Discovery of patterns of scientific and technological development and knowledge transfer

Anthony F.J. van Raan[1] (Keynote Speaker), Ed C.M. Noyons
Centre for Science and Technology Studies (CWTS)
University of Leiden

Abstract

This paper addresses a bibliometric methodology to discover the structure of the scientific 'landscape' in order to gain detailed insight into the development of R&D fields, their interaction, and the transfer of knowledge between them. This methodology is appropriate to visualize the position of R&D activities in relation to interdisciplinary R&D developments, and particularly in relation to socio-economic problems. Furthermore, it allows the identification of the major actors. It even provides the possibility of foresight. We describe a first approach to apply bibliometric mapping as an instrument to investigate characteristics of knowledge transfer.

1 Introduction

In this paper we discuss the creation of 'maps of science' with help of advanced bibliometric methods. This 'bibliometric cartography' can be seen as a specific type of data-mining, applied to large amounts of scientific publications. As an example we describe the mapping of the field neuroscience, one of the largest and fast growing fields in the life sciences. The number of publications covered by this database is about 80,000 per year, the period covered is 1995-1998. Current research is going on to update the mapping for the years 1999-2002. This paper addresses the main lines of the methodology and its application in the study of knowledge transfer.

2 Basic Principles of Bibliometric Mapping

Each year about a million scientific articles are published. How to keep track of all these developments, particularly the relations with other fields? Are there *cognitive structures* 'hidden' in this mass of published knowledge, at a 'meta-level'?

A research field can be defined ('delineated') by various approaches: on the basis of classification codes, selected terms in a (discipline-) specific database, selected sets of journals, a database of field-specific publications, or any combination of these approaches. In this paper we take *neuroscience*, a large field within the life sciences, as an example.

We delineate this field with the Neuroscience Citation Index of the Institute for Scientific Information (ISI)[2]. This is an appropriate database that covers over 80,000 publications annually. We collected the titles and abstracts of all these publications, for a series of successive years (1995-1998, we are currently adding 1999-2002), thus operating on several hundreds of thousands publications. With a specific computer-linguistic algorithm we parse the abstracts *of all*

1 Corresponding author, vanraan@cwts.leidenuniv.nl

2 The Institute for Scientific Information in Philadelphia, the publisher of the Science Citation Index (SCI) and all other related citation indexes.

these publications. These completely automated grammatical procedures yield all nouns and noun-phrases (standardized) that are present in the entire set of publication abstracts.

An additional algorithm creates a frequency-list of these many thousands of parsed nouns and noun-phrases while filtering out general, trivial words (Noyons 1999). We consider the most frequent nouns/noun phrases as the most characteristic concepts of the field (this can be 100 to 1,000 concepts, say N concepts).

The next step is to *encode* each of the yearly 80,000 publications with these concepts. In fact this code is a binary string (yes/no) indicating which of the N concepts is present in title or abstract. This encoding is as it were the 'genetic code' of a publication. Like in genetic algorithms, we now compare the encoding of each publication with that of any other publication. So we calculate 'genetic code similarity' (here: *concept-similarity*) of all 80,000 publications pair-wise. The more concepts two publications have in common, the more these publications are related on the basis of concept-similarity and thus can be regarded as belonging to the same subfield, research theme or research specialty. In a biological metaphor: the more specific DNA-elements two living beings have in common, the more they are related. Above a certain similarity threshold, they will belong to a particular species.

The above procedure allows clustering of *information carriers* -the publications- on the basis of similarity in *information elements* - the concepts ('co-publication' analysis). Alternatively, the more specific concepts are mentioned together in different publications, the more these concepts are related. Thus, information elements are clustered ('co-concept' analysis). Both approaches, the co-publication and the co-concept analysis are related by simple matrix algebra rules. In practice, the co-concept approach (Callon et al 1983; Noyons and Van Raan 1998) is most suited for science mapping, i.e., the 'organization of science according to concepts'.

Intermezzo: For a super market 'client similarity' on the basis of shopping lists can be translated into a clustering of either the clients (information carriers, where the information elements are the products on their shopping lists) or of the products. Both approaches are important: the first gives insight into groups of clients (young, old, male, female, different ethnic groups, etc.), and the second is important in the organization of the super market.

In main lines the clustering procedure is as follows. We first construct a matrix composed by co-occurrences of the N concepts in the set of publications for a specific period of time, e.g., 1997-1998. We normalize this 'raw co-occurrence' matrix in such a way that the similarity of concepts is no longer based on the pair-wise co-occurrences, but on the co-occurrence 'profiles' of the two concepts in relation to all other concepts.

