Science mapping with asymmetric paradigmatic proximity

Jean-Philippe Cointet, David Chavalarias

Science mapping with asymmetric co-occurrence analysis: methodology and case study on the complex systems community

David Chavalarias[1],[2] and Jean-Philippe Cointet[1],[3]

[1] CREA CNRS-Ecole Polytechnique, 1 rue Descartes, 75005 Paris, France
david.chavalarias@polytechnique.edu, http://chavalarias.com
[2] Institut des Systèmes complexes de Paris Ile-de-France
[3] TSV (Social and Political Transformations related to Life Sciences and Life Forms), INRA, 65 Boulevard de Brandebourg, 94205 Ivry-sur-Seine Cedex France
cointet@shs.polytechnique.fr

Summary. We propose new innovative methods in order to reconstruct paradigmatic fields thanks to simple statistics over a scientific content database. We first define an asymmetric paradigmatic proximity between concepts which provides hierarchical structure over the set of concepts. We propose to implement overlapping categorization to describe paradigmatic fields as sets of concepts that may have several different usage. Concepts can also be dynamically clustered providing a high-level description of the evolution of the paradigmatic fields. We apply our set of methods on a case study from the Complex Systems Community through the mapping of the dynamics of more than 400 *Complex Systems Science concepts* indexed in a database of of several millions of journal papers.

Keywords: Mapping and visualisation of knowledge ; publication analysis ; co-word analysis ; paradigmatic evolution ; paradigmatic proximity, asymmetric similarity measure.

Introduction

Modern acceptation of paradigm has been provided by 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. A paradigm enables the scientific community to converge toward a consensus concerning the 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 have co-occurred in one paper at least. 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 dynamics. Following Actor Network Theory which would describe science dynamics as the enrollment and juxtaposition of heterogeneous actants (LATOUR, 2005), we treat concepts found in publication as "stigmergic markers" that make possible the stabilization of scientific work. In this complex network, the meaning of every concept is modified each time a new association is built.

We thus focus onto the conceptual side of this complex affiliation network in order to describe the scientific field as an overlapping set of paradigmatic fields.

We shall then define paradigmatic field as a strongly co-occurring set of concepts which corresponds 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.

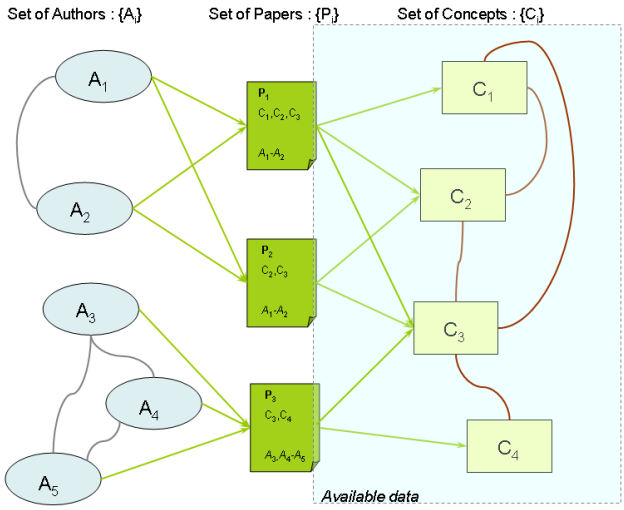

Fig. 1. scientific knowledge production scheme : a set of authors $\{A_i\}$ produce publications $\{P_i\}$ which are made of concepts $\{C_i\}$. We defined paradigmatic field as strongly cooccurring set of concepts.

The aim of this paper is to present tools for automatic bottom-up identification of paradigmatic fields directly from articles database. The strength of our approach is that is does not require other information than the one already available in most existing database to dynamically reconstruct the multi-scale structure of paradigmatic

fields. Rough statistics about occurrences and co-occurrences of words in indexed documents are sufficient. In particular, it does not require a real access to the content of each articles nor a particular linguistic treatment on words.

A simple measure of *paradigmatic proximity* henceforth noted P_p is defined between a set of given concepts and is used to perform paradigmatic field detection. This bottom-up approach also aims at describing paradigmatic fields evolution through mere statistics on concepts occurrences and co-occurrences, over a 25 years period. First explanatory results are given.

