Science mapping with asymmetrical paradigmatic proximity

David Chavalarias, Jean-Philippe Cointet

Science mapping with asymmetrical paradigmatic proximity

Jean-Philippe Cointet*†‡& David Chavalarias*†§

{*jean-philippe.cointet,david.chavalarias*}*@polytechnique.edu*

October 16, 2007

Abstract

We propose a series of methods to represent the evolution of a field of science at different levels: namely micro, meso and macro levels. We use a previously introduced asymmetric measure of paradigmatic proximity between terms that enable us to extract structure from a large publications database. We apply our set of methods on a case study from the Complex Systems Community through the mapping of more than 400 *Complex Systems Science concepts* indexed from a database as large as several millions of journal papers. We will first recapitulate the main properties of our asymmetric proximity measure. Then we show how salient paradigmatic fields can be embedded into a 2-dimensional visualization into which the terms are plotted according to their relative specificity and generality index. This meso-level helps us producing macroscopic maps of the field of science studied featuring the former paradigmatic fields.

Keywords: Mapping and visualisation of knowledge ; semantic network ; co-word analysis ; paradigmatic proximity, asymmetric proximity measure.

Introduction

Scientific activity can be interpreted as a complex process (Hull, 1988) that derives from a large scale interaction network made of a great number of heterogeneous actors: scientists, engineers, natural objects, journals, public and private laboratories, etc... (Latour, 1988)... The main core of this large-scale intertwining system can be coarsely resumed to scientists publishing articles through new collaborations that synthetize a state of knowledge at a given time. Following Latour & Woolgar (1986), we consider that texts are among the major products of scientifc and that they make possible the coordination of millions of people distant in space and time. As such the publications is one of the main communication medium for scientists. this is a stigmergic media as it is not directed toward a specific recipient, information are public and each new associations between concepts shall modify even slightly the scientific landscape.

*Both authors have equally contributed to this work

†CREA (Center for Research in Applied Epistemology), CNRS/Ecole Polytechnique, 1 rue Descartes, 75005 Paris, France.

‡TSV (Social and Political Transformations related to Life Sciences and Life Forms), INRA, 65 Boulevard de Brande- bourg, 94205 Ivry-sur-Seine Cedex France.

§ISCPIF (Complex Systems Institute, Paris Ile-de-France), 57-59 rue Lhomond, 75005 Paris, France.

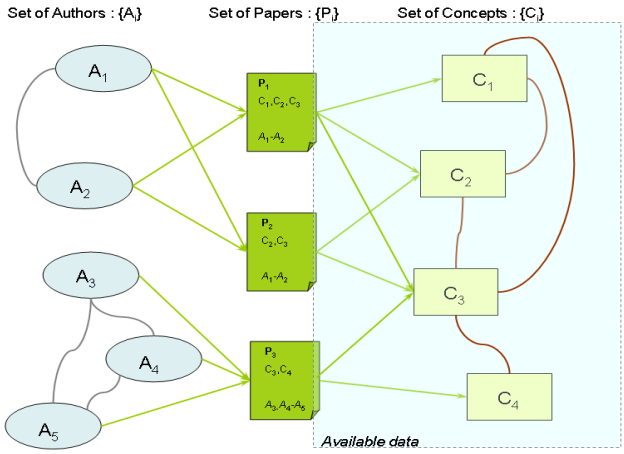

Figure 1: **scientific knowledge production scheme**: a set of authors $\{A_i\}$ produce publications $\{P_i\}$ which are made of concepts $\{C_i\}$. We defined paradigmatic field as strongly cooccurring set of concepts.

The figure 1 represents the scientific knowledge production process. Scientists $\{A_i\}$ publish papers $\{P_i\}$ which can be described by a set of concepts $\{C_i\}$. Following Kuhn's observation that "a paradigm is what the members of a scientific community share, and, conversely, a scientific community consists of men who share a paradigm" (Kuhn, 1969) we identify the scientific communities (on the left of the diagram) with the paradigms, made of co-occurring sets of concepts, on the right.

We shall define a paradigmatic field as a set of concepts that reflects the structure of the activity of scientific communities. We are then looking for characteristic patterns of words co-occurrences corresponding for example to hierarchical structures (in our example, see figure 2 "knwoledge discovery" is embedeed in the "complex systems" field). Concepts may also be part of close though distinct paradigmatic fields (as illustrated figure 2, "knwoledge discovery" may be used by machine learning community ("genetic algortihm") but also by data mining community ("Mining technology")). We will thus have to develop overlapping paradigmatic fields detection.

