Mapping Research Topics through Word-reference Co-occurrences[#]

Gaston Heimeriks[*] and Peter Van den Besselaar[*,**]

**gaston.heimeriks@niwi.knaw.nl, peter.van.den.besselaar@niwi.knaw.nl*
Social Sciences department, NIWI, Royal Netherlands Academy of Arts and Sciences (The Netherlands)

***p.a.a.vandenbesselaar@uva.nl*
Amsterdam School of Communication Research, ASCoR, University of Amsterdam,
PO BOX 95110, 1090 HC Amsterdam (The Netherlands)

Abstract
Mapping of science and technology can be done at different levels of aggregation, using a variety of methods. In this paper, we propose a method in which title words are used as indicators for the content of a research topic, and cited references are used as the context in which words get their meaning: co-occurrences of *word-reference combinations*. In this way we can use words without neglecting differences and changes in meaning. As we will show, the method has several advantages, such as high coverage of publications and the use of the same words in different contexts. Applying the method in information science produces knowledge maps that are an adequate representation of research topics in the context of the entire field.

Introduction
Science mapping aims at revealing the structure and dynamics of science using attributes of communications, most importantly scientific publications. Mapping, however, can be done at several levels of granularity. For example, one may want to map the development of research fields, subfields or topics (the research front). In the cartography of science, a wide variety of units of analysis can be distinguished. Examples include ideas, concepts, themes and paradigms. These concepts are represented and conveyed through words, terms, documents and collections by individual authors, groups of authors, specialties and scientific communities. The definition of proximity determines the structure that is analyzed. Mapping science and technology has a long tradition that started with the co-citation analysis developed by Small (1973). Since then, many methods have been proposed, such as journal-journal citations for mapping research fields; author and article co-citation for mapping the more fine-grained structure of research fields, and co-word analysis and bibliographical coupling for mapping micro-level research topics within fields. In this paper we will introduce a method for mapping research topics, based on co-occurrences of word-reference combinations. These co-occurrences are used to cluster similar papers as representations of research fronts. The method is different from the current methods used in the literature, and we will show its practical and theoretical advantages. We will apply the method in the field of information science, and use the preliminary results to evaluate the method.

Mapping research fields
First of all, it is necessary to define some concepts, which indicate the various levels at which we can map structure and dynamics of research. The highest (broadest) level is the discipline, such as sociology or physics. The next level is the research field, such as science and technology studies, or particle physics. A more detailed level still is the research subfield, such as scientometrics, and this again is subdivided into research topics, which is the smallest unit we focus on. Within scientometrics we consider patent studies, or co-word analysis as research topics.

At the higher level of research fields, cartography is generally based on sets of scholarly journals. A good example is science and technology studies, which can be defined by journal-journal citations as a network around five scholarly journals. At a lower level, this network is composed of three research subfields, almost without a shared knowledge base, and with only very weak ties

[#] Authors are in alphabetical order. The research underlying this paper is partly funded by the European Commission in two projects: the SOEIS project (SOE1-CT97-1060) and the EICSTES project (IST-1999-20350).

Proceedings of ISSI 2005, pages 575–584.
Edited by P. Ingwersen and B. Larsen.

between the subfields (Van den Besselaar 2000, 2001).[1] At this level of aggregation, many studies have shown that journal-journal citations can be used for mapping the structure of research fields. Underlying this method is the idea that researchers in a field share a common knowledge base, and this is reflected in the choice of references. Through their local citing behavior, researchers reproduce the identity of the field at an aggregated level – and journal-journal citations can be used to map this identity in terms of sets of journals. A research field can be defined as a network of journals dealing with similar research questions and methodologies and referring to a largely overlapping set of literature. As a consequence of this last characteristic, we expect journals belonging to the same research field to exhibit similar aggregated citation patterns. If that is the case, the analysis of journal-journal citations may result in an operational definition of a research field in terms of a set of journals with similar citation patterns. The method has been developed and used for delineating disciplinary research fields, but elsewhere we showed that the method is also well suited for delineating interdisciplinary fields (Van den Besselaar & Heimeriks, submitted). As the method is described in detail in another paper (Van den Besselaar & Leydesdorff, 1996), we will not go into detail here. In this paper it forms the context for mapping the more fine-grained structure of science, at the level of research topics.