Although the context here is the one of scientific knowledge production, the same method may be applied to get global insights of any kind of electronic database (patents, blogs or webpages, etc...)

1 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 worldwide every day. 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 do intelligent 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 lexical statistics of the articles database being surveyed.

Citation-based analysis can be of two kinds. On one hand "Bibliometric coupling" builds a similarity measure between documents according to the frequency with which two documents are cited together (SMALL, 1973; LEYDESDORFF & VAUGHAN, 2006), on the other hand "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 about the number of co-occurrences of a concept with another. A classical statistics in co-word analysis which has been extensively documented in literature (M. CALLON, 1983; CALLON ET AL., 1991; NOYONS & VAN RAAN, 2002) is the similarity index measured as the ratio between the number of co-occurrences between concepts a and b divided by the product of the number of total occurrences of a and b . 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 & SOERGEL, 1991; SUN, 2004). Many approaches also propose to use both concepts occurrences and references to help producing knowledge maps (PETER VAN DEN BESSELAAR, 2006).

In our paper we claim that co-word analysis is a fruitful way to analyze massive scientific database. We show that it is possible to exhibit hierarchical structure in the basic original information with the sole help of statistics extracted from 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 related to our methodology.

2 What can indexed scientific databases tell us about paradigmatic fields ?

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 favorite search engine and get within a second the total number of papers dealing with this concept. To be more selective, you can refine your request to "A AND B". At this point we can associate each concept with the set of articles that mention this concept. At this step, we have at least the two elementary tools of set theory : the number of articles that mention concept A (set size) and the number of articles that contain both concept A and concept B ($A \cap B$). As we shall see, these two simple notions enable to define measures of paradigmatic proximity that are highly relevant to characterize paradigmatic fields. 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.

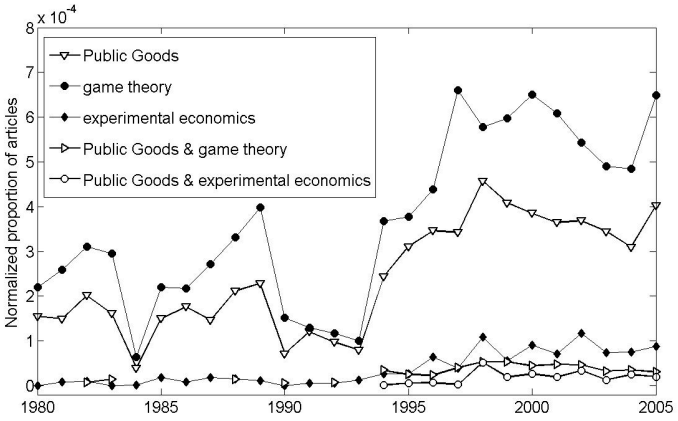

Fig. 2. Comparative dynamics of occurrences and co-occurrences of concepts related to *Public Goods*

Let's illustrate our point with 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... 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. Then, if we switch the reference concept, $P(Public\ goods|experimental\ economics)$ 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 these kinds of situation.

We can summarize the different possible scenario that might be encountered as follows:

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.

3 Paradigmatic proximity definition

Classical scientometric 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 for the time window t and n_{ij}^t the number of co-occurrences for the same time range.

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

1. $P_p(i,j) = 0$ if $n_{ij}^t = 0$
2. $lim_{\frac{n_{ij}^t}{n_i^t} \to 0} (P_p(i,j)) = 0$
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 concepts to be independent of the total number of articles in the database to be an homogeneous function of n^t_{ij}, n^t_i, n^t_j i.e. $f(\lambda x, \lambda y, \lambda z) = f(x, y, z)$. From this property we deduce that we can write f as a function of n^t_{ij}/n^t_i and n^t_{ij}/n^t_j

When $n_{ij^t} \rightarrow 0$ we expect our distance to be null. Hence if we write the Taylor development of P_p in 0 we should have : $P_p(x, y) = \alpha_0 + \alpha_{11}x + \alpha_{12}y + \alpha_{21}x^2 + \alpha_{22}y^2 + \alpha_{23}xy + \alpha_{31}x^3 + \alpha_{32}y^3 + \alpha_{33}xy^2 + \alpha_{34}x^2y + ...$ From assumption 2) we can deduce that $\alpha_0 = 0$, $\alpha_{11} = \alpha_{12} = 0$ and so on.... Hence P_p can be written as the sum of crossed products : $P_p(\frac{n^t_{ij}}{n^t_i}, \frac{n^t_{ij}}{n^t_j}) = \Sigma^{\infty}_{i=1}\Sigma^{i-1}_{j=1}\alpha_{ij}(n^t_{ij}/n^t_i)^j(n^t_{ij}/n^t_j)^{i-j}$.