The aim of this paper is to elaborate on the key benefits we derived from an asymmetric proximity measure previously introduced (Chavalarias & Cointet, 2008). We will then present methods and tools for automatic bottom-up identification of multi-scale structure of paradigmatic fields and apply these onto a case study concerning complex systems science.

This measure is based on mere statistics on occurrences and co-occurrences of words usages in a scientific database. In the first part of this paper, We shall explain the main properties of this proximity measure and show its advantages compared to other classical measures. We shall then describe the clustering method we use to detect paradigmatic fields and define a two dimensional space that help us representing them in a convenient manner. The last part of this paper will deal with large-scale representation of a corpus. We will finally propose a method to represent in an understandable way a large set of keywords which structure organizes an entire corpus.

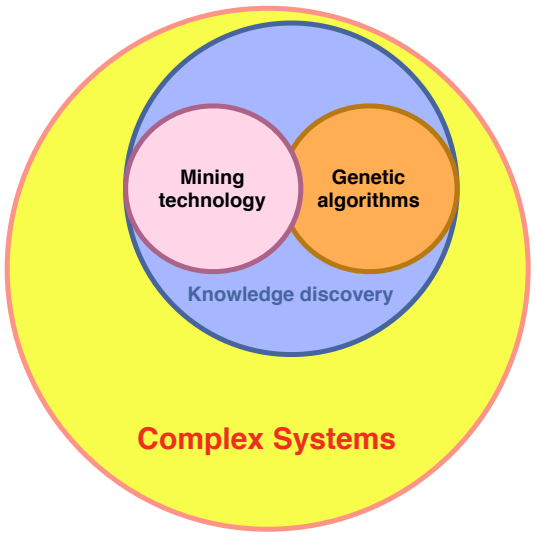

Figure 2: **Schematic example of organisation of paradigmatic field**: Typical concepts organization. Concepts area corresponds to their occurrences and overlapping areas to co-occurring concepts.

1 Context and rationale

Retrieving structure from a set of concepts occurrences and co-occurrences is one of the main objective of scientometry. Doyle (1961) was among the first to point to the fact that navigation in large scientific database was ineffective due to the lack of relevance of traditional document retrieval techniques. One method that has been proposed and largely commented is "co-word analysis". A classical statistics in co-word analysis which has been extensively used (M. Callon, 1983; Callon *et al.*, 1991; Noyons & van Raan, 2002) is the similarity index given by the ratio between the number of co-occurrences between the two concepts a and b divided by the product of the number of total occurrences of a and b.

This similarity index is entirely symmetric which means that given a concept a and another concept b, a is at the same distance from b than b from a. This a priori symmetric constraint can happen to be problematic when comparing concepts of different "frequencies". Let's consider a case where every occurrence of b is accompanied by an occurrence of a for example which may be the case if b refers to the sub-field of a more generic field a (for example "mining technology" can be described as a subfield of "knowledge discovery"). In this case, we would like to be able to derive this hierarchical relation directly from our proximity measure. This is impossible with the classical proximity index which will not enable us to exhibit this kind of highly hierarchical structure.

In order to retrieve this kind of organisation we proposed (Chavalarias & Cointet, 2008) an alternative proximity measure between concepts i and j (n_i^t and n_j^t being the number of occurrences of i (respectively j) at time t, n_{ij}^t is the number of co-occurrences of i and j): paradigmatic proximity ($alpha$ being a tunable real positive parameter)

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

that have the following properties:

1. $P_p(i,j) = 0$ if $n_{ij}^t = 0$
2. $lim_{\frac{n_{ij}^t}{n_i^t} \to 0}(P_p(i,j)) = 0$

