Scientometrics, Vol. 26. No. 2 (1993) 231-242

THE USE OF PATENT TITLES FOR IDENTIFYING THE TOPICS OF INVENTION AND FORECASTING TRENDS

J.P. COURTIAL, M. CALLON, A. SIGOGNEAU

Centre de Sociologie, Ecole des Mines de Paris, 62 Bd Saint-Michel, 75006 Paris (France)

(Received April 3, 1992)

Co-word analysis applied to patents through WPIL normalized title words appears to give a useful picture of a given field: we obtain both qualitative (themes) and quantitative information (weight of themes). It also gives information about the strategic aspects of the themes. Furthermore, in some cases, it is an indication of the future of certain themes that may help forecasting and management studies. Finally, it provides information about what could be a real technology growth process, in relation to the so-called translation model used in co-word analysis.

Introduction

The discovery of scientometric laws could be very useful for an improved understanding of the growth process of science and technology and, hopefully, for science and technology development forecasts.

Over the last decades, a few scientometric laws have been suggested. Very general laws such as growth curves in a specific field (logistical laws) or the number of authors publishing x papers and more (Lotka's law) or the number accounting for a constant ratio – for example, the first third, then the second and last third – of the total number of papers in a field (Bradford's law). It is possible to use these laws – for instance logistical laws – in order to calculate the state of maturity of a field, but this is not yet the case. An exception lies in a recent paper by *Trofimenko* which makes predictions in a research field by calculating, for example, the number of key-words present in one article as compared to the number of words present in many articles: an expanding field has more scattered words than a mature field.[1] One may, in any case, notice that these laws are unidimensional, refering mostly to the distribution of variables (authors, words and so on). However this latest example can be regarded as a step towards linkage, properties, and, consequently, network or bidimensional properties. Looking at co-citation network structure, *Small* has recently suggested some properties for such networks.[2]

Elsevier, Amsterdam – Oxford – New York – Tokyo
Akadémiai Kiadó

However most common laws of scientometrics can be regarded as consequences of Zipf's law.[3] A synthesis of these results is given in a recent book.[4] *Mandelbrot* suggested linking Zipf's law with a fractal structure called a regular tree.[5] A regular tree is one having the same number, N, of branches at each junction, and a constant ratio, r, for the weight (or frequency) of each node to the weight of the preceding node in the tree (generally $r < 1$). This definition leads to a fractal dimension:

$$D = \log N / \log(1/r)$$

If $r = 1/N$, which corresponds to the case of a "perfect" thesaurus structure – in which each key-word is always associated with all the macro-terms (all specific key-words having the same occurrence), then $D = 1$.

Van Raan also suggested a fractal structure for co-citation clusters when looking at a size distribution.[6] This dimension has a value approaching 1 as the co-citation cluster level – from 1 to 4, according to the ISI's technique[7] – rises.

Both these fractal properties suggest a process similar to the "diffusion-limited aggregation" process for knowledge production.[8] We can thus link these properties to classical models of knowledge growth like Kuhn's paradigmatic model[9] which forms the theoretical background for most citation and co-citation studies: science grows where pioneer work opens new perspectives on nature.

For many years, *Callon* et al. have looked at this process as one which is interactional between authors, and problems, refered to as "the translation model", thus producing socio-cognitive networks of associated problems or "problematic networks".[10] Two problems A and B are linked through the translation model as far as, in the mind of the association author, generally a scientist, solving A implies solving B (and vice versa). This is not an associative property of A and B: the association often disappears when the two problems are solved (and, consequently, is not present in expert mind). The relation only expresses the fact that scientific articles statistically refering to A also refer to B. Or that technical publications refer simultaneously to the same couple of problems.

By assimilating problems to key-words and using co-word analysis (Leximappe program), we can calculate, for a set of scientific articles indexed by key-words and corresponding to a given field, (a) association values for each key-word pair, (b) clusters of about ten most tightly related – through at least one pathway – words, (c) centrality and density indexes for each cluster. Centrality is the sum of link values from words belonging to the cluster to other words, external to the cluster. Density is the medium value of internal links, within the cluster. Centrality, for a cluster, is an indication of the number of links to other clusters. It gives an idea of the strategic location of the cluster within the network. Density is an indication of the weight of the internal links. A high density cluster is made of problems tightly associated to each other. It corresponds to a set of very similar papers, refering to a very stable and well known subject area. We can compare a central cluster to somebody being related to many other people, as in the case with a leader. We can compare a high density cluster to somebody having strong relations with other people. Generally, as it has been checked with many former studies,[11,12] central and dense clusters correspond to the most important subject areas within a field. We have already pointed out that properties as: a relation should be found between the number of external links (centrality) and the value of internal links (density) for a convenient calculation of clusters.[13,14] This is, of course, in conformity with aggregation processes corresponding to fractal models.

