Co-occurrence and Knowledge Mapping to Identify Hot Topics and Key Players in the Field of Mobility and Transport

Clemens Widhalm[1]
Ute Gigler
Alexander Kopcsa
Edgar Schiebel

Proceedings of the 8th Conference of the
International Society for Scientometrics and Informetrics ISSI,
Vol 2, pp 751-758,
July 2001, Sydney

[1] Email: clemens@widhalm.co.at ; Internet: www.widhalm.co.at

Contents

1. Introduction

The article will demonstrate different approaches for bibliometric analysis of a research field as it is documented in international scientific literature and in project descriptions of the 4th European Framework Programme by applying a number of methods including co-occurrence analysis and visualisation in 2-dimensional knowledge maps. Emphasis will be placed on how specific topics can be located in the scientific landscape and how actors can be identified and grouped together either because of their actual collaboration or because of their thematic specialisation.

The analysis will focus on the aggregate situation of international research, but will also examine specific topics such as propulsion systems. The objective is to identify specific features within the patterns of research fields, patterns of experts and patterns of institutions relevant in the context of mobility and transport and specifically in the areas of alternative fuels and fuel cells.

The research also aims at demonstrating how to analyse the role of key centres of excellence and at determining the relative importance of specific countries with respect to certain topics.

2. Database

2.1 Information on data

For this analysis, we used a wide range of databases to offer a comprehensive look at the topics mobility and transport. Literature databases with different emphases were selected. INSPEC (The Institution of Electrical Engineers), Compendex and National Technical Information Service NTIS (US Department of Commerce) focus on technical and engineering aspects; the database Transport Research Information System TRIS concentrates on transport-related topics, whereas Enviroline (Congressional Information Service, Inc.) and Environmental Bibibliography emphasise environmental and sustainability topics.

In order to obtain literature to various topics in the fields mobility and transport from the databases described above, search terms were logically connected as follows. One of the four terms, transport, traffic, mobility, and infrastructure, had to occur together with a term on a list of more detailed subtopics in the areas of information and communication technologies (ICT) and services, sustainability, or innovative technologies. The query was restricted to title words and to a period of 9 years (1992-2000), in order to obtain a representative number of articles (on the order of 3500). The articles downloaded included bibliographic information such as title, author, institution, source, publication year, keywords, and the abstract, but not the full text.

The literature data were completed with about 800 project descriptions of the 4th Framework Programme in the field of transport research.

3. Methodology

3.1 Automatic indexing

To represent the contents of the articles described above in a way that is suitable for further analysis the titles and abstracts had to be indexed automatically. For this purpose a stemming procedure based on the context-sensitive longest-match principle and a phrase recognition algorithm were applied to titles and abstracts (Widhalm et al. 1999).

3.2 Standardisation of data

In order to be able to apply the bibliometric analysis with the software BibTechMon™ for the network analysis all the data had to be unified with respect to different spelling of the same names of authors or institutions or keywords. In addition, synonyms and abbreviations were standardised. After standardisation 1600 terms and phrases (keywords) were selected for further analysis.

3.3 Co-occurrences

Bibliometric methods are used to structure electronically stored information in internal or external databases. The structuring is based on the calculation and often visualisation of relations between objects, such as documents, keywords, authors or institutions. The relations are derived from indicators that can be defined through different models (Van Raan, 1992).

In order to find out systematically which keywords are closely related to one another and therefore form a topic, a multidimensional bibliometric measure is needed. It is assumed that keywords of one topic are more often used together in documents than keywords belonging to different themes. So as a measure for each pair of keywords the number of co-occurrences in a document (article or project description) is counted (compare Callon et al. 1983, Rip and Courtial 1984, Leyersdorf 1989, Kostoff 1993).

For statistical reasons the Jaccard Index was used to normalise the elements of the respective co-occurrence matrix. This index provides better information about the "intensity" correlation of keywords:

$$J_{ij} = \frac{c_{ij}}{c_{ii} + c_{jj} - c_{ij}}$$

where c_{ij} is the co-occurrence of keywords *i* and *j*, c_i is the total number of occurrences of keyword *i*. J_{ij} therefore is a normalised measure for the intensity of correlation.