This similarity matrix is input for a cluster analysis. In most cases, we use a standard hierarchical agglomerative cluster algorithm including statistical criteria to find an optimal number of clusters. The identified clusters of concepts represent 'subfields'. These subfields are labeled with the four most frequent concepts in a cluster.

The clusters resulting from the mapping procedure are tested for internal coherence. We calculated the average linkage between all concept-pairs within a cluster, and the standard deviation. This internal coherence measure indicates the robustness of the identified cluster. We refer to Noyons and Van Raan (2002).

Each subfield represents a sub-set of publications on the basis of the above discussed concept-similarity profiles. If any of the concepts is in a publication, this publication will be attached to the relevant subfield. Thus, publications may be attached to more than one subfield. The overlap between subfields in terms of joint publications is used to compile a further co-occurrence matrix, now based on subfield publication overlap. This matrix is used to calculate a similarity measure of subfields by comparing their co-occurrence profile with others.

To construct a map of the field, the subfields (clusters) are positioned by multidimensional scaling. Thus, subfields with a high similarity (with a similar 'cognitive orientation' within the field) are positioned in each other's vicinity, and subfields with low similarity are distant from

each other. The size of a subfield (represented by the surface of a circle) indicates the share of publications in relation to the field as a whole. Particularly strong relations between two individual subfields are indicated by a connecting line (see discussion in Section 4).

A similar mapping procedure can be applied to documents other than publications, for instance patents. Thus, maps of technology can be constructed. In this paper we confine ourselves to the mapping of neuroscience.

3 Maps as Analytical Instrument

The above procedure generates the *cognitive structure* of the field neuroscience. As discussed above, it is entirely based on the total of relations between all publications. The fascinating point is that the discovered structure is not the result of any pre-arranged classification system or whatsoever. Nobody prescribes this structure. It emerges solely from the internal relations through concept-similarities of the whole ensemble of publications together. In other words, what we make visible by our mathematical methods, is a *self-organized cognitive ecology of science*.

A detailed discussion of science maps is given by Noyons (1999). Our mapping procedure depends partly on expert input. Special internet-facilities enable experts to comment on the concepts used to generate the structure of the field.

The maps are put in a digital form on a (protected, in cases of confidentiality) part of the CWTS website[3]. Thus we make the maps easily accessible for users in order to explore the field or to validate the results[4]. We also provide information 'behind' the map (actors, and their output and impact indicators) by an interface that can be used via standard graphical internet browsers (e.g., MS Internet Explorer and Netscape Communicator).

This advanced bibliometric mapping has many interesting analytical potentials. First, it visualizes the landscape of a scientific field 'embedded in its surroundings', i.e., in its interdisciplinary relations. We found that a major part of the landscape relates to socio-economic problems (Van Raan, 2001). For neuroscience obvious examples are: Altzheimer disease, Parkinson's disease, aging, stroke. Second, by making these maps for a series of years, we are able to observe trends and changes in structure (see our website for examples). Extrapolation of these trends enables foresight of near-future developments.

Third, bibliometric maps allow localization of major actors. Thus we are creating a strategic map: who is where in science, and, more precisely, what is the position of these actors in terms of interdisciplinary relations of the different fields? In addition to that, we may assess an actor's scientific influence ('impact') in the field by applying standard CWTS bibliometric analysis. In this way, the two major pillars of bibliometric methodology, concept-based mapping and citation-based impact analysis, are combined. This combined approach is very useful in the identification of scientific 'centers of excellence'. We will not further address this search for excellence and refer to Van Raan (2000). However, citation analysis is crucial in this paper, not as an instrument to assess impact, but to identify communication patterns.

In this paper we focus on the application of bibliometric mapping in the analysis of knowledge transfer. This is a first and still experimental approach, to begin with an analysis on a not-too-large scale (science as a whole), but within a major field, in this case neuroscience.

In order to develop this map-based knowledge transfer analysis more systematically, we distinguish two types of intra-field (i.e., between subfields) relations: (1) *conceptual linkages* (which is the basis if the map structure, as explained above), and (2) *communication linkages*,

3 http://www.cwts.leidenuniv.nl

4 An important aspect of the mapping methodology is the retrieval rate: which part of the publications covered by the dataset used as a starting point for the mapping procedure, can be found back in the map? For this field, neuroscience, we reach a retrieval rate of at least 80%.

based on the extent to which publications in a specific subfield cite publications in other subfields.

We hypothesize that these two linkage modalities are basic elements of scientific development and that the bibliometric mapping allows the visualization and further analysis of the patterns involved. Furthermore, we regard conceptual linkage as the source of potential knowledge transfer, and communication linkages as the realization of knowledge transfer. Thus, comparison of these two types of linkage may reveal differences in potential versus 'already existing' knowledge transfer.

