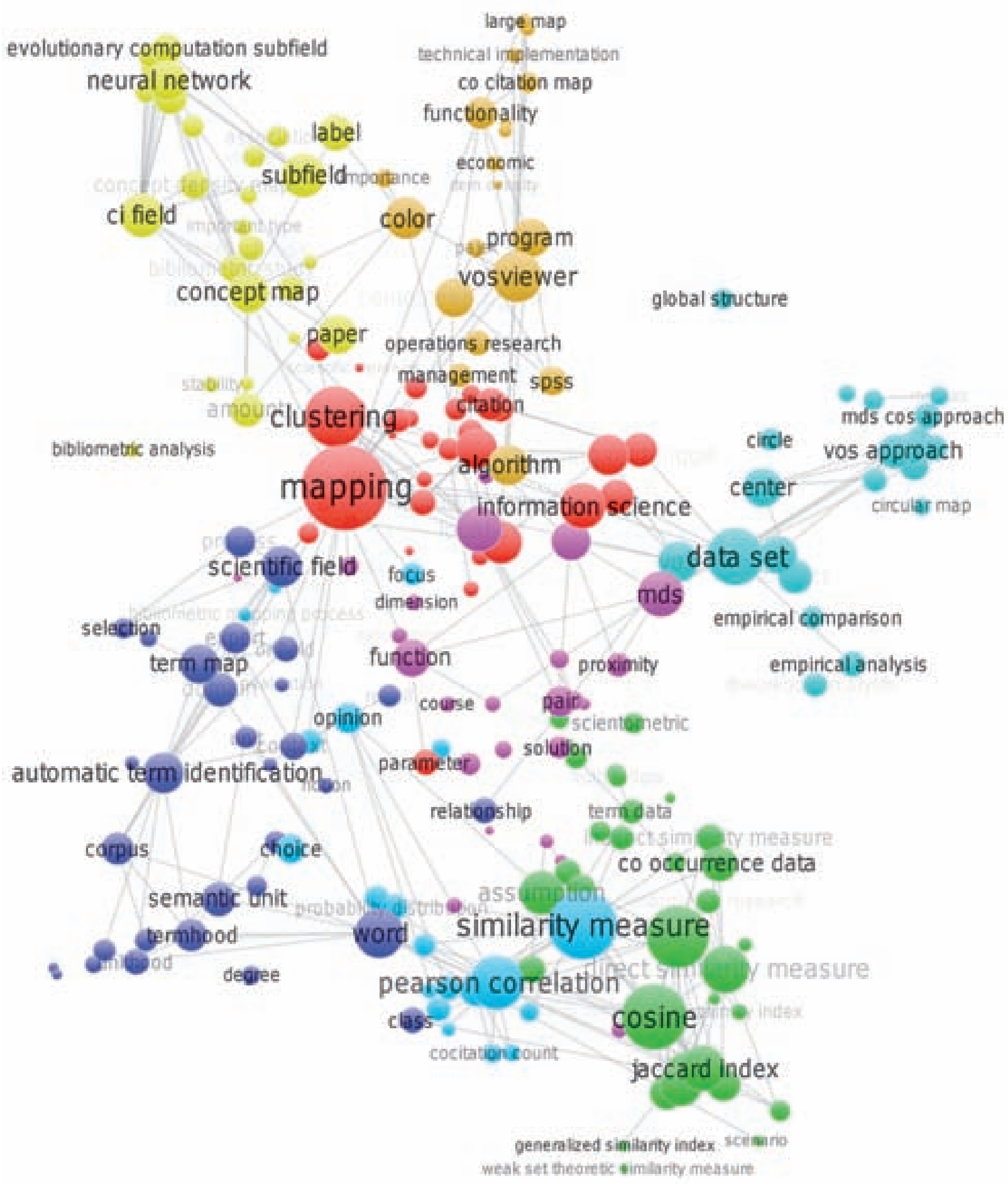
NEES JAN VAN ECK

Methodological Advances in Bibliometric Mapping of Science



Methodological Advances in Bibliometric Mapping of Science

Methodological Advances in Bibliometric Mapping of Science

Methodologische ontwikkelingen in het
bibliometrisch karteren van de wetenschap

Thesis

to obtain the degree of Doctor from the
Erasmus University Rotterdam
by command of the
rector magnificus

Prof.dr. H.G. Schmidt

and in accordance with the decision of the Doctorate Board.

The public defence shall be held on

Thursday, 13 October 2011 at 13.30 hrs

by

Nees Jan van Eck
place of birth, Utrecht.

ERASMUS UNIVERSITEIT ROTTERDAM

Doctoral Committee

Promotor:	Prof.dr.ir. R. Dekker
Copromotor:	Dr.ir. J. van den Berg
Other members:	Prof.dr. P.J.F. Groenen
	Prof.dr.ir. U. Kaymak
	Prof.dr. A.F.J. van Raan

Erasmus Research Institute of Management - ERIM
The joint research institute of the Rotterdam School of Management (RSM)
and the Erasmus School of Economics (ESE) at the Erasmus University Rotterdam
Internet: http://www.erim.eur.nl

ERIM Electronic Series Portal: http://hdl.handle.net/1765/1

ERIM PhD Series Research in Management, 247
ERIM reference number: EPS-2011-247-LIS
ISBN 978-90-5892-291-5

Design: B&T Ontwerp en advies www.b-en-t.nl

This publication (cover and interior) is printed by haveka.nl on recycled paper, Revive®.
The ink used is produced from renewable resources and alcohol free fountain solution.
Certifications for the paper and the printing production process: Recycle, EU Flower, FSC, ISO14001.
More info: http://www.haveka.nl/greening

The illustration on the cover was produced using the VOSviewer software for bibliometric mapping (www.vosviewer.com). The illustration shows the relations between the most important terms in this thesis.

FSC

SIKS Dissertation Series No. 2011-32
The research reported in this thesis has been carried out in cooperation with SIKS, the Dutch Research School for Information and Knowledge Systems.

Acknowledgements

This thesis presents the research that I have been working on during my time as a PhD student. Completing this thesis would not have been possible without the help, support, and encouragements of a number of people. I would like to take this opportunity to express my gratitude to them.

First of all, I want to thank my promotor Rommert Dekker and my copromotor Jan van den Berg. Before we got to know each other, I had never really thought about doing a PhD. Their trust in my capabilities made me seriously consider it. I am grateful to them for giving me the opportunity to start working as a PhD student, and I want to thank them for their enthusiasm and their fresh ideas. They have been an important source of motivation and inspiration to me. I also want to thank Jan for his ongoing commitment after his move to Delft University of Technology.

I am especially thankful to my colleague and dear friend Ludo Waltman. He has been involved in all studies presented in this thesis. I always very much enjoy working together with him. Besides our excellent working relation, I also want to thank Ludo for his close friendship. We have had so much fun throughout the years, and we share many incredible memories! I will never forget our conference trip to Hawaii and our cycling holiday to the Mont Ventoux.

I also would like to thank my fellow PhD students. Without our joint coffee breaks, the PhD trips, and the occasional drinks, being a PhD student would have been much less enjoyable. A special thanks goes to Martijn Kagie for his friendship. We have had a lot of fun during all the breaks and evenings drinking cola and beer together.

Finally, I want to thank my family. I am deeply grateful to my parents for their endless love and support. Without them, I would not have been where I am today. I also want to thank my sister Marleen for her interest and encouragements. Last but certainly not least, I want to thank my wife Emma for her unconditional love, support, and patience at all times.

Nees Jan van Eck
August 2011

Table of Contents

Chapter 1

Introduction

1.1 Bibliometric Mapping of Science

Given the sheer volume of scientific literature currently available and the rapid growth of this literature, it is often difficult to have a comprehensive, up-to-date, and unbiased overview of all relevant literature on the topics that one is interested in. More and more attention is therefore being paid to computerized methods and tools that help to identify and structure the scientific literature relevant to one's interests. Such methods and tools typically provide some kind of visual representation of the identified literature. These visual representations, often referred to as maps, are the topic of study of this thesis.

More specifically, we refer to the topic of this thesis as bibliometric mapping of science. Bibliometrics is the scientific field that quantitatively studies all kinds of bibliographic data, such as the titles, keywords, authors, and cited references of articles and books.[1] Accordingly, bibliometric mapping of science is about quantitative methods for visually representing scientific literature based on bibliographic data. Bibliometric mapping results in bibliometric maps. As will be discussed later on, there are many different types of bibliometric maps, each of them providing somewhat different information and serving a somewhat different purpose. However, the general aim of a bibliometric map is to provide an overview of the structure of the scientific literature in a certain domain or on a certain topic. A bibliometric map can for example be used to identify the main

[1] Bibliometrics is closely related to scientometrics and informetrics. For a discussion of these three terms, we refer to Hood and Wilson (2001).

research areas within a scientific field, to get insight into the size of the different areas, and to see how the areas relate to each other. Bibliometric maps are especially useful when one has to deal with a relatively large body of literature and when one's interest is not only in the individual elements (e.g., the individual documents, authors, or keywords) that can be identified in this body of literature but also in the way in which the various elements are interrelated.

Bibliometric maps can be used in a number of different contexts. Researchers can use bibliometric maps to get an overview of the field in which they are active or to perform a high-level exploration of the literature on a certain topic. In the context of science policy and research management, bibliometric maps can be used to support decision making by governments, funding agencies, and universities (e.g., Franklin & Johnston, 1988; Healey, Rothman, & Hoch, 1986; Noyons, 2001, 2004). Bibliometric maps can also be of value to journal editors, scientific publishers, and librarians, who may for example use these maps to explore how a journal is positioned relative to other related journals. Other possible applications of bibliometric maps are in science teaching (e.g., Börner et al., 2009; Klavans & Boyack, 2009) and in the history, philosophy, and sociology of science (e.g., Small, 2003).

To construct a bibliometric map, one needs to have access to a bibliographic database of the domain of interest. Such a database contains bibliographic records of a large number of documents. These records indicate for example the title, the abstract, and the authors of a document and the source and the year in which a document was published. The cited references of a document are sometimes indicated as well. Currently, two broad multidisciplinary bibliographic databases are available, namely Web of Science and Scopus, which are provided by Thomson Reuters and Elsevier, respectively. In addition, various bibliographic databases are available for specific disciplines. Examples include Chemical Abstracts for chemistry, Inspec for engineering, computer science, and physics, and MEDLINE for medical and life sciences. A disadvantage of some of these disciplinary databases is that they do not contain the cited references of a document. This can be a serious limitation for bibliometric mapping purposes. In this thesis, we mainly use the Web of Science database. Only in Chapter 8 we use different databases, namely Scopus and IEEE Xplore.

Below, we will first give an overview of different types of bibliometric maps (Section 1.2). We will also discuss the value or the utility of bibliometric maps (Section 1.3). We will then focus on the main contribution of this thesis, which is the introduction of a new methodology for bibliometric mapping (Section 1.4). Finally, we will give an outline of the thesis (Section 1.5).

1.2 Types of Bibliometric Maps

There are infinitely many ways in which scientific literature can be visually represented. It is therefore difficult to give a comprehensive and systematic overview of the various types of bibliometric maps that have been proposed in the literature. Accordingly, the overview presented in this section focuses on the most important types of maps that have been studied in the field of bibliometrics. Bibliometric maps are also sometimes studied in other fields, such as artificial intelligence, information retrieval, and information visualization, but the literature from these fields will not be considered here. Also, bibliometric maps that focus specifically on showing developments over time (e.g., Garfield, 2009; Garfield, Pudovkin, & Istomin, 2003) will not be considered. For overviews of the bibliometric mapping literature from various different perspectives, we refer to Börner (2010), Börner, Chen, and Boyack (2003), C. Chen (2003a, 2006b), Morris and Van der Veer Martens (2008), and White and McCain (1997). An overview from a historical perspective is provided by De Bellis (2009). Furthermore, two journals published a special issue on bibliometric mapping (C. Chen, 2003b; Shiffrin & Börner, 2004).

Bibliometric maps can be categorized in many different ways. In this section, two categorizations of bibliometric maps are discussed, namely a categorization based on the unit of analysis and the measure of relatedness (Subsection 1.2.1) and a categorization based on the type of visualization (Subsection 1.2.2). To illustrate the discussion, several examples of the different types of bibliometric maps will be shown.

1.2.1 Unit of Analysis and Measure of Relatedness

The unit of analysis is the type of object shown in a bibliometric map. The most commonly used units of analysis are documents, authors, journals, and words or terms. However, many other units of analysis can be used as well (e.g., countries, research institutes, and scientific fields). The mapping of documents (and clusters of documents) was pioneered by Henry Small at the Institute for Scientific Information since the 1970s (e.g., Griffith, Small, Stonehill, & Dey, 1974; Small & Griffith, 1974; Small & Sweeney, 1985; Small, Sweeney, & Greenlee, 1985, see Figures 1.1 and 1.2). The mapping of authors and, to a lesser extent, journals was pioneered by researchers at Drexel University since the 1980s (e.g., McCain, 1990, 1991; White & Griffith, 1981; White & McCain, 1998, see Figure 1.3). Early work into the mapping of words was done by a group of primarily French researchers (e.g., Callon, Courtial, Turner, & Bauin, 1983; Callon, Law, & Rip, 1986; Rip & Courtial, 1984, see Figure 1.4) and somewhat later also by researchers at the Centre for Science and Technology Studies of Leiden University (e.g., Peters & Van Raan, 1993b; Tijssen & Van Raan, 1989, see Figure 1.5).

To construct a bibliometric map, one needs to know not only the objects to be shown in the map, but also the relatedness of the objects. In other words, one needs to know for each pair of objects how strongly the objects are related to each other. This means that one needs to have a measure of the relatedness of objects. There are many different ways in which the relatedness of objects can be measured. We will discuss the most commonly used approaches. It should be noted that for different units of analysis the measures of relatedness that can be used are also somewhat different.

When dealing with documents, authors, or journals, the relatedness of objects is often measured using citation relations. There are three basic approaches. These approaches use, respectively, direct citation relations, co-citation relations, and bibliographic coupling relations. In the direct citation approach, the relatedness of two objects is measured by the number of citations going from one object to the other. Unlike other approaches, the direct citation approach yields an asymmetric measure of relatedness. Bibliometric mapping typically requires a symmetric measure of relatedness, and this may explain why the direct citation approach does not seem very popular. A much more popular approach is the co-citation approach (Small, 1973). In this approach, the relatedness of two objects is measured by the number of times the objects are cited together. For example, if there are three documents that cite both document A and document B, then document A and document B have three co-citations. The third approach is the bibliographic coupling approach (Kessler, 1963a, 1963b). This approach works in exactly the opposite way as the co-citation approach. In the bibliographic coupling approach, the relatedness of two objects is measured by the number of references the objects have

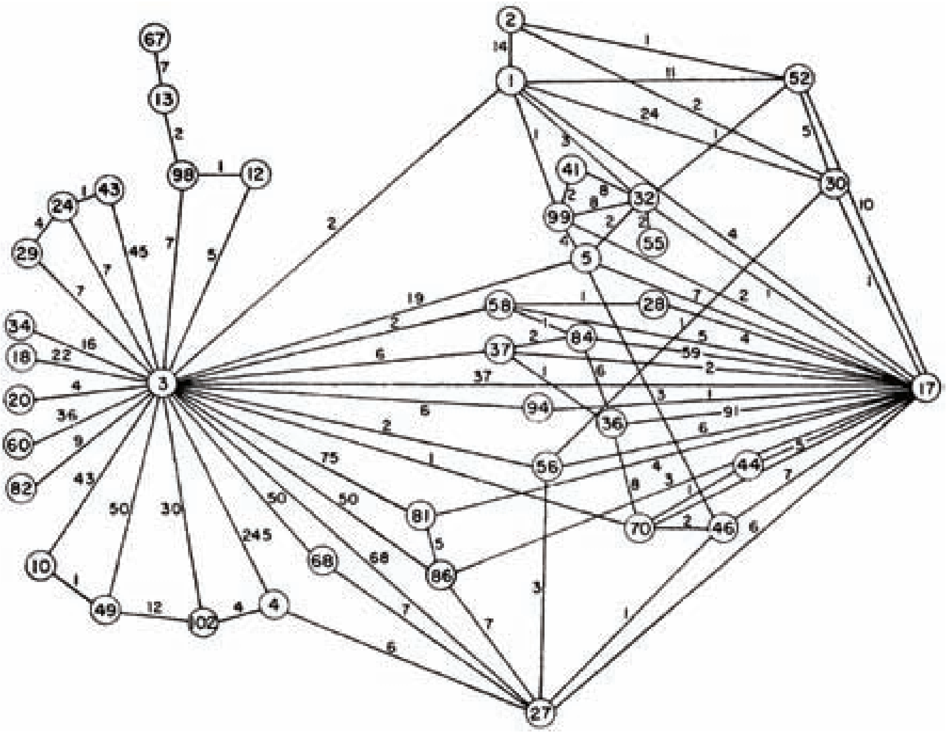


Figure 1.1: One of the first document cluster maps. The map shows 41 clusters of documents and their co-citation relations in the Science Citation Index in 1972. Cluster 3 is by far the largest cluster and contains publications in biomedicine. Three other relatively large clusters are clusters 1, 2, and 17, which contain publications in, respectively, nuclear structure physics, particle physics, and chemistry. For the contents of the remaining clusters, see Griffith et al. (1974, Table 1). Reprinted from Griffith et al. (1974, Figure 1) with kind permission of Sage Publications.

in common. For example, if there are three documents that are cited both by document A and by document B, then document A and document B have a bibliographic coupling strength of three. An overview of studies in which bibliographic coupling is used is provided by Jarneving (2007).