3.4 Cluster analysis of keywords

Cluster analysis on the basis of the Jaccard matrix J_{ij} leads to groups of keywords that are named as themes. As an example two of the defined themes – *fuels* and *traffic management and control* (TMC) are monitored as to their temporal development in terms of number of respective keywords used in publications per year (see fig. 1).

Traffic management and control (TMC) seems to have been a dominant topic in the last several years, although the peak publication activity has already passed. This might indicate that this topic has been discussed extensively in scientific journals until 1997 and has since developed from a predominately scientific topic to a more widely applied one. As a result, publications in scientific journals have decreased (compare Widhalm 2000). *Fuels,* however, is a topic of increasing interest.

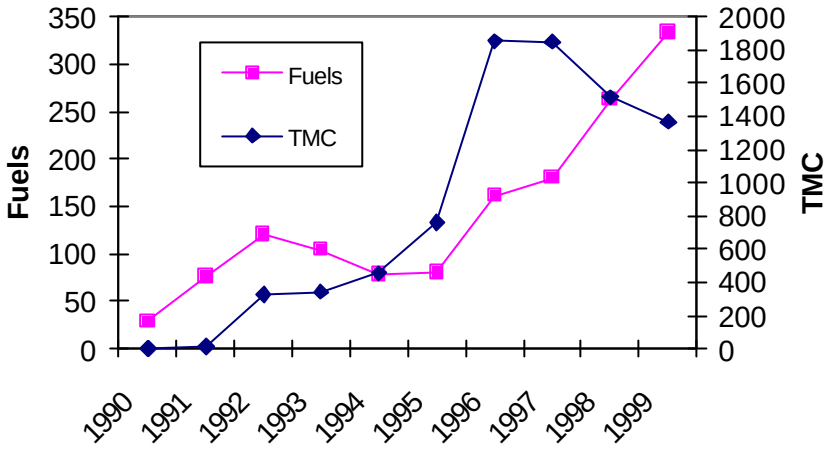


Fig. 1: Temporal development of *fuels* and *TMC* in terms of number of respective keywords used in publications per year

A closer look at the development of the use of some of the individual keywords of the cluster *fuels* (see fig. 2) illustrates that the discussion on *alternative fuels* or *fuel cells* started in the *transport sector* and only recently became important in the *passenger car* sector. Figure 2 also shows that the discussion on *alternative fuels* like *methanol* is decreasing, while *electric vehicles* and *fuel cells* are topics of increasing importance.

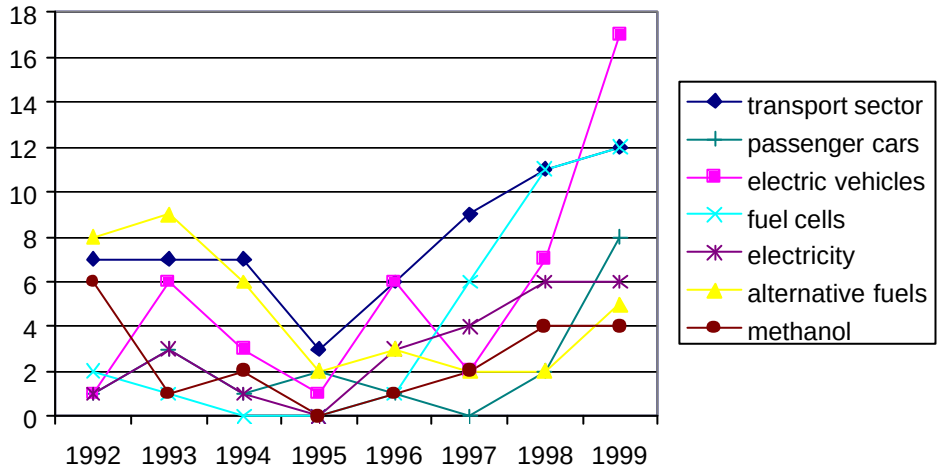


Fig. 2: Temporal development of some keywords of the cluster *fuels* in terms of number of respective keywords used in publications per year