4 Results

The result of the neuroscience mapping is shown in Figure 1a. The clusters represent subfields and research themes according to List 1 in which the (at most) four most frequent concepts of the cluster are given to label the cluster. In addition, we have indicated as an example the relatively strongest *conceptual linkages* between cluster 10, brain infarction research (stroke) with other research subfields of neuroscience, particularly subfields 3 (Etiology), 11 (Subarachnoid hemorrhage), 15 (Magnetic resonance imaging, MRI), and 21 (Ischemia).

In Figure 1b we present the relatively strongest *communication linkages* (citation-based) between brain infarction research and other research subfields of neurosciences, and now these linkages are particularly with subfield 3 (Etiology) and 20 (Animal model).

We observe that brain infarction research is an example of reasonably similar but still different conceptual (words) and communication (citations) linkages, as is illustrated by comparison of Fig. 1a and Fig. 1b. For instance, we see more conceptual linkages than communication linkages. A first step to explain these findings is the analysis of the research fields involved in the different subfields. Although we deal with subfields of neuroscience, publications in these neuroscience subfields may belong to other fields than neuroscience only. For instance, it is clear that brain infarction research will involve the field of cardiovascular system. This means, that publications on brain infarction research, may appear in cardiovascular journals.

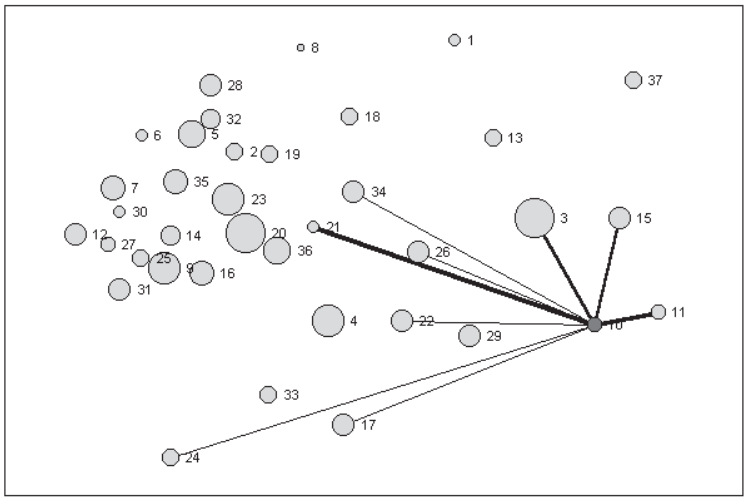

Figure 1a:Conceptual linkages between brain infarction research with other subfields of neuroscience. Two-dimensional representation based on the similarities between identified clusters of concepts (subfields). For the list of subfields with corresponding number we refer to List 1. The size of the subfields represents the number of publications in a specific subfield.

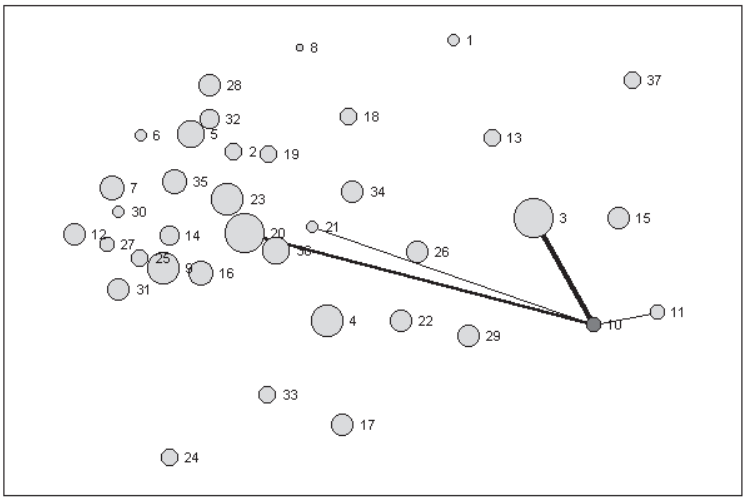

Figure 1b: Communication linkages between brain infarction research and other subfields of neuroscience.