Another way to measure the relatedness of documents, authors, or journals is to use relations based on words or terms rather than relations based on citations. For example,

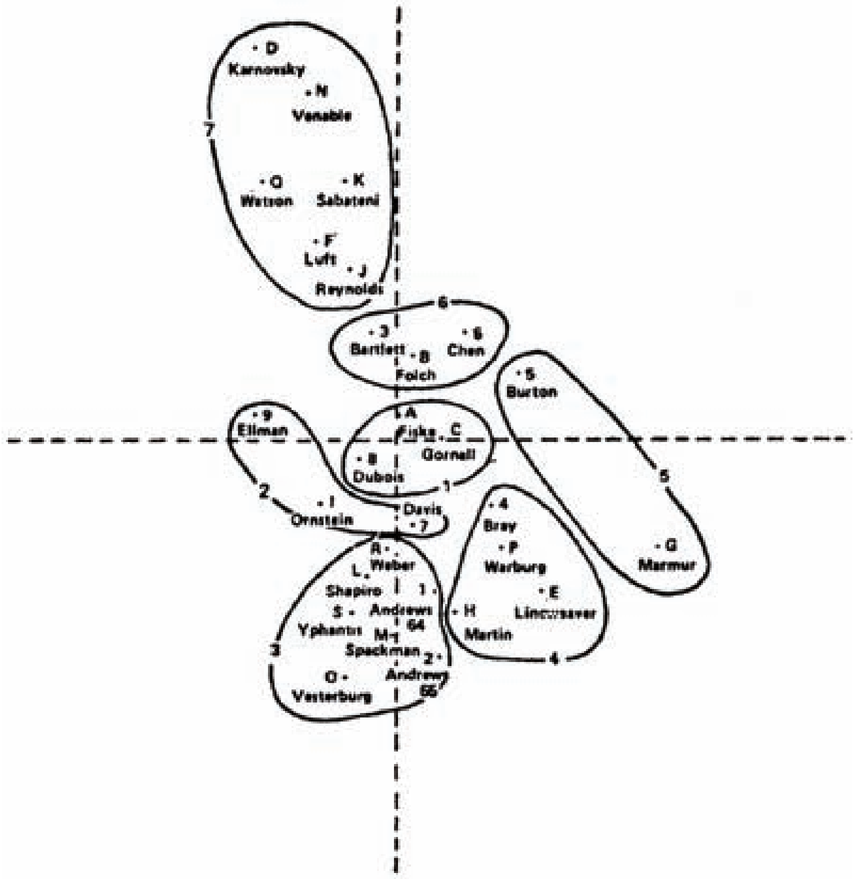


Figure 1.2: One of the first document maps. The map shows 28 biomedical methods publications and their co-citation relations in the Science Citation Index in 1972. The map was constructed using multidimensional scaling. Reprinted from Griffith et al. (1974, Figure 2 upper part) with kind permission of Sage Publications.

the relatedness of two documents can be measured by the number of words that occur in both documents. In many cases, the full text of a document is not available, and only words in the title and sometimes also in the abstract of a document are considered. An alternative is to use the keywords assigned to a document. In general, the use of word relations is more difficult than the use of citation relations. This is because not

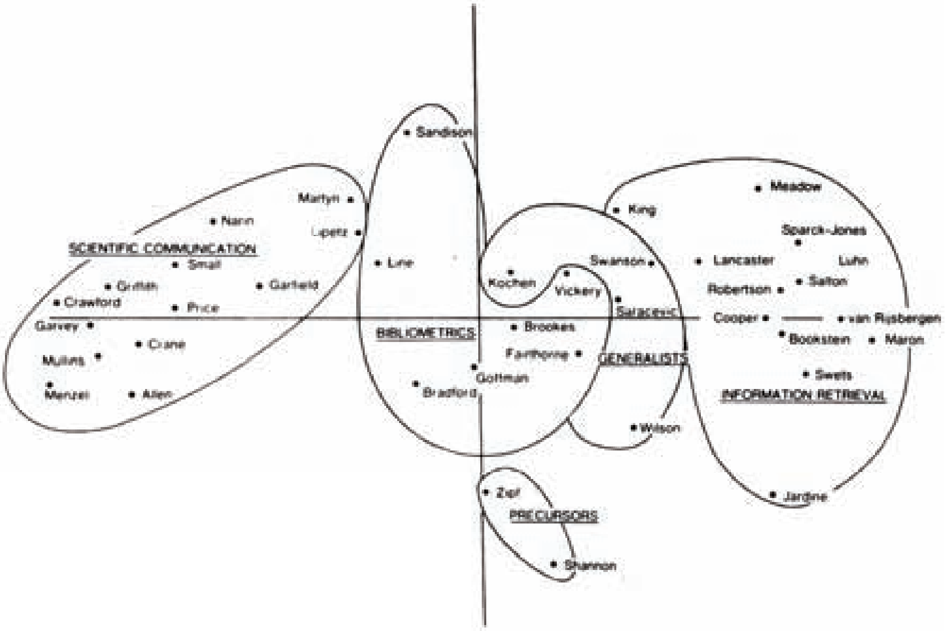


Figure 1.3: One of the first author maps. The map shows 39 information science authors and their co-citation relations in the Social Sciences Citation Index in the period 1972–1979. The map was constructed using multidimensional scaling. Reprinted from White and Griffith (1981, Figure 1) with kind permission of John Wiley and Sons.

all words are equally informative. Uninteresting words therefore need to be filtered out. Also, different words may need to be given different weights. In the bibliometric mapping literature, measuring the relatedness of documents, authors, or journals using word relations does not seem to be a frequently used approach. In recent work, however, some attention is paid to the combined use of citation relations and word relations (e.g., Janssens, Glänzel, & De Moor, 2008).

When words or terms are the unit of analysis, relatedness is typically measured using co-occurrence relations. If two words both occur in the same document, the words are said to co-occur in the document. The relatedness of two words can be measured by the number of co-occurrences of the words, that is, the number of documents in which the words co-occur.

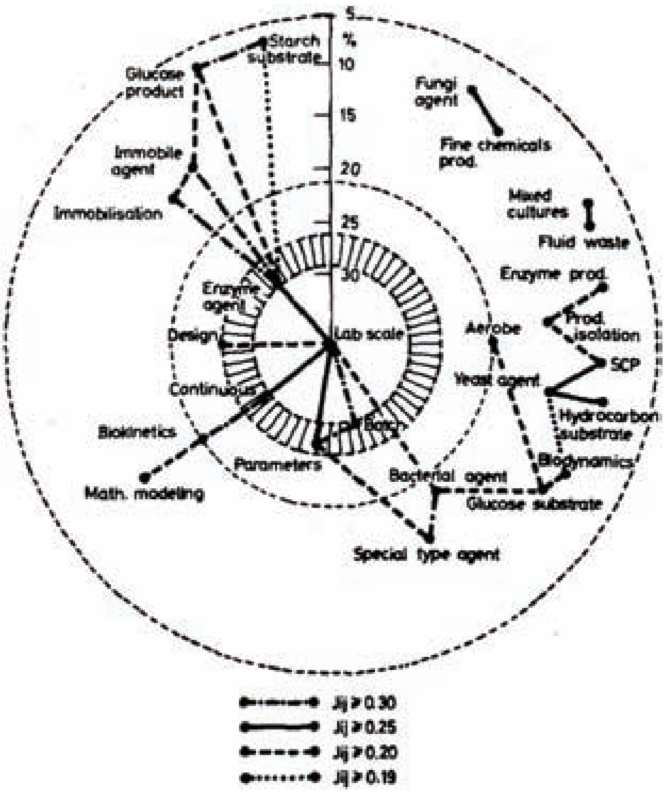


Figure 1.4: One of the first word maps. The map shows 26 biotechnology keywords and their co-occurrence relations in the journal Biotechnology and Bioengineering in the period 1970–1974. Reprinted from Rip and Courtial (1984, Figure 1) with kind permission of Springer Science and Business Media.

Another approach to measuring the relatedness of objects is to use co-authorship relations. This approach can be used when dealing with authors, research institutes, or countries. For example, the relatedness of two authors can be measured by the number of documents they have co-authored.

We have now discussed the most commonly used approaches for measuring the relatedness of objects. For each unit of analysis, multiple measures of relatedness are available. For example, the relatedness of authors can be measured using co-citation relations, bibliographic coupling relations, word relations, or co-authorship relations. It

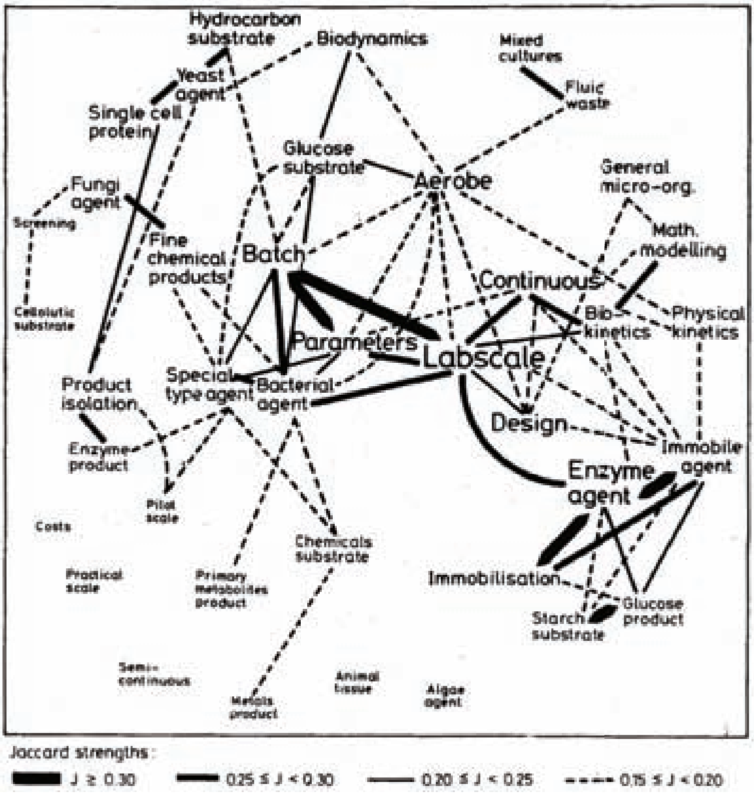


Figure 1.5: A ‘second generation’ word map. The map is based on the same data as the map in Figure 1.4. The map was constructed using multidimensional scaling. Reprinted from Tijssen and Van Raan (1989, Figure 3) with kind permission of Springer Science and Business Media.

is clear that each of these measures captures a somewhat different aspect of the way in which authors relate to each other. It can be useful to add together multiple measures of relatedness (e.g., Small, 1997). In this way, more data is used and a more accurate overall measure of relatedness may be obtained.

Finally, we want to make two terminological remarks. First, in the bibliometric mapping literature, the term ‘co-occurrence’ is sometimes used to indicate not only the co-occurrence of two words in a document but more generally any type of relation as mentioned above (e.g., co-citation, bibliographic coupling, or co-authorship). In this thesis, the term ‘co-occurrence’ is often used in this broader sense. Second, co-occurrence relations define a network, for example a co-citation network of documents, a bibliographic coupling network of journals, or a co-authorship network of authors. In this thesis, such networks are sometimes referred to as bibliometric networks.

1.2.2 Visualization

A bibliometric map is a visual representation of a bibliometric network. Hence, a bibliometric map visualizes a set of objects and the relations among the objects. Many different types of visualizations can be used. We will now discuss some important types of visualizations.

A fundamental distinction is between distance-based visualizations and graph-based visualizations. In distance-based visualizations, the distance between two objects reflects the relatedness of the objects. The smaller the distance between two objects, the stronger the relation between the objects. In graph-based visualizations, on the other hand, the distance between two objects need not reflect the relatedness of the objects. Instead, relations between objects are typically indicated by drawing lines between objects. In the bibliometric mapping literature, both distance-based and graph-based visualizations are used. In early research, distance-based visualizations are predominant, for example in the work of Henry Small and colleagues on the mapping of documents (e.g., Griffith et al., 1974; Small et al., 1985, see Figure 1.2) and in the work done at Drexel University on the mapping of authors (e.g., McCain, 1990; White & Griffith, 1981; White & McCain, 1998, see Figure 1.3). The most popular technique for distance-based visualization is multidimensional scaling (e.g., Borg & Groenen, 2005; T. F. Cox & Cox, 2001). In more recent research, both distance-based and graph-based visualizations can be found. Graph-based visualizations are typically produced using the graph-drawing techniques of Fruchterman and Reingold (1991) or Kamada and Kawai (1989). These techniques are available in computer programs for social network analysis, such as Pajek (De Nooy, Mrvar, & Batagelj, 2005). Graph-drawing techniques are sometimes used in combination with the pathfinder network technique for graph pruning (Schvaneveldt, 1990; Schvaneveldt, Dearholt, & Durso, 1988). An example of a graph-based visualization is shown in Figure 1.6. This example is taken from White

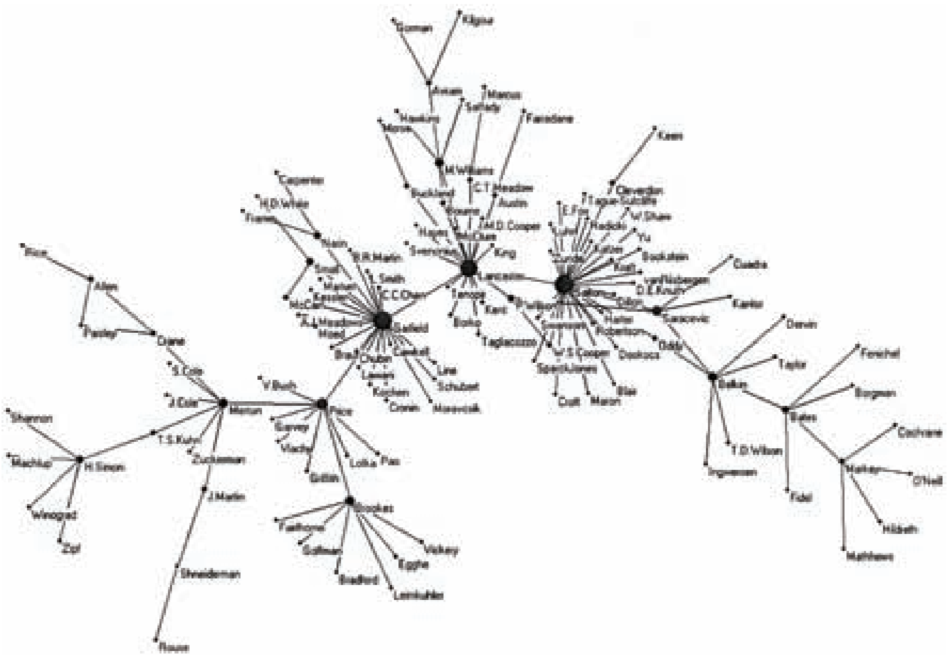


Figure 1.6: A graph-based author map. The map shows 121 information science authors and their co-citation relations in the period 1972–1995. The map was constructed using the pathfinder network technique for graph pruning and the technique of Kamada and Kawai (1989) for graph drawing. Reprinted from White (2003b, Figure 1) with kind permission of John Wiley and Sons.