3.5 Visualisation in knowledge maps

The matrix J_{ij} contains information about correlation of terms, nevertheless, it is not easy to interpret. For this reason a visualisation method had to be carried out that is capable of transforming the matrix into an intuitive readable 2-dimensional map. For this purpose a mechanical spring model is applied (Kopcsa and Schiebel 1998) as follows. Keywords in this model are mass points with a mass and size proportional to the total frequency J_{ii}. The masspoints are connected with each other by forces (springs) correlated to J_{ij}. The masspoints are positioned randomly as a starting position and then will move driven by the forces defined above. This is done by iteration of the respective n-dimensional differential equation system. So the masspoints will be positioned on a 2-dimensional map in correlation to their co-operational relation to each other as defined by the Jaccard matrix. Through this model the keywords are positioned according to their correlation; intensively correlated terms will be pictured in close proximity to one another.

Thus the map provides information based on the size of the objects (correlated to the total frequency) and the relative position to one another (see also Van Raan, A. F. J., 1992). To enrich the content of such maps additional information is appended by colouring the objects to visualise other parameters of interest as for instance the belonging to a theme as defined by the cluster analysis.

4. Discussion of knowledge maps

The purpose of this paper is not to provide a comprehensive discussion of the topics mobility and transport, but rather to present a few examples on how to facilitate interpretation of a large number of articles using co-occurrence analysis and visualisation in maps. Different knowledge maps consisting of different types of objects (keywords, institutions and authors) will be presented in order to demonstrate the possibilities offered by this method which supersedes traditional analysis.

4.1 Network of keywords

The keywords shown in the network below are coloured to indicate their belonging to a theme as found by the cluster analysis. Immediately it can be seen, which of the themes have close relations to one another. Sustainable development (Sust) appears to be a topic overlapping with transport modes (Mod); both are located between traffic/transport/urban (TTUP) planning on the one side and the topic of *fuels* on the other. This corresponds with the fact that intermodality in transport (e.g. park and ride, public and individual transport) is pushed forward by urban planning authorities to improve the traffic situation in urban regions, while alternative propulsion systems (fuel cells or alternative fuels like ethanol, ...) should reduce air pollution caused by traffic.

Traffic and transport planning (TTUP) on the other side is closely connected with safety (Saf) for all participants of traffic and only can be effective in connection with traffic management and control systems (TMC), the latter being connected with necessary hardware networks/telecommunication (Net).

As an example we have a closer look (see fig. 4) at the *fuel* cluster (b) (north east section in fig. 3). The map shows a concentration of two subclusters –*fuel cell* (a) and *alternative fuels* (c).

name	elem.	color
air traffic / navigation	31	
eu	52	
fuels	80	
networks/telecommunication	164	
sustainable development	**99**	
traffic / transport / urban planning	120	
traffic management/controll	270	
transport modes	123	
Industry	88	
safety	64	