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

$$P^{\alpha,\beta}_p(i, j) = (n^t_{ij}/n^t_i)^{\alpha}(n^t_{ij}/n^t_j)^{\beta}$$

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

- $1 >> \alpha > 0$ will favor 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 proportional to the probability that an article contain both concepts i and j in the database ($\frac{n_{ij}}{N}$) over the probability that an article would contain both concepts i and j if co-occurrences of i and j where random ($\frac{n_i}{N}.\frac{n_j}{N}$). The classical similarity 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 limit the parameter space, we will reduce our investigations to a parameterized expression of $P^{\alpha,\beta}_p$ noted P^{α}_p with $\alpha > 0$. Given the remarkable symmetrical proximity for $\alpha = \beta = 1$ the condition we choose is that $P^{\alpha}_p(i, j) = P_p\frac{1}{\alpha}(j, i)$ i.e. if a concept j is qualified as more specific from the point of view of i (case 3), then changing α for $\frac{1}{\alpha}$ 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 :

$$P^{\alpha}_p(i, j) = (n^t_{ij}/n^t_i)^{\alpha}(n^t_{ij}/n^t_j)^{1/\alpha}$$

As we shall see, this distance will enable to describe 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.

4 Methodology

The case study presented here focuses on a set of concepts coming from two sources : a set of key-words for complex systems associated with European projects in IST Cordis's database from FP6 and FP7 (765 key-words generously provided by the Arc System team lead by Joseph Frohlish -

see appendix for the collection protocol) ; a set of key-words collected near colleagues (about one hundred).

As for the scientific articles database, we established a partnership with Scirus, Elsevier's free science-specific search engine (www.scirus.com) in order to collect the number of occurrences and co-occurrences per year of these concepts from 1975 to 2005 in the full text of the articles. The database gathered more than $20.000.000$ indexed scientific covering a wide range of scientific content platforms[4].

To collect necessary statistics in a reasonable time we first had to restrain our set of concepts to 448 key-words (which are given in appendix). 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 key-words" (each concept has been mentioned by at least one author as an article key-words for the year considered). 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 key-words the year considered.

This database enables to compute the paradigmatic proximity for any time window from 1975 to 2005. If we choose a time range between years $Y1$ and $Y2$, we thus have the following extended formulation of paradigmatic proximity:

$$P_p^{\alpha}(i,j,[Y_1...Y_2]) = (\frac{\sum_{t=Y_1...Y_2}(n_{ij}^t)}{\sum_{t=Y_1...Y_2} n_i^t})^{\alpha}(\frac{\sum_{t=Y_1...Y_2} n_{ij}^t}{\sum_{t=Y_1...Y_2} n_j^t})^{\frac{1}{\alpha}}$$

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 448 concepts. There could be some more relevant concepts for the reader that will not be found because of the database incompleteness.

5 Case Study

5.1 Paradigmatic neighborhoods

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

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

[4] BioMed Central, Crystallography Journals Online, Institute of Physics Publishing, MEDLINE/PubMed, Project Euclid, ScienceDirect, Scitation, Society for Ind. & App. Mathematics and Pubmed Central

This neighborhood structure defined for each value of a outline relations of specification or generalization. On the example of public goods (cf. Figure 3), we can see that as α increases, concepts 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 searched.

Fig. 3. Two kinds of neighborhood of concept *Public Goods*. Inferior part: $\alpha = 10$, the 10 closest concepts that specify *Public Goods* ; superior part: $\alpha = 0.1$, the 10 closest concepts that are more generic than *Public Goods*.

