Scientific Information in Continuous Characteristics Spaces

Andrea Scharnhorst, Paul Wouters
Networked Research and Digital Information (Nerdi)
andrea.scharnhorst@niwi.knaw.nl, paul.wouters@niwi.knaw.nl
Werner Ebeling
Institute of Physics, Humboldt-University Berlin
ebeling@physik.hu-berlin.de

Abstract

Scientific information is characterized as the flow of information created by scientists, which is evaluated by peer reviewers. A final evaluation is connected with the use of the newly generated information either by other scientists, engineers or the general scientific community. This evaluation can be estimated by measures such as the relative number of citations in the scientific literature and the relative number of citations in patents.

We model the dynamics of the evolution of scientific knowledge as hill-climbing in an adaptive landscape over a continuous characteristics space. A problem is described by a large number of attributes, features or characteristics covering problem-inherent aspects and disciplinary parameters. These quantities span a characteristics space, which is a real Euclidean vector space, analogous to the phenotype space in biology.

We also define a population density $x(\vec{q},t)$ that describes the density of the newly generated information at time t in the point $\vec{q}$ and a real-valued multimodal fitness function/functional $V(x(\vec{q},t))$ which expresses the evaluation in a qualitative way. The evolutionary dynamics including competition and mutations/innovations is modelled by reaction-diffusion equations of Fisher-Eigen type.

What is scientific information?

Scientific information is part of the total information created by mankind. It expresses the flow of knowledge and data created by scientists. Information can have several aspects. We are interested in what *use* the information is considered to have. Considering the “usefulness” of information implies a valuation or evaluation process of such information (Ebeling, Schweitzer, & Freund, 1998). The valuation of information by different users in different contexts might be quite different. The creation of scientific information can be seen as an evolutionary process, which means that information is newly created on a continual basis and is evaluated.

In the evaluation process in science, different phases can be differentiated. Informal communications with colleagues, and self-evaluation in the process of knowledge production are usually followed by formally institutionalized forms of evaluation. In the peer review process in science, peers of the researchers act as evaluators (Burnham, 1990; Chubin & Hackett, 1990; Cole, Cole, & Simon, 1981; Godlee & Jefferson, 1999; Kronick, 1990; Wouters, 1997). This peer review process has three main forms[1]. The first form is used by scientific journals, which invite peers to

[1] Apart from these three forms, we can distinguish other forms of quality control, e.g. the checks on the validity of data in data archives. In the framework of this paper on modelling the peer review, these differences are not so important.

Andrea Scharnhorst, Paul Wouters and Werner Ebeling, *Scientific information in continuous characteristics spaces*, In: Guohua Jiang, Ronald Rousseau, Yishan Wu (Eds), Proceedings of the 9th International Conference on Scientometrics and Informetrics – ISSI 2003, Dalian University of Technology, Dalian, China, 2003, 280-288.

judge manuscripts before publication. The second is the quality control of research proposals submitted to funding agencies. The third main form of peer review is implemented in in-house quality control processes in large techno-scientific institutions, e.g. NASA or CERN. The usefulness of information is a dynamic variable. Therefore, its value can also change.

Scientific information becomes visible in scientific communications. Scientific communications are codified in different forms (Cole et al., 1981). Publication in a scientific journal is one form of such a codification; publication in the form of a patent is another. If one takes scientific publications as a starting point, the extent to which a specific publication has been cited can be seen as a form of evaluation by the scientific communities involved. Despite the ambiguity that citations have, the number of citations that a certain paper receives can be related to the relative usefulness of the paper within its context. This is certainly not true for every single reference, but in a number of scientific disciplines it often holds for bigger ensembles of publications and citations (Cronin & Atkins, 2000). The number of citations a publication receives changes over time. The time-dependency of citation rates might be taken as a sign of a changing valuation (called "successive citations" by Vlachý (Vlachý, 1986)).

In this article, we will introduce a simple model of the complex dynamic valuation process. First, we consider the valuation as immediate, without time-delay effects. We are however quite aware of the multidimensionality of valuation in science and will discuss some variants for further operationalization (Wouters, 2000).

