Epistemic fields identification and dynamics visualisation from a scientific content database

David Chavalarias, Jean-Philippe Cointet

Detecting epistemic fields dynamics from a scientific content database

David Chavalarias[1] & Jean-Philippe Cointet[2]

Abstract
Massive collections of scientific publications are now available on-line thanks to multiple public platforms. These databases usually cover large-scale scientific production over several decades and for a broad range of thematic areas. Today researchers are used to perform queries on these databases with keywords or combination of keywords in order to find articles associated to a precise scientific field. This full text indexation performed for millions of articles represents a huge amount of public information. But instead of being used to characterize articles, can we revert the standpoint and use this information to characterize concepts neighborhood and their evolution? In this paper we give a yes answer to this question looking more precisely at the way concepts can be dynamically clustered to shed light on the way paradigm are structured. We define an asymmetric paradigmatic proximity between concepts which provide hierarchical structure to the scientific database upon which we test our methods. We also propose overlapping categorization to describe paradigms as sets of concepts that may have several usages.

Keywords
Mapping and visualisation of knowedge ; publication analysis ; coword analysis ; paradigmatic evolution ; paradigmatic proximity.

[1] CREA, Ecole Polytechnique, 1, rue Descartes, 75005 Paris, France david.chavalarias@polytechnique.edu

[2] CREA & TSV (Social and Political Transformations related to Life Sciences and Life Forms), INRA.

Introduction

Modern acceptation of paradigm has been provided by T. Kuhn (1970) as "an entire constellation of beliefs, values and techniques, and so on, shared by the members of a given community". He contended that, a paradigm enables a group of scientists to focus its efforts on a well-defined range of problems. Once they belong to a paradigmatic field, scientists no longer need to explain extensively the meaning of each concept used. A paradigm enables the scientific community to reach a consensus concerning de definition of important problems and identification of techniques needed to solve them, and last but not least for our purpose, which set of concepts shall be used to share their breakthrough. In the following we will call such sets *paradigmatic fields*.

The figure 1 represents a schematic view of scientific knowledge production. Authors $\{A_i\}$ publish papers $\{P_i\}$ that contain informative sets of concepts $\{C_i\}$. Some of these publications have been co-authored while some concepts may be strongly co-occurring with others. On this scheme, we linked authors that have co-authored an article, and concepts that co-occurred in at least one paper. Thus we can isolate different paradigmatic fields - a paradigmatic field being defined as a strongly co-occurring set of concepts which can be defined in graph theory as a dense subset of the conceptual network. Our example features two overlapping paradigmatic field, the first one is made of the set of concepts $\{C_1,C_2,C_3\}$, the second one is made of $\{C_3,C_4\}$. We will voluntarily disregard the collaboration side (on the left) in the following to concentrate on the conceptual network we built.

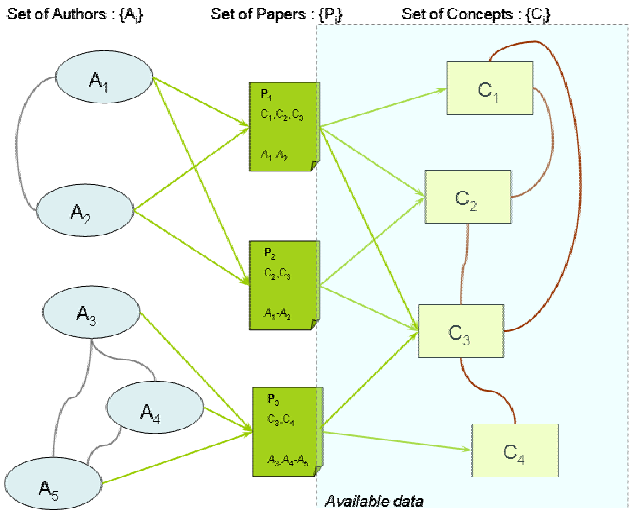

Figure 1: **Methodology scheme**: scientific knowledge production scheme: a set of authors $\{A_i\}$ produce publications $\{P_i\}$ in their scientific communities. This structure reappears on the right side of the scheme on the paradigmatic field made of overlapping community of concepts $\{C_i\}$

Our assumption is that paradigmatic fields found in public sphere of knowledge production provide a direct insight into the very structure of science and researchers communities: there is a deep correlation between the social community structure (left part of the figure) and the conceptual community structure (right part).