Mapping research topics

Various methods for the mapping of the fine-structure of research topics are discussed in the literature. Co-citation analysis and co-word analysis are among the most widely used techniques. Marshakova (1973) and Small (1973) independently developed co-citation analysis by noting that if two references are cited together in a publication, the two references are themselves related. The greater the number of times they are cited together, the greater their co-citation strength. Author co-citation analysis results in clusters of authors, who are linked through co-occurrences in reference lists. This method maps research fields in terms of clusters of authors rather than topics per se, and these have to be derived from the known research interests of these authors. As most researchers cover various topics in their active life, the clusters generally do not exhibit a micro picture of the (changing) research topics that dominate a discipline. The resulting clusters can be conceived as representations of research foci, but substantial identification needs additional methods. An alternative to author co-citation is article co-citation, used for mapping the fine structure of a research field.

Earlier, Kessler (1963) suggested a technique known as bibliographic coupling. Bibliographi-cal coupling is measuring similarity between papers by the number of references two papers have in common. He showed that a clustering based on this measure yields meaningful groupings of papers as "a number of papers bear a meaningful relation to each other when they have one or more references in common". The major difference between bibliometric coupling and co-citation is that while coupling measures the relationship between source documents, co-citation measures the relations between cited documents. The latter is based on conscious behavior: an author purposefully decides to relate two articles together, whereas the former is used merely as an association with hindsight between two articles.

Another approach to map research topics is word-analysis and co-word analysis, resulting in clusters of words jointly appearing in titles, abstracts, or in full texts. Co-word analysis does not lead to clusters of authors, but should give more of a direct access to the research topics in terms of concepts. Different research topics are expected to use different words, and sets of co-occurring words may indicate the specific research topics within larger specialties. The literature presents many meaningful examples of co-word mappings (Bhattacharya & Basu, 1998), and various directions for improving methods for co-word analysis have been proposed (Noyons & Van Raan, 1998).

The methods discussed here have also been criticized. For example Leydesdorff (1997) argues that “words and co-words cannot map the development of the sciences”, because words are not specific enough, and do have different meanings in different (textual) contexts. Also, co-citation analysis has been criticized for loss of relevant papers, inclusion of non-relevant papers, over-

[1] *Social Studies of Science, Scientometrics, Research Policy, Science & Public Policy, Science, Technology & Human Values*. By the way, the selection of the core journal(s) should be based on some consensus among specialists in the field.

representing theoretical papers, time lag, and subjectivity in threshold setting (King, 1987). Apart from that, the behavioral foundations of the methods are also weak, as they both refer only to a single behavioral dimension: referencing strategies, and choice of words (Rip, 1988).

In some cases, a combination of methods was used. For example, Braam et al (1991a, 1991b) combined co-citation analysis with co-word analysis. In this way, one uses the information about the stock of knowledge in research fields (the references) as well as the current research front represented by the concepts (the words) used by the authors. The co-occurrence of sets of title-words and sets of cited references is expected to identify a research topic in a better way than title-words alone. The way researchers draw on earlier works, and their sharing of a set of exemplars is considered to be reflected in the referencing practices of the specialty members. On the other hand, the shared interest in a set of research problems and concepts is expected to be reflected in the word patterns. The congruence in both mapping approaches is presupposed in many scientometric studies, but also criticized. Braam et al. (1991a; 1991b) showed that the mapping of science by combining and comparing co-citation and co-word analysis is a useful tool to map the subject matter of research specialties in a given period. However, the analysis is based on a sequential application of citation relations and words.

Method

We use papers as the units of analysis within the set of journals that together define the field of information science (IS). One can view IS, like other scientific fields, as an evolving communication network of researchers as nodes, and communications between these nodes as links. Scientific publications in journals allow us to map these communication systems. Knowledge is produced by combining and extending existing papers, and new knowledge is related to previous research by cited references. Consequently, the development of scientific disciplines can be observed in the form of an evolving journal system. Our understanding is that researchers in a field use a common knowledge base, which is reflected in the references they use. Journal-journal citations can be used to map this common identity of researchers in the same field in terms of sets of journals: a scientific field can be defined as a network of journals dealing with similar research questions and methodologies, and referring to a common set of publications.

The idea behind our approach is the following: Researchers simultaneously select words to describe their research subject and refer to specific literature to indicate the tradition in which they do their work. The words acquire their specific meaning within the context of the cited references. In this way we can account for different ways of using words, and for changes over time in the meanings of words. In other words, we expect the cited references to provide more of a context for the words. Referring to a similar set of literature is an indicator for a research topic; word clusters can be interpreted as representations of research topics. The advantage of using word-reference pairs is that it combines two relevant attributes of documents in determining the fine-grained structure of the specialty under study. This combined indicator reflects the subject of the research topic through the title words, and its position within the specialty through the references (Van den Besselaar & Heimeriks 2000).