However it might be interesting to check particular aspects of our model which make it different from general laws observed in natural processes. It is possible for instance to link the property already observed (i.e. the relation between centrality and density) to actor translation strategies? In this paper we (a) have attempted to observe the property in a technology network with patent titles using co-word analysis and (b) suggest a linkage between this property and actor translation strategies.

Normalized title words as key-words

For co-word analysis, a list of key-words corresponding to each document must be supplied. A statistical calculation is applied to these lists. Generally data bases patents have no key-words (except, for example, for the French patent data base FPAT), but only a few classification codes. These classification codes are not very useful for innovation content description, namely for details, specific uses, features and so on.

Fortunately, the WPIL patent data base (DERWENT) provides a normalized title for each patent family of the base (a patent family is made with all patents corresponding to the same invention as contained in a single priority patent: a family normally includes all foreign applications related to a domestic prior application). The normalized title is a normalized version of a family title given by WPIL editors. Generally the applicant himself gives only a short, meaningless title recorded in the standard data bases. WPIL provides improved titles based on the whole text of the priority document. Furthermore, WPIL tries to use thesaurus terms. The normalized version is obtained by suppressing tool-words, normalizing words and so on, through informatic tools. For instance, the title, "Appts. for high intensity treatment of solid particles – comprises means of adding fuel, particles and oxidising gas to combustion chamber" will give: "APPARATUS HIGH INTENSITY TREAT SOLID PARTICLE COMPRISE ADD FUEL PARTICLE OXIDATION GAS COMBUST CHAMBER". In some cases, WPIL add additional words which are useful to understand the context.

It is thus interesting to try to use the normalized title as a list of key-words, a key-word being defined as a string between two space characters. (Using word processing techniques, it is even possible to improve this list of uniterms by joining a set of two succeeding words – for instance, joining "ice cream" to make "ice*cream"). The advantage of WPIL titles (also called interpreted titles) is that they account for uses, specific features and a host of useful information for describing technology networks – information often absent from classification codes.

We obtain with these titles both meaningful results and network properties as in the case of genuine key-words (scientific articles).

Data

Data are patent family normalized titles from the WPIL data base. As already explained, the WPIL data base records all patent documents (applications, granted patents, ...) published all over the world and groups them into families (a WPIL record is such a family) corresponding to the same priority patent. It should be noted that although most patent families are made up of more than one member, it is possible for a family to be made up of only one recent application, for instance published in Japan. Thus, the number of records can vary greatly according to national patent used.

The data used are patent from 11 subfields of the WPIL data base in the food products section of the International Patent Classification (IPC). Each subfield corresponds to a specific IPC code. This code has been defined in order to have between 300 and 700 patents for a period of one or two years. For each subfield, two periods of time have been used in order to make comparisons over time. Table 1 gives the list of subfields and the number of patents.

Table 1
Number of WPIL records per IPC codes

Subject area	IPC code	First period		Second period	
		Year	Invetions	Year	Inventions
Alcoholic beverage	C12+ sauf C12M,N,P,Q	1985	437	1989	315
Oven cooking: pasta	A21+	1985	714	1989	603
Meat and fish	A22+	1985	555	1989	389
Food	A23B+	1985	524	1989	505
Dairy products	A23C+	1985	329	1989	337
Fat	A23D+	85-83	352	88-89	356
Pet food	A23K+	1985	498	1989	475
Non alcoholic beverage	A23L-002/00	82-85	212	87-89	253
Canned food	A23L-003/00	82-85	306	87-89	277
Sugar and starch	C13+	84-85	444	88-89	276
Leather and skin	C14+	84-85	531	88-89	522
Total			4942		6457

For each set of patents we downloaded normalized titles (the IT field) and used these titles as lists of words, as if they were a list of document descriptors.

Co-word analysis

Before going any further, it might be useful to briefly recall the principles of co-word analysis (leximappe program), already described in detail in many publications. Co-word analysis is a kind of objects classification tool, but it takes into account only positive (not equal to zero) links between objects.