The aim of this paper is to present tools for automatic bottom-up identification of paradigmatic fields linked with scientific communities' structures from an article database. The strength of our approach is that is does not require other information than the one already available in most existing database to reconstruct the multi-scale structure of paradigmatic fields. In particular, it does not require an access to the content of each articles (full text, abstracts or titles) nor a particular linguistic treatment on words. Rough statistics about occurrences and co-occurrences of words in untagged documents are sufficient.

A simple measure of *paradigmatic proximity* henceforth noted P_p is defined between key-words and is used to perform paradigmatic field detection. This bottom-up approach also aims at describing paradigmatic fields evolution through mere statistics on key-words occurrences and co-occurrences, over a 25 years period. First explanatory results are given.
Although the angle here is the one of scientific knowledge production, the same method can be applied to get global insights of any kind of electronic database, in particular blogs or webpages.

Context and rationale
Scientometric research deals with study of science or technology using quantitative data. One of its prominent objective is the development of information systems that may help science studies practitioners or searchers to navigate into the outstanding mass of scientific papers published every day around the world. That is the reason why a great number of methods for automatically designing conceptual maps have been proposed. Doyle (1961) was one of the first to point to the fact that traditional document retrieval techniques are ineffective in finding relevant documents due to a lack of semantic understanding of relevance. Since then, several methods have been proposed to inject "intelligence" into scientific database management. The two main methods developed have been "citation-based analysis" and "co-word analysis". These methods are generally bottom-up which means that they do not need any supplementary information than the ones enclosed in the very articles they are trying to map.
Citation-based analysis can be of two kinds. In the "bibliometric coupling" a similarity measure is built between two documents according to the frequency with which they are cited together (Small, 1973), other methods called bibliographical coupling link preferentially document which share the same set of references (Salton, 1963). Co-word analysis usually tries to map concepts landscape using exclusively statistics on the number of co-occurrences of a word with another. A classical statistic in co-word analysis is the similarity index measured as the ratio between the number of co-occurrences between two words divided by the product of number of total occurrences of a and b (Callon, Bauin, 1983, Callon, Courtial and Laville 1991). Once this data has been collected clustering algorithms like kohonen maps algorithms are used to provide smarter navigation tools in articles databases thanks to conceptual mapping of a wide research area (Lin and Soergel, 1991, Sun, 2004). Many approaches also propose to use both words occurrences and references to help producing knowledge maps (Besselaar and Heimeriks, 2006).

In our paper we claim that co-word analysis is a fruitful way to analyze large scientific database. We show that it is possible to exhibit hierarchical structure in the basic original information with the sole help of statistics on our original database. We explain our intuitive idea of paradigmatic proximity in the next section and explicit its formal expression in section 4. Our method is then tested on a very large scientific database (see section 5) before some preliminary results are given in the static and dynamical cases (section 6). We finally describe few perspectives that open our methodology.

What can online search engine tell us about scientific concepts?
It is now part of everyday life. When you want to find an article related to a concept A you enter a request in your favourite search engine and get within a second the total number of papers dealing with this concept as well as the first n links to the most relevant articles. To be more precise, you can refine your request to "A AND B". Now, if we associate each concept with the set of articles that mention this concept in full text (or title, abstract, etc.), the above means that we have at least the two elementary tools of set theory: the set of articles that mention concept "a" (A of size |A|) and the set of articles that contain both concept "a" and concept "b" ($A \cap B$). These elements of set theory also enable to use conditional probability ; if it is known that a publication contains word "b", then the probability that it also contains "a" is defined to be the conditional probability of A given B: $p(A|B)=|A \cap B|/|B|$.

As we shall see, these simple notions enable to define measures of paradigmatic proximity that are highly relevant to characterize paradigmatic fields and their inner structure. Moreover, since articles can be clustered by year of publication, it is possible to get the dynamics of the paradigmatic proximity that happens to be relevant to track the evolution of paradigms.