List 1: Most frequent concepts in the neuroscience subfields

1 Multiple sclerosis / myelin basic protein / experimental autoimmune encephalomyelitis / lewis rat
2 Astrocytes / Glial cell / TNF Alpha / acidic protein
3 Etiology / differential diagnosis / neurological deficit / spinal cord injury
4 Schizophrenia / Ethanol / Alcohol / normal control
5 Retina / skeletal muscle / neuronal cell / molecular mechanism
6 NGF / nerve growth / neurotrophic factor / pc12 cell
7 Ca2+ / inhibitory effect/ protein kinase
8 Amyotrophic lateral sclerosis / motor neuron disease
9 h 3 / Dopamine / Antagonist / Agonist
10 Stroke / ischemic stroke / stroke patient / cerebral infarction
11 subarachnoid hemorrhage / middle cerebral artery / internal carotid artery
12 Peptide / Hormone / Secretion / Male Rat
13 CSF / HIV / AIDS / human immunodeficiency virus
14 Glutamate / NMDA / NMDA Receptor / glutamate receptor
15 MRI/ computed tomography / Functional MRI
16 Acetylcholine / Neurotransmitter / Uptake / Norepinephrine
17 Depression / Placebo / Anxiety
18 Alzheimers Disease / a beta / amyloid precursor protein / beta amyloid
19 Apoptosis / cell death / neuronal death / neurodegenerative disease
20 Animal model / electrical stimulation / Fiber / Pathophysiology
21 Ischemia / cerebral ischemia / neuronal damage / neuroprotective effect
22 Dementia / Aging / cognitive function / cognitive impairment
23 Axon / Immunoreactivity / Immunohistochemistry / Adult Rat
24 Hart rate / blood pressure / sympathetic nervous system / heart rate variability
25 Gaba / synaptic transmission / gamma aminobutyric acid / synaptic plasticity
26 PET / cerebral blood flow / white matter
27 Hypothalamus / c fos / paraventricular nucleus / locus coeruleus
28 Gene/CDNA / polymerase chain reaction / expression pattern
29 Seizure / EEG / Epilepsy / temporal lobe
30 nitric oxide synthase / l arginine / neuronal nitric oxide synthase
31 Stress / substance p / neuropeptide y / tyrosine hydroxylase

32 spinal cord / Peripheral nerve / sensory neuron / dorsal root ganglion
33 Memory / Learning / working memory / memory impairment
34 Pathogenesis / Parkinsons Disease / basal ganglion / oxidative stress
35 MRNA / rat brain / gene expression / olfactory bulb
36 Hippocampus / Cortex / Cerebellum / Striatum
37 Tumor / Brain Tumor / radiation therapy / primitive neuroectodermal tumor

For the classification of journals into fields we use the ISI classification system. Thus, on the basis of the journals in which the publications of the different subfields have been published, we make a frequency list of the research fields involved.

Below we present the ranking of the first ten research fields involved in brain infarction research with the number of publications in 1997-1998. These results clearly show the interdisciplinary 'make up' of the different subfields. The infarction focus of subfield 10 is clearly visible in the fields immediately following both general neuro-fields (neuroscience and clinical neurology): cardiovascular system and vascular diseases.

Sf 10	*Brain infarction research*
1305	NEUROSCIENCE
1042	CLINICAL NEUROLOGY
710	CARDIOVASCULAR SYSTEM
575	VASCULAR DISEASES
206	MEDICINE, GEN. & INTERNAL
184	SURGERY
111	PHARMACOLOGY & PHARMACY
104	RADIOLOGY & NUCL MEDICINE
101	PSYCHIATRY
86	HEMATOLOGY

As discussed above, the strongest *conceptual* relations of subfield (Sf) 10 (Brain infarction research) are with subfields 3 (Etiology), 11 (Subarachnoid hemorrhage), 15 (MRI) and 21 (Ischemia). An analysis of the research fields involved in these subfields gives the following results:

Sf 3	*Etiology*	*Sf 11*	*Subarachnoid hemorrhage*
5963	NEUROSCIENCE	1491	NEUROSCIENCE
3577	CLINICAL NEUROLOGY	1056	CLINICAL NEUROLOGY
1616	SURGERY	650	SURGERY
1082	OPHTHALMOLOGY	382	CARDIOVASCULAR SYSTEM
865	PEDIATRICS	307	VASCULAR DISEASES
856	MEDICINE, GEN. & INTERNAL	305	RADIOLOGY & NUCLEAR
786	PSYCHIATRY	91	MEDICINE, GEN. & INTERNAL
653	OTORHINOLARYNGOLOGY	78	ANESTHESIOLOGY
627	RADIOLOGY & NUCLEAR MEDICINE	62	PEDIATRICS
575	ANESTHESIOLOGY	58	PHARMACOLOGY & PHARMACY