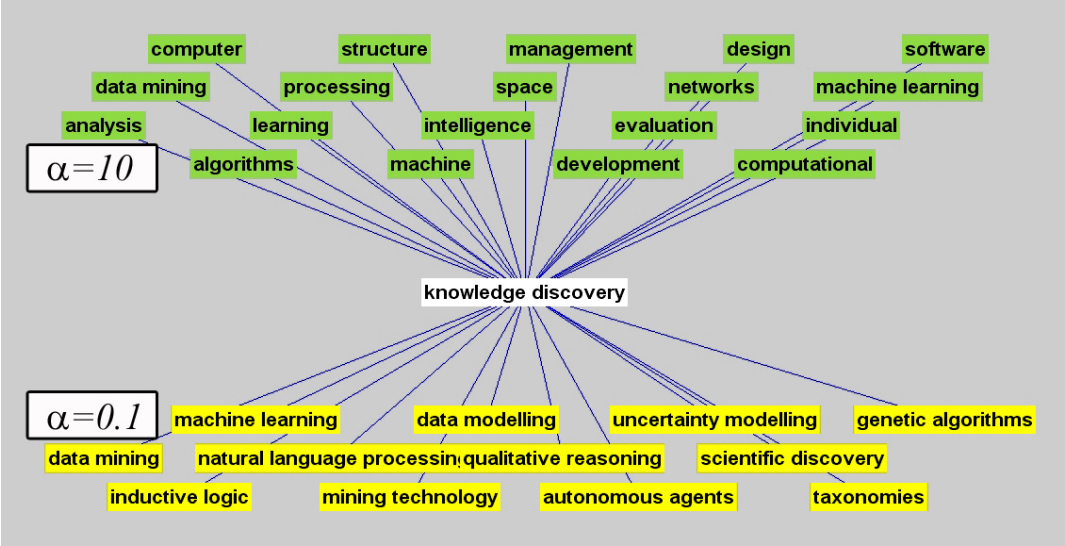

Figure 3: **specific and generic neighborhood of term *Knowledge Discovery***. bottom: $\alpha = 10$, the 10 closest concepts that specify *Knowledge Discovery* ; top: $\alpha = 0.1$, the 10 closest concepts that are more generic than *Knowledge Discovery*.

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. On the contrary, growing n_i^t of n_j^t, n_{ij}^t being constant, will decrease the value of $P_p(i,j)$.
5. Under the assumption that the sample is representative $P_p(i,j)$ is independent of the total number of articles in the database

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

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

As we shall see, this neighborhood will enable us to describe the way a concept belongs to a sub-field of a target concept or on the contrary how a target concept is part of more generic fields.

The interpretation of the results obtained with this neighborhood measure is straightforward. At low value of α, the nearest neighbors of a given target concept will tend to be more specific. When rising up α, we access to more generic expressions. The figure 3 illustrates this property around the target term "knowledge discovery" extracted from the case study exposed below. As α increases, concepts in the neighborhood of "knowledge discovery" 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 knowledge discovery. On the contrary, values of α below 1 will tend to reframe the target word in its broader context.

2 Case study

Our case study focuses on a set of concepts coming from two sources : a set of key-words associated to complex systems European projects extracted from IST Cordis database of FP6 and FP7 (765 key-words generously provided by the Arc

System team lead by Joseph Frohlish) and a set of key-words collected near colleagues (about one hundred). We established a partnership with Scirus, Elsevier's free science-specific search engine[1] 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 is made of more than $20.000.000$ publication covering a wide range of scientific content platforms[2].

Due to the numerous access to the database, data collection was pretty slow and we had to restrain our set of concepts to 448 key-words [3]. Since co-occurrences extraction were very demanding in terms of server availability, we also decided to send a query for a co-occurrence of two terms only when the two queries on single terms gave a non zero result in the "authors key-words" field (each concept has been mentioned by at least once as an article key-word for the year considered). Consequently our database is made of all queries results for single terms in full text from 1975 to 2005, and every query results on full text co-occurrences for pairs of concepts that both appeared at least once as author key-words the year considered.

This rough statistics enables us 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 given this time range:

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

We will now give some examples of application of our paradigmatic proximity measure at different scales. We propose to illustrate our measure advantages at three different levels: micro, meso and macro levels. It should not be forgotten that the clusters and thematic fields that we will have extracted from our database are limited to the 448 initial concepts chosen.

3 multilevel science mapping

A classical objective in bibliometric literature is to draw knowledge maps (Buter & Noyons, 2002; Marshakova-Shaikevich, 2005). Clustering methods like Kohonen maps algorithms have been 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).

Here our approach is restrained to the mere occurrences and co-occurrences statistics but we apply our asymmetric paradigmatic proximity in order to detect more detailed structure than classical flat maps from our set of key-words as we are now able to distinguish between different level of specificity/generality.

We propose to represent our initial set of concepts at three distinct levels of aggregation. First we define a micro-level or local level based on the proximity measure measure from a sole target concept to other concepts in the initial terms list. Meso-level enables us to define paradigmatic fields which define consistent and relevant set of concepts. The macro-level built upon the former level provides an intelligible map of the complete scientific landscape.