Modeling the dynamics of the evolution of scientific information

State space and occupation landscape

First, we will investigate the dynamics of the evolution of scientific information. The production of scientific knowledge will be described as hill-climbing in an adaptive landscape over a continuous characteristics space (Scharnhorst, 2001). The characteristics space is thought of as a reservoir of all possible problems. It is an abstract problem space. One question is how this problem space can be made visible.

In the last decades, scientometrics has used co-word and co-citation analyses to produce maps of articles, topics and authors (Callon, Courtial, & Penan, 1993; Noyons & Van Raan, 1998; Small, 1997). The process of constructing such maps proceeds from determining the similarities between different articles to the ordination of these articles. The latter task is quite sophisticated because the problem space will usually be multidimensional which, nevertheless, has to be projected on a two-dimensional plane. In these maps, a certain article or author can be given a specific, unambiguous location. In this way, a scientific landscape can be made visual (see for a good overview of the different mapping techniques (Chen, 2003) and see also http://www.cs.sandia.gov/projects/VxInsight.html). Let us only sketch one approach which seems to be particularly suitable for our modelling approach. In the vector space model developed by Salton and others (Salton, Yang, & Wong, 1975), a document or paper is represented by a number of terms in the document. The number of all unique terms determines the dimensionality of the space. Each document corresponds to a vector in this space. For each document, the terms are counted and a weighted frequency of occurrence determines the values of the components of the vector. In this way, each document occupies a certain point in this abstract and multidimensional problem space.

We will further interpret this problem space, which is a real Euclidean vector space, analogous to the space of phenotypic properties in biology. Each document then corresponds to an "individual". In biological evolution both the individual and the population may be seen as the target of evolution. From an evolutionary point of view we are particularly interested in the developments of scientific specialties and disciplines. Therefore, we consider groups of documents rather than individual documents. More specifically, we introduce an occupation function. Mathematically described as a population density function, the occupation function represents the frequency in which documents or papers appear in certain areas of the problem space.

The occupation function is not a homogeneous function but forms mountains in certain areas and leaves valleys in others. Recently, a procedure of data mining and data visualization has been developed at the Sandia National Laboratories which produces maps with similar landscape characteristics. (VxInsight, see, e.g., (Davidson, Hendrickson, Johnson, Meyers, & Wylie, 1998)). According to this procedure, articles are located in a certain area of a plane. The plane is the two-dimensional presentation of a multidimensional space. The frequency of papers in a certain area at this plane forms the third dimension.

Fitness and the two-landscape picture

Our model is based on the idea that evolution is hill-climbing in an adaptive landscape over a continuous characteristics space. The idea of the evolutionary landscape (adaptive or fitness landscape) goes back to theoretical biology and was originally created by Wright (Wright, 1932, 1988). The concept was developed further by Conrad (Conrad & Ebeling, 1992), and later mathematically developed by several authors (Ebeling & Feistel, 1990; Feistel & Ebeling, 1982, 1989). Recently, the concept of fitness landscapes has found fruitful applications in the description of socio-economical processes (Ebeling & Scharnhorst, 2000; Scharnhorst, 2000).

We will assume that the properties and dynamics of scientific research can be described by a number of attributes, features, or characteristics covering behavioral characteristics, and thematic dimensions. Furthermore, we assume that these characteristics are metrical and can be expressed by quantitative variables. The quantities are real numbers $\{q_1, q_2, ..., q_i, ...\}$. They span a characteristics space which is a real Euclidean vector space *Q*. First, we define the population density function $x(\vec{q})$. This function describes the extent to which a certain area $dq^{'}$ is occupied by problems.

As discussed earlier in the paper, scientific information gains value by being used. The usage is coupled to a valuation process. Independent of the informal or formal procedure of such a valuation, we can assume that each location in the problem space is linked to a valuation. We will assume that favourable combinations of parameters stand for solvable problems.

Formally, the valuation will be expressed by assigning a real value to each point in the characteristics space. In this way a second landscape is defined over the characteristics space. In the following we will speak of the valuation or fitness landscape and the occupation landscape (Fig. 1).