Two words can be associated in many documents. This is taken as an association index between words. (The usual index is the probability of obtaining the second word of the pair when the first one appears, multiplied by the same probability calculated in the other direction). After calculating links between word, co-word analysis:

- orders links in decreasing order;
- selects from this list words having most important links with about nine other words: this gives clusters of the most tightly linked words through a pathway (single link clustering, not the most tightly linked words all together); when a cluster of ten such words appears (when reading the word-pair list in decreasing order), the succeeding pairs made up of one of these words are deleted from the

list (but the links are maintained between clusters for further calculation as external links);[15]

- calculates for each cluster centrality and density weights: the density is the mean value of the internal links; the centrality is the sum of values of the external links.

Usually, using median values for centrality and density (values which divide cluster list into two sets of equal size) we classify the clusters into four groups according to their degree of centrality and density. We also display the clusters on a "strategic diagram" (or graph) made by plotting centrality and density rank values along two axes. In the case of key-words from scientific publications, in due course, we observed typical moves on the strategic diagram: for instance from the lower right quadrant to the upper right quadrant of the strategic diagram. This corresponds to an increase in density values for clusters (research themes) first central, thus to a decrease in the centrality/density ratio.

For normalized patent title words, since this is a less controlled kind of information than key-words, we only studied changes over a period of time for the case of clusters belonging to the upper right quadrant in the second period of time (clusters externally and internally linked over median values). For these clusters we look at earlier clusters (i-e having in common most words or the more important – regarding weight of links – words) in the first period of time. We compare the change in the centrality/density ratio (rank values) in time.

Results

By applying co-word analysis to patent family titles we obtain through clusters, easily understandable technological themes.

An example of co-word theme is given, in pasta and oven cooking field A21, in the second period of time, by the following word list: flour, fat, sugar, egg, wheat, add, water, preparation, rye, mixture. All words are linked to flour. Most important links are, in decreasing order, with following words: fat, sugar, egg, wheat. This means that a lot of patent refer to pasta composition. All these patent content words flour, fat, sugar, egg and wheat. Some of them content, in addition to some of these words, other words like: add (i-e specific features to add to pasta) or rye and so on.

However, unlike usual classification technics (which require a maximum of links within all word pairs belonging to the same class in order to obtain a high level of homogeneity), themes obtained from single link clustering are not always exact subfields or subject areas within the field: a co-word theme may be made of an aggregation of words through only a common associated word. This means that a new theme may appear in this area that has not yet been clearly identified. Indeed, co-word themes have different meanings within technology regarded as a changing network: if some themes correspond to real stable, technology subfields, others correspond to technological orientations with more than one current specific content. This is of utmost importance for the study of the dynamics of technology.

For each theme (or cluster) we furthermore have an indication of its relative weight through the number of articles belonging to the theme. Links between themes also give information about the logic of technology. For instance, the theme "flour" described above is linked to another important theme, "dough". Changes in theme content over time are also an important indication. The theme flour was already present in the first period of time. However, new words have appeared in the second period of time: fat and sugar. This means that novelty concerning pasta flour patent lies on fat and sugar related problems. Looking back to patents containing these words point out that dietary problems related to fat and sugar are a new general purpose for patents.

The location of themes on the strategic diagram is also another indicator for the general importance of corresponding patents. Themes on the upper right quadrant are of strategic importance. In the case of the theme "dough", which belong to this quadrant for both period of time and correspond to most patents, patents corresponding to the theme in the first period of time are cited 0.9 times (medium value). In general patents of the field are cited 0.6 times. This indicates that citation criteria is related to strategic positions of themes. In other words, looking to location of themes within the strategic diagram is a way, without waiting for citations, to identify patents that will be more cited.

Furthermore, Table 2 and Fig. 1 (x axis = centrality/density value for the first period of time, y axis = variation in centrality/density value between the second period of time and the first period of time) indicate, for a list of 25 themes, changes in the centrality/density ratio over time. These themes are present in both periods of time, and, for the second period of time, in upper right quadrant of the strategic diagram. It is clear that, if this ratio below 1, it will increase and vice-versa. This is a well established property (for 100% of themes below 1 (10 themes) and 80% (12/15) of themes above 1).