Operationally, a research topic is defined as a set of papers that are similar in terms of word-reference combinations. So, the similarity measure is the number of word-reference combinations two papers share. If we define the boundary of a topic in a restrictive way, the maps of the research fields become more detailed; if we take a more relaxed criterion, the research topics will be of a broader nature. The case study will illustrate this.

Case

Most mappings of information science use author co-citation analysis. McCain (1990) presented a comprehensive technical review of mapping (information science) authors in intellectual spaces. And White and McCain (1998) again used author co-citation to map information science for the period 1972-1995. We will also use information science as the case study, and we will use our method of word-reference co-occurrences instead of co-citations. At the level of the research field, information science can be defined as a journal system. However, previous analysis showed that this journal system has differentiated into three smaller fields (Van den Besselaar & Heimeriks 2000). Over the years, information science has consisted of three different but linked journal sets, representing somewhat distinct research foci:

- A set of journals around the *Journal of the American Society of Information Science* (JASIS) and the *Journal of Documentation*.
- A set of journals around *Scientometrics* and the *Journal of Information Science*.
- A set of journals on libraries and library research.

As clusters 1 and 2 have become more strongly related over time, we will take them together as the starting point of our mapping exercise.

In this paper we will study the more fine-grained picture of information science, at the level of specific research topics. We focus here only on the cognitive dimension of the cartography of information science: what is the research about and how does this change over the years? However, as the research topics are represented by journal articles, the method may also be used to make a social map of the research field: who is conducting this research? [2]

Summarizing, we distinguish three levels: At the highest level, we have the *research fields* operationalized in terms of journal sets. These fields can be empirically delineated by journal-journal citations. Within research fields one can identify the research front as a changing set of *research topics*. Empirically these topics consist of sets of related publications. In between, we distinguish subfields which may either consists of a single or a small number – more specialized – journals with a shared knowledge base, or consist of a set of related research topics.

In this study, we will focus on the research topics within the field of information science that is defined in terms of the journal set around JASIS in the period 1986 to 2002. The questions addressed in this study are:

- Can we identify topics within the field of information science?
- Do these research topics cluster in subfields?
- How complete is the map?
- Is the reconstruction stable in time?
- Can we trace the development of the field in the period 1986-2002?

Data

As stated earlier, research fields are constituted by sets of journals. The analysis of the topology of research topics is based on the documents published in the set of journals that define the research field. Table 1 summarizes the journals and papers included in the analysis.

Table 1: Data

Research field: Information Science	**Years:**	**Nr of documents***	**Included**
Journals included:			
• Journal of the American Society of Information Science	1986	325	77%
• Journal of Documentation	1992	422	81%
• Information Processing and Management	1996	368	85%
• Annual Review of Information Science and Technology	2000	409	70%
• Proceedings Asist	2002	464	61%
• Journal of Information Science			
• Scientometrics			

* Articles and reviews in the mentioned journals

As customary in this type of analysis, we only used articles and reviews, as they form the set of publications that constitute the field. We omitted the more marginal publications such as letters, notes, book reviews, meeting reports, editorials, and the like. The bibliographical information about these documents has been downloaded in the *dialog* format, for all of the years under study, using the CD-Rom version of the SCI and the SSCI. Special software was used for further processing, to transform the data into a format appropriate for analysis.

[2] However this is not within the scope of this paper.

Frequency lists of title words (excluding a list of stop words) and of cited references were calculated. We then created a database with all possible combinations of title words and cited references, and we linked those to all publications. Of course, only combinations that occur in more than one publication are included, because the relationship between publications is defined as sharing the same word-reference combination. This means that we did not have to set a threshold, as is necessary in co-citation analysis.

Studying large and complex networks, like the relationships between all publications a research field, requires a relational analysis that concentrates on the emerging clusters of papers. We use the so-called cosine algorithm for determining the association (proximity) of two papers. As the number of nodes is often very large, we used network visualizations using BibTechMon©. This is a software tool for analyzing and visualizing large networks in various dimensions, and it is based on a 'mechanical spring model' (Kopcsa 2000). It enables a transformation into a two-dimensional map. These relational maps provide information about the cliques and cohesive subgroups into which a network can be divided. Here we determine the boundaries through visual inspection of the network representations.[3] How to read the maps: The nodes represent documents. The size of the node is proportional to the total number of relationships (word-reference combinations) it has with other nodes. The more links between two nodes, the closer they are in the graph. The thickness of the line between two nodes indicates the number of links. Overlapping nodes share many links.

