the American Society for Information Science and Technology*, 57 (5), 601–613.

Liu, X., Yu, S., Janssens, F., Glänzel, W., Moreau, Y. & De Moor, B. (2010), Weighted Hybrid Clustering by Combining Text Mining and Bibliometrics on Large-Scale Journal Database. *Journal of the American Society for Information Science and Technology*, 61(6), 1105–1119.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65.

Sen, S.K. & Gan, S.K. (1983), A mathematical extension of the idea of bibliographic coupling and its applications. *Annals of Library Science and Documentation*, 30, 78–82.

Shibata, N., Kajikawa, Y., Takeda, Y. & Matsushima, K. (2008), Detecting emerging research fronts based on topological measures in citation networks of scientific publications. *Technovation* 28 (11), 758–775.

Small, H. (1973), Cocitation in scientific literature - new measure of relationship between 2 documents. *Journal of the American Society for Information Science*, 24 (4), 265–269.

Zhang, L., Glänzel, W. & Liang, L. (2009), Tracing the role of individual journals in a cross-citation network based on different indicators. *Scientometrics*, 81 (3), 821–838.

Zitt, M. & Bassecoulard, E. (1994), Development of a method for detection and trend analysis of research fronts built by lexical or cocitation analysis. *Scientometrics*, 30 (1), 333–351.