(2003b). For other examples of graph-based visualizations, we refer to Bollen et al. (2009), de Moya-Anegón et al. (2007), and Leydesdorff and Rafols (2009).

Another distinction that can be made is between visualizing all individual objects of interest and visualizing clusters of objects. Visualization is mostly done at the level of individual objects. In some cases, however, a more useful picture emerges when visualization is done at the level of clusters of objects. Examples of visualizations at the cluster level are shown in Figures 1.1 and 1.7. These examples are taken from Griffith et al. (1974) and Noyons and Van Raan (1998). Of course, when visualization is done at the level of individual objects, it is still possible to indicate a clustering of the objects. This can be accomplished by marking off areas in a map that correspond with clusters

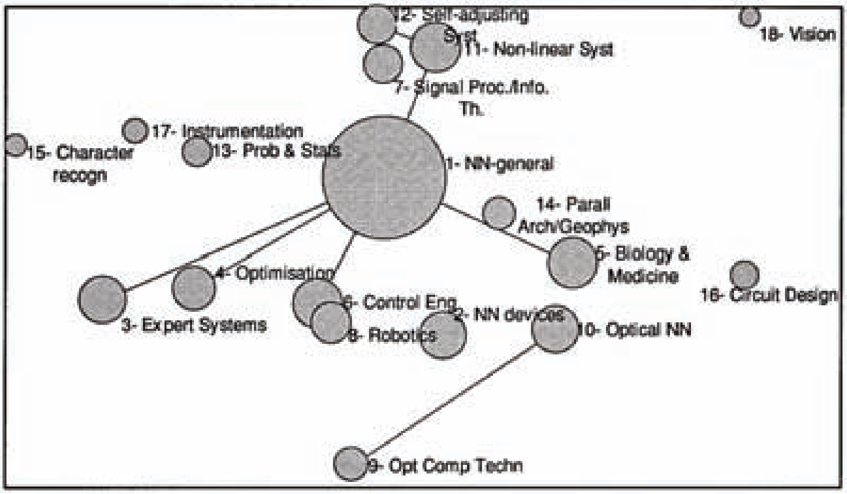


Figure 1.7: A document cluster map. The map shows 18 clusters of neural network publications in the period 1992–1993. A publication can belong to multiple clusters. The relatedness of two clusters is measured by the number of shared publications. The map was constructed using multidimensional scaling. Reprinted from Noyons and Van Raan (1998, Figure 2b) with kind permission of John Wiley and Sons.

(e.g., Griffith et al., 1974; White & Griffith, 1981, see Figures 1.2 and 1.3) or by coloring objects based on the cluster to which they belong (e.g., Leydesdorff & Rafols, 2009).

A third distinction is between interactive and non-interactive visualizations. A non-interactive visualization just provides a static picture of a bibliometric network. An interactive visualization, on the other hand, offers additional possibilities, such as the possibility to zoom in on areas of interest or the possibility to request additional information on objects and their relations. It is clear that interactive visualizations need to be presented on a computer, while non-interactive visualizations can also be presented on paper. Interactive visualizations usually have the advantage that they provide more information than non-interactive visualizations. However, interactive visualizations also require more user involvement, which in some cases may be a disadvantage. For examples of interactive visualizations, we refer to Boyack, Wylie, and Davidson (2002), Buter and Noyons (2001), C. Chen (2006a), and Small (1999).

1.3 The Value of Bibliometric Maps

In this section, we discuss the value or the utility of bibliometric maps. Our focus is not on specific applications of bibliometric maps, but rather on the general use of bibliometric maps to study a certain domain of interest.

Bibliometric mapping has various limitations, and due to these limitations bibliometric maps always need to be interpreted in a careful manner. There are two main types of limitations that should be kept in mind when interpreting a bibliometric map:

- *Limitations imposed by the data.* The availability of data will always be limited, and the data that is available will always contain a certain amount of noise. Noise in the data may for example arise from all kinds of relatively arbitrary decisions researchers make when choosing the references they cite or the terminology they use.

- *Limitations imposed by the map.* A map provides a simplified representation of reality, and simplification generally implies some loss of information. For example, in the case of distance-based bibliometric maps, there is a loss of information because objects are put in a Euclidean space and because this space has only a small number of dimensions (typically two).

Due to the above limitations, a bibliometric map should never be assumed to provide a perfectly valid representation of the domain of interest.

Given the various limitations of bibliometric mapping, one may wonder what the value of a bibliometric map is. In our view, there are at least three ways in which a bibliometric map can be of value to an analyst who interprets the map:

- A bibliometric map may confirm some of the ideas an analyst has. In this case, the confidence of the analyst in his ideas will increase. However, given the limitations of bibliometric mapping, a map in itself can never make an analyst fully confident of his ideas.

- A bibliometric map may contradict some of the ideas an analyst has. In this case, the confidence of the analyst in his ideas will decrease. Of course, the analyst should not lose all his confidence. Some of the suggestions made by the map may not be valid, and therefore the ideas of the analyst could still be correct.

- A bibliometric map may suggest new insights to an analyst. In this case, the map provides the analyst with new ideas. Given the limitations of bibliometric mapping, the analyst should have only a limited amount of confidence in these ideas. It could be that some of the ideas are not correct.

This list makes clear that in order to see the value of a bibliometric map, it should be recognized that the knowledge someone has of a certain domain will typically be incomplete and uncertain and in some cases even partially incorrect. Although a bibliometric map will not provide a perfectly valid representation of the domain of interest, such a map can be of significant value by extending the (uncertain) knowledge someone has, by decreasing the amount of uncertainty in someone's knowledge, and by uncovering elements in someone's knowledge that may not be correct.

In summary, a bibliometric map makes all kinds of suggestions concerning the structure and the properties of a certain domain. Not all suggestions made by a map will be perfectly valid. An analyst should therefore treat a map as just one piece of evidence, in addition to other pieces of evidence, such as the analyst's own knowledge, the opinions of experts, and the results of possible other quantitative analyses. Each piece of evidence should have its own weight. The weight that is given to a bibliometric map may depend strongly on the amount of data on which the map is based. The larger the amount of data, the more confidence one may have in the suggestions made by the map and, consequently, the more weight one may give to the map. Different pieces of evidence will sometimes contradict each other. In that case, an analyst may decide to collect additional evidence, for example by consulting additional experts or by performing additional quantitative analyses. One way to perform an additional quantitative analysis is by producing an additional bibliometric map. Compared with the original map, the new map may be based on a different data source or may use a different unit of analysis, a different measure of relatedness, or a different type of visualization. As the amount of evidence increases, one gradually obtains a more reliable picture of the domain of interest.

1.4 A New Methodology for Bibliometric Mapping

Bibliometric mapping of science can be studied from different perspectives. Broadly speaking, bibliometric mapping studies take either a methodological point of view or an application point of view (or a combination of both). Methodological research focuses on the technical issues in producing bibliometric maps, the proper interpretation of bibliometric maps, and the validation of bibliometric maps. Application oriented research is concerned with the use of bibliometric maps for all kinds of purposes, for example to assist researchers to get an overview of their field or to support science policy makers to make well-founded decisions. This thesis has a strong focus on the technical aspects of bibliometric mapping. The main contribution of the thesis consists of introducing a new methodology for bibliometric mapping.[2] In this section, we will give an overview of this new methodology.

The process of bibliometric mapping can be divided into a number of relatively independent steps. Different divisions are possible. For the purpose of this thesis, we divide the bibliometric mapping process into the following six steps:

(1) Selection of the objects of interest.

(2) Calculation of the relatedness of objects.

(3) Normalization of the relatedness scores.

(4) Construction of a map.

(5) Presentation of the map.

(6) Evaluation of the map.

These steps are performed sequentially. However, bibliometric mapping is an iterative process. Going through the above steps only once usually does not yield a satisfactory bibliometric map. In step 6, it often turns out that one needs to go back to one of the earlier steps in order to revise the choices made in that step. All subsequent steps then need to be redone. A number of iterations are typically required to obtain a satisfactory bibliometric map.

[2] From now on, we use the term ‘methodology’ in a narrow sense, namely to refer to the technical aspects of bibliometric mapping.

Why do we need a new methodology for bibliometric mapping? As we will argue in this thesis, existing methods and techniques for bibliometric mapping, especially the methods and techniques that are commonly used in steps 3, 4, and 5 of the bibliometric mapping process, have important shortcomings. The most popular approaches for normalizing relatedness scores (step 3) lack a solid mathematical justification. Popular multidimensional-scaling-based approaches for constructing bibliometric maps (step 4) do not always yield satisfactory results, especially not in the case of larger data sets. And the presentation of bibliometric maps (step 5) is often done using very simple static pictures and without offering any possibility for interaction. The methodology introduced in this thesis aims to provide improved methods and techniques for bibliometric mapping, especially for steps 3, 4, and 5 of the bibliometric mapping process.

We will now consider the six steps of the bibliometric mapping process in more detail. For each step, we will discuss to what extent the methodology introduced in this thesis enhances existing methods and techniques for bibliometric mapping.

1.4.1 Step 1: Selection of the Objects of Interest

In this step, one delineates the domain that one wants to study, one chooses the unit of analysis, and one selects the objects to be shown in the map. Delineation of the domain can be done by identifying relevant documents based on keywords, classification codes, or the journal in which a document was published. The choice of the unit of analysis is determined by the type of map that one wants to have. Depending on the unit of analysis, selection of the objects to be shown in the map can be done in different ways. In the case of documents, one could for example select the documents with the largest number of citations. In the case of authors, one could select the authors who have published the largest number of documents. In the case of words or terms, the selection of the objects to be shown in the map is usually more difficult. In general, simply selecting the most frequently occurring words or terms does not work well. Many frequently occurring words or terms have a general meaning and are therefore not very relevant. In this thesis, a new technique for automatic term identification is introduced (see Chapter 2). This technique aims to automatically select the most relevant terms to be shown in a term map. Basically, a term is considered relevant if it is strongly associated with a single topic within the domain of study.

1.4.2 Step 2: Calculation of the Relatedness of Objects

The most commonly used measures of the relatedness of objects were discussed in Subsection 1.2.1. The bibliometric mapping methodology introduced in this thesis can be used with all these measures. Measures of relatedness can be calculated in two different ways, namely using a full counting method or using a fractional counting method (Small & Sweeney, 1985). In the bibliometric mapping literature, the full counting method is almost always used. The importance of the fractional counting method was pointed out by Small and Sweeney (1985). They argued that the fractional counting method can be used to make co-citation counts from different scientific fields comparable with each other. In this thesis, both the full counting method and the fractional counting method are used. However, the fractional counting method should be seen as the preferred choice in the bibliometric mapping methodology introduced in this thesis.

1.4.3 Step 3: Normalization of the Relatedness Scores

Relatedness scores usually need to be normalized in order to correct for differences in the size of objects. For example, it is only natural that two large journals with lots of publications have more co-citations with each other than two small journals with just a few publications. Such a difference in co-citations does not imply that the two large journals should be regarded as more strongly related to each other than the two small journals. Co-citation counts should first be normalized before such conclusions can be drawn. The normalization of relatedness scores has received a significant amount of attention in the literature (e.g., Ahlgren, Jarneving, & Rousseau, 2003; Klavans & Boyack, 2006a; Peters & Van Raan, 1993a). The methods used to normalize relatedness scores are often referred to as similarity measures. This terminology is also used in this thesis. Sometimes relatedness scores are normalized in an indirect way. Two objects are then considered to be related if they have similar relations with other objects. The indirect normalization approach was popularized by the work done at Drexel University on the mapping of authors based on co-citation relations (e.g., McCain, 1990; White & Griffith, 1981; White & McCain, 1998).

In this thesis, we study both the direct and the indirect approach to normalizing relatedness scores (see Chapters 3 and 4). In the literature, normalization methods are mainly studied in a somewhat informal, empirical manner. In our view, however, the most appropriate way to study normalization methods is by analyzing their mathematical properties. We therefore take a strictly mathematical point of view in this thesis. More specifically, we formulate a number of properties that we believe a reasonable normalization method should satisfy, and we derive which normalization methods indeed satisfy these properties and which do not. Although both the direct and the indirect approach to normalizing relatedness scores are studied in this thesis, only the direct approach should be seen as part of the bibliometric mapping methodology that is introduced in the thesis. We argue (see Chapter 4) that of the various direct normalization methods that we study, the so-called association strength method is the most satisfactory one. This method is preferable over other more commonly used methods, such as the cosine method and the Jaccard method.

1.4.4 Step 4: Construction of a Map

In this step, a spatial representation of the objects of interest is created based on the normalized relatedness scores of the objects. This usually means that for each object a location in a two-dimensional space is calculated. In many cases, the objects are also clustered, that is, the objects are divided into a number of non-overlapping groups.

The focus of this thesis is on distance-based maps. Graph-based maps are not considered. Hence, in the bibliometric maps in this thesis, the distance between two objects is supposed to provide an indication of the relatedness of the objects. As discussed in Subsection 1.2.2, the most popular technique for constructing distance-based maps is multidimensional scaling. In this thesis, an alternative to multidimensional scaling is introduced (see Chapter 5). This alternative is referred to as the VOS mapping technique, where VOS stands for *visualization of similarities*. It is argued that the VOS mapping technique yields more satisfactory maps than popular multidimensional-scaling-based approaches to bibliometric mapping. Maps constructed using these multidimensional-scaling-based approaches are shown to suffer from certain artifacts. Maps constructed using the VOS mapping technique do not have this problem.

In addition to the VOS mapping technique, this thesis also introduces the VOS clustering technique (see Chapter 6). The VOS mapping technique and the VOS clustering technique are based on the same underlying mathematical principle, and therefore these two techniques together provide a unified framework for mapping and clustering. The VOS clustering technique can be used to cluster the objects in a bibliometric map. The technique can serve as an alternative to other clustering techniques, such as the commonly used technique of hierarchical clustering. It is shown that the VOS clustering technique is closely related to modularity-based clustering, which is a popular clustering technique in the physics literature (Newman, 2004a, 2004b; Newman & Girvan, 2004). An advantage of the combined use of the VOS mapping technique and the VOS clustering technique is that mapping and clustering are performed in a consistent way. In the literature, mapping and clustering techniques are often used together, but the techniques are typically based on different principles, which may lead to mapping and clustering results that are not consistent with each other.

1.4.5 Step 5: Presentation of the Map

In the literature, the presentation of bibliometric maps often receives relatively little attention. However, in many cases the value of a bibliometric map can be enhanced significantly by choosing an appropriate way of presenting the map. For example, the size of objects can be varied in order to indicate differences in the importance of objects, colors can be used to discern different types of objects, and labels can be displayed in such a way that they do not overlap each other. Also, in some cases, especially when visualization is done at the level of clusters of objects rather than at the level of individual objects, the choice of good labels needs special attention. Another way to improve the presentation of a bibliometric map may be by allowing the map to be explored interactively. This requires special computer software.

In this thesis, a new computer program for displaying and exploring bibliometric maps is introduced (see Chapter 7). The program is called VOSviewer and is freely available at http://www.vosviewer.com. The VOSviewer software has extensive visualization capabilities. Bibliometric maps can be displayed in various different ways, each emphasizing a different aspect of a map. Colors can be used to indicate clusters of objects. A special labeling algorithm guarantees that labels do not overlap each other. Zoom, scroll, and search functionality is provided to support the interactive exploration of a map. The VOSviewer software can also be employed to construct bibliometric maps using the VOS mapping and clustering techniques.

1.4.6 Step 6: Evaluation of the Map

This is a non-technical step in which one needs to determine whether the bibliometric map that one has obtained is satisfactory or not. If the bibliometric map is not considered satisfactory, one needs to go back to one of the earlier steps of the bibliometric mapping process and one needs to revise the choices made in that step. There can be various reasons for not being satisfied with a bibliometric map. For example, it may turn out that the domain of interest has not been properly delineated. There may also be too many or too few objects in the map, in which case the map does not provide the right level of detail. Another possibility is that due to the limited availability of data the relatedness of objects has not been measured with sufficient accuracy. The map then does not give a proper representation of the domain of interest. Also, if a clustering technique has been used, the number of clusters may turn out to be too large or too small. The clustering of the objects may then be of little value. In practice, one often needs to go through the various steps of the bibliometric mapping process a number of times in order to obtain a satisfactory bibliometric map. Because the focus of this thesis is on the technical aspects of bibliometric mapping, no special attention is paid to the evaluation step of the bibliometric mapping process.