5.2 Identification of paradigmatic fields

Once we have defined a similarity measure, and a neighborhood, we can try to draw knowledge map which is a common goal in scientometric literature (BUTER & NOYONS, 2002; MARSHAKOVA-SHAIKEVICH, 2005). Looking at the bottom part of figure 3, we can see that it seems to coexist two distinct spheres of knowledge that use the concept "public goods". The first usage is rather "game theory" oriented, as the other is rather used as a political science concept. For example, *Public Goods* is linked to *procedural invariance* and to *collective action* but there no studies mentioning both *collective action* and *procedural invariance*. These two concepts belong to two different spheres of knowledge production and might be associated to two different meanings. Contextual information enables us to exhibit automatically these multiple usages.

To automatically exhibit these multiple usages and identify set of keywords reflecting scientific activity, we need a broader view of the keywords' landscape taking into account the relations between the different keywords' neighbors. Given an α value, we need to automatically categorize our data according to the values of the paradigmatic proximity P_p^α. Since a word can have several meanings and can be used in several scientific communities, the categorization algorithm should make it possible for a keyword to belong to several different clusters. One successful method in line with this requirement is the k-clique percolation algorithm (PALLA ET AL., 2005) that operates on graphs of keywords to detect communities. Hence we generate a keywords graph based on our proximity measure by fixing a threshold s and linking each keyword i to its set of neighbors: $V_{s,\alpha}(i)$. To avoid linking very generic words

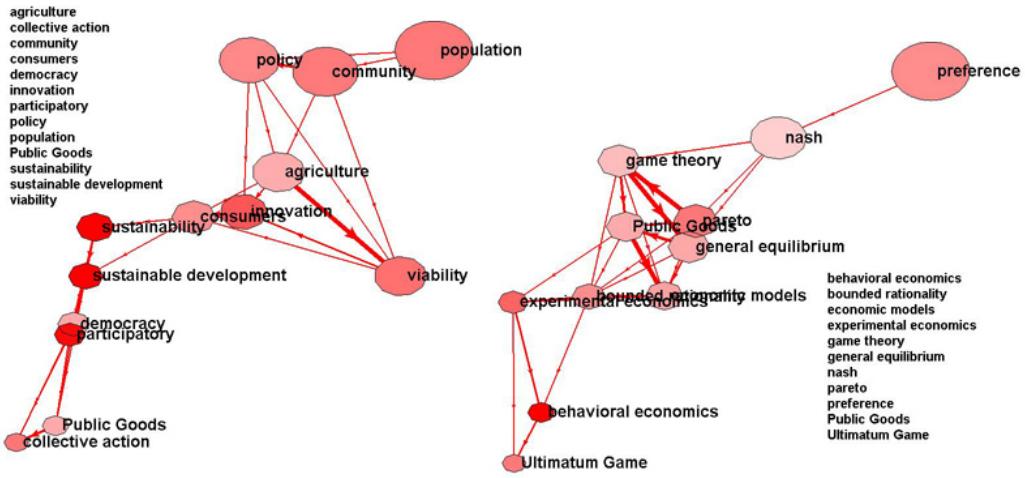

Fig. 4. Two paradigmatic fields mentioning the term *public goods*. *Public Goods* belongs to two spheres of knowledge production, one game theory oriented, the other political sciences oriented. Within a cluster, from left to right i_s decreases, from top to down i_n decreases. The size of points is a growing function of the mean frequency of the corresponding terms in the database on the period considered (here 2002-2005). For each plot, sizes are normalized so that the biggest points have always the same size. Colors indicates the activity of the fields, i.e. the relative and normalized mean increase in frequency of the concepts in the database between the period 1999-2002 and 2002-2005. Blue colors correspond to negative growth rates, red to positive growth rates. A full red point means that the concept's occurrences has increased of at leat 150% compared to the period 1999-2002). Arrows are proportional the proximity measure between two words, the direction of the arrow is from more general to more specific. For clarity, only couples with highest proximity have been plotted.

to any words we limit the maximum number of neighbors to 30, taking the 30 closest when neighborhood size is superior. This enables to build a non-directed graph on keywords. Then we can apply the k-clique percolation algorithm which outlines communities of keywords that qualify distinct spheres of knowledge production. The output of this algorithm is clusters of keywords such that within each cluster, there is a k-clique percolation. Clusters are general and are a property of α and s. Hence there is no need for additional data like a predetermined keyword to extract these clusters. To recover the asymmetric aspect of our paradigmatic measure, we then recompute the distance between words within a cluster C and define two quantities characterizing a words w within this cluster : the neighborhood index i_n and the specificity index i_s defined as follow :