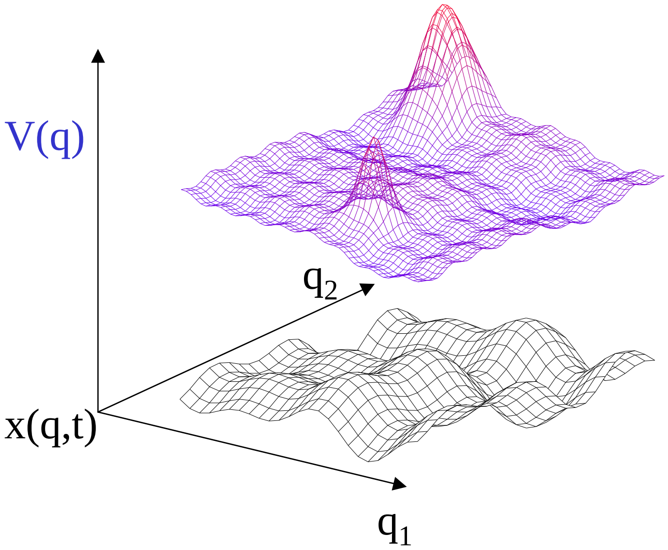

Figure 1: Schematic representation of a the two-dimensional characteristics space. Over this space an occupation landscape $x(\vec{q})$ and a valuation landscape $V(\vec{q},t)$ are defined. Usually, one assumes that the population occupies areas of high valuation step by step. This way the shape of the occupation landscape will converge towards the shape of the valuation landscape.

The introduction of such a formal valuation entails another problem. What appears to be favourable, solvable or important will clearly change over time. Therefore, the valuation landscape has to be thought of as changing in time. The valuation landscape is the outcome of a reflexive action based on the production of scientific information (an operation of second-order type) (Leydesdorff, 2001). We may assume that the change of the valuation follows a slower dynamic than the production process itself. This is usually the case, because the valuation will be based on the criteria that determine “quality” within the specific discipline. These types of criteria tend to change more slowly than empirical or theoretical contributions. The exception is the rapid change in overall shape of knowledge during a paradigm shift or scientific revolution. By definition, however, this type of transformation may be the exception to the rule of normal science (Kuhn, 1962). For the time being we will therefore assume that the valuation function $V(\vec{q})$ remains constant in time.

In more advanced mathematical settings, one may include an additional time-dependency of $V(\vec{q},t)$. Then, one can further differentiate between exogenously and endogenously caused changes. In the first case the function $V(t)$ is explicitly time-dependent. In the second case, the valuation turns into a mathematical functional[2] $V(x(\vec{q}))$ and the time-dependency is implicitly caused by the temporal changes in the occupation function.

The valuation landscape has a complicated structure with many hills and valleys. The hills correspond to the positive, favourable combinations of the parameters, the valleys to the unfavourable combinations. The population density will attempt to follow the structure of the valuation landscape, i.e., it will develop maxima at places where the fitness is large and minima where the fitness has a deep valley. However, this is a process which takes time and, therefore, at a given time not all of the fitness hills will be populated. In the course of evolution more favourable hills may be occupied. However, due to the high dimensionality of the space, total occupation will never be reached. The evolution never reaches its final solution. Progress is always relative and limited. Innovation corresponds to the occupation of hills that were not populated so far.

The evolutionary dynamics based on the mix of competition and mutation/innovation is modelled by reaction-diffusion equations of Fisher-Eigen or Lotka-Volterra types. We use a so-called continuous approach in which the potential goes beyond the widespread applications of discrete replicator dynamics (Bruckner, Ebeling, & Scharnhorst, 1990; Nowakowska, 1984; Wagner--Döbler & Berg, 1993).

The mathematical model

Now let us give a mathematical description of the ideas given above. We will assume that the characteristics space consist of *d* dimensions. These dimensions build the axes of an abstract vector space *Q*. Usually, we may assume that the problem dimensions can be expressed using characteristic terms. Each problem represented by a document or scientific paper can be characterized by a set of numbers $\{q_1, q_2, ..., q_i, ...\}$. The number q_i stands for the frequency with which a certain term occurs. The terms stand for problem properties and the q_i are the components of a vector $\vec{q}$. As a rule, we have many characteristics, i.e., $d >> 1$ is a large number.