1.4.7 Summary of the New Bibliometric Mapping Methodology

We have now discussed the six steps of the bibliometric mapping process. In Table 1.1, the implementation of these steps in the new bibliometric mapping methodology introduced in this thesis is summarized.

1.5 Outline of the Thesis

The thesis consists of nine chapters. The chapters roughly follow the steps of the bibliometric mapping process discussed in the previous section. Chapter 2 introduces a new technique for automatic term identification. This technique can be used to automatically select the terms to be shown in a term map. Chapters 3 and 4 are concerned with the mathematical analysis of methods for normalizing relatedness scores of objects. Chapters 5 and 6 focus on techniques for constructing bibliometric maps.

Table 1.1: The six steps of the bibliometric mapping process and their implementation in the new methodology introduced in this thesis.

Step of the bibliometric mapping process	Implementation in the new methodology
1. Selection of the objects of interest	Automatic term identification technique (only for term maps; see Chapter 2)
2. Calculation of the relatedness of objects	Fractional counting method (Small & Sweeney, 1985)
3. Normalization of the relatedness scores	Association strength normalization method (see Chapter 4)
4. Construction of a map	VOS mapping technique (see Chapter 5) VOS clustering technique (see Chapter 6)
5. Presentation of the map	VOSviewer software (see Chapter 7)
6. Evaluation of the map	

Chapter 5 presents the VOS mapping technique and compares this technique with the technique of multidimensional scaling. Chapter 6 introduces the VOS clustering technique and proposes a unified framework for mapping and clustering of bibliometric networks. Chapter 7 is concerned with the presentation of bibliometric maps. This chapter introduces the VOSviewer software for displaying and exploring bibliometric maps. Chapter 8 presents an application of bibliometric mapping. In this application, bibliometric mapping is used to study the field of computational intelligence. Finally, Chapter 9 summarizes the thesis and suggests some directions for future research.

Chapters 2 to 8 have all been published in the international peer-reviewed scientific literature. Chapters 2 to 7 have appeared in bibliometrics journals. Chapters 2 and 7 were published in *Scientometrics*, Chapters 3, 4, and 5 in the *Journal of the American Society for Information Science and Technology*, and Chapter 6 in the *Journal of Informetrics*. Chapter 8 has appeared in a computer science journal, the *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*.

Chapter 2

Automatic Term Identification for Bibliometric Mapping[*]

Abstract

A term map is a map that visualizes the structure of a scientific field by showing the relations between important terms in the field. The terms shown in a term map are usually selected manually with the help of domain experts. Manual term selection has the disadvantages of being subjective and labor-intensive. To overcome these disadvantages, we propose a methodology for automatic term identification and we use this methodology to select the terms to be included in a term map. To evaluate the proposed methodology, we use it to construct a term map of the field of operations research. The quality of the map is assessed by a number of operations research experts. It turns out that in general the proposed methodology performs quite well.

2.1 Introduction

Bibliometric mapping is a powerful tool for studying the structure and the dynamics of scientific fields. Researchers can utilize bibliometric maps to obtain a better understanding of the field in which they are working. In addition, bibliometric maps can provide valuable insights for science policy purposes (Noyons, 1999, 2004).

[*] This chapter is based on Van Eck, Waltman, Noyons, and Buter (2010).

Various types of bibliometric maps can be distinguished, which each visualize the structure of a scientific field from a different point of view. Some maps, for example, show relations between authors or journals based on co-citation data. Other maps show relations between words or keywords based on co-occurrence data (e.g., Rip & Courtial, 1984; Peters & Van Raan, 1993b; Kopcsa & Schiebel, 1998; Noyons, 1999; Ding, Chowdhury, & Foo, 2001). The latter maps are usually referred to as co-word maps. In this chapter, we are concerned with maps that show relations between terms. We refer to these maps as term maps. By a term we mean a word or a phrase that refers to a domain-specific concept. Term maps are similar to co-word maps except that they may contain any type of term instead of only single-word terms or only keywords.

When constructing a bibliometric map, one first has to select the objects to be included in the map. In the case of a map that contains authors or journals, this is usually fairly easy. To select the important authors or journals in a field, one can usually simply rely on citation counts. In the case of a term map, things are not so easy. In most cases, it is quite difficult to select the important terms in a field. Selection of terms based on their frequency of occurrence in a corpus of documents typically yields many words and phrases with little or no domain-specific meaning. Inclusion of such words and phrases in a term map is highly undesirable for two reasons. First, these words and phrases divert attention from what is really important in the map. Second and even more problematic, these words and phrases may distort the entire structure shown in the map. Because there is no easy way to select the terms to be included in a term map, term selection is usually done manually based on expert judgment (e.g., Noyons, 1999; Van Eck & Waltman, 2007a). However, manual term selection has serious disadvantages as well. The most important disadvantage is that it involves a lot of subjectivity, which may introduce significant biases in a term map. Another disadvantage is that it can be very labor-intensive.

In this chapter, we try to overcome the problems associated with manual selection of the terms to be included in a term map. To do so, we propose a methodology that aims to automatically identify the terms that occur in a corpus of documents. Term selection using the proposed methodology requires less involvement of domain experts than manual term selection. Consequently, we expect term maps constructed using the proposed methodology to be more objective representations of scientific fields. An additional advantage of the proposed methodology is that it makes the process of term selection less labor-intensive.

The general idea of the methodology that we propose can be explained briefly as follows. Given a corpus of documents, we first identify the main topics in the corpus. This is done using a technique called probabilistic latent semantic analysis (Hofmann, 2001). Given the main topics, we then identify in the corpus the words and phrases that are strongly associated with only one or only a few topics. These words and phrases are selected as the terms to be included in a term map. An important property of the proposed methodology is that it identifies terms that are not only domain-specific but that also have a high discriminatory power within the domain of interest. This is important because terms with a high discriminatory power are essential for visualizing the structure of a scientific field. Suppose, for example, that we want to construct a term map of the field of statistics. *sample* and *chi-square test* are both statistical terms. However, *sample* is a quite general statistical term, while *chi-square test* is more specific and, consequently, more discriminatory. Because of the relatively high discriminatory power of *chi-square test*, inclusion of this term in a term map may help to reveal the structure of the field of statistics. Inclusion of *sample*, on the other hand, probably does not provide much additional insight into the structure of the field. Hence, to visualize the structure of a scientific field, terms with a high discriminatory power play an essential role.

The organization of this chapter is as follows. We first provide a brief overview of the literature on automatic term identification. After discussing the literature, we propose a new methodology for automatic term identification. We then experimentally evaluate the proposed methodology, focusing in particular on its performance in the context of bibliometric mapping. Evaluation is done by applying the proposed methodology to the field of operations research and by asking a number of experts in this field to assess the results that are obtained. We end this chapter with a discussion of the conclusions of our research.

2.2 Overview of the Automatic Term Identification Literature

In this section, we review the literature on automatic term identification (also known as automatic term recognition or automatic term extraction). More extensive literature reviews are provided by Kageura and Umino (1996), Cabré Castellví, Estopà Bagot, and Vivaldi Palatresi (2001), Jacquemin (2001), and Pazienza, Pennacchiotti, and Zanzotto (2005). We note that there are almost no studies on automatic term identification in the context of bibliometric mapping. Exceptions are the work of Janssens, Leta, Glänzel, and De Moor (2006), Noyons (1999), and Schneider (2006), in which automatic term identification receives some attention. Kostoff and Block (2005) are concerned with automatic term identification in a bibliometric context, but not specifically for mapping purposes. In the literature reviewed in the rest of this section, automatic term identification is studied for purposes other than bibliometric analysis.

We first discuss the notions of unithood and termhood (for the original definitions of these notions, see Kageura & Umino, 1996). We define unithood as the degree to which a phrase constitutes a semantic unit. Our idea of a semantic unit is similar to that of a collocation (Manning & Schütze, 1999). Hence, a semantic unit is a phrase consisting of words that are conventionally used together. The meaning of the phrase typically cannot be fully predicted from the meaning of the individual words within the phrase. We define termhood as the degree to which a semantic unit represents a domain-specific concept. A semantic unit with a high termhood is a term. To illustrate the notions of unithood and termhood, suppose that we are interested in statistical terms. Consider the phrases *many countries*, *United States*, and *probability density function*. Clearly, *United States* and *probability density function* are semantic units, while *many countries* is not. Hence, the unithood of *United States* and *probability density function* is high, while the unithood of *many countries* is low. Because *United States* does not represent a statistical concept, it has a low termhood. *probability density function*, on the other hand, does represent a statistical concept and therefore has a high termhood. From this it follows that *probability density function* is a statistical term.

In the literature, two types of approaches to automatic term identification are distinguished, linguistic approaches and statistical approaches. Linguistic approaches are mainly used to identify phrases that, based on their syntactic form, can serve as candidate terms. Statistical approaches are used to measure the unithood and termhood of phrases. In many cases, linguistic and statistical approaches are combined in a single hybrid approach.

Most terms have the syntactic form of a noun phrase (Justeson & Katz, 1995; Kageura & Umino, 1996). Linguistic approaches to automatic term identification typically rely on this property. These approaches identify candidate terms using a linguistic filter that checks whether a sequence of words conforms to some syntactic pattern. Different researchers use different syntactic patterns for their linguistic filters (e.g., Bourigault, 1992; Dagan & Church, 1994; Daille, Gaussier, & Langé, 1994; Justeson & Katz, 1995; Frantzi, Ananiadou, & Mima, 2000). Each syntactic pattern covers a specific class of noun phrases, such as the class of all noun phrases consisting of nouns only or the class of all noun phrases consisting of nouns and adjectives only.

Statistical approaches to automatic term identification are used to measure the unithood and termhood of phrases. We first discuss some statistical approaches to measure unithood (for a much more extensive discussion of such approaches, see Manning & Schütze, 1999). We note that measuring unithood is only necessary for the identification of multi-word terms. The simplest approach to measure unithood relies on the idea that phrases that occur more frequently are more likely to be semantic units (e.g., Dagan & Church, 1994; Daille et al., 1994; Justeson & Katz, 1995). This approach uses frequency of occurrence as a measure of unithood. However, it is much more surprising to observe a phrase consisting of words that individually all have a low frequency of occurrence than it is to observe a phrase consisting of words that individually all have a high frequency of occurrence. Frequency of occurrence as a measure of unithood does not take this into account. As an alternative to frequency of occurrence, measures based on, for example, (pointwise) mutual information (e.g., Church & Hanks, 1990; Damerau, 1993; Daille et al., 1994) or a likelihood ratio (e.g., Dunning, 1993; Daille et al., 1994) can be used. Frantzi et al. (2000) propose another alternative to frequency of occurrence, to which they refer as the C-value. In addition to frequency of occurrence, the C-value takes into account that longer phrases are less likely to be observed than shorter phrases. The C-value also pays special attention to nested terms, which are terms that are part of other longer terms. Because the C-value does not indicate whether phrases are domain-specific, we regard it as a measure of unithood. (This contrasts with Frantzi et al., who regard the C-value as a measure of termhood.)

There also exist a number of statistical approaches to measure the termhood of semantic units. We now discuss some of these approaches. The NC-value (Frantzi et al., 2000) and the SNC-value (Maynard & Ananiadou, 2000) are extensions of the C-value. These extensions measure not only unithood but also termhood. To measure the termhood of a semantic unit, the NC- and the SNC-value use contextual information, that is, information on the words that occur in the vicinity of a unit. For example, the presence of certain words or certain word classes (such as verbs and prepositions) in the vicinity of a unit increases the units termhood. Other statistical approaches to measure termhood rely on the idea that a semantic unit is likely to represent a domain-specific concept if the unit occurs relatively more frequently in a specific domain than in general or if within a specific domain the distribution of the units occurrences is in some way biased (Kageura & Umino, 1996). Drouin (2003) uses this idea by only taking into account semantic units having the property that each word individually occurs significantly more frequently in a domain-specific corpus than in a general corpus. This approach seems to improve the identification of single-word terms, but it does not seem to work very well for multi-word terms. Matsuo and Ishizuka (2004) propose an approach based on the idea that the occurrences of a term usually have a biased distribution. They use a corpus consisting of a single document. Basically, they first select a number of frequently occurring phrases and count the co-occurrences of these phrases with all other phrases. Based on the co-occurrence frequencies, they then measure, using a chi-square value, whether the distribution of the occurrences of a phrase is biased. The chi-square value obtained in this way can be regarded as a measure of the termhood of a phrase. The approach that we propose in this chapter is based on a somewhat similar idea as the approach of Matsuo and Ishizuka. One important difference is that our approach exploits the property that in many cases a corpus consists of a large number of documents, each of which is concerned with a somewhat different topic. This property turns out to be very useful to determine whether the occurrences of a semantic unit are biased towards one or more topics.

In the field of information retrieval, researchers study the problem of determining which words and phrases in a document are important for indexing purposes and which are not (e.g., Kim & Wilbur, 2001). This problem is quite similar to the problem of automatic term identification (Kageura & Umino, 1996) or, more specifically, to the problem of measuring termhood. Although it is not our intention here to review the information retrieval literature, we do want to mention the work of Bookstein and Swanson (1974) and Harter (1975). This is because the approach that we propose in this chapter to measure termhood is based on a similar idea as their work. This is the idea that in a corpus of documents the occurrences of a term tend to cluster together while the occurrences of a general word or phrase tend to be randomly distributed. Our approach relies on this idea, but it applies the idea at the level of topics rather than at the level of individual documents.

Finally, we note that the problem of automatic term identification also receives some attention in the field of machine learning. In this field, an interesting statistical approach that can be used to measure both unithood and termhood is proposed by Wang, McCallum, and Wei (2007). This approach relies on a probabilistic model of the process of generating a corpus of documents. Terms can be identified by estimating the parameters of this model. The approach that we propose in this chapter is related to the approach of Wang et al. because it makes use of a somewhat similar probabilistic model.

2.3 Methodology

In this section, we propose a three-step methodology for automatic term identification. An overview of the proposed methodology is provided in Figure 2.1. Consider some domain or some scientific field, and suppose that we want to identify terms that belong specifically to this domain or this field. Our methodology assumes the availability of a corpus that is partitioned into a number of segments, each of which is concerned with a particular topic or a particular combination of topics within the domain of interest. Such a corpus may for example consist of a large number of documents or abstracts. In the first step of our methodology, a linguistic filter is applied to the corpus in order to identify noun phrases. In the second step, the unithood of noun phrases is measured in order to identify semantic units. In the third and final step, the termhood of semantic units is measured in order to identify terms. Termhood is measured as the degree to which the occurrences of a semantic unit are biased towards one or more topics. Compared with

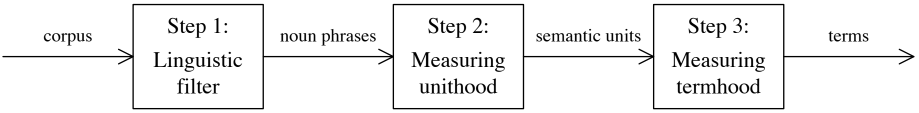

corpus
Step 1: Linguistic filter
noun phrases
Step 2: Measuring unithood
semantic units
Step 3: Measuring termhood
terms

Figure 2.1: Overview of the proposed methodology.

alternative approaches to automatic term identification, such as the ones discussed in the previous section, the innovative aspect of our methodology mainly lies in the third step, that is, in the measurement of termhood. We now discuss the three steps of our methodology in more detail.

2.3.1 Step 1: Linguistic Filter

In the first step of our methodology, we use a linguistic filter to identify noun phrases. We first assign to each word occurrence in the corpus a part-of-speech tag, such as noun, verb, or adjective. The appropriate part-of-speech tag for a word occurrence is determined using a part-of-speech tagger developed by Schmid (1994, 1995). We use this tagger because it has a good performance and because it is freely available for research purposes.[1] In addition to a part-of-speech tag, the tagger also assigns a so-called lemma to each word occurrence in the corpus. The lemma assigned to a word occurrence is the root form (or the stem) of the word. The words *function* and *functions*, for example, both have *function* as their lemma. In all further stages of our methodology, we use the lemmatized corpus instead of the original corpus. In this way, differences between, for example, uppercase and lowercase letters and singular and plural nouns are ignored.