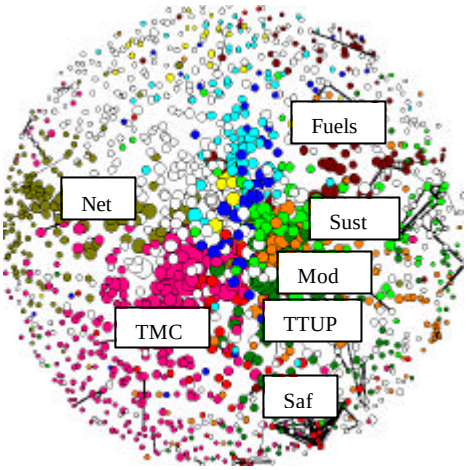
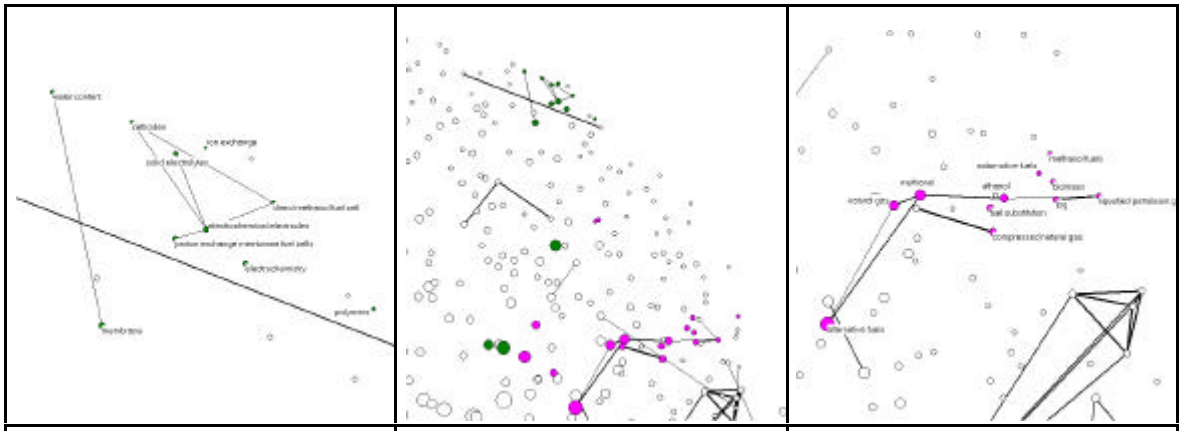


Fig. 3: Network of keywords and its thematic clusters (description see text above)



a) fuel cell subcluster | b) fuel cluster | c) alternative fuel subcluster

Fig. 4 a-c: Sections of network of keywords

The separation of the cluster *fuels* allows a closer look at possible differences. If the development of publications per year using keywords of the respective clusters is compared, the extremely strong increase of the *fuel cells* publications since 1995 is evident. The *alternative fuel* topic lost importance in the beginning of the 90s while growing again together with the other topic.

A question of interest might be which regions are responsible for the different developments as shown in fig. 5.

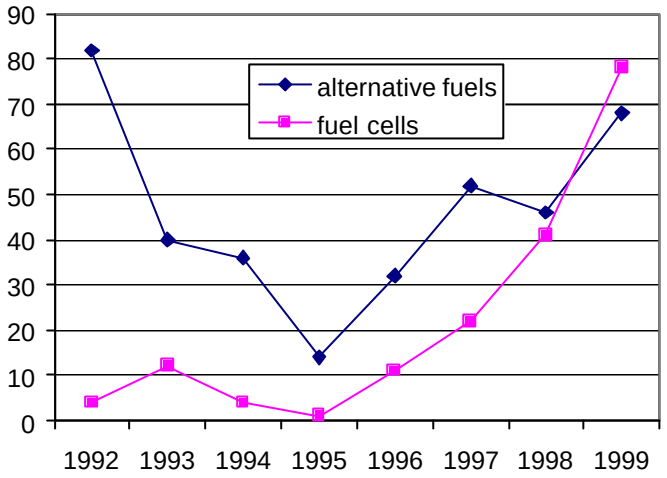


Fig. 5: Temporal development of subclusters *alternative fuels* and *fuel cells* in terms of number of respective keywords used in publications per year

Since each document provides information about institutions responsible for the publication, we will examine the most prominent institutions for the fuel subclusters (see table 1 a-b).