[1] www.scirus.com

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

[3] The set of keywords can be consulted in **http://iscpif.fr/ScienceMaps**

3.1 Micro scale : paradigmatic neighborhoods

Micro scale is the most straightforward level. A simple representation has already been exposed figure 3. Here the approach is local which means that we only refer to concept-centered distances. Given a target concept, we simply gather the nearest concepts for a given value of α. If α is below 1 we will preferentially exhibit more generic terms than the target concept, if α is above 1 we will retrieve the sub-field of our target concept. In the special case where $\alpha = 1$ we exclude very generic and very specific neighbors and our paradigmatic measure equals the equivalence index (e-coefficient) introduced by Callon *et al.* (1991).

3.2 Meso scale: identification of paradigmatic fields

Looking at the bottom part of figure 3, we can see that several distinct spheres of knowledge seem to co-exist that share the use of the concept "knowledge discovery". One is more *machine learning* oriented while the other more *data mining* oriented. Contextual information enables us to exhibit automatically these multiple practices.

Identifying set of keywords reflecting this complex scientific activity practices require a broader view of the keywords landscape structure taking into account every relations between the different keywords neighbors. We wish to automatically categorize our data according to the values of the paradigmatic proximity P_p^α given for each pair of concepts. Since a concept can be polysemic, it may be used by several scientific communities, the categorization algorithm should then be able to categorize a concept in different clusters. One successful method in line with this requirement is the k-clique percolation algorithm recently introduced by Palla *et al.* (2005) that operates on graphs to detect possibly overlapping communities.

We proceed by generating a conceptual graph based on our proximity measure by setting a threshold s and linking each keyword i to its set of neighbors: $V_{s,\alpha}(i)$. 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 made of clusters of concepts such that within each cluster, one can perform a k-clique percolation (with $k \geq 3$). Clusters are a general property of the graph (though they may depend on α and s), they do not depend on a predetermined target concept.

The very latest enhancements of the k-clique percolation algorithm (Illés Farkas & Vicsek, 2007; Gergely Palla & Vicsek, 2007) propose an extension of the algorithm to directed and weighted networks. This would allow us to treat directly our proximity matrix without having to define a threshold value s. Though the non-direceted and non-weighted algorithm already provides convincing results retrieving relevant overlapping set of concepts.

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 word w within this cluster : the genericity index i_n and the specificity index i_s defined as follows :

The specificity index provides the extent to which 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 :

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

The genericity index defines 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

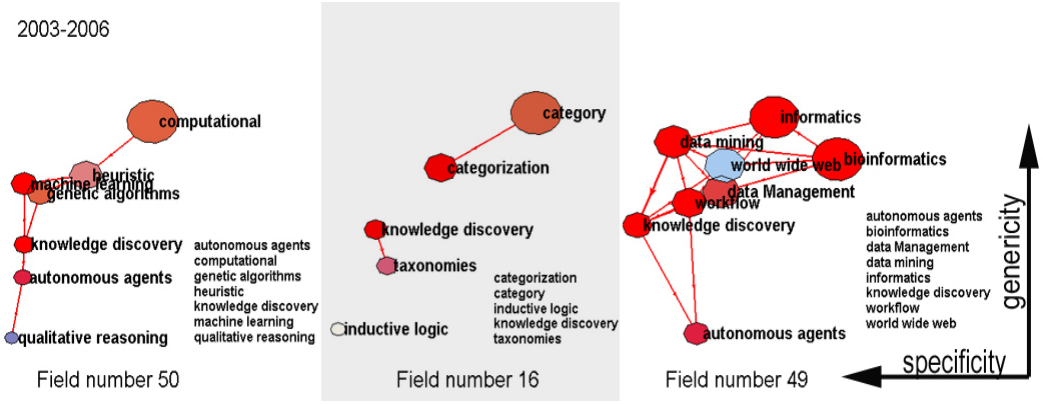

Figure 4: **Three paradigmatic fields mentioning the term *Knowledge Discovery*.** *Knowledge Discovery* belongs to several spheres of knowledge production. Here we can see one *machine learning* oriented (left), the other *data mining* oriented (right). Within a cluster, from left to right i_s decreases, from top to down i_g decreases. The points size is a growing function of the 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 indicate the activity of the fields, i.e. the relative and normalized 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 to the proximity measure between two words, the direction of the arrow is oriented from more general to more specific. For clarity, only couples with highest proximity have been plotted.

the mean of "out-paradigmatic measures" from target word w and is defined by :

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

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

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

This enables to draw an arrow between two words in C representing the strength of the asymmetry. This quantity provides the direction of the arrow. Its width is proportional to the maximal proximity between the two words (with a log factor). Last, we map the size of the terms according to the number of occurrences of words in the corpus (with a logarithmic scaling).