Any point in *Q* is a potential state of evolution. At a given point in time, the set of all considered problems corresponds to the set of all occupied points in *Q*. Assuming that the set of

[2] A function which depends on another function instead of on a variable number.

state points is dense, we may introduce the density function $x(\vec{q})$. The population density $x(\vec{q})$ is a real, non-negative function. As a rule, $x(\vec{q})$ has a complex structure, it has many peaks and valleys, and in many parts of *Q* it is simply zero, which means that these combinations of different problem dimensions are (not yet or no more) realized at given time. In the continuous description, the density function $x(\vec{q})$ takes over the role of the discrete set of occupation numbers N_i.

Populations are groups of elements with similar properties; we may think of them as a broad peak of the density function. Evolution means change of $x(\vec{q})$ in time. The change of problem properties corresponds to a trajectory in the space *Q*. In the continuous description, the trajectory corresponds to a re-shaping of the form of the function $x(\vec{q})$, and can be visualized as the growing of a new mountain.

The laws of temporal change may be very complicated. As a first approximation, we restrict ourselves to simple mathematics. We will assume that the evolutionary dynamics based on a mix of competition and mutation/innovation is determined by some function which we call valuation or fitness function. The valuation function $V(\vec{q},t)$ plays the same role as a state function in non-linear dynamics. Their extremes correspond to stationary behaviour, the maxima to stable behaviour, and the minima to instable situations.

The population density $x(\vec{q})$ represents the density of publications appearing in a certain time unit (for example a month or the quarter of a year) and $V(\vec{q},t)$ expresses the valuation of new work in the field $\vec{q}$ at time *t* by the referees. One can measure this valuation in different ways. One possibility is to look at the citations of a certain paper.

In general the dynamics of the system is described by the following equation

$$\frac{\partial}{\partial t}x(\vec{q},t)=w(q;\{x\})x(\vec{q},t)+Mx(\vec{q},t) \qquad \textbf{(1)}$$

Here *w* is a function of $\vec{q}$ (or/and some times a functional of $x(\vec{q})$). The function *w* determines if the population density at a certain point $\vec{q}$ increases or decreases. Therefore, we can speak here of a generalized fitness function. We will discuss later how *w* and $V(\vec{q},t)$ are related to each other. The second term in Eq. (1) describes processes of re-location of problems (papers).

Evolutionary dynamics in the phenotype space

In order to come to a concrete dynamic we recall the general concept that evolution is hill-climbing in an adaptive landscape over a continuous characteristics space. In other words, any local element in the space attempts to move to domains of the space where the fitness is higher. In this paper, we consider scientific papers as proxies for, or in other words representations of, scientific problems. The evolution of knowledge can be seen as resulting from both the publication of new papers and the relocation of already published papers. Mathematically, these two processes can be represented in the same way. The only difference is that in the latter case, papers can also disappear (from their previous position in the problem space Q)[3]. The re-location of written papers implies that their characteristics $\vec{q}$ change over time. The location of a paper at a certain place in the problem space will depend on the perception of this paper by the scientific community. If this perception changes, then the location of the paper also changes. In this way, papers can diffuse in the problem space. Their movement creates a reflexive, second-order dynamic inside the problem space compared with the first-order dynamic of placing new papers in the space. In this paper we will not differentiate further between these two dynamics.

The production of new papers is assumed to be oriented to the valuation landscape. So, in areas in which the fitness is higher, more problems will tend to be considered, and correspondingly more papers will appear.

An evolutionary dynamic which formulates this idea in mathematical terms may be based upon reaction-diffusion equations of Fisher-Eigen or Lotka-Volterra type. The general approach for the dynamic is given in Eq. (1) above (Ebeling, Engel, Esser, & Feistel, 1984; Feistel & Ebeling, 1982, 1989). The Fisher-Eigen equation is the simplest possibility to model an evolutionary process that includes selection between competing units. We see the production and use of scientific information as an evolutionary process in which information - represented by scientific papers – is produced, evaluated and selected among the scientific community.

The basic assumption that leads to the Fisher-Eigen model is that the local growth rate w (the general fitness) is proportional to the difference between the local fitness value $V(\vec{q})$ and the social average $\langle V(\vec{q})\rangle$. So, we have

$$w = V(\vec{q}) - \langle V(\vec{q})\rangle \tag{2}$$

The social average is defined as

$$\langle V(\vec{q})\rangle = \frac{\int V(\vec{q}\,')x(\vec{q}\,')d\vec{q}\,'}{\int x(\vec{q}\,')d\vec{q}\,'} \tag{3}$$

The resulting equation for $x(\vec{q})$ reads as:

$$\frac{\partial}{\partial t}x(\vec{q},t) = \left(V(\vec{q}) - \langle V(\vec{q})\rangle\right)x(\vec{q},t) \tag{4}$$

[3] Of course, papers may literally disappear because of inadequate archiving. We are however not addressing these issues.