Results

Figure 1 shows the set of related publications in information science in 1986, produced by the visualization software mentioned above. It is clear from the visualization that the field of information science does consist of quite a few clusters of papers. In 1986, the journals contained 711 papers, of which 325 are relevant for producing the map (articles and reviews). In total, 250 of the 325 documents shared a word-reference combination with at least one other document. Visual inspection allows us to divide the papers in the field of Information science into subfields, which in turn consist of several topics. The core of the 1986 map is identified as the subfield of information retrieval (subfield 1) in the form of a dense cluster of papers on topics like information retrieval (1.3), information searching (1.1) and general information science (1.2). Also several topics within the subfield of scientometrics (subfield 2) could be identified. E.g. the map shows a cluster of documents around the topics of bibliometrics (2.1). The underlying database of documents enabled us to retrieve the titles, authors, journal names, etc. of the documents that cluster in a research topic. The topic of bibliometrics, as expected, is dominated by publications in the journal *Scientometrics*. More distantly related to IR and Information searching are the clusters around Socio-economic topics (Information Society (3.1) and Information Management (3.2)) and Information Systems (4). Several unrelated clusters of papers appear as well; of those, Colon classification (2.2) and patentometrics (2.3) (both of them within the subfield of scientometrics) are the largest.

In 1992, the journals contained 879 documents, of which 422 belong to the categories of articles and reviews. More than 80% share a word-reference combination with another publication. The map of 1992 (which we do not show here) revealed that the subfields information retrieval and general information science again form the core of information science. The subfield of information retrieval is a large densely connected set of papers around topics such as information retrieval from electronic databases and user studies. More isolated topics are present in the periphery of the map (e.g. scientometric distributions). In general, the subfield of scientometrics is better represented than in 1986 with (in addition to scientometric distributions) topics such as scientific collaboration and scientometric indicators.

In 1996, again, information retrieval is at the heart of the discipline. This map consisted of 368 nodes (out of 417 articles and reviews). The role of scientometrics and bibliometrics again increased, and this is once more primarily based on papers published in the journal Scientometrics. In later years, the other journals also increasingly published about bibliometrics. Within the subfield of scientometrics, the most important topics are citation analysis, impact-factors and (performance) indicators.

[3] In the future we plan to apply formal network criteria to determine the boundaries of the clusters.