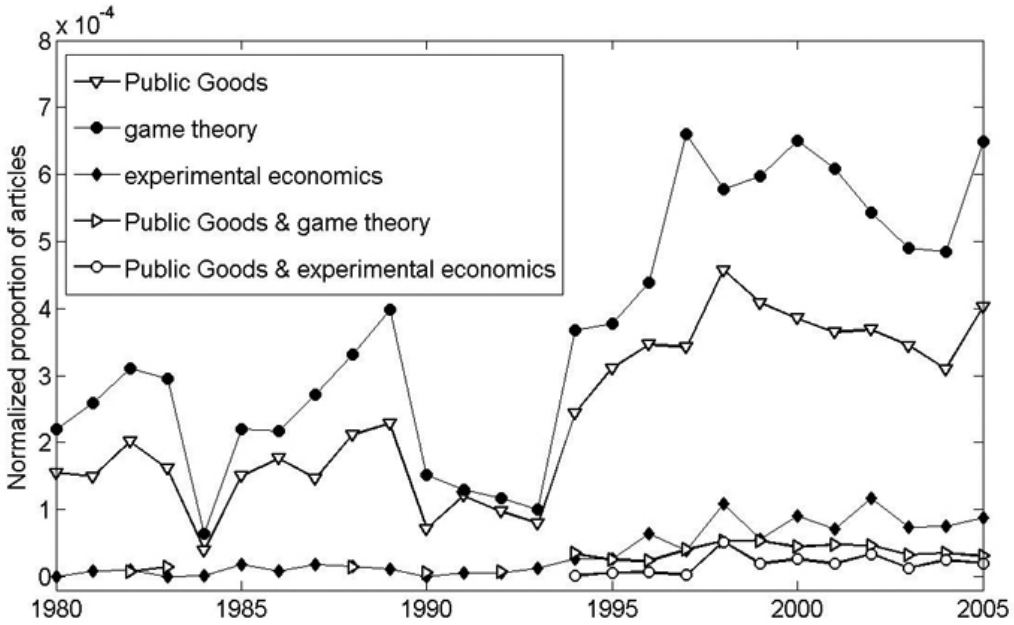

Figure 2: Comparative dynamics of occurrences and co-occurrences of concepts related to *Public Goods*. *Game theory* and *Experimental economics* are both relevant concepts for the study of public goods. But the concept “experimental economics” is more specific than “game theory”

Let's take an example. On the figure 2 we plotted together occurrences and co-occurrences of "*Public Goods*", "*Game theory*" and "*Experimental economics*". "Game theory" and "Experimental economics" are both relevant concepts for the study of public goods. But the concept "Experimental economics" is more specific than "Game theory". Specific terms in game theory related to public goods would have been "Ultimatum game", "Prisoner's dilemma", etc... It means that researchers using "Prisoner's dilemma" for example shall certainly belong to the wider community of game theorists. This notion of degree of specificity in word usage is important and suggests that we might want to have a parameter to tune the desired specificity. But this notion is not clear if we look only at co-occurrences from the point of view of "public goods": P(experimental economics|Public goods) and P(game theory|Public goods) are of the same order of magnitude. On the contrary, P(Public goods|experimental economics) is much higher than P(Public goods|game theory). This means that the concept of public goods is widely used in experimental economics studies but is less central in game theory. If we want to define a paradigmatic proximity that could distinguish "game theory" from "Experimental economics" we should thus use the both kind of conditional probabilities. This notion of degree of specificity is important and suggests that we might want to have a parameter to tune the desired specificity.
Moreover, whereas a majority of papers in experimental economics deals with public goods, the reverse is not true and there are probably scientists working on public goods that never worked on experimental economics studies. The paradigmatic proximity should thus be *asymmetric* to reflect this kind of situations.
We can summarize the different kinds of situations that might be encountered :

1. **P(A|B) high, P(B|A) high :** A and B are in the same paradigm and have about the same degree of specificity,
2. **P(A|B) low, P(B|A) high :** B is general relatively to concept A (e.g. A:public good and B:game theory),
3. **P(A|B) high, P(B|A) low :** B belongs to a sub-domain relatively to A (e.g. A:Game theory and B:public good),
4. **P(A|B) low, P(B|A) low :** A and B are weakly relevant to each other,

We will now try to define a paradigmatic proximity such that it could be possible to discriminate the three first cases and eliminate the last one.