The neighborhood index tells to what extent the cluster C is a good neighborhood for the word w with respect to the paradigmatic proximity P_p^α considered. It is the mean of "out-paradigmatic measures" from word w and is defined by :

$$i_n(w) = \frac{1}{card(C)} \sum_{w' \in C} P_p^\alpha(w, w')$$

The specificity index tells to what extent the word w is specific to the cluster C with respect to the paradigmatic proximity P_p^α considered (*i.e.* is w relevant for the terms in C ?). It is the mean of "in-paradigmatic measures" from word $w' \in C$ to w and is defined by :

$$i_s(w) = \frac{1}{card(C)} \sum_{w' \in C} P_p^\alpha(w', w)$$

These two indexes enable to plot an intuitive 2D embedding a cluster. We assign to each word the coordinate (i_s, i_n). We then compute for each couple (w, w') the asymmetry measure with respect to the paradigmatic proximity P_p^α defined as

$$a(w, w') = Proxm^\alpha(w, w) - Proxm^\alpha(w', w)$$

This enables to draw an arrow between two words in C representing the strength of the asymmetry. Last, we map the size of the font so that is reflects the number of occurrences of words in the corpus.

To illustrate this, we present here two cliques that share the concepts *public goods* in the period 2003-2005 (cf. fig. 4). As we noted above, this concept indeed belongs to several communities in our concepts set, one more game theory oriented, the other more political sciences oriented.

On these graphes we can see that in the part of the "political sciences sphere", defined here by the general terms *population*, *policy* and *community*, the expression *public goods* is related to the domain of *sustainable development* and *viability*, particularly in *agricultural* issues. Moreover, it is a sub-domain of "political sciences" quite specific since it is at the left edge of the cluster.

On the contrary, in the part of the "game theory sphere" defined here by the general terms *game theory*, *Nash* and *trust*, the expression *public goods* is quite in the middle of the cluster indicating that public good studies are a quite important domain in game theory. As we can see, in this context, public good studies address the questions of *pareto* states, *altruism* and *selfishness* and more specifically researches are done on *experimental economics* with *social dilemma*, the *prisoner's dilemma* game and *ultimatum game* as well as on modeling with again *prisoner's dilemma* game and *ultimatum game* and more specifically *spatial games*.

It should be emphasized here that this global visualization is complementary but clearly different from neighborhoods visualization. In this case, only neighbors that satisfies global conditions may appear. Thus the detected fields outline trends in science according to a given degree of specificity tuned by α. The whole list of concepts as well as other examples of reconstructed paradigmatic fields can be found on http://chavalarias.com.

5.3 Dynamics of paradigmatic fields

Dynamical science mapping is another challenge that aim at describing dynamical patterns in science evolution (GARFIELD, 2004; BRAAM ET AL., 1991) . Static visualization based paradigmatic proximity is only partially informative. Our temporal time series enable us to study evolution of paradigmatic proximity and paradigmatic

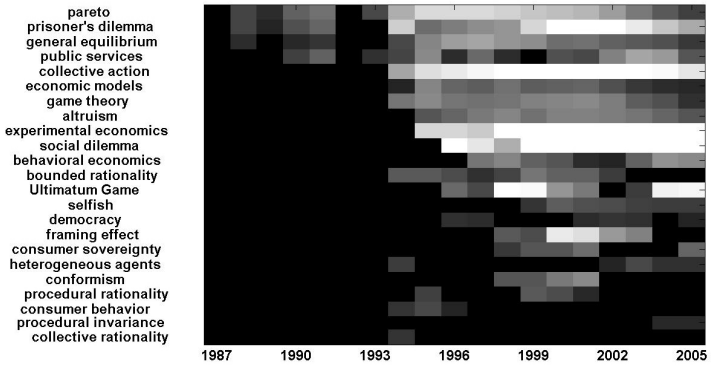

Fig. 5. Dynamical view of the evolution of the paradigmatic fields around *Public goods* from 1987 to 2005 (each year corresponds to the aggregation of a 3-years time-window) for $\alpha = 1$. A black box means that the concept was below the threshold at the considered year. The lighter the square, the higher the paradigmatic proximity.