To illustrate this, we present here three cliques that share the concepts *Knowledge discovery* 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 *machine learning* oriented, the other more *data mining* oriented while a third one refers more generally to the field of *categorization*.

It should be emphasized here that this meso-scale visualization is complementary but clearly different from micro-scale visualization. In this case, only neighbors that satisfy global conditions may be gathered together. Thus the detected

fields outline trends in science according to a given degree of specificity tuned by α. Other example of automatically reconstructed paradigmatic fields can be found on http://iscpif.fr/ScienceMaps.

Given this meso-scale approach, it is now possible to draw a global map of this set of concepts related to complex systems to have a macro-scale overview.

4 Macro-scale : science mapping

The next step now is to give an insight of the articulation of the different paradigmatic fields we have identified in the meso-scale approach in order to provide a global view of the scientific landscape defined by our initial set of concepts. For a given period of time, we have defined paradigmatic fields as relevant set of terms. These terms may belong to several paradigmatic fields. A natural procedure is to consider a weighted graph made whose nodes are the paradigmatic fields and the edges are defined according to the overlap between paradigmatic fields. Since within each paradigmatic fields we can qualify a term by its contextual specificity and genericity index, this function could in principle take into account theses two local indexes. We will however consider here the simplest function *i.e.* and consider that the weight of a link between two paradigmatic fields will be equal to the number of common terms. These two indexes are nevertheless useful to label automatically the paradigmatic fields by their most generic and specific terms.

Another information are worth displaying on this global map: the activity of each paradigmatic fields. Again, complex indexes can be built using the dynamic data and the 2D embedding. We will take here a simple index defined as the averaged increase of the normalized occurrences for a field C between a period T and the immediately preceding period of same length T_-:

$$A_C = \frac{1}{card(C)} \sum_{i \in C} \frac{p_i^T}{p_i^{T_-}}$$

where p_i^T is defined by $p_i^T = \frac{n_i^T}{\sum_j n_j^T}$

The figure 5 display the map of science on the period 2002-2005 that provides all these informations. To keep the overall figure understandable we limited the number of paradigmatic fields to display plotting only those that had between 6 and 20 terms.

The size of a point is a monotonous mapping of the number of terms in the cluster and colors indicate the mean activity of the fields. Blue colors correspond to a negative growing rate, yellow, red and brown to a positive growing rate. We can notice here that all fields are growing and that the fields are organized in large thematic areas that are themselves connected. They roughly are : biology, cognitive science, game theory, environmental sciences, social sciences, statistical physics, maths, artificial intelligence, computer sciences and information technologies. An interactive map that allow to zoom in the general map to navigate the paradigmatic fields can be found online: http://sciencemaps.iscpif.fr.

5 Perspective

We have already sketched methods to provide high-level description of our set of concepts. The next step may be to further study the dynamical properties of this mutli-level description.