Paradigmatic proximity definition

Classical scientometrics statistics uses number of concepts occurrences and co-occurrence in a given time window. Starting from an article database with *N* articles, for given concepts *i* and *j*, let's note n_i^t and n_j^t the number of occurrences of *i* and *j* on the time window *t* and n_{ij}^t the number of co-occurrences.

From the above, there are some properties that we wish our paradigmatic proximity *Prox* to compel:

1. $Prox(i,j)=0$ *if* $n_{ij}^t=0$
2. $lim_{[(nij^t)/(ni^t)]\to 0}(P_p(i,j))=0$
3. $P_p(i,i)=1$
4. $P_p(i,j)$ is growing with n_{ij}^t as larger co-occurrences sets illustrate higher paradigmatic proximity. $P_p(i,j) = f(n_{ij}^t)$, *f* being a growing function.
5. $P_p(i,j)$ should depend on n_i^t and n_j^t, so that if one of them is growing $P_p(i,j)$ will decrease. It follows that $P_p(i,j)=f(n_{ij}^t,n_i^t,n_j^t)$, *f* being a growing function according to its first coordinate and a decreasing function according to the two others.
6. Last, we will have to estimate the paradigmatic proximity on a representative sample of the set of articles in the fields (typically a collection of journals). Under the assumption that the sample is representative we want the estimation to be independent of the sample's size. This means that we also wish that semantic proximity between two words to be independent of the total number of articles in the database to be an homogeneous function of n_{ij}^t,n_i^t,n_j^t i.e. $f(\lambda n_{ij}^t, \lambda n_i^t, \lambda n_j^t)=f(n_{ij}^t,n_i^t,n_j^t)$. From this property we deduce that we can write f as a function of n_{ij}^t/n_i^t and n_{ij}^t/n_j^t. From condition 5 we deduce that *f* is a growing function according to its two new reduced coordinate.

From property 2 we expect our distance to be null when $n_{ij}^t \to 0$. Hence if we write the Taylor development of *Prox* in 0 we should have : $Prox(x,y) = \alpha_0 + \alpha_{11}x + \alpha_{12}y + \alpha_{21}x^2 + \alpha_{22}y^2 + \alpha_{23}xy + \alpha_{31}x^3 + \alpha_{32}y^3 + \alpha_{33}xy^2 + \alpha_{34}x^2y + \ldots + o(\Sigma_{j=1}^{i-1} x^j y^{n-j})$ we can deduce *that* $\alpha_0 = 0$, $\alpha_{11} = \alpha_{12} = 0$ and so on.... Hence *Prox* can be written as the sum of crossed products:

$$P_p([(n_{ij}^t)/(n_i^t)],[(n_{ij}^t)/(n_j^t)]) = \Sigma_{i=1}^{\infty}\Sigma_{j=1}^{i-1} \alpha_{ij} (n_{ij}^t/n_i^t)^j (n_{ij}^t/n_j^t)^{i-j}.$$

The simplest class of functions that fit this Taylor development in 0 as well as all the above conditions are the Cobb-Douglas functions $f_{\alpha,\beta}(x,y)=x^\alpha y^\beta$. From 6, *f* is growing and thus $\alpha>0$ and $\beta>0$. We thus decide to define the paradigmatic proximity by:

$$P_p^{\alpha,\beta}(i,j)=(n_{ij}^t/n_i^t)^\alpha(n_{ij}^t/n_j^t)^\beta \ : \alpha>0, \beta>0$$

From this expression, it is straightforward to see that given a concept *i* and looking for the closest concepts *j*:

- $1 >> \alpha > 0$ will favors concepts j such that P(j|i) is low,
- $\beta >> 1$ will favor concepts j such *that P(i|j)* is low,

For $\alpha = 1$ and $\beta = 1$, the paradigmatic proximity has an intuitive interpretation : it is the probability that an article contain both concepts *i* and *j* in the database ($[(n_{ij})/N]$) over the probability that an article would contain both concepts *i* and *j* if co-occurrences of *i* and *j* where random ($[(n_i)/N][(n_j)/N]$). The classical equivalence index is thus a particular case of our paradigmatic measure for $\alpha = \beta = 1$.