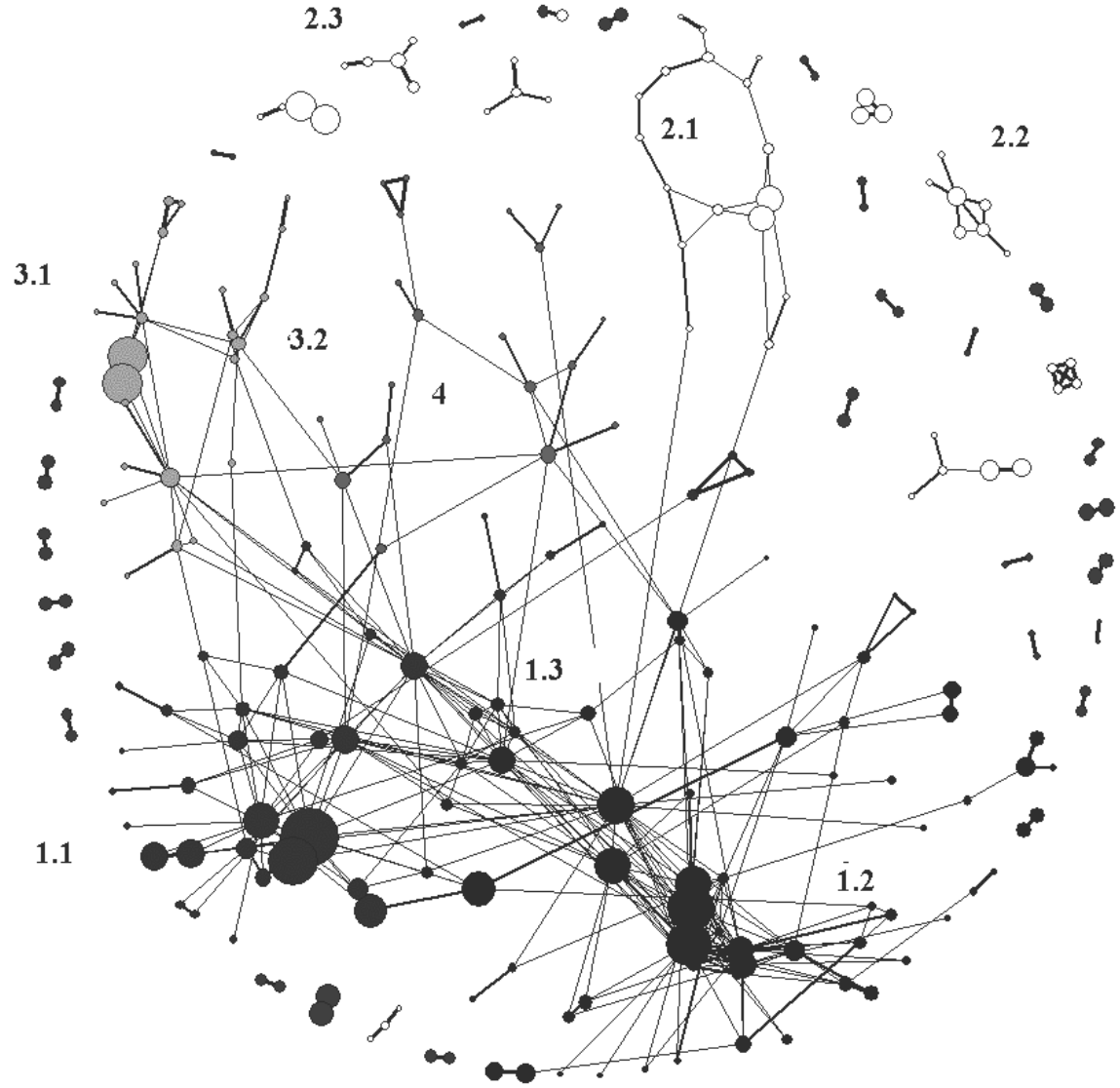

Figure 1. Information Science 1986.

In 2000, the network consisted of 285 nodes (out of a potential set of 409). Apart from the older topics, a few new research topics emerged, such as digital libraries and webometrics. The cluster of publications on Information searching is strongly related to the topics of online searching and digital libraries, which in turn, are strongly related to webometrics. The resulting map is much more densely interlinked than in previous years, possibly because the cluster of papers around the topic of webometrics is positioned in between the traditional subfields of IR and scientometrics. The subfield of scientometrics consisted of several connected clusters of papers around topics in the field such as scientific Web publications and citations analyses. The largest topic within the sub field of scientometrics is related to impact analysis.

In 2002, scientometrics gained even more importance. The field of scientometrics – as represented in the map – consisted of several connected clusters of papers around topics in the field. Again the topic of scientific publications in the web is located close to the field of webometrics. Inspecting the content of the papers, we find that the nature of the clusters can easily be identified. Scientometrics proved to be an important subfield, and on the map it consists of several connected clusters of papers, each representing a research topic within scientometrics. The following topics can be discerned: scientific publications on the web (1.1), impact analysis (1.2) and case studies of scientific fields mostly based on citation analysis (1.3). Also webometrics (2.1) became more

important in 2002, and it is situated near the scientometric studies of web publications (1.1). Between webometrics and information retrieval (3.1) we find web users research (2.2). Other research topics are information behavior (3.2), library services (4) and studies about creativity (7). Finally, we see technical topics such as middleware (8) and data-mining (6), which seem to be developing into important topics in information science.