In the framework of this model, the form of the fitness landscape is fixed, only the reference level $\langle V(\vec{q}) \rangle$ changes in time. The general fitness landscape *w* is shifted around its zero-level, but not changed in its shape.

If the fitness itself depends on the population density we may use the Lotka-Volterra type dynamic. Here the value of *w* is determined as a functional of the density $x(\vec{q})$ (Ebeling, Karmeshu, & Scharnhorst, 2001). In a separate paper, we will give an *approach* for modelling the peer review process in the framework developed so far.

If the population density is concentrated in certain regions of *Q* ("islands") then these "islands" can be related to the original classified populations. The "selective value" $V(\vec{q})$ is linked to the net reproduction or growth rate. This is a dynamic definition of the valuation landscape. As we discussed earlier, citations of papers or patents might be taken as representations of the valuation landscape.

In contrast to discrete models, the vanishing, merging, division, and emergence of new problems or problem fields are expressed by changes in the shape of the function $x(\vec{q})$, without having to consider changes in the taxonomy of the model. This results in a greater mathematical complexity of the model. As mentioned above, the population density follows the shape of the fitness landscape. If we assume $V(\vec{q})$ to be a random function, then the shape of $x(\vec{q})$ is sensitive to statistical properties of this function given by the probability density functional $P[V(\vec{q})]$.

The Fisher-Eigen model in Eq. (4) describes a selection process. This becomes evident if we consider the temporal evolution of populations without mutations. With increasing time, the population is concentrated in islands which correspond to particularly high values of the random function $V(\vec{q})$. Regions of low density surround these islands of high density. This means, that the selection process leads to a concentration of the distribution around the maxima. This way, we can explain how observable clustering and grouping of problems (papers) in certain areas appear. These clusters may develop into specialties and scientific fields.

Summary

Evolution is described as a sequence of self-organization processes in which innovations play a central role. In this paper we consider a model of the continuous dynamics on fitness landscapes.

The landscape picture and continuous dynamical evolution models seem to be particularly interesting as a description of search processes in social systems. The main objective of the present paper is to link the concept of an adaptive landscape to the process of generating scientific information. The main factor controlling this process is the valuation of the newly generated information, initially by reviewers, and finally by the users of that information. In this paper, we do not differentiate between the different phases of the evaluation cycle. We introduce the concept of a problem space as an abstract characteristics space analogous to the space of phenotypic properties in biological evolution.

Scientific information is codified, for example, in scientific papers. Scientific papers can be placed in an abstract space by various methods, and maps of scientific problem areas can be obtained. The published papers form an occupation landscape in this problem space. We discuss how changes in this occupation can be understood, assuming a co-evolution between the occupation landscape and a valuation landscape. This assumption is based on an understanding of scientific information as being useful for others. The use of information is always linked with a valuation.

In introducing a simple mathematical model (Fisher-Eigen) of the interaction between valuation and occupation, we show how the emergence of a structured problem space comprising specialties and fashionable problems mixed with empty areas can be explained.

References

Bruckner, E., Ebeling, W., & Scharnhorst, A. (1990). The application of evolution models in Scientometrics. Scientometrics, 18(1-2), 21-41.

Burnham, J. C. (1990). The evolution of editorial peer review. Journal of the American Medical Association, 263, 1323-1329.

Callon, M., Courtial, J. P., & Penan, H. (1993). La scientométrie. Paris: Presses Universitaires de France.

Chen, C. (2003). Mapping scientific frontiers: the quest for knowledge visualization. London et al.: Springer.

Chubin , D., & Hackett, E. J. (1990). Peerless science: Peer review and U.S. science policy. Albany, US: State University of New York Press.

Cole, S., Cole, J. R., & Simon, G. (1981). Chance and consensus in peer review. Science, 214, 881-886.

Conrad, M., & Ebeling, W. (1992). M.V. Volkenstein, evolutionary thinking and the structure of fitness landscapes. BioSystems, 27, 125-128.