In this article, we will focus on the relations of paradigmatic proximity qualified by "specificity" or "generalization", i.e. on cases 2 and 3. To reduce the parameter space, we will reduce our investigations to a parameterized expression of $P_p^{\alpha,\beta}$ noted P_p^α with $\alpha > 0$. Given the remarkable

symmetrical proximity for $\alpha = \beta = 1$ the condition we choose is that $P_p^{\alpha}(i,j)=P_p[1/(\alpha)](j,i)$ i.e. if a concept j is qualified as more specific from the point of view of *i* (case 3), then changing α for $[1/(\alpha)]$ will enable to detect concept *i* as a general neighbor from the point of view of *j* (case 2) the values of paradigmatic proximities being the same in both cases.
We will thus further consider the sub-class of function:

$$P_p^{\alpha}(i,j)=(n_{ij}^t/n_i^t)^{\alpha}(n_{ij}^t/n_j^t)^{1/\alpha}$$

As we shall see we can describe with this distance the way a concept belongs to a sub-field of a target concept or on the contrary how a target concept belongs to a sub-field of another concept.

We will now use this paradigmatic proximity measure to explore a given set of concepts with two different approaches. The first one can be defined as concept-centered. We will study neighborhoods of concepts in function of α (specific or generic paradigmatic proximity). At low value of α, we catch the most precisely expressions near our target concept. When rising up α, we access to more generic expressions. The second approach is a global mapping of the scientific field treated. We designed methods to describe dynamics of high level properties such as community structure.

Case Study

Methodology
The case study presented here focuses on a set of concepts coming from two data sources: a set of keywords for complex systems field associated with European project in IST FP6 and FP7 Cordis's database (765 keywords generously provided by Joseph Fröhlish's Arc System team - see appendix for the collection protocol); a set of words collected among our colleagues favourite keywords (about one hundred). We then got a convention with online platform of one of the major scientific publisher in order to collect the number of occurrences and co-occurrences per year of theses concepts from *1975* to *2005* in the full text of articles. The database gathered more than *20.000.000* indexed articles.
The computational resources provided by our partners enabled to perform about one query per second on their database. To get our database in a reasonable time we first restrained our set of concepts to about 448 keywords (given in appendix)[3]. This reduced the foreseen query time to 70 days. Since co-occurrences are very demanding in terms of server availability, we also decided to do a query on a co-occurrence only if the two queries on single terms gave a non zero result for authors keywords. This reduced to 7 days the data collection process. Consequently our database is built on all query results for single terms in full text from 1975 to 2005, and every query results on full text co-occurrences for couples of concepts that both appeared at least once as author keywords the year considered.
This database enables us to compute the paradigmatic proximity for any time-window from 1975 to 2005. In case of a computation on a time-window between $year_1$ and $year_2$, we thus have the following extended formulation of paradigmatic proximity:

$$Prox^{\alpha}(i,j,[Y_1 \ldots Y_2])=\left(\frac{\sum_{t=Y_1\ldots Y_2}(n_{ij}^t)}{\sum_{t=Y_1\ldots Y_2} n_i^t}\right)^{\alpha}\left(\frac{\sum_{t=Y_1\ldots Y_2} n_{ij}^t}{\sum_{t=Y_1\ldots Y_2} n_j^t}\right)^{[1/(\alpha)]}$$

[3] The list of keywords will be available on http://bibliography.free.fr/epistemics.htm with the agreement of ARC Systems.

We will now give some examples of application of our paradigmatic proximity measure. It should not be forgotten that the clusters and thematic fields that we will exhibit are conditional to our database of 450 concepts. There could be some more relevant concepts for the reader that will not be found because of database incompleteness.

Paradigmatic neighborhoods

Our paradigmatic proximity enables to define neighborhood of a target concept "*i*" given a threshold *s* and an α value by :

$$V_{s,\alpha}^{t}(i)=\{j / Prox^{\alpha}(i,j,t) > s\}$$

This neighborhood structure defined for each value of α outline relations of specification or generalization. On the example of *public goods* (*cf.* Fig. 3), we can see that as α increases, words in a the neighborhood of *public goods* become more specific and closer to the concepts used by specialists of the fields. We thus get concepts that sharply qualify areas of investigations about *public goods*). Note that such a visualisation could also be used to navigate in a concept map with specific tools to zoom in or zoom out according to the specificity or generality of concepts sought.