Table 1: Number of keywords used by institution in subclusters

a) *alternative fuels*:

192	USA	department of energy, washington, dc.
82	USA	iowa state university, ames, midwest transportation center
44	USA	california state dept. of transportation, sacramento.
36	USA	booz-allen and hamilton, inc., usa
36	USA	federal transit administration, washington, dc.
14	USA	transportation research board, nw , washington, dc, usa
14	EU	swedish transport and communications research board, stockholm (sweden)

b) *fuel cells*:

15	USA	department of energy, washington, dc.
11	EU	dept. of chem. & process eng., newcastle upon tyne univ., uk
6	USA	los alamos nat. lab., nm, usa
6	USA	arthur d. little inc., cambridge, ma, usa
5	Asia	hong kong univ of science & technology
5	EU	forschungszentrum julich gmbh, germany
5	USA	iowa state university, ames, midwest transportation center
4	EU	kommunikationsforskningsberedningen, stockholm, sweden

While the US clearly dominates literature on *alternative fuels*, the EU (European Union) plays a considerable role in the topic *fuel cells*. The following table illustrates the number of keywords used in the two subclusters:

Table 2: Number of keywords used by institutions of a region in subclusters

Region	*alternative fuels*	*fuel cells*
EU	114	45
USA	470	61

4.2 Network of institutions

Focusing on institutions(= e.g. author's affiliation) provides a different way of grouping articles. In this case, institutions will not be grouped together based on their co-operation on articles or research projects (conf. Widhalm et al. 2001), but based on their affiliation with similar contents in articles or research projects. So two institutions are correlated closely to one another if their respective publications are represented by common keywords. If this focus on institutions is used, institutions with similar interests are grouped together, even if they never at all co-operated with each other. In fig. 6 we find those institutions dealing with *alternative fuels* (AF) quite concentrated in the lower section, while those publishing on *fuel cells* (FC) are spread over a wider area mainly in the upper section. Two central institutions dealing with both topics are also shown (compare table 1). The size of the circles corresponds with the number of keywords used by an institution in all its publications together.

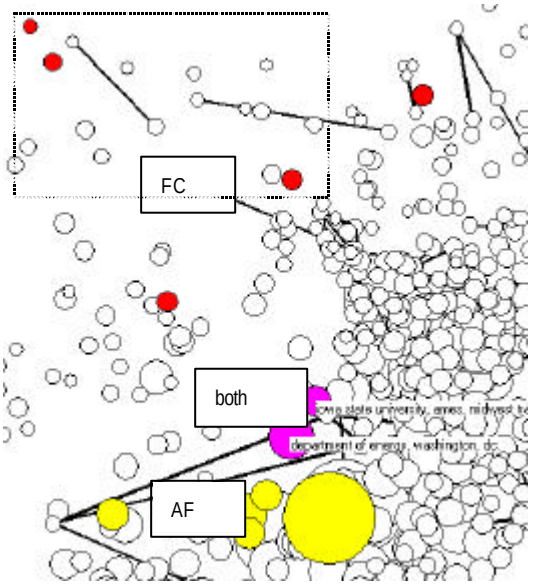


Fig. 6: Network of institutions with experts on fuels (description see text above)

Looking for more institutions relevant for fuel cells we just calculate those using similar keywords measured by respective maximum J_{ij} to those five already found and as a result get those marked with flags in fig. 7 (zoom of fig. 6). If looked up in the database it can be verified that those two published using keywords like fuel cells, fuel substitution or electric vehicles.

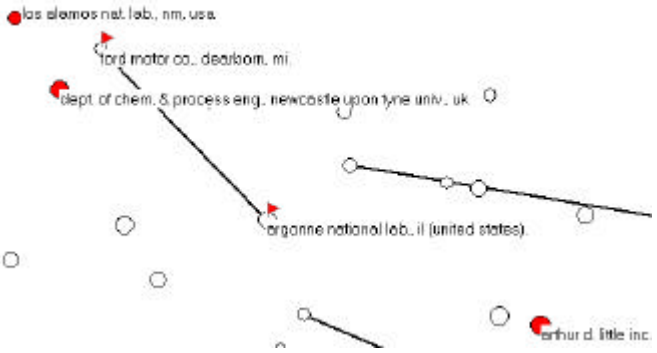


Fig. 7: Network of institutions with experts on fuel cells (description see text above)

The above example illustrates the advantage of using bibliometric tools for finding actors for specific topics in a systematic way.