After the corpus has been tagged and lemmatized, we apply a linguistic filter to it. The filter that we use identifies all word sequences that meet the following three criteria:

(1) The sequence consists of nouns and adjectives only.

(2) The sequence ends with a noun.

(3) The sequence occurs at least a certain number of times in the corpus (ten times in the experiment discussed later on in this chapter).

[1] See http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/.

Assuming an English language corpus, the first two criteria ensure that all identified word sequences are noun phrases. Notice, however, that our filter does not identify all types of noun phrases. Noun phrases that contain a preposition, such as the phrase *degree of freedom*, are not identified (for a discussion of such noun phrases, see Justeson & Katz, 1995). We emphasize that the choice of an appropriate linguistic filter depends on the language of the corpus. The filter that we use works well for the English language but may not be appropriate for other languages. For all noun phrases that are identified by our linguistic filter, the unithood is considered in the second step of our methodology.

2.3.2 Step 2: Measuring Unithood

In the second step of our methodology, we measure the unithood of noun phrases. Unithood is only relevant for noun phrases consisting of more than one word. For such noun phrases, unithood determines whether they are regarded as semantic units. The main aim of the second step of our methodology is to get rid of noun phrases that start with uninteresting adjectives such as *first*, *many*, *new*, and *some*.

The most common approach to measure unithood is to determine whether a phrase occurs more frequently than would be expected based on the frequency of occurrence of the individual words within the phrase. This is basically also the approach that we take. To measure the unithood of a noun phrase, we first count the number of occurrences of the phrase, the number of occurrences of the phrase without the first word, and the number of occurrences of the first word of the phrase. In a similar way as Dunning (1993), we then use a so-called likelihood ratio to compare the first number with the last two numbers. We interpret this likelihood ratio as a measure of the unithood of the phrase. In the end, we use a cutoff value to determine which noun phrases are regarded as semantic units and which are not. (In the experiment discussed later on in this chapter, noun phrases are regarded as semantic units if the natural logarithm of their likelihood ratio is less than -30.) For all noun phrases that are regarded as semantic units (which includes all single-word noun phrases), the termhood is considered in the third step of our methodology.

2.3.3 Step 3: Measuring Termhood

In the third step of our methodology, we measure the termhood of semantic units. As mentioned earlier, we assume that we have a corpus that is partitioned into a number of segments, each of which is concerned with a particular topic or a particular combination of topics within the domain of interest. A corpus segment may for example consist of a document or an abstract, or it may consist of the set of all documents or all abstracts that appeared in a journal during a certain period of time. We use the following mathematical notation. There are K semantic units of which we want to measure the termhood. These units are denoted by $u_1, \ldots, u_K$. The corpus is partitioned into I segments, which are denoted by $s_1, \ldots, s_I$. The number of occurrences of semantic unit u_k in corpus segment s_i is denoted by n_{ik}. Finally, there are J topics to be distinguished. These topics are denoted by $t_1, \ldots, t_J$.

The main idea of the third step of our methodology is to measure the termhood of a semantic unit as the degree to which the occurrences of the unit are biased towards one or more topics. We first discuss an approach that implements this idea in a very simple way. We assume that there is a one-to-one relationship between corpus segments and topics, that is, each corpus segment covers exactly one topic and each topic is covered by exactly one corpus segment. Under this assumption, the number of corpus segments equals the number of topics, so $I = J$. To measure the degree to which the occurrences of semantic unit u_k, where $k \in \{1, \ldots, K\}$, are biased towards one or more topics, we use two probability distributions, namely the distribution of semantic unit u_k over the set of all topics and the distribution of all semantic units together over the set of all topics. These distributions are denoted by, respectively, $P(t_j|u_k)$ and $P(t_j)$, where $j \in \{1, \ldots, J\}$. Assuming that topic t_j is covered by corpus segment s_j, the distributions are given by

$$P(t_j|u_k) = \frac{n_{jk}}{\sum_{j'=1}^{J} n_{j'k}} \tag{2.1}$$

and

$$P(t_j) = \frac{\sum_{k=1}^{K} n_{jk}}{\sum_{j'=1}^{J} \sum_{k=1}^{K} n_{j'k}}. \tag{2.2}$$

The dissimilarity between the two distributions indicates the degree to which the occurrences of u_k are biased towards one or more topics. We use the dissimilarity between the two distributions to measure the termhood of u_k. For example, if the two distributions are identical, the occurrences of u_k are unbiased and u_k most probably does not represent a domain-specific concept. If, on the other hand, the two distributions are very dissimilar, the occurrences of u_k are strongly biased and u_k is very likely to represent a domain-specific concept. The dissimilarity between two probability distributions can be measured in many different ways. One may use, for example, the Kullback-Leibler divergence, the Jensen-Shannon divergence, or a chi-square value. We use a somewhat different measure. Based on this measure, the termhood of u_k is calculated as

$$\text{termhood}(u_k) = \sum_{j=1}^{J} p_j \log p_j, \tag{2.3}$$

where $0 \log 0$ is defined as 0 and where

$$p_j = \frac{P(t_j|u_k)/P(t_j)}{\sum_{j'=1}^{J} P(t_{j'}|u_k)/P(t_{j'})}. \tag{2.4}$$

It follows from (2.4) that $p_1, \ldots, p_J$ define a probability distribution over the set of all topics. In (2.3), termhood(u_k) is calculated as the negative entropy of this distribution. Notice that termhood(u_k) is maximal if $P(t_j|u_k) = 1$ for some j and that it is minimal if $P(t_j|u_k) = P(t_j)$ for all j. In other words, termhood(u_k) is maximal if the occurrences of u_k are completely biased towards a single topic, and termhood(u_k) is minimal if the occurrences of u_k do not have a bias towards any topic.

The approach discussed above relies on the assumption of a one-to-one relationship between corpus segments and topics. For most corpora, this assumption is probably not very realistic. For example, if each segment of a corpus consists of a single document or a single abstract, there will most likely be some segments that are concerned with more or less the same topic. Or the other way around, if each segment of a corpus consists of a set of documents or abstracts that all appeared in the same journal, there will most likely be some segments (particularly segments corresponding to multidisciplinary journals) that are concerned with more than one topic. Below, we extend our approach in such a way that it no longer relies on the assumption of a one-to-one relationship between corpus segments and topics.

2.3.4 Identifying Topics

In order to allow for a many-to-many relationship between corpus segments and topics, we make use of probabilistic latent semantic analysis (PLSA) (Hofmann, 2001). PLSA is a quite popular technique in machine learning, information retrieval, and related fields. It was originally introduced as a probabilistic model that relates occurrences of words in documents to so-called latent classes. In the present context, we are dealing with semantic units and corpus segments instead of words and documents, and we interpret the latent classes as topics.

When using PLSA, we first have to determine an appropriate value for the number of topics J. This value is typically much smaller than both the number of corpus segments I and the number of semantic units K. In this chapter, we manually choose a value for J. PLSA assumes that each occurrence of a semantic unit in a corpus segment is independently generated according to the following probabilistic process. First, a topic t is drawn from a probability distribution $P(t_j)$, where $j \in \{1, \ldots, J\}$. Next, given t, a corpus segment s and a semantic unit u are independently drawn from, respectively, the conditional probability distributions $P(s_i|t)$, where $i \in \{1, \ldots, I\}$, and $P(u_k|t)$, where $k \in \{1, \ldots, K\}$. This then results in the occurrence of u in s. It is clear that, according to the generative process assumed by PLSA, the probability of generating an occurrence of semantic unit u_k in corpus segment s_i equals

$$P(s_i, u_k) = \sum_{j=1}^{J} P(t_j) P(s_i|t_j) P(u_k|t_j). \tag{2.5}$$

The probabilities $P(t_j)$, $P(s_i|t_j)$, and $P(u_k|t_j)$, for $i = 1, \ldots, I$, $j = 1, \ldots, J$, and $k = 1, \ldots, K$, are the parameters of PLSA. We estimate these parameters using data from the corpus. Estimation is based on the criterion of maximum likelihood. The log-likelihood function to be maximized is given by

$$L = \sum_{i=1}^{I} \sum_{k=1}^{K} n_{ik} \log P(s_i, u_k). \tag{2.6}$$

We use the EM algorithm discussed by Hofmann (1999, Section 3.2) to perform the maximization of this function.[2] After estimating the parameters of PLSA, we apply Bayes' theorem to obtain a probability distribution over the topics conditional on a semantic unit. This distribution is given by

$$P(t_j|u_k) = \frac{P(t_j)P(u_k|t_j)}{\sum_{j'=1}^{J} P(t_{j'})P(u_k|t_{j'})}. \tag{2.7}$$

In a similar way as discussed earlier, we use the dissimilarity between the distributions $P(t_j|u_k)$ and $P(t_j)$ to measure the termhood of u_k. In this case, however, $P(t_j|u_k)$ is given by (2.7) instead of (2.1) and $P(t_j)$ follows from the estimated parameters of PLSA instead of being given by (2.2). We again use (2.3) and (2.4) to calculate the termhood of u_k.

2.4 Experimental Evaluation

In this section, we experimentally evaluate our methodology for automatic term identification. We focus in particular on the performance of our methodology in the context of bibliometric mapping.

2.4.1 Application to the Field of Operations Research

We apply our methodology to the field of operations research (OR), also known as operational research. The OR field was chosen because some of us have some background in this field and because we have easy access to a number of OR experts who can help us with the evaluation of our results. We note that sometimes a distinction is made between OR on the one hand and management science on the other hand (e.g., Eto, 2000, 2002). For our purpose, however, such a distinction is not important. In this chapter, the term OR therefore also includes management science.

We start with a discussion of how we put together our corpus. We first selected a number of OR journals (for a recent bibliometric study of OR journals, see Kao, 2009). This was done based on the subject categories of Thomson Reuters. The OR field is covered by the category *Operations Research & Management Science*. Since we wanted

[2] A MATLAB implementation of this algorithm is available on request.

Table 2.1: Overview of the selected journals.

Journal	Number of documents	Coverage
European Journal of Operational Research	2705	97.2%
Journal of the Operational Research Society	830	96.9%
Management Science	726	98.9%
Annals of Operations Research	679	95.3%
Operations Research Letters	458	93.0%
Operations Research	439	97.7%
Naval Research Logistics	327	98.5%
Omega-International Journal of Management Science	277	97.1%
Interfaces	257	98.4%
Journal of Operations Management	211	98.1%
Journal of the Operations Research Society of Japan	158	96.8%
Asia-Pacific Journal of Operational Research	140	99.3%
OR Spectrum	140	97.9%
RAIRO-Operations Research	92	93.5%
Military Operations Research	53	98.1%
Total	7492	97.0%

to focus on the core of the field, we selected only a subset of the journals in this category. More specifically, a journal was selected if it belongs to the category *Operations Research & Management Science* and possibly also to the closely related category *Management* and if it does not belong to any other category. This yielded 15 journals, which are listed in the first column of Table 2.1. We used the database of the Centre for Science and Technology Studies, which is similar to the Web of Science database of Thomson Reuters, to retrieve all documents, except those without an abstract, that were published in the selected journals between 2001 and 2006. For each journal, the number of documents retrieved from the database is reported in the second column of Table 2.1. Of each of the documents retrieved, we included the title and the abstract in our corpus.

After putting together the corpus, we applied our methodology for automatic term identification. In the first step of our methodology, the linguistic filter identified 2662 different noun phrases. In the second step, the unithood of these noun phrases was measured. 203 noun phrases turned out to have a rather low unithood and therefore could not be regarded as semantic units. Examples of such noun phrases are *first problem*, *good* *use*, and *optimal cost*. The other 2459 noun phrases had a sufficiently high unithood to be regarded as semantic units. In the third and final step of our methodology, the termhood of these semantic units was measured. To do so, each title-abstract pair in the corpus was treated as a separate corpus segment. For each combination of a semantic unit u_k and a corpus segment s_i, it was determined whether u_k occurs in s_i $(n_{ik} = 1)$ or not $(n_{ik} = 0)$. Topics were identified using PLSA. This required the choice of the number of topics J. Results for various numbers of topics were examined and compared. Based on our own knowledge of the OR field, we decided to work with $J = 10$ topics. The output of our methodology consisted of a list of 2459 semantic units together with their termhood values. For the interested reader, this list is available online.[3]

2.4.2 Evaluation Based on Precision and Recall

The evaluation of a methodology for automatic term identification is a difficult issue. There is no generally accepted standard for how evaluation should be done. We refer to Pazienza et al. (2005) for a discussion of the various problems. In this chapter, we evaluate our methodology in two ways. We first perform an evaluation based on the well-known notions of precision and recall. We then perform a second evaluation by constructing a term map and asking experts to assess the quality of this map. Since our methodology for automatic term identification is intended to be used for bibliometric mapping purposes, we are especially interested in the results of the second evaluation.

We first discuss the evaluation of our methodology based on precision and recall. The main aim of this evaluation is to compare the performance of our methodology with the performance of two simple alternatives. One alternative is a variant of our methodology. This variant assumes a one-to-one relationship between corpus segments and topics, and it therefore does not make use of PLSA. The other alternative is a very simple one. It uses frequency of occurrence as a measure of termhood.

In the context of automatic term identification, precision and recall are defined as follows. Precision is the number of correctly identified terms divided by the total number of identified terms. Recall is the number of correctly identified terms divided by the total number of correct terms. Unfortunately, because the total number of correct terms in the OR field is unknown, we could not calculate the true recall. This is a well-

[3] See http://www.neesjanvaneck.nl/term_identification/.

known problem in the context of automatic term identification (Pazienza et al., 2005). To circumvent this problem, we defined recall in a slightly different way, namely as the number of correctly identified terms divided by the total number of correct terms within the set of all semantic units identified in the second step of our methodology. Recall calculated according to this definition provides an upper bound on the true recall. However, even using this definition of recall, the calculation of precision and recall remained problematic. The problem was that it is very time-consuming to manually determine which of the 2459 semantic units identified in the second step of our methodology are correct terms and which are not. We solved this problem by estimating precision and recall based on a random sample of 250 semantic units. Two experts, who both have knowledge of the OR field, independently determined which of these 250 units are correct terms and which are not. Units on which the experts did not agree were discussed until agreement was reached.

To identify terms, we used a cutoff value that determined which semantic units were regarded as terms and which were not. Semantic units were regarded as terms if their termhood value was greater than the cutoff value. Obviously, a lower cutoff value leads to a larger number of identified terms and, consequently, to a higher recall. However, a lower cutoff value usually also leads to a lower precision. Hence, there is a trade-off between precision and recall. By varying the cutoff value, the relation between precision and recall can be obtained. In Figure 2.2, the graphs labeled *PLSA* and *No PLSA* show this relation for, respectively, our methodology and the variant of our methodology that does not make use of PLSA. The third graph in the figure shows the relation between precision and recall for the approach based on frequency of occurrence. It is clear from the figure that our methodology outperforms the two simple alternatives. Except for very low and very high levels of recall, our methodology always has a considerably higher precision than the variant of our methodology that does not make use of PLSA. The low precision of our methodology for very low levels of recall is based on a very small number of incorrectly identified terms and is therefore insignificant from a statistical point of view. The approach based on frequency of occurrence has a very bad performance. For almost all levels of recall, the precision of this approach is even lower than the precision that would have been obtained if terms had been identified at random. Unfortunately, there is no easy way to compare the precision/recall performance of our

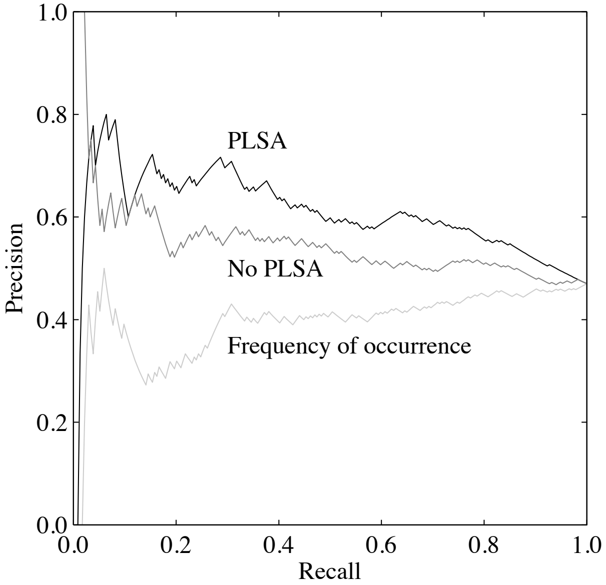


Figure 2.2: The relationship between precision and recall for our methodology and for two simple alternatives.

methodology with that of other approaches proposed in the literature. This is due to the lack of a generally accepted evaluation standard (Pazienza et al., 2005). We refer to (Cabré Castellví et al., 2001) for an overview of some precision/recall results reported for other approaches.