Figure 3: Neighborhood for the word *Public Goods* for different values of α. Inferior part: α = 10, the 10 closest concepts which are specific to *Public Goods* ; superior part: α = 0.1, the 10 closest concepts that are more generic than *Public Goods*.

Identification of paradigmatic fields

Once we have defined a proximity measure and neighborhoods, the next issue is to draw a concept map in the line with scientometric studies (Buter and Noyons, 2002, Marshakova-Shaikevich, 2005). Looking at the bottom part of figure 3, we can see that there seem to be two distinct spheres of knowledge that use the concept *public goods*. One usage is rather *game theory* oriented, as the other is rather used as a political sciences concept. For example, *Public Goods* is linked to *procedural invariance* and to *collective action* but paper mentioning both *collective action* and *procedural invariance* do not exist. These two concepts belong to two different spheres of knowledge production.

To automatically exhibit these multiple usages and identified set of concepts reflecting scientific activity, we need a broader view of the conceptual landscape taking into account the relations between the different concept's neighborhoods.

Given an α value, we need to categorize our data on the basis of the values of the paradigmatic proximity $Prox^{\alpha}$. Since a word can have several meanings and can be used in several scientific communities, the categorization algorithm should be able to assign a word to several different clusters. One successful method in line with this requirement is the k-cliques percolation algorithm (Palla, Derenyi, Farkas and Vicsek, 2005) that operates on graphs of concepts to detect communities. One method to generate a concepts graph based on our proximity measure, is to define a threshold *s* and to link each concept *i* to its neighbors in $V_{s,\alpha}(i)$. To avoid linking very generic words to any words we fixed the maximum number of neighbors to *30*, taking the *30* closest when neighborhood size was superior. This enables to build a non-directed graph on concepts. Then we can apply the k-clique percolation algorithm which outlines communities of concepts that qualify distinct spheres of

knowledge production. We illustrate this overlapping categorization displaying the to paradigmatic fields identified around *public goods*[4] in the period 2003-2005 (cf. fig 4). We observe that it indeed belong to two communities in our concepts set.

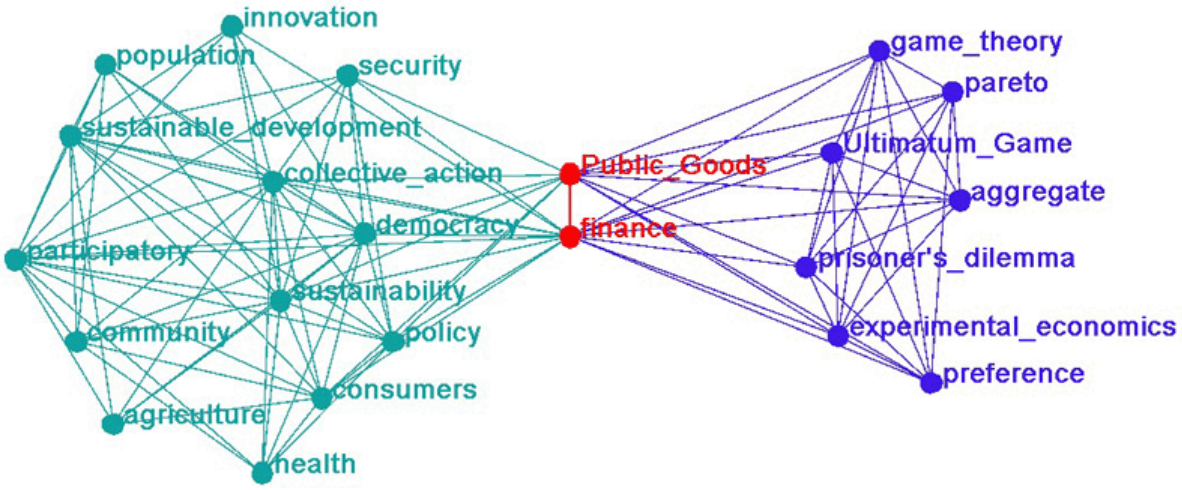

Figure 4: The two paradigmatic fields of *public goods* in 2003-2005 (conditionally to our set of 450 concepts). *Public Goods* (as well as *finance*) belongs to two spheres of knowledge production, one game theory oriented, the other political sciences oriented.