fields through time. Several questions arise : is it possible to reconstruct the historical evolution of major paradigmatic shifts, can we detect automatically emerging approaches and sub-fields ? The simplest way to take into account the dynamical dimension of our data is to represent the evolution of paradigmatic neighborhoods through time. Given a target concept and a threshold s, we can plot for each time-window t the set of concepts belonging to the neighborhood $V^t_{s,\alpha}(i)$ as given figure 3. We can thus provide dynamical evolution of a target concept's evolution as illustrated 5. We can see on this example that public goods studies appraised as game theory issue developed a lot these last years. Among emerging close concepts in the fields we find *heterogenous agents* and *procedural rationality*. These observations fit well with what we actually observe in evolution of public goods studies.

6 Perspective

We have already sketched methods to provide high-level description of our set of concepts. 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.

Conclusion

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 concepts or combination of concepts

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 paradigms are structured. The innovative aspect of this method is the use of an asymmetric similarity measure the enable to have a structure view of each paradigms identified as well as is open opportunity for a multi-scale browsing of a set of concepts from the more general to the more specific.

Acknowledgements

This study was supported by the CREA - Ecole Polytechnique, the IST-FET coordinated action ONCE-CS (once-cs.net) and the Paris Ile-deFrance Institute for Complex Systems (iscpif.fr). We warmly thank Scirus (scirus.com) for their partnership and Craig Scott for his kind help with the data processing, as well as Arc System research for their keywords list.

References

BRAAM, R. R., MOED, H. F., VAN RAAN, A. F. J. (1991), Mapping of science by combined cocitation and word analysis. II. dynamical aspects, *Journal American Society Information Science*, 42(4):252–266.

BUTER, R., NOYONS, E. (2002), Using bibliometric maps to visualise term distribution in scientific papers, in: *Sixth International Conference on Information Visualisation (IV'02)*, pp. 697–702.

CALLON, M., COURTIAL, J., LAVILLE, F. (1991), Co-word analysis as a tool for describing the network of interaction between basic and technological research: The case of polymer chemistry, *Scientometric*, 22(1):155–205.

DOYLE, L. B. (1961), Semantic road maps for literature searchers, *J. ACM*, 8(4):553–578.

GARFIELD, E. (2004), Historiographic mapping of knowledge domains literature, *Journal of Information Science*, 30(2):119–145.

KUHN, T. S. (1970), *The Structure of Scientific Revolutions*, UCP, Chicago, second edition.

LATOUR, B. (2005), *Reassembling the Social: An Introduction to Actor-network-theory (Clarendon Lectures in Management Studies)*, Oxford University Press.

LEYDESDORFF, L., VAUGHAN, L. (2006), Co-occurrence matrices and their applications in information science: Extending aca to the web environment, *J. Am. Soc. Inf. Sci. Technol.*, 57(12):1616–1628.

LIN, X., SOERGEL, D. (1991), A self organizing semantic map for information retrieval, *Proc. 14th International SIGIR Conference*:262–269.

M. CALLON, J. C., S. BAUIN (1983), From translation to problematic networks: an introduction to coword analysis, *Social Science Information*, 22:191–235.

MARSHAKOVA-SHAIKEVICH, I. (2005), Bibliometric maps of field of science, *Infometrics*, 41(6):1534–1547.

NOYONS, E., VAN RAAN, A. (2002), *Dealing with the data flood. Mining data, text and multimedia.*, J. Meij (ed.), The Hague: STT/Beweton, pp. 64–72.

PALLA, G., DERENYI, I., FARKAS, I., VICSEK, T. (2005), Uncovering the overlapping community structure of complex networks in nature and society, *Nature*, 435:814.

PETER VAN DEN BESSELAAR, G. H. (2006), Mapping research topics using word-reference co-occurrences: A method and an exploratory case study, *Scientometrics*, 68(3):377–393.

SALTON, G. (1963), Associative document retrieval techniques using bibliographic information, *J. ACM*, 10(4):440–457.

SMALL, H. G. (1973), 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.

SUN, Y. (2004), Methods for automated concept mapping between medical databases, *J. of Biomedical Informatics*, 37(3):162–178.