Sf 15	*Magnetic resonance imaging (MRI)*	*Sf 21*	*Ischemia*
2404	NEUROSCIENCE	1409	NEUROSCIENCE
1925	CLINICAL NEUROLOGY	394	CLINICAL NEUROLOGY
886	RADIOLOGY & NUCLEAR MEDICINE	214	CARDIOVASCULAR SYSTEM
683	SURGERY	175	PHARMACOLOGY & PHARMACY
329	PSYCHIATRY	157	VASCULAR DISEASES
328	PEDIATRICS	155	BIOCHEMISTRY & MOLEC BIOLOGY
213	MEDICINE, GEN. & INTERNAL	153	ENDOCRINOLOGY & METABOLISM
134	CARDIOVASCULAR SYSTEM	147	HEMATOLOGY
121	ONCOLOGY	130	SURGERY
107	VASCULAR DISEASES	63	ANESTHESIOLOGY

The strongest *communication* linkages of subfield 10 are with subfields 3 and 20. The fields involved in subfield 3 (Etiology) are already presented above, for subfield 20 (Animal model) we find:

Sf 20	*Animal model*
9000	NEUROSCIENCE
2815	PHARMACOLOGY & PHARMACY
1655	CLINICAL NEUROLOGY
1437	BIOCHEMISTRY & MOL. BIOLOGY
1254	PHYSIOLOGY
763	PSYCHIATRY
488	CELL BIOLOGY
427	PSYCHOLOGY
422	ENDOCRINOLOGY & METABOLISM
393	SURGERY

We observe a preference for the conceptual linkages in both (cardio)vascular research (subfields 11, 15, 21) as well as surgery (subfields 3, 11, 15), whereas at the communication side we see less cardiovascular research and more orientation toward surgery (subfield 3) and toward the quite general subfield 20, Animal model.

This first observation suggests that formal communication (in terms of citation patterns) is more inclined to clinical practice, and less to the more scientifically based interaction with other medical fields such as cardiovascular research. It is more 'general', and less specific.

Similar analysis of the conceptual and communication linkages of other subfields tends to confirm that the communication (citation-based) linkages are more clinically and more generally oriented than the more scientifically oriented conceptual linkages. Such 'mismatches' of the two types of knowledge linkages are very interesting as they may point to (significant) differences in 'potential knowledge transfer' and 'realized knowledge transfer'.

But our above observations represent first and preliminary results, we are currently investigating this phenomenon more systematically and will report on further results during the conference.

5 Concluding Remarks

Bibliometric mapping is a powerful methodology to visualize the cognitive landscape of a research field. In this paper we present our current work on a method to analyse knowledge transfer by distinguishing two types of intra-field research relations, conceptual linkages and communication linkages. On the basis of first results of this approach we are convinced that further systematic work along these lines will lead to a better understanding of the process of knowledge transfer.

6 References

Callon, M., J.-P. Courtial, W.A. Turner, and S. Bauin (1983). From translations to problematic networks: an introduction to co-word analysis. *Social Science Information* 22, 191-235.

Noyons, E.C.M. and A.F.J. van Raan (1998). Monitoring Scientific Developments from a Dynamic Perspective: Self-Organized Structuring to Map Neural Network Research. *Journal of the American Society for Information Science (JASIS),* 49, 68-81.

Noyons, E.C.M. (1999), *Bibliometric mapping as a science policy and research management tool*. Thesis Leiden University. Leiden: DSWO Press.

Noyons, E.C.M., M. Luwel and H.F. Moed (1999). Combining Mapping and Citation Analysis for Evaluative Bibliometric Purpose. A Bibliometric Study on Recent Development in Micro-Electronics. *J. of the American Society for Information Science (JASIS),* 50, 115-131.

Noyons, E.C.M. and A.F.J. van Raan (2002). Science mapping from publications. In: *Dealing with the data flood*, J. Meij (ed.), The Hague: STT/Beweton, ISBN 90-804496-6-0 (also available on CD-Rom), p. 64-72.

Van Raan, A.F.J. (2000). The Pandora's Box of Citation Analysis: Measuring Scientific Excellence, the Last Evil? In: B. Cronin and H. Barsky Atkins (eds.), *The Web of Knowledge. A Festschrift in Honor of Eugene Garfield*, p. 301-319. Medford (New Jersey): ASIS Monograph Series, 2000. ISBN 1-57387-099 -4.

Van , A.F.J. (2001). Mapping R&D related to socio-economic problems. In: *Proceedings of the Quality of Life Impact Workshop of the European Commission*, June 2000. Brussels: European Commission.

7 Contact Information

Anthony F.J. van Raan
Centre for Science and Technology Studies (CWTS)
University of Leiden
Wassenaarseweg 52
P.O. Box 9555
2300 RB Leiden

e-mail: vanraan@cwts.leidenuniv.nl