It shoud be emphasized that this this is visualisation is complementary to the one of neighborhoods. Here only neighbors that satisfies global conditions appear. These fields outline tends in science, with the degree of specificity tuned by α. Paradigmatic neighborhoods also contains concepts that are "new" in the field and could be used to detect future trends if we could anticipate the dynamics of their evolution.

Dynamics of paradigmatic neighborhoods

Dynamical science mapping is another challenge that allow to describe dynamical pattern in science evolution (Braam, Moed and Van Raan, 1991, Garfield, 2004). Static visualization based paradigmatic proximity is only partially informative. In the example of figure 3, we learn that *Public goods* belong to the class of *social dilemma.* This is not very informative since this will always be the case. What is really informative about use of concepts is the way the paradigmatic proximity varies over time outlining long trends, fads or new approaches and new concepts in the underlying communities. Can we detect automatically emerging approaches and sub-fields? The simplest visualization of this kind of evolution is to look at evolution of paradigmatic neighborhoods through time. Given a target word and a threshold *s*, we can plot for each time-window *t* (here we took the tree past years of the year indicated on the *Ox* axis) the set of words belonging to the neighborhood $V_{s,\alpha}^{t}(i)$ as given fig. 5. We can thus visualize the dynamical evolution of a target concept's neighborhood as illustrated in figure 5.

[4] Other examples as well as the whole graph generated by Cfinder is available on http://bibliography.free.fr/epistemics.htm.

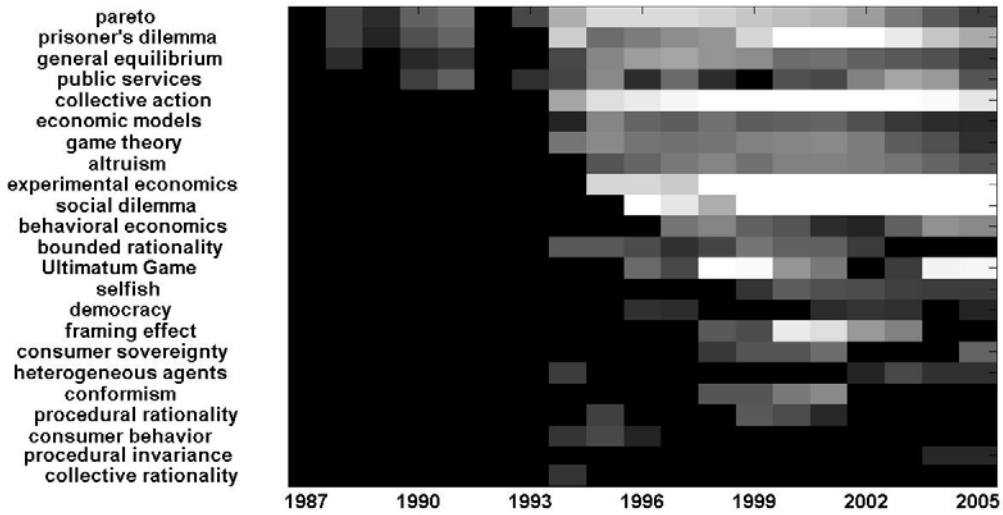

Figure 5: Dynamical view of the evolution of the neighborhood of *Public goods* from 1987 to 2005 for $\alpha = 10$ (a 3-years time-window for each point). Black means that the concepts were below the threshold the year considered. The lighter the square, the higher the paradigmatic proximity. We can see that public goods studies developed a lot these last years, mostly in their game theoretical aspect. Among emerging concerns in the fields we find the relation with heterogeneous agents and the question of procedural rationality. These observations fit well with what we actually observe in public goods studies.

Perspectives and conclusions

In this paper we proposed a method to use articles database indexation to characterize paradigmatic neighborhood of concepts, their evolution and we sketched methods to provide high-level descriptions of our set of concepts that we called.paradigmatic fields. More precisely we looked at the way concepts can be dynamically clustered to shed light on the way paradigm are structured.