Table 2

Subject area – Theme	c/d (1985)	Δ(c/d) × 10
		85-89
Can-Acid	0.286	797
Meat-Sausage	0.385	615
Fat-Fatty	0.429	238
Milk-Fat	0.545	677
Sugar-Control	0.625	946
Pasta-Acid	0.7	100
Leather-Fat	0.727	273
Leather-Group	0.857	143
Can-Gas	1	375
Pet food-New	1	62
Pasta-Oven	1.059	227
Can-Fresh	1.083	167
Leather-Sulphate	1.1	-450
Pasta-Flour	1.133	-344
Fat-Spread	1.167	-167
Alc Bev-Sugar	1.167	-547
Milk-	1.2	-300
Meat-Fish	1.2	-200
Leather-Acid	1.211	189
Alc Bev-Acid	1.273	-106
Meat-	1.333	-458
Sugar-Beet	1.385	-485
Milk-Cheese	1.571	-238
Pasta-Dough	1.818	-356
Can-Food	2	-700
Pet-	11	-9400

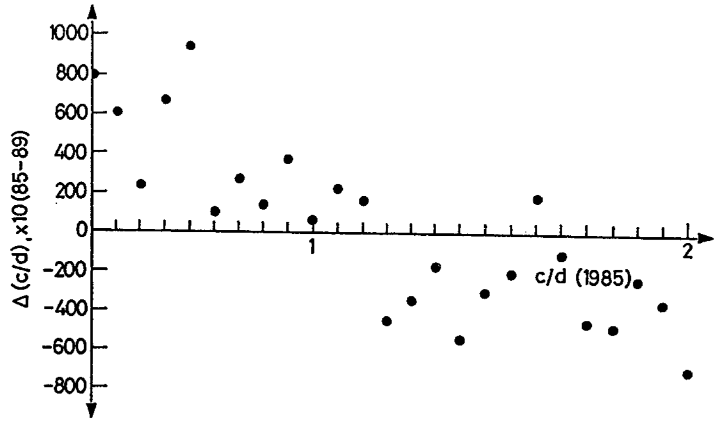

Fig. 1. Changes in centrality/density ratio (Y axis), from year 85 to year 89, according to year 85 value (X axis)

Discussion

If we want to use this general property – checked with 11399 patents for 11 different subfields – as a predictive one, we must follow themes that will appear in the next period of time in the upper right quadrant of the strategic diagram. This is not easy to do as the content of themes changes over time: if we define a theme by the list of words it contains, some themes have two or more successors. As it has been described in an earlier study on polymers, a theme often splits into two or more themes or merges with another: another way of expressing the fact that technologies are constantly evolving. It is thus necessary to have a look in the first period of time at a large list of themes, not necessary first belonging to the upper right quadrant.

Nevertheless, the goal of this study is not to give predictions but rather to check the possibility of using normalized title words for patents. Undoubtedly, minor improvements to the method are necessary and now possible. For instance, the results actually depend in some way on the size chosen for the themes (number of words). It could be possible to avoid this by working at the word level, calculating centrality and density for each word within the network of associated words (and calculating co-words themes only after having identified typical changes at the level of the words).

The main purpose of this account is to point out some properties of the dynamics of technology. Results suggest that technology located well within a technological network will increase in development (density), whereas a well developed technology, but in a less strategic position within the network, will improve its external links. Patents on a new flour composition – a strategic position for a lot of doughs – will develop improvements. On the other hand, patents on a high-frequency bread oven will for instance develop baking-control software programs in order to obtain bread whose particular quality corresponds to customer tastes.

It should also be noticed that this model doesn't account for the whole technology dynamic. Completely new themes may appear as Table 3 indicates. From one field to another the amount of completely new patents – having nothing in common with preceding ones – differs greatly. This seems to be a field property. In the case for instance of non-alcoholic beverages, the ratio of patents corresponding to completely new products is very high.

Table 3

Subject area	Total	Changing themes	New themes
Alcoholic Beverage	7	2	5
Oven cooking: pasta	6	4	2
Meat and fish	4	3	1
Food	6	2	4
Dairy products	4	3	1
Fat	4	2	2
Pet food	7	2	5
Non alcoholic beverage	7	-	7
Canned food	5	2	3
Sugar and starch	4	2	2
Leather and skin	6	4	2
Total	60	26	34

Co-word model and others

We now turn back to those specific aspects of the translation model that could be related to these properties. Inasmuch as words can be problem indicators (scientific as well as technological ones), translation theory implies specific kinds of links between words: (a) an agreement about problem definitions among actors implies a specific structure of "local" linkages; most actors agree about the same links; only a small number of actors suggest other links; (b) an agreement about both problem definitions and problem links – which *Callon* calls a high network "convergence" – implies a specific structure for all links.

This kind of structure for word chains has been pointed out through a mathematical tool called "similitude analysis" in french (namely extraction of maximal trees within a network[16]). Similitude analysis extracts skeletons of word links. When this analysis produces a single skeleton, data structure corresponds to a common definition of problems or, more generally, interests at stake. We have already demonstrated (e.g. in the case of dietary fibers) that the innovation process leads to word skeletons organized around the word defining the innovation.