2.4.3 Evaluation Using a Term Map

We now discuss the second evaluation of our methodology for automatic term identification. This evaluation is performed using a term map. The evaluation therefore focuses specifically on the usefulness of our methodology for bibliometric mapping purposes.

A term map is a map, usually in two dimensions, that shows the relations between important terms in a scientific field. Terms are located in a term map in such a way that the proximity of two terms reflects their relatedness as closely as possible. That is, the smaller the distance between two terms, the stronger their relation. The aim of a term map usually is to visualize the structure of a scientific field.

In order to evaluate our methodology, we constructed a term map of the OR field. The terms to be included in the map were selected based on the output of our methodology. It turned out that, out of the 2459 semantic units identified in the second step of our methodology, 831 had the highest possible termhood value. This means that, according to our methodology, 831 semantic units are associated exclusively with a single topic within the OR field. We decided to select these 831 semantic units as the terms to be included in the term map. This yielded a coverage of 97.0%, which means that 97.0% of the title-abstract pairs in the corpus contain at least one of the 831 terms to be included in the term map. The coverage per journal is reported in the third column of Table 2.1.

The term map of the OR field was constructed using a procedure similar to the one used in our earlier work (Van Eck & Waltman, 2007a). This procedure relies on the association strength measure (Van Eck & Waltman, 2009) to determine the relatedness of two terms, and it uses the VOS technique (Van Eck & Waltman, 2007b; Van Eck, Waltman, Dekker, & Van den Berg, 2010) to determine the locations of terms in the map. Due to the large number of terms, the map that was obtained cannot be shown in this chapter. However, a simplified version of the map is presented in Figure 2.3. This version of the map only shows terms that do not overlap with other more important terms. The complete map showing all 831 terms is available online.[4] A special computer program called VOSviewer (Van Eck & Waltman, 2010) allows the map to be examined in full detail. VOSviewer uses colors to indicate the different topics that were identified using PLSA.

The quality of the term map of the OR field was assessed by five experts. Two of them are assistant professor of OR, one is associate professor of OR, and two are full professor of OR. All experts are working at Erasmus University Rotterdam. We asked each expert to examine the online term map and to complete a questionnaire. The questionnaire consisted of one multiple-choice question and ten open-ended questions. The main results of the questionnaire are discussed below. The full results are available on request.

In the multiple-choice question, we asked the experts to indicate on a five-point scale

[4] See http://www.neesjanvaneck.nl/term_identification/.

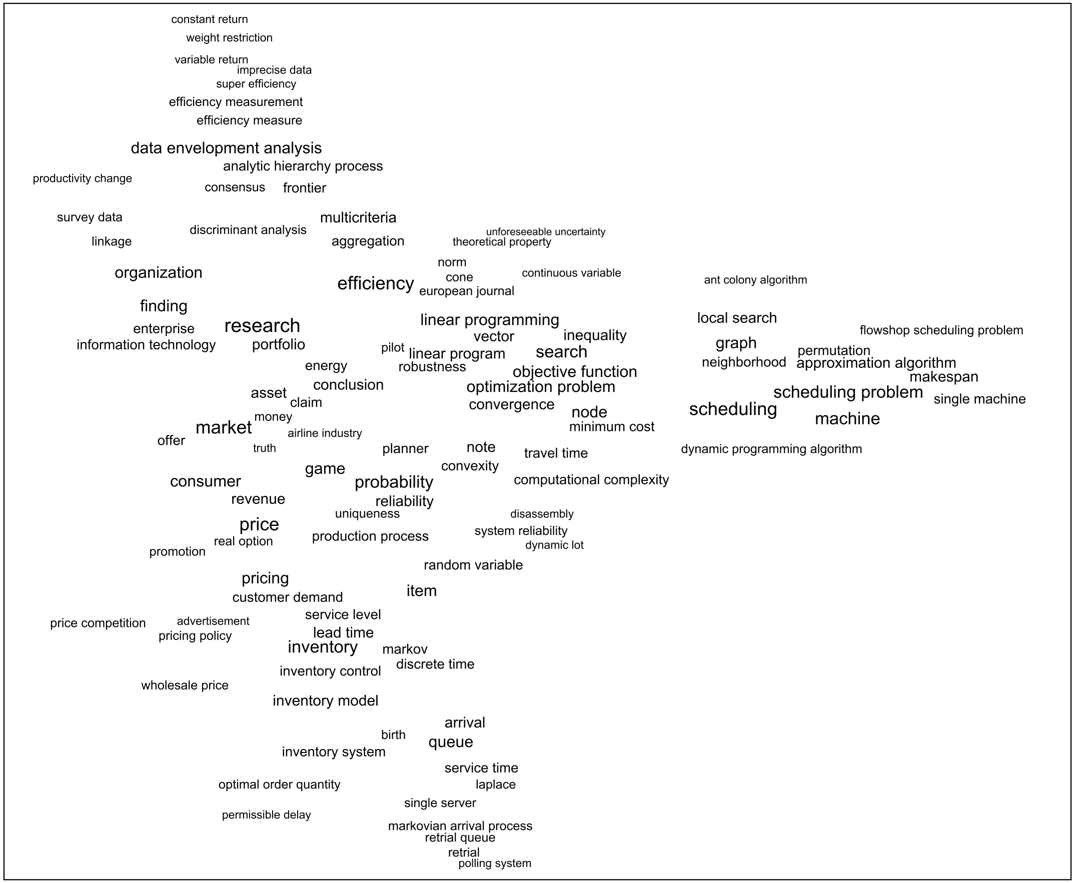


Figure 2.3: Simplified version of the term map of the OR field.

how well the term map visualizes the structure of the OR field. Four experts answered that the map visualizes the structure of the field quite well (the second highest answer on the five-point scale). The fifth expert answered that the map visualizes the structure of the field very well (the highest answer on the five-point scale). Hence, overall the experts were quite satisfied with the map. The experts could also easily explain the global structure of the map, and for them the topics shown in the map (indicated using colors) generally had an obvious interpretation. We also asked the experts whether the map showed anything unexpected to them. One expert answered that he had not expected scheduling related terms to be located at the boundary of the map. Two other experts turned out to be surprised by the prominent position of economics related terms such as *consumer*, *price*, *pricing*, and *revenue*. None of these three experts regarded the unexpected results as a weakness of the map. Instead, two experts stated that their own perception of their field may not have been correct. Hence, it seems that these experts may have learned something new from the map.

The experts also indicated some weak points of the term map. Some of these points were related to the way in which the terms shown in the map were selected. Other points were of a more general nature. The most serious criticism on the results of the automatic term identification concerned the presence of a number of rather general terms in the map. Examples of such terms are *claim*, *conclusion*, *finding*, *item*, and *research*. There were three experts who criticized the presence of terms such as these. We agree with these experts that some of the terms shown in the map are too general. Although the number of such terms is not very large, we believe that it is highly desirable to get rid of them. To achieve this, further improvement of our methodology for automatic term identification would be necessary. We will come back to this below.

Another point of criticism concerned the underrepresentation of certain topics in the term map. There were three experts who raised this issue. One expert felt that the topic of supply chain management is underrepresented in the map. Another expert stated that he had expected the topic of transportation to be more visible. The third expert believed that the topics of combinatorial optimization, revenue management, and transportation are underrepresented. It seems likely that in many cases the perceived underrepresentation of topics was not due to our methodology for automatic term identification but was instead caused by the way in which the corpus used by our methodology was put together. As discussed earlier, when we were putting together the corpus, we wanted to focus on the core of the OR field and we therefore only included documents from a relatively small number of journals. This may for example explain why the topic of transportation is not clearly visible in the map. Thomson Reuters has a subject category *Transportation Science & Technology*, and it may well be that much transportation related OR studies are published in journals that belong to this category (and possibly also to the category *Operations Research & Management Science*). The corpus that we put together does not cover these journals and hence may contain only a small portion of the transportation related OR studies. It is then not surprising that the topic of transportation is difficult to see in the map.

The remaining issues raised by the experts are of a more general nature, and most likely these issues would also have been raised if the terms shown in the term map had been selected manually. One of the issues had to do with the character of the OR field. When asked to divide the OR field into a number of smaller subfields, most experts indicated that there are two natural ways to make such a division. On the one hand, a division can be made based on the methodology that is being used, such as decision theory, game theory, mathematical programming, or stochastic modeling. On the other hand, a division can be made based on the area of application, such as inventory control, production planning, supply chain management, or transportation. There were two experts who noted that the term map seems to mix up both divisions of the OR field. According to these experts, one part of the map is based on the methodology-oriented division of the field, while the other part is based on the application-oriented division. One of the experts stated that he would be interested to see an explicit separation of the methodology and application dimensions.

A final issue, which was raised by two experts, had to do with the more detailed interpretation of the term map. The experts pointed out that sometimes closely related terms are not located very close to each other in the map. One of the experts gave the terms *inventory* and *inventory cost* as an example of this problem. In many cases, a problem such as this is probably caused by the limited size of the corpus that was used to construct the map. In other cases, the problem may be due to the inherent limitations of a two-dimensional representation. The best solution to this kind of problems seems to be not to show individual terms in a map but to only show topics (e.g., Noyons

& Van Raan, 1998; Noyons, 1999). Topics can then be labeled using one or more representative terms.

2.5 Conclusions

In this chapter, we have addressed the question how the terms shown in a term map can be selected without relying extensively on the judgment of domain experts. Our main contribution consists of a methodology for automatic identification of terms in a corpus of documents. Using this methodology, the process of selecting the terms to be included in a term map can be automated for a large part, thereby making the process less labor-intensive and less dependent on expert judgment. Because less expert judgment is required, the process of term selection also involves less subjectivity. We therefore expect term maps constructed using our methodology to be more objective representations of scientific fields.

We have evaluated our methodology for automatic term identification by applying it to the OR field. In general, we are quite satisfied with the results that we have obtained. The precision/recall results clearly indicate that our methodology outperformed two simple alternatives. In addition, the quality of the term map of the OR field constructed using our methodology was assessed quite positively by five experts in the field. However, the term map also revealed a shortcoming of our methodology, namely the incorrect identification of a number of general noun phrases as terms. We hope to remedy this shortcoming in future work.

Finally, we would like to place the research presented in this chapter in a broader perspective. As scientific fields tend to overlap more and more and disciplinary boundaries become more and more blurred, finding an expert who has a good overview of an entire domain becomes more and more difficult. This poses serious difficulties for any bibliometric method that relies on expert knowledge. Term mapping is one such method. Fortunately, advanced computational techniques from fields such as data mining, machine learning, statistics, and text mining may be used to take over certain tasks in bibliometric analysis that are traditionally performed by domain experts (for an overview of various computational techniques, see Leopold, May, & Paaß, 2004). The research presented in this chapter can be seen as an elaboration of this idea in the context of term mapping.

We acknowledge, however, that our research is only a first step towards fully automatic term mapping. To produce accurate term maps, the output of our methodology for automatic term identification still needs to be verified manually and some amount of expert knowledge is still required. In future work, we intend to take even more advantage of the possibilities offered by various kinds of computational techniques. Hopefully, this allows the dependence of term mapping on expert knowledge to be reduced even further.

Chapter 3

Appropriate Similarity Measures for Author Cocitation Analysis*

Abstract

We provide a number of new insights into the methodological discussion about author cocitation analysis. We first argue that the use of the Pearson correlation for measuring the similarity between authors' cocitation profiles is not very satisfactory. We then discuss what kind of similarity measures may be used as an alternative to the Pearson correlation. We consider three similarity measures in particular. One is the well-known cosine. The other two similarity measures have not been used before in the bibliometric literature. We show by means of an example that the choice of an appropriate similarity measure has a high practical relevance. Finally, we discuss the use of similarity measures for statistical inference.

3.1 Introduction

In the past few years, there has been a lot of discussion about the way in which author cocitation analysis (ACA) should be performed. Ahlgren et al. (2003) questioned the appropriateness of the Pearson correlation for measuring the similarity between authors' cocitation profiles.[1] Their paper caused quite some debate. In particular, White (2003a)

*This chapter is based on Van Eck and Waltman (2008).

[1] The cocitation profile of an author is a vector in which each element indicates the number of times the author has been cocited with some other author.

argued that the objections of Ahlgren et al. against the Pearson correlation are mainly of theoretical interest and have little practical relevance, and Bensman (2004) defended the use of the Pearson correlation for statistical inference. Leydesdorff and Vaughan (2006), however, went even further than Ahlgren et al. and asserted that cocitation data should be analyzed directly, without first calculating a similarity measure. This is a point of view with which we do not agree (Waltman & Van Eck, 2007; see also Leydesdorff, 2007). Leydesdorff and Vaughan further argued that it is preferable to analyze citation data rather than cocitation data. Schneider and Borlund (2007a) pointed out that from a statistical perspective the common practice of calculating similarity measures based on cocitation data rather than citation data is quite unorthodox. In addition, they also mentioned some drawbacks of the use of the Pearson correlation as a similarity measure. Despite the objections that have been raised against the use of the Pearson correlation, many researchers still rely on it when measuring the similarity between cocitation profiles (e.g. Liu, 2005; McCain, Verner, Hislop, Evanco, & Cole, 2005; de Moya-Anegón, Herrero-Solana, & Jiménez-Contreras, 2006; Zhao, 2006; Zuccala, 2006; Miguel, de Moya-Anegón, & Herrero-Solana, 2008; Eom, 2008).

In this chapter, our aim is to provide a number of new insights into the methodological discussion about ACA. First of all, we agree with Schneider and Borlund (2007a) that from a statistical perspective calculating similarity measures based on cocitation data rather than citation data is a somewhat unconventional procedure. While the procedure is unconventional, we do not believe that it has any fundamental statistical problems. In our opinion, a statistically valid analysis can be performed using either citation data or cocitation data (although the two types of data may require different similarity measures). In this chapter, like in most of the literature on ACA, we focus our attention on the use of cocitation data. Following Ahlgren et al. (2003), we believe that the use of the Pearson correlation to measure the similarity between authors' cocitation profiles is problematic. Below, we will discuss some shortcomings of the Pearson correlation, most of which have not been mentioned before in the bibliometric literature. Because of these shortcomings, the Pearson correlation is, in our opinion, not a very satisfactory similarity measure for cocitation profiles. We will also discuss what kind of similarity measures may be used as an alternative to the Pearson correlation. Using a well-known author cocitation study by White and McCain (1998) as an example, we will show that the choice of an appropriate similarity measure is not merely of theoretical interest but also has a high practical relevance. Finally, we will comment on the use of similarity measures, in particular the Pearson correlation, for statistical inference. We note that although we concentrate on ACA in this chapter, our observations apply equally well to other kinds of cocitation analysis, such as journal cocitation analysis (McCain, 1991).

3.2 Shortcomings of the Pearson Correlation

Suppose that we have a bibliographic data set and that we are interested in analyzing the cocitations of a set of n authors in this data set. Typically, the analysis is performed as follows (see McCain, 1990 for a detailed discussion and White & Griffith, 1981 and White & McCain, 1998 for well-known examples). First, for each pair of two authors i and j ($i \neq j$), the number of cocitations in the data set, denoted by c_{ij}, is counted. Next, the cocitation counts are used to calculate similarities between the authors. Traditionally, this is done using the Pearson correlation as similarity measure for cocitation profiles. The similarity between authors i and j then has a value between -1 and 1 and is calculated as

$$r(i,j) = \frac{\sum_{k \neq i,j} (c_{ik} - \bar{c}_i)(c_{jk} - \bar{c}_j)}{\sqrt{\sum_{k \neq i,j} (c_{ik} - \bar{c}_i)^2 \sum_{k \neq i,j} (c_{jk} - \bar{c}_j)^2}},$$

where $\bar{c}_i$ and $\bar{c}_j$ denote the averages of, respectively, the cocitation counts c_{ik} and the cocitation counts c_{jk} (for $k \neq i, j$).[2] As a final step, the similarities between the authors are analyzed using multivariate statistical techniques such as multidimensional scaling and hierarchical clustering.