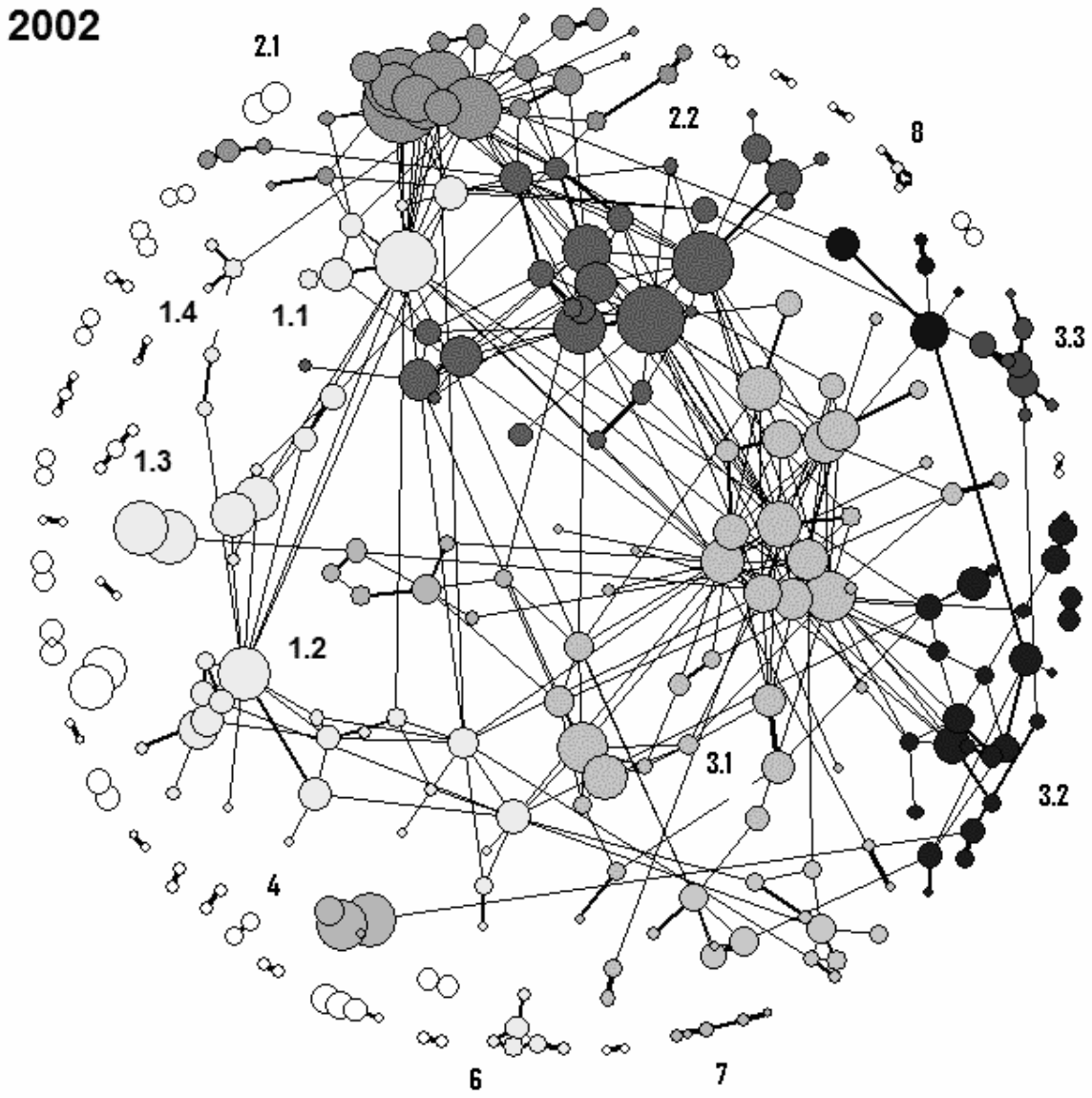

Figure 2. The network of topics of Information Science 2002.

Summarizing, a number of topics is present in all years (such as information retrieval, information seeking behavior, and scientometrics) while others have disappeared (e.g. research on scientometric distributions such as the Bradford law). New topics emerge as a recombination of existing topics and in interaction with new developments such as the increasing importance of ICT. The core of information science consists of information retrieval, although its relative importance has declined over the years. Scientometrics and bibliometrics have become more important. The maps also suggest that new web related research topics (webometrics, and search and retrieval on the web) occupy a position between the older topics of scientometrics and information retrieval. The emergence of web research seems to create a bridge between the two 'poles' in the field. In order to gain more insight in the development of the field, we prepared a table of the most important subfields in each year under

study.[4] The rank indicates the number of papers in each year: on rank 1, we have the research topic with the largest number of papers, and so on.

Table 2 shows the rise of scientometrics within the field of information science. However, within scientometrics the focus seems to shift quickly. New topics emerge from a recombination of previous topics and in reaction to developments such as the increasing role of ICT**.** In the earlier years under study, topics such as Colon classification and Bradford and Lotka distributions were represented as well connected sets of papers. These clusters are no longer present in later years, while new topics such as Web Impact Factor emerged.

Table 2: Most important *subfields* in Information Science

Rank	Subfields 1986	Subfields 1992	Subfields 1996	Subfields 2000	Subfields 2002
1	Information Retrieval	Information Retrieval	Information Retrieval	Scientometrics	Information Retrieval
2	Information Searching	Scholarly communi-cation / bibliometrics	Scientometrics	CMC and digital libraries	Scientometrics
3	Scientometrics	Scientometrics	Information Searching	Information Retrieval	Socio-Economics
4	Socio-Economics	Application of Models	Socio-Economic	Empirical studies IR (also web)	Webometrics
5	Information Systems	Research policy & socio-economic issues		Web information uses	Knowledge Management

In order to gain more insight in the development of the field over time, we also created a map of all articles and reviews in the period under study, 1986-2002 (figure 3). The figure shows 1235 nodes (of a total of 2037 articles and reviews in this period) as well as 4851 relations between them. Also here it is visible that information science has a bi-polar network, based on the clusters information retrieval and bibliometrics. The recent emergence of webometrics is positioned between the two traditional poles. It is also clear that the field has changed over time, as the years (grey tones) are not evenly distributed over the map.