However, if we assimilate a well-defined problem area to a cluster of most associated words, a converging network of well-defined problems leads to an equilibrium between internal and external links: actors enter a process of network (corresponding to each actor) merging. They "negociate" problem definitions and linkages through a dynamic convergence. Thus, the transformation of clusters according to the process of convergence corresponds to the properties we have discovered: increase in centrality for high density but low centrality problems and increase in density for low density but high centrality problems.

It should be noted that, as oppose to fractal model, the number of external links for such well defined clusters may, after levelling off, decrease: a new subfield, now able to be independent, is born. At the complete end of the process, we may have clusters belonging to the left upper quadrant of the strategic diagram.

Conclusions

The properties we discovered could go with classic models for science and technology changes like fractals (or order out of chaos) as well as interactional models among actors like the translation model. It is a general feature of a lot of these models to tend to reach local equilibrium between, for instance, external and internal cluster links. In order to validate more typical aspects of the translation model, it would be necessary to go on studying the specific aspects of the interaction between actors, by calculating for instance the curve of the frequencies of laboratory themes or answering questions like these: do laboratory profiles, in terms of activity, gradually correspond to co-word themes? Do we suddenly observe a rapid change in words associations in relation to the thematic merging of two big laboratories? What difference rythm of change for co-word themes can we observe when a central but underdeveloped area is made up of many laboratories or is made up of one or two big laboratories?

Co-word analysis applied to patents through WPIL normalized title words appears to give a good picture of a given field: we obtain both qualitative information (themes and their word change over time) and quantitative information (weight of themes). It also supplies information about the strategic aspects of the themes, in relation with future patent citation scores. Furthermore, in some cases where completely new products or processes independant from technological networks, are not frequent, it gives an indication of the future of themes that may help forecast and management studies.

Notes and References

1. A.P. TROFIMENKO, Scientometrics analysis of the topical content of scientific research and its particularities, *Scientometrics*, 18 (1990) 409-435.

2. H. SMALL, E. GREENLEE, A cocitation study of AIDS research, *Communication Research*, 16 (1989) 000.
3. S. NARANAN, Power law relations in science bibliography. A self consistant interpretation, *Journal of Documentation* 27 (2) (1971) 83-97.
4. J.P. COURTIAL, *Intoduction à la Scientométrie*, Paris, 1990, Ed. Anthropos, Diff. Economica.
5. B. MANDELBROT, *The Fractal Geometry of Nature*, W.H. Freeman and Co, San Francisco, 1982.
6. A.F.J. VAN RAAN, Fractal geometry of information space as represented by cocitation clustering, *Scientometrics*, 20 (1990) 493-449.
7. H. SMALL, E. GREENLEE, Clustering the science citation index using co-citations, I-A comparisons of methods, *Scientometrics*, 7 (1985) 391-409.
8. B. SAPOVAL, *Les Fractales, Fractals*, Paris, Ed. B. Aumont, Aditech, 1990.
9. T.S. KUHN, *The Structure of Scientific Revolutions*, Chicago University Press, 1970.
10. M. CALLON, J. LAW, A. RIP, *Mapping The Dynamics of Science and Technology*, London, MacMillan, 1986.
11. W.A. TURNER, G. CHARTRON, F. LAVILLE, B. MICHELET, Packaging Information for Peer Review: new coword analysis technics, A.G. VAN RAAN (Ed.), *Handbook of Quantitative Studies of Science and Technology*, Elsevier Publ., North Holland, 1988, p. 291-323; M. CALLON, J.P. COURTIAL, F. LAVILLE, Co-word analysis as a tool for describing the network of interactions between basic and technological research: the case of polymer chemistry, *Scientometrics*, 22 (1991) 155-205.
12. M. CALLON, J.P. COURTIAL, F. LAVILLE, Co-word analysis as a tool for describing the network of interactions between basic and technological research: the case of polymer chemistry, *Scientometrics*, 22 (1991) 155-205.
13. J.P. COURTIAL, Qualitative models, quantitative tools and network analysis, *Scientometrics*, 15 (1989) 527-534.
14. J.P. COURTIAL, B. MICHELET, A mathematical model of development in a research field, *Scientometrics*, 19 (1990) 123-138.
15. Another possibility could be in keeping words from the association list as far as clusters are built: we may obtain many more clusters slightly different from each other and thus a more precise picture of the field.
16. J.B. KRUSKAL, On the shortest spanning subtree of a graph and the travelling salesman problem, *Proceedings of the American Mathematical Society*, 7 (1956) 48-50.