We will now discuss some shortcomings of the Pearson correlation as a similarity measure for cocitation profiles. In the examples that we give, there are $n = 6$ authors. Hence, when comparing two authors, each author's cocitation profile consists of four cocitation counts. Consider first the comparison between two authors, author 1 and author 2, with cocitation profiles $[1\ 2\ 3\ 4]$ and $[10\ 20\ 30\ 40]$, respectively. These cocitation

[2] We have not defined c_{ij} for $i = j$. In the above equation, the Pearson correlation is therefore applied to cocitation profiles of length $n - 2$ rather than length n. The diagonal elements of the cocitation matrix can also be handled in other ways (Ahlgren et al., 2003; White, 2003a), but this is not important for the present discussion.

profiles indicate that authors 1 and 2 have, respectively, 1 and 10 cocitations with author 3, 2 and 20 cocitations with author 4, and so on. Although author 2 has ten times as many cocitations as author 1, the relative frequencies with which authors 1 and 2 are cocited with each of the four other authors are exactly equal. That is, authors 1 and 2 both have 10% of their cocitations with author 3, 20% of their cocitations with author 4, and so on. Since the similarity between two authors should not be influenced by each author's total number of cocitations, authors 1 and 2 should be regarded as perfectly similar. The Pearson correlation does indeed indicate a perfect similarity between the authors, as it has a value of 1 for the above two cocitation profiles. Now consider what happens when author 2's cocitation profile is changed into $[11\ 12\ 13\ 14]$. The Pearson correlation still has a value of 1, which again indicates a perfect similarity between the authors. However, whereas author 1 still has 10% of his cocitations with author 3, 20% of his cocitations with author 4, and so on, author 2 now has his cocitations more or less equally distributed. The cocitation profiles of authors 1 and 2 are therefore quite different, and the Pearson correlation incorrectly indicates a perfect similarity between the authors.

Another interesting example is obtained when authors 1 and 2 have cocitation profiles $[11\ 12\ 13\ 14]$ and $[14\ 13\ 12\ 11]$, respectively. In this example, author 1 has approximately the same number of cocitations with each of the four other authors as author 2. As a consequence, we would expect the similarity between authors 1 and 2 to be quite high. However, the Pearson correlation has a value of -1, and hence the similarity between the authors is as low as possible. As a final example, suppose that authors 1 and 2 have cocitation profiles $[10\ 1\ 0\ 0]$ and $[0\ 0\ 1\ 10]$, respectively. There is then no author with whom authors 1 and 2 have both been cocited. We would therefore expect the similarity between authors 1 and 2 to be as low as possible. However, the Pearson correlation has a value of -0.43, which indicates a low similarity between the authors but not the lowest possible similarity. Comparing the last two examples, we believe that the Pearson correlation gives counterintuitive results. In the first example, authors 1 and 2 have the lowest possible similarity, even though they have both been cocited with all four other authors. In the second example, on the other hand, authors 1 and 2 do not have the lowest possible similarity, even though there is no author with whom they have both been cocited. In other words, in the second example authors 1 and 2

are regarded as more similar than in the first example, even though they have much less similar distributions of their cocitations.

Based on the above examples, we believe that an appropriate similarity measure for cocitation profiles should at least satisfy the following two conditions:

(1) The similarity between two authors is maximal if and only if the authors' cocitation profiles differ by at most a multiplicative constant.

(2) The similarity between two authors is minimal if and only if there is no author with whom the two authors have both been cocited.

The above examples have shown that the Pearson correlation satisfies neither of these conditions. In our opinion, the Pearson correlation is therefore not a very satisfactory similarity measure for cocitation profiles.

From a theoretical point of view, the shortcomings of the Pearson correlation can be explained as follows. In general statistical usage, the Pearson correlation is a measure of the strength of the linear relationship between two random variables. Consequently, when applied to cocitation profiles, the Pearson correlation measures the strength of the linear relationship between the cocitation counts of two authors. The important point is that a strong linear relationship between the cocitation counts of two authors need not imply a high similarity between the authors and, the other way around, that a high similarity between two authors need not imply a strong linear relationship between the cocitation counts of the authors. For example, there is a perfect linear relationship between the cocitation counts $[1\ 2\ 3\ 4]$ and $[11\ 12\ 13\ 14]$, but as we discussed above, we would not regard authors with these cocitation counts as very similar. On the other hand, we would regard authors with the cocitation counts $[10\ 10\ 11\ 11]$ and $[10\ 11\ 10\ 11]$ as very similar, even though there is no linear relationship at all between their cocitation counts (see Schneider & Borlund, 2007a, for a similar example). In summary, the Pearson correlation measures linear relatedness, and because linear relatedness is not the same as similarity, the use of the Pearson correlation as a similarity measure can be problematic.

3.3 Alternatives to the Pearson Correlation

In addition to the Pearson correlation, the cosine is a relatively popular similarity measure for cocitation profiles (see Anderberg, 1973 and Schneider & Borlund, 2007a for a discussion of the relationship between the Pearson correlation and the cosine). Using the cosine, the similarity between authors i and j has a value between 0 and 1 and is calculated as

$$\cos(i,j) = \frac{\sum_{k \neq i,j} c_{ik} c_{jk}}{\sqrt{\sum_{k \neq i,j} c_{ik}^2 \sum_{k \neq i,j} c_{jk}^2}}. \tag{3.1}$$

Unlike the Pearson correlation, the cosine satisfies the two conditions introduced in the previous section (see Proposition 3.1 in Appendix 3.A). Both the Pearson correlation and the cosine have the property that multiplying an author's cocitation profile by an arbitrary constant has no effect on the author's similarity with other authors (Anderberg, 1973). This is called the property of coordinate-wise scale invariance by Ahlgren et al. (2003). It is an indispensable property for any similarity measure for cocitation profiles, since it guarantees that the similarity between two authors is not influenced by each author's total number of cocitations. In other words, it guarantees that the similarity between two authors depends only on the relative frequencies with which the authors are cocited with other authors.

Because of the property of coordinate-wise scale invariance, the similarity between two authors calculated using a measure such as the Pearson correlation or the cosine does not change when the authors' cocitation profiles are normalized to sum to one. That is, the values of the Pearson correlation and the cosine do not change when the c_{ik}s and c_{jk}s in the equations provided above are replaced by p_{ik}s and p_{jk}s that are given by

$$p_{ik} = \frac{c_{ik}}{\sum_{k' \neq i,j} c_{ik'}} \qquad \text{and} \qquad p_{jk} = \frac{c_{jk}}{\sum_{k' \neq i,j} c_{jk'}}.$$

Interestingly, these p_{ik}s and p_{jk}s have a natural interpretation in probabilistic terms. p_{ik} (p_{jk}) can be interpreted as the probability that a randomly drawn cocitation of author i (j) is a cocitation with author k. Under this interpretation, the normalized cocitation profile of author i (j) is a probability distribution that indicates the probability of author i (j) being cocited with each of the other authors. Hence, when we are comparing the cocitation profiles of two authors, what we are in fact doing is comparing the probability distributions of each of the authors' cocitations.

The interpretation of cocitation profiles as probability distributions is especially interesting because it provides new insights into the question of what might be useful similarity measures for ACA. It can now be seen that a natural approach to this question is to have a look at some well-known similarity measures for probability distributions. We first note that the use of the Pearson correlation or the cosine to measure the similarity between probability distributions is very uncommon. For the Pearson correlation this is not surprising, since the Pearson correlation does not satisfy two basic requirements that one would expect to be satisfied by any reasonable similarity measure for probability distributions. These are the requirements that the value of the similarity measure is maximal if and only if two distributions are identical and that it is minimal if and only if two distributions are non-overlapping. The cosine, however, does satisfy these requirements. (This follows from Proposition 3.1.) We therefore do not see any theoretical objections against the use of the cosine as a similarity measure for probability distributions, even though it is rather unusual to use the cosine in this way. Perhaps the most popular similarity measure for probability distributions is the Kullback-Leibler divergence (Kullback & Leibler, 1951) from the field of information theory. However, this similarity measure has difficulties with zero probabilities and hence with zero cocitation counts. As a consequence, the measure is not very useful for ACA. The Jensen-Shannon divergence (Lin, 1991), which is closely related to the Kullback-Leibler divergence, does not have these difficulties and is therefore more interesting from the point of view of ACA.[3] Based on the Jensen-Shannon divergence, the similarity between authors i and j can be calculated as

$$\mathrm{JS}(i,j) = 1 - \frac{1}{2}\left(\sum_{k \neq i,j} p_{ik} \log \frac{p_{ik}}{\bar{p}_k}\right) - \frac{1}{2}\left(\sum_{k \neq i,j} p_{jk} \log \frac{p_{jk}}{\bar{p}_k}\right), \tag{3.2}$$

where the logarithm has base 2 and where $0 \log 0$ and $0 \log(0/0)$ are defined as 0. Fur-

[3] Both the Kullback-Leibler divergence and the Jensen-Shannon divergence are in fact measures of the *dissimilarity* between probability distributions. For the present discussion, the difference between similarity and dissimilarity measures is not important, and we therefore refer to all measures as similarity measures. Leydesdorff (2005) also studies the use of information-theoretic similarity measures in ACA, in particular in the context of clustering algorithms.

thermore, $\bar{p}_k = (p_{ik} + p_{jk})/2$. Another well-known similarity measure for probability distributions is the Bhattacharyya distance (Bhattacharyya, 1943). This is a popular similarity measure in pattern recognition and related fields. Using the Bhattacharyya distance, the similarity between authors i and j is calculated as

$$\mathrm{B}(i,j) = \sum_{k \neq i,j} \sqrt{p_{ik} p_{jk}}. \tag{3.3}$$

JS(i,j) and B(i,j) both have a value between 0 and 1. They have a value of 1 if and only if the probability distributions given by the p_{ik}s and p_{jk}s are identical, and they have a value of 0 if and only if these distributions are non-overlapping (see Propositions 3.2 and 3.3 in Appendix 3.A). It follows from this that JS(i,j) and B(i,j) both satisfy the two conditions introduced in the previous section. In addition to the similarity measures mentioned above, there are a number of other similarity measures that are sometimes used to compare probability distributions. In the rest of this chapter, however, we focus our attention on the above-mentioned similarity measures.

3.4 Practical Relevance

White (2003a) argues that theoretical shortcomings of the Pearson correlation are problematic only if there is a substantive difference between results based on the Pearson correlation and results based on theoretically sound similarity measures. We agree with this reasoning. However, contrary to White, we believe that such substantive differences do indeed exist. To show the existence of these differences, we take a well-known author cocitation study by White and McCain (1998) as an example. Among other things, White and McCain provide a multidimensional scaling map of the similarities between the top 100 authors in the field of information science in the period 1988–1995. They use the Pearson correlation to calculate the similarities between the authors. Using the ALSCAL program in SPSS, we replicated the analysis of White and McCain and obtained the map shown in Figure 3.1. This map is almost identical to the one provided by White and McCain. (Where the two maps are different, this is most likely due to slight differences in the way in which the cocitation data was collected and preprocessed.) In addition to the map in Figure 3.1, we constructed three more maps. In these maps, the similarities between the authors were calculated based on the cosine, the Jensen-Shannon divergence, and the Bhattacharyya distance. The maps obtained using the cosine and the Jensen-Shannon divergence are shown in Figures 3.2 and 3.3, respectively. The map obtained using the Bhattacharyya distance turned out to be almost identical to the map obtained using the Jensen-Shannon divergence and is therefore not shown.

Comparing the map in Figure 3.1 with the maps in Figures 3.2 and 3.3, it is immediately apparent that there is a substantive difference between results based on the Pearson correlation and results based on theoretically sound similarity measures such as the cosine and the Jensen-Shannon divergence. In the map in Figure 3.1, there is a clear division of the authors into two clusters, a cluster of domain analysis authors and a cluster of information retrieval authors. The clusters are located on opposite sides of the map, and only a small number of authors are located in between the clusters. Hence, based on the map in Figure 3.1, information science appears to be a field consisting of two subfields, domain analysis and information retrieval, that are almost completely separated from each other. Now consider the maps in Figures 3.2 and 3.3. In these maps, the clustering of authors is either much less pronounced than in the map in Figure 3.1 or there is no clustering at all. Although a number of typical domain analysis authors are located in the far left part of the maps in Figures 3.2 and 3.3 and a number of typical information retrieval authors in the far right part, many authors are located somewhere in between the extremes. Consequently, based on these maps, information science appears to be a fairly unified field with a substantial number of connections between its two main subfields, domain analysis and information retrieval. This is a very different picture of the information science field than the picture that emerges from the map in Figure 3.1. So, contrary to some earlier research (Leydesdorff & Zaal, 1988), we find that different similarity measures can lead to quite different interpretations.

There are two remarks that we would like to make. Both remarks are based on a paper by White (2003b) in which he uses pathfinder networks to perform an author cocitation study of the information science field. First, White (2003b, p. 427) does not seem to be completely satisfied with the maps of the information science field provided in White and McCain (1998). In particular, he expresses some concerns about the "empty centers" that appear in these maps (also visible in the map in Figure 3.1). He further notes that the appearance of "empty centers" is not confined to information sci-

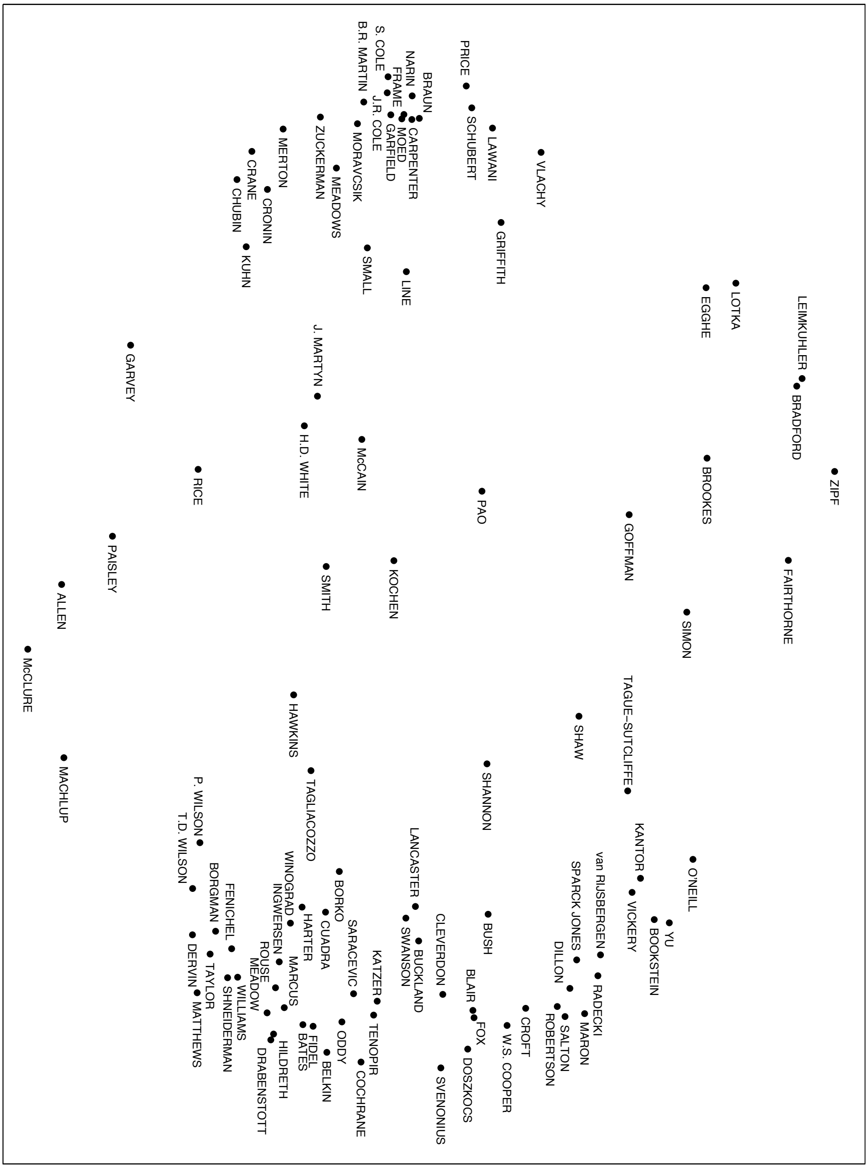


Figure 3.1: Map obtained using the Pearson correlation.