Cronin, B., & Atkins, H. B. (Eds.). (2000). The Web of knowledge. A Festschrift in honor of Eugene Garfield. Medford, New Jersey: Information Today, Inc.

Davidson, G. S., Hendrickson, B., Johnson, D. K., Meyers, C. E., & Wylie, B. N. (1998). Knowledge mining with VxInsight: discovery through interaction. Journal of Intelligent Information Systems, 11(3), 259-285.

Ebeling, W., Engel, A., Esser, B., & Feistel, R. (1984). Diffusion and reaction in random media and models of evolution processes. Journal of Statistical Physics, 37, 369-384.

Ebeling, W., & Feistel, R. (1990). Evolution of complex systems. Dordrecht: Kluwer.

Ebeling, W., Karmeshu, & Scharnhorst, A. (2001). Dynamics of economic and technological search processes in complex adaptive landscapes. Advances in Complex Systems, 4(1), 71-88.

Ebeling, W., & Scharnhorst, A. (2000). Evolutionary models of innovation dynamics. In D. Helbing & H. J. Herrmann & M. Schreckenberg & D. E. Wolf (Eds.), Traffic and granular flow '99 - social, traffic, and granular dynamics (pp. 43-56). Berlin: Springer.

Ebeling, W., Schweitzer, F., & Freund, J. (1998). Komplexe Strukturen: Entropie und Information. Stuttgart, Leipzig: B.G. Teubner.

Feistel, R., & Ebeling, W. (1982). Models of Darwinian processes and evolution principles. BioSystems, 15, 291-299.

Feistel, R., & Ebeling, W. (1989). Evolution of complex systems. Berlin: VEB Deutscher Verlag der Wissenschaften.

Godlee, F., & Jefferson, T. (1999). Peer review in health sciences. London: BMJ Books.

Kronick, D. A. (1990). Peer review in 18th-century scientific journalism. Journal of the American Medical Association, 263, 1321-1322.

Kuhn, T. S. (1962). The structure of scientific revolutions. Chicago, London: University of Chicago Press.

Leydesdorff, L. (2001). A sociological theory of communication: the self-organization of the knowledge-based society. Parkland: Universal Publishers.

Nowakowska, M. (1984). Theories of research (Vol. I-II). Seaside: Intersystems Publications.

Noyons, E. C. M., & Van Raan, A. F. J. (1998). Advanced mapping of science and technology. Scientometrics, 41(1-2), 61-67.

Salton, G., Yang, C., & Wong, A. (1975). A vector space model for automatic indexing. Communications of the ACM, 18(11), 613-620.

Scharnhorst, A. (2000). Evolution in adaptive landscapes - examples of science and technology development (Discussion paper FS II 00 - 302). Berlin: Wissenschaftszentrum für Sozialforschung Berlin.

Scharnhorst, A. (2001). Constructing knowledge landscapes within the framework of geometrically oriented evolutionary theories. In M. Matthies & H. Malchow & J. Kriz (Eds.), Integrative systems approaches to natural and social dynamics - System Science 2000 (pp. 505-515). Berlin : Springer.

Small, H. (1997). Update on science mapping: creating large document spaces. Scientometrics, 38(2), 275-293.

Vlachý, J. (1986). Scientometric analyses in physics - where we stand. Czechoslovak Journal of Physics, 36(1), 1-13.

Wagner--Döbler, R., & Berg, J. (1993). Mathematische Logik von 1847 bis zur Gegenwart. Berlin, New York: de Gruyter.

Wouters, P. (2000). Garfield as alchemist. In B. Cronin & H. B. Atkins (Eds.), The Web of knowledge (pp. 65-71). Medford, New Jersey: Information Today, Inc.

Wouters, P. F. (1997). Citation cycles and peer review cycles Proceedings of the Erasmus Workshop on Quantitative Approaches to Science & Technology Studies, May 2124 1996, Amsterdam The Netherlands. Scientometrics, 38(1), 39-55.

Wright, S. (1932). The roles of mutation, inbreeding, crossbreeding and selection in evolution. Proceedings of the Sixth International Congress on Genetics, 1(6), 356-366.

Wright, S. (1988). Surfaces of selective value revisited. The American Naturalist, 131(1), 115-123.