The most densely connected set of papers within the field is the large cluster of publications within the subfield of Information retrieval (IR). Examples of stable topics within IR are library studies and information seeking. The subfield of Scientometrics, in comparison to IR, shows a lower level of overlap between different years. This confirms the previous observation of a rapidly shifting focus within scientometrics. Some topics have lost importance, such as Bradford and Lotka Distributions, while other new topics have emerged (university website hyperlinks). The related topics of webometrics and web search and retrieval occupy positions between the more traditional scientometrics and information retrieval subfields.

Conclusions

Co-occurrences of (title) words and co-occurrences of cited references have been used for mapping the content of research fields. Both methods have produced interesting results, but also suffer from various problems. We introduced an alternative a method based on a combined use of title words and cited references, and this technique seems more promising. The advantage of the method introduced here is that it combines both attributes of documents in determining the fine-grained structure of the specialty under study. The combined indicator reflects the subject of the research topic (through the title words) and its position within the specialty (through the set of co-occurring references). The resulting maps give an accurate representation of the research topics (dense areas of related papers) and also show the relationship between the different topics within the field of information science. The representation of

[4] Note that 'important' refers to the number of papers that share a set of title words and cited references. The nature of the clusters of papers changes from year to year.

the field is strikingly stable over time; a bipolar structure characterizes information science over the years, and it is dominated by information retrieval and bibliometrics.[5]

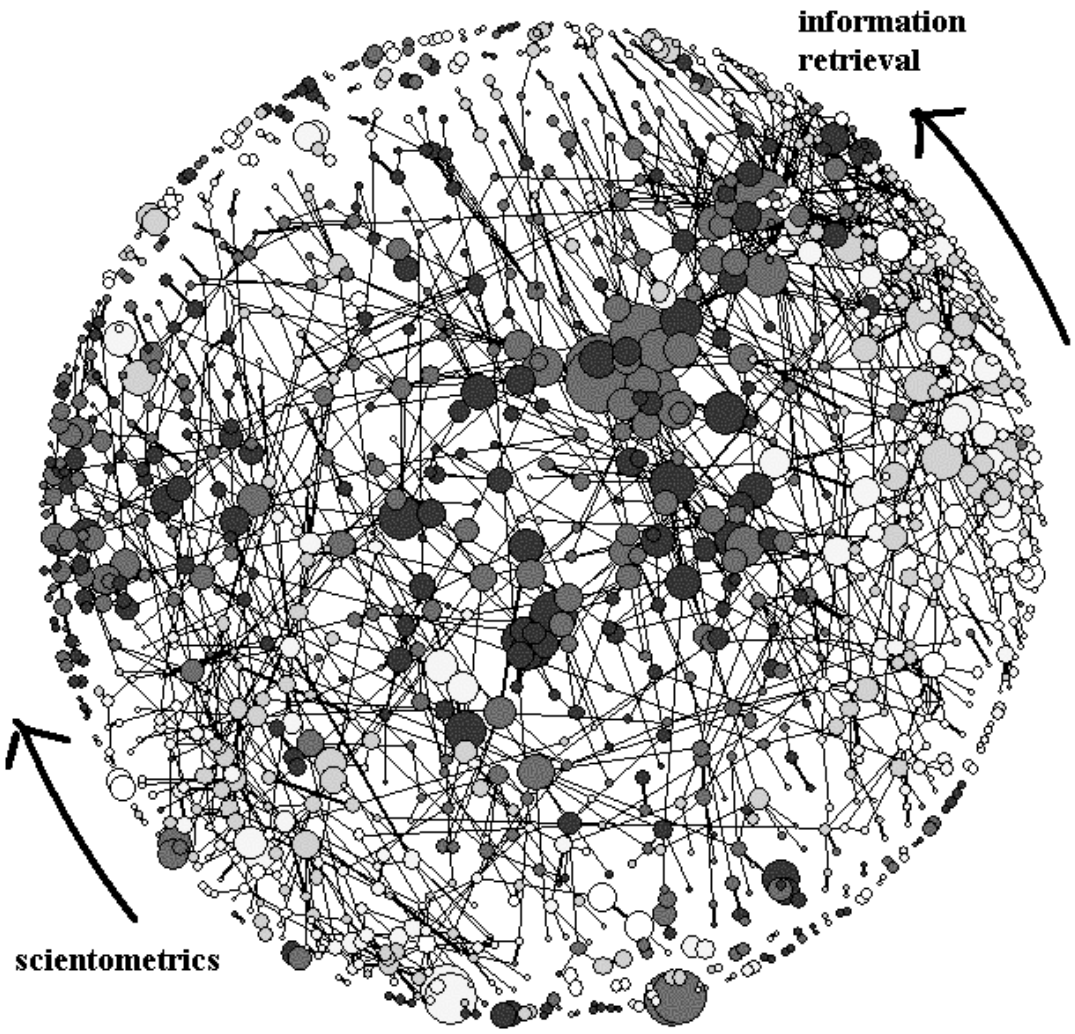

Figure 3. The changing network of topics in Information Science

(1986: white; 1992: light grey; 1996: medium grey; 2000: dark grey; 2002: black).