The next step may be to integrate time related data in this high-level description, in order to have a dynamical evolution of the paradigmatic fields and not only paradigmatic neighborhoods. We could then describe evolution of coherent paradigmatic groups as described figure 3. One question is how this evolution is coupled to the dynamical view illustrated figure 5.

Another challenge is to reintroduce the directionality in the high-level description we developed. The sets of concepts clusterer with the k-clique percolation algorithm is all flat and links are non weighted. Yet our original data exhibited asymmetric relations between concepts according to the value of α. It would be interesting to describe cohesive paradigmatic fields not only as clouds of concepts but as a three dimensional structure by adding the dimension illustrated figure 3.

Another interesting perspective would be to extend the definition of word neighbourhood to a group of word neighborhoods. Hence we can imagine to define paradigmatic distance between of concepts in our corpus and a selected group of key words of special interest for a user.

References

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

Buter R. & Noyons E.. Using bibliometric maps to visualise term distribution in scientific papers. In *Sixth International Conference on Information Visualisation (IV'02)*, pages 697-702, 2002.

Callon, M., Courtial, J. P. and Laville, F.: Co-Word Analysis as a Tool for Describing the Network of Interactions between Basic and Technological Research: the Case of Polymer Chemistry. Scientometrics, 22 (1), 155-205, 1991.

Callon J. C. M. & Bauin S. From translation to problematic networks: an introduction to coword analysis. *Social Science Information*, 22:191-235, 1983.

Doyle. L. B. Semantic road maps for literature searchers. *J. ACM*, 8(4):553-578, 1961.
Garfield E. Historiographic mapping of knowledge domains literature. *Journal of Information Science*, 30(2):119-145, 2004.
Kuhn T. S. *The Structure of Scientific Revolutions*. 1970.
Lin X. & Soergel D. A self organizing semantic map for information retrieval. In *Proc. 14th International SIGIR Conference*, pages 262-269. Chicago, 1991.
Marshakova-Shaikevich I. Bibliometric maps of field of science. *Infometrics*, 41(6), 2005.
Palla G, Derenyi I., Farkas I. & Vicsek T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435:814, 2005.
Peter van den Besselaar G. H. Mapping research topics using word-reference co-occurrences: A method and an exploratory case study. *Scientometrics*, 68(3), 2006.
Salton G. Associative document retrieval techniques using bibliographic information. *J. ACM*, 10(4):440-457, 1963.
Small H. G., Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of American Society for Information Science*, 24(4):265-269, 1973.
Sun Y., Methods for automated concept mapping between medical databases. *J. of Biomedical Informatics*, 37(3):162-178, 2004.

Appendix : Collection of concept list from Cordis FP6 and FP5 IST database

This is a summary of how the 765 keywords related to complex systems were gathered by the ARC System team (http://www.systemsresearch.ac.at).
A set of documents were retrieved from FP6 and FP7 database by employing the following search terms :

Search term	number of projects
ADAPTATION + COMPLEX	129
SELF-ORGANIZATION	20
EMERGENCE	149
RANDOMNESS	9
CHAOS	48
COMPLEXITY	670
FRACTALS	24
POWER LAW(S)	5
AGENT-BASED MODELS	1
SIMULATION + COMPLEX	449
CELLULAR AUTOMATA	6
AUTO-CATALYSIS	1
COEVOLUTION	9
GENETIC ALGORITHM	23
ARTIFICIAL LIFE	1
NONLINEAR DYNAMICS	9
SELF-REPRODUCTION	0
UNIVERSAL TURING MACHINE	9
COMPUTABILITY	5
CRITICALITY	27
INTRACTABILITY	1
SINGULARITY	10
UNCERTAINTY	305

1525 projects were found that were reduced by hand to 900 relevant projects (this procedure being, unfortunately, not fully reproducible). Using the database fields 'title', 'subject index' and 'general description', these 900 projects were fed into our automated keyword generator (a feature of BibTechMon(R)), and after a manual standardisation procedure, we ended up with 765 keywords representing the content of the 900 relevant projects.