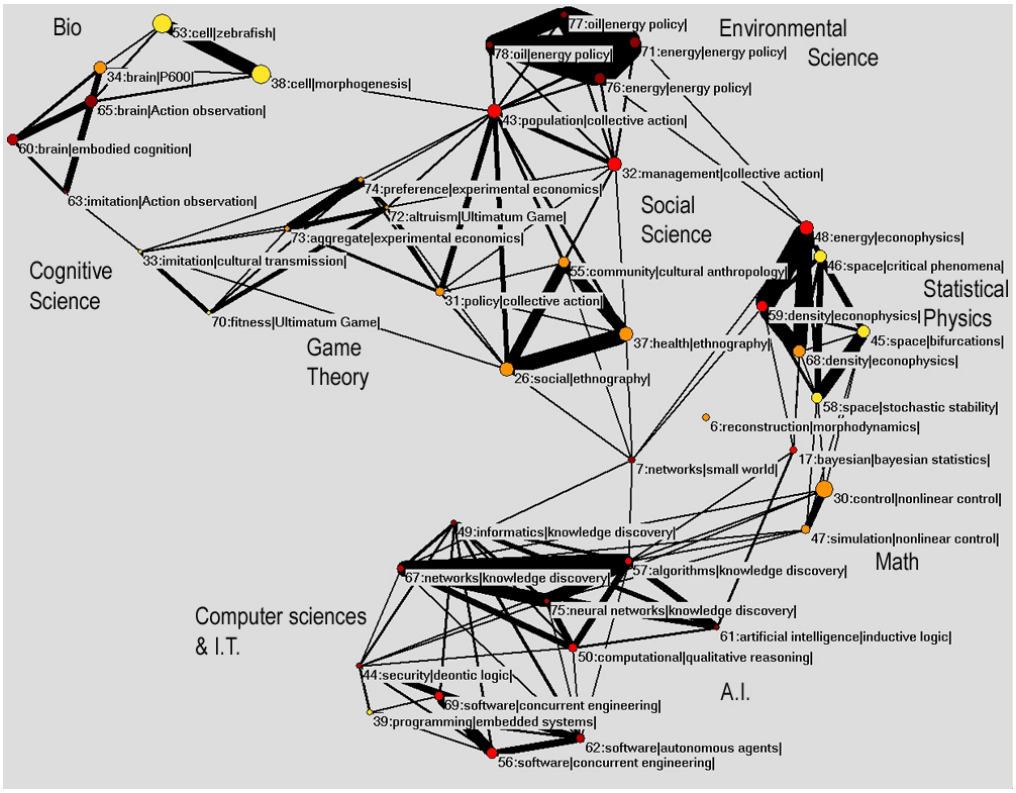

Figure 5: sciencemap

Conclusion

We have tried to show the way science mapping could benefit from an asymmetric proximity measure between concepts. Hierarchical and overlapping complex structures have been exhibited and represented in convenient space. Another advantage of this measure is the possibility to perform multi-scale browsing over a set of concepts from the more general to the more specific, which may happen to be of great help as well for scientists than for other audiences (science policy maker, etc...). Besides the methods exposed are not entirely specific to scientific world ; one can imagine to draw these kinds of knowledge map from any other sources of content (such as press content, blogs, etc...).

Acknowledgements This study was supported by the CREA - Ecole Polytechnique, the IST-FET coordinated action ONCE-CS (once-cs.net) and the Paris IledeFrance 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

Buter, R.K., & Noyons, E.C.M. 2002. Using Bibliometric Maps to Visualise Term Distribution in Scientific Papers. *Pages 697–702 of: Sixth International Conference on Information Visualisation (IV'02)*.

Callon, M., Courtial, J.P., & 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.

Chavalarias, D., & Cointet, J.P. 2008. Bottom-up scientific field detection for dynamical and hierarchical science mapping - methodology and case study. *Scientometric*, **75**(1).

Doyle, Lauren B. 1961. Semantic Road Maps for Literature Searchers. *J. ACM*, **8**(4), 553–578.

Gergely Palla, Illés J Farkas, Péter Pollner Imre Derényi, & Vicsek, Tamás. 2007. Directed network modules. *New Journal of Physics*, **9**(6), 186.

Hull, D. 1988. *Science as a process: an evolutionary account of the social and conceptual development of science*. Chicago: University of Chicago Press.

Illés Farkas, Dániel Ábel, Gergely Palla, & Vicsek, Tamás. 2007. Weighted network modules. *New Journal of Physics*, **9**(6), 180.

Kuhn, Thomas S. 1969. *The Structure of Scientific Revolutions, Postscript*. Second edn. Chicago: University of Chicago Press.

Latour, Bruno. 1988. *Science in Action: How to Follow Scientists and Engineers Through Society*. Harvard University Press.

Latour, Bruno, & Woolgar, Steve. 1986. *Laboratory Life: the social construction of scientific facts*. Princeton, New Jersey: Princeton University Press.

Lin, X., & Soergel, D. 1991. A Self Organizing Semantic Map for Information Retrieval. *Proc. 14th International SIGIR Conference*, 262–269.

M. Callon, S. BAUIN, J.P. COURTIAL. 1983. From Translation to Problematic Networks: an Introduction to Coword Analysis. *Social Science Information*, **22**, 191–235.

Marshakova-Shaikevich, Irina. 2005. Bibliometric maps of field of science. *Infometrics*, **41**(6), 1534–1547.

Noyons, E.C.M., & van Raan, A.F.J. 2002. *Dealing with the data flood. Mining data, text and multimedia.* The Hague: STT/Beweton: J. Meij (ed.). Pages 64–72.

Palla, Gergely, Derenyi, Imre, Farkas, Illes, & Vicsek, Tamas. 2005. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, **435**, 814.

Peter van den Besselaar, Gaston Heimeriks. 2006. Mapping research topics using word-reference co-occurrences: A method and an exploratory case study. *Scientometrics*, **68**(3), 377–393.

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