Various problems of older methods disappear (or diminish). Firstly, threshold setting is not a problem, as all papers that share at least a one word-reference combination with at least one other paper are included. Secondly, a higher percentage of source documents is included in the analysis than co-word and co-citation mappings normally have. Thirdly, the proposed method is therefore also less susceptible to an overrepresentation of theoretical papers.

The method has at least two theoretical advantages. First, it does not depend on a single behavioral mechanism but on two: citing strategies *and* word choice. Our hypothesis is that it therefore reflects to a greater extent the cognitive development of a research field, and not the social relations (as may be the case with citation relations only). Secondly, it uses word patterns, but only in the context of the cited references. Therefore, the method does account for variety of use and changes in the meaning of words.

[5] Elsewhere we present maps of Artificial Intelligence, an example of a research field showing more turbulence (Heimeriks et al, submitted).

References

Bhattacharya, S., & Basu, P. K. (1998). Mapping a research area at the micro level using co-word analysis. *Scientometrics*, 43, 359-372.

Braam, R. R., H.F. Moed, & Van Raan, A. F. J. (1991). Mapping of science by combined co-citation and word analysis, I: Structural Aspects. *Journal of the American Society for Information Science*, 42, 233-251. (a)

Braam, R. R., Moed, H. F., & Van Raan, A. F. J. (1991). Mapping of science by combined co-citation and word analysis, II: Dynamical Aspects. *Journal of the American Society for Information Science*, 42, 252-266. (b)

Heimeriks, G. J., Van Someren, M., & Van den Besselaar, P. (submitted). *Structure and development of AI: a bibliometric analysis.*

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, 14, 10-25.

King, J. (1987). A review of bibliometric and other science indicators and their role in research evaluation. *Journal of Information Science*, 13, 261-276.

Kopcsa A., S. E. (2000). Science and technology mapping: A new iteration model for representing multidimensional relationships. *Journal of the American Society for Information Science*, 49, 7-17.

Leydesdorff, L. (1997). Why words and co-words cannot measure the development of the sciences. *Journal of the American Society of Information Science*, 48, 418-27.

Marshakova, I. V. (1973). A system of document connection based on references. *Scientific and Technical Information Serial of VINITI*, 6(2), 3-8.

McCain, K. W. (1990). Mapping authors in intellectual space: a technical review. *Journal of the American Society for Information Science*, 41, 433-443.

Noyons, E. C. M., & Van Raan, A. F. J. (1998). Monitoring scientific developments from a dynamic perspective; Self-organized structuring to map neural network research. *Journal of the American Society for Information Science*, 49, 68-81.

Rip, A. (1988). Mapping of Science, Possibilities and limitations. In A. F. J. Van Raan (Ed.), *Handbook of quantitative studies of science and technology*. Amsterdam: Elsevier Science. 253-273

Small, H. (1973). Co-citation in Scientific literature: a new measure of the relationship between publications. *Journal of the American Society for Information Science*, 24, 265-269.

Van den Besselaar, P. (2000). Communication between science and technology studies journals: A case study in differentiation and integration in scientific fields. *Scientometrics,* 47, 169-93.

Van den Besselaar, P. (2001). The cognitive and the social structure of Science and Technology Studies. *Scientometrics,* 51, 441-460.

Van den Besselaar, P & Heimeriks, G.J. (2000). *Codification and self-organization in the European STI system*. (Deliverable 2.6 – SOEIS project). University of Amsterdam.

Van den Besselaar, P., & Heimeriks, G. J. (submitted). *Disciplinary and Interdisciplinary Identities*.

Van den Besselaar, P., & Leydesdorff, L. (1996). Mapping change in scientific specialties; a scientometric case study of the development of artificial intelligence, *Journal of the American Society of Information Science*, 47, 415-436.

White, H. D., & McCain, K. W. (1998). Visualizing a discipline: an author co-citation analysis of information science, 1972-1995. *Journal of the American Society for Information Science*, 49, 327-356.