4.3 Network of authors

In the same way as keywords can be grouped together because of their co-occurrence in publications it is possible to look at the authors of papers grouped together because of their co-authorship in articles. Typical patterns can be seen on fig. 8, such as isolated authors (IA) or author teams (Teams) on the one hand, and as authors serving as linkages (Link) between various teams. Further investigations on the basis of an author's network can be performed in the same way as for institutions and will therefore not be demonstrated here.

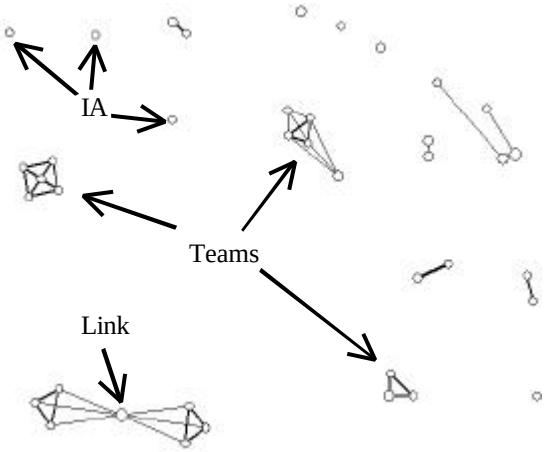


Fig. 8: Network of authors with typical patterns of co-operation

5. Conclusions

To gain insight into a research area just hierarchical and one-dimensional structuring of the respective data is not sufficient. The complexity of a research network is multidimensional and therefore multidimensional indicators like co-occurrence and the Jaccard index need to be used for an adequate characterisation. Unfortunately, the interpretation of multidimensional indicators is extremely difficult, and this is where the main advantage of the methodology and the tool BibTechMon™ used in this article lies. As a reasonable compromise, a graphical mapping methodology based on a spring model is presented where the multidimensional problem is reduced to a 2-dimensional knowledge-map with minimum aberrations to the original system.

As example for different possible network representations, those for keywords or authors because of their co-occurrence in publications, and for institutions because of their affiliation to similar contents in their publications are introduced. The advantages of each of these views on a research topic are shown dealing with the global theme on mobility and transport as well as focusing on subthemes like alternative fuels or fuel cells.

The analysis of the topics above using the visualisation tool BibTechMon™ provides insights on hot topics and the respective keyplayers, centres of excellence and experts. The interconnection of research fields can be studied by analysing correlations with respect to content or relations of common key players. Hence, this systematic tool supports the identification of possible partners or competitors.

6. References

Callon, M., J. P. Courtial, W. A. Turner and S. Bauin (1983): From Translation s to Problematic Networks: an Introduction to Co-Word Analysis. Social Sciences Information, **22**, 191-235

Kopcsa, A. and E. Schiebel (1998): ‚Science and Technology Mapping: A New Iteration Model for Representing Multidimensional Relationships', Journal of the American Society for Information Science JASIS, Jan, 1998, No. 1/ Vol49, p.7ff.

Kostoff R. (1993): Co-Word Amalysis; Kluwer Academic Publishers, 63-78

Leyersdorf, K. (1989): Words and Co-Words as Indicators of Intellectual Organisations; Research Policy, **18**, 209-223

van Raan, A. F. J. (1992): Advanced bibliometric methods to assess research performance and scientific development: basic principles and recent practical applications. Research Evaluation, **3**, 3, 151-166

Rip, A. and P. Courtial (1984): Co-word maps of biotechnology: An example of cognitive scientometrics, Scientometrics, 6, 381-400

Widhalm C., A. Kopcsa, E. Schiebel, H.G. Müller and N. Balling (1999): Konzeptive Entwicklung eines Einlesesystems und einer Strategie zur automatisierten Schlagwortgenerierung; OEFZS-S-0051, confidental.

Widhalm C.: Datenrecherche Magneto- Elektronik, Austria Innovativ 3/2000, S.26/27, 2000

Widhalm C., M. Topolnik, A. Kopcsa and E. Schiebel (2001): Evaluating Patterns of Co-operation: Application of a bibliometric visualisation tool to the 4th Framework Programme and the Transport Research Programme; Research Evaluation, accepted.