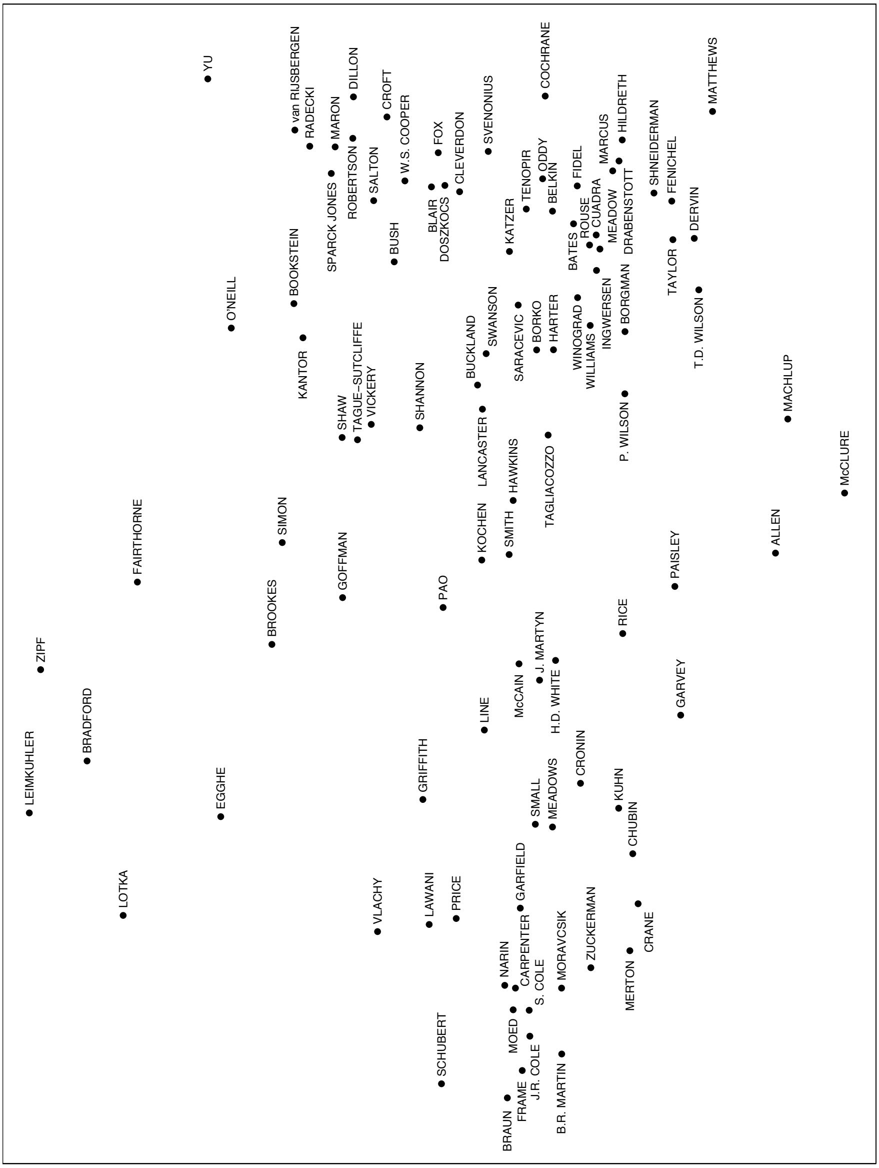


Figure 3.2: Map obtained using the cosine.

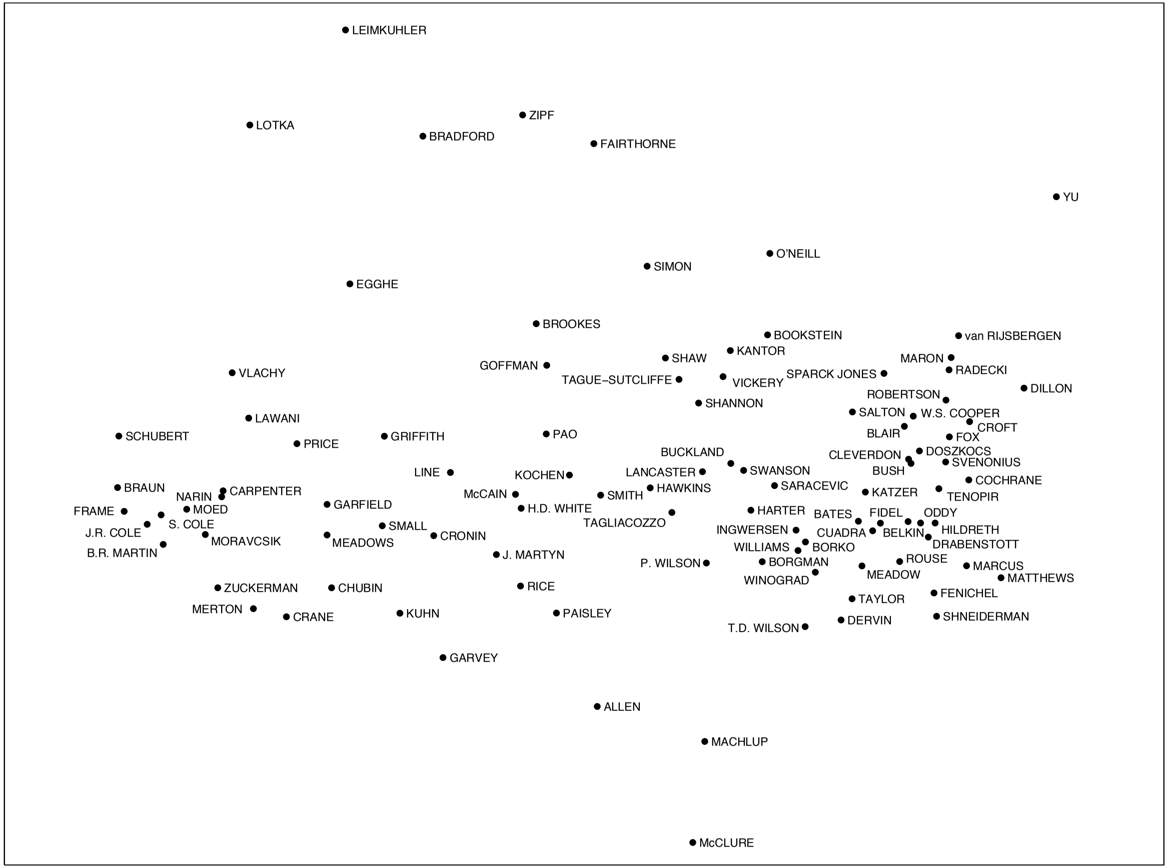


Figure 3.3: Map obtained using the Jensen-Shannon distance.

ence but also happens when mapping other heterogeneous fields. White seems to prefer maps based on pathfinder networks because such maps do not have “empty centers”. Interestingly, our results seem to indicate that the issue of the “empty centers” can simply be resolved by using a theoretically sound similarity measure, such as the cosine or the Jensen-Shannon divergence, instead of the Pearson correlation. Our second and related remark is concerned with White’s statement that “the ‘empty center’ should be recognized as a metaphor growing out of the (multidimensional scaling) mapping algorithm” (White, 2003b, p. 427). Our results point in a different direction. The issue of the “empty centers” seems to be caused by the use of the Pearson correlation as a similarity measure for cocitation profiles rather than by the use of multidimensional scaling as a mapping technique for author similarities.

3.5 Statistical Inference

Under certain assumptions, it is possible to use the Pearson correlation for statistical inference. For example, as discussed in almost every statistical textbook, a t test can be used to test the hypothesis that the population correlation equals zero. Statistical packages such as SPSS typically report the p value of this test and use it to indicate whether a Pearson correlation is significant. More elaborate possibilities for statistical inference are obtained by applying the Fisher transformation to the Pearson correlation (e.g. Snedecor & Cochran, 1989). Using the Fisher transformation, one can determine a confidence interval for the population correlation and one can test various hypotheses, such as the hypothesis that the population correlation equals a particular value (not necessarily zero) and the hypothesis that two sample correlations are estimates of the same population correlation.

According to some researchers, the Pearson correlation has an advantage over other similarity measures because of the possibility of using it for statistical inference. Recently, the use of the Pearson correlation for statistical inference in ACA was defended by Bensman (2004). In addition, Leydesdorff customarily takes into account the significance of Pearson correlations in his work on ACA (e.g. Leydesdorff & Vaughan, 2006; Leydesdorff, 2007). In our opinion, however, there are three reasons why the possibility of statistical inference does not give the Pearson correlation an advantage over other similarity measures.

First, it is well-known that the distributional assumptions underlying the use of the Pearson correlation for statistical inference are not met in ACA (e.g. Ahlgren et al., 2003; White, 2003a). For example, the t test for the significance of the Pearson correlation between two random variables assumes that at least one of the two variables is normally distributed (e.g. Snedecor & Cochran, 1989). Since cocitation counts have discrete distributions that are typically highly skewed (e.g. Ahlgren et al., 2003; White, 2003a), this assumption is violated in ACA. Bensman (2004) claims that the Pearson correlation is distributionally robust and that a violation of the assumption of normality therefore does not make much difference. When checking Bensman's claim in the statistical literature, there turns out to be no consensus on this issue (see Kowalski, 1972, for an overview of the relevant literature). A number of early Monte Carlo studies (e.g. Pearson, 1931) conclude that the distribution of the sample correlation is quite insensitive to violations of the assumption of normally distributed variables, especially when the population correlation equals zero. A number of other studies contradict this conclusion, sometimes even for population correlations equal to zero. In particular, a Monte Carlo study by Kowalski (1972), which seems to be one of the most recent studies of the robustness of the Pearson correlation that is available, indicates that the distribution of the sample correlation can be sensitive to violations of the assumption of normally distributed variables even when the population correlation equals zero. It follows from this result that the t test for the significance of the Pearson correlation between two variables may not be very accurate when the variables are both non-normally distributed, as is the case in ACA.[4] In our opinion, it is therefore better not to use the t test in ACA.[5]

Second, even if an appropriate statistical test is used, it is not clear what it means

[4] Another Monte Carlo study is performed by Duncan and Layard (1973). Their results seem to indicate that in many cases standard tests for the significance of a Pearson correlation perform quite well when variables are non-normally distributed. This seems to contradict the results of Kowalski. However, the results of Duncan and Layard apply only to two-tailed tests at the 5% significance level. We performed some Monte Carlo simulations ourselves (with lognormally distributed variables) and found that especially the performance of one-tailed tests and tests at low significance levels (e.g., 1%) can be problematic. This confirms the results of Kowalski.

[5] As pointed out by White (2004), instead of a t test, a randomization test (also called a permutation test) can be used to test the significance of a Pearson correlation. For a description of such a test, we refer to Edgington (1995) and to the statistical textbook by Stout, Marden, and Travers (2000). Unlike a t test, a randomization test does not assume normally distributed variables.

to know that the Pearson correlation between the cocitation counts of two authors is significantly greater than zero (or significantly different from zero). On the one hand, a positive correlation is not necessary for a high similarity between two authors. As we discussed earlier, we would regard authors with the cocitation counts $[10\ 10\ 11\ 11]$ and $[10\ 11\ 10\ 11]$ as very similar, even though the correlation between their cocitation counts equals zero. On the other hand, a positive correlation is also not sufficient for a high similarity between two authors. We would not regard authors with the cocitation counts $[1\ 2\ 3\ 4]$ and $[11\ 12\ 13\ 14]$ as very similar, even though the correlation between their cocitation counts equals one. So, a positive correlation is neither necessary nor sufficient for a high similarity. Conversely, a correlation of zero is neither necessary nor sufficient for a low similarity (or for no similarity at all). It is therefore not clear why one would be interested to know whether the correlation between the cocitation counts of two authors is significantly greater than zero.

Third, all similarity measures can be used for statistical inference, not only the Pearson correlation. One way to do this is to use a statistical technique called bootstrapping. Bootstrapping is a generally applicable computer-intensive technique that can be used to calculate standard errors and confidence intervals and to test hypotheses. It replaces traditional statistical analysis by a considerable amount of computation and can be applied to problems for which a theoretical analysis either is too complicated or requires very demanding assumptions. Bootstrapping is a popular statistical technique in many scientific fields, but in bibliometric research there seem to be almost no studies in which it has been used (see Van Eck & Waltman, 2007a, for an exception). For introductions to bootstrapping, we refer to Efron and Tibshirani (1986, 1993) and to the statistical textbook by Stout et al. (2000).

3.6 Conclusion

We have argued that the Pearson correlation has some shortcomings as a measure of the similarity between cocitation profiles. As a consequence, the use of the Pearson correlation in ACA is, in our opinion, not very satisfactory. The cosine does not have the shortcomings of the Pearson correlation, and we therefore regard it as a more appropriate similarity measure for cocitation profiles. The interpretation of cocitation profiles as probability distributions suggests other similarity measures that may be useful for ACA. In particular, similarity measures based on the Jensen-Shannon divergence or the Bhattacharyya distance may be considered. In an author cocitation study of the field of information science, the Pearson correlation gives results that are quite different from results obtained using theoretically sound similarity measures. This shows that the choice of an appropriate similarity measure is not merely of theoretical interest but also has a high practical relevance. We have further argued that the possibility of statistical inference does not give the Pearson correlation an advantage over other similarity measures.

There is one final remark that we would like to make. Our objections against the use of the Pearson correlation in ACA apply only to situations in which the Pearson correlation is interpreted as a similarity measure. The Pearson correlation is usually interpreted in this way. This is for example the case when Pearson correlations are used as input to multidimensional scaling or hierarchical clustering. Sometimes, however, the Pearson correlation is not interpreted as a similarity measure but as a measure of linear relatedness. This is in particular the case when Pearson correlations are used in the context of factor analysis. Our objections against the use of the Pearson correlation do not apply to such situations.

3.A Appendix

In this appendix, we provide proofs of some results mentioned in the chapter.

Proposition 3.1. The cosine similarity between two authors, which is defined in (3.1), has a value of at most 1, and its value is 1 if and only if the authors' cocitation profiles differ by at most a multiplicative constant. The cosine similarity between two authors has a value of at least 0, and its value is 0 if and only if there is no third author with whom the authors have both been cocited.

Proof. Consider two authors i and j. It is an immediate consequence of Cauchy's inequality that

$$\left(\sum_{k\neq i,j} c_{ik}c_{jk}\right)^2 \leq \sum_{k\neq i,j} c_{ik}^2 \sum_{k\neq i,j} c_{jk}^2,$$

with equality if and only if the cocitation profiles differ by at most a multiplicative constant. Consequently,

$$\sum_{k \neq i,j} c_{ik} c_{jk} \leq \sqrt{\sum_{k \neq i,j} c_{ik}^2 \sum_{k \neq i,j} c_{jk}^2},$$

again with equality if and only if the cocitation profiles differ by at most a multiplicative constant. It now follows that

$$\cos(i,j) = \frac{\sum_{k \neq i,j} c_{ik} c_{jk}}{\sqrt{\sum_{k \neq i,j} c_{ik}^2 \sum_{k \neq i,j} c_{jk}^2}} \leq 1$$

and, more specifically, that $\cos(i,j) = 1$ if and only if the cocitation profiles differ by at most a multiplicative constant. This proves the first part of the proposition. The second part of the proposition is trivial. □

Proposition 3.2. $\mathrm{JS}(i,j)$ defined in (3.2) has a value of at most 1, and its value is 1 if and only if the probability distributions given by the p_{ik}s and p_{jk}s are identical. $\mathrm{JS}(i,j)$ has a value of at least 0, and its value is 0 if and only if the probability distributions given by the p_{ik}s and p_{jk}s are non-overlapping.

Proof. Note that

$$\sum_{k \neq i,j} p_{ik} \log \frac{p_{ik}}{\bar{p}_k} \geq 0,$$

with equality if and only if $p_{ik} = \bar{p}_k$ for all $k \neq i,j$. This follows from the observation that the left-hand side denotes the Kullback-Leibler divergence between two probability distributions (or, equivalently, it follows from Gibbs' inequality) (e.g. MacKay, 2003). Similarly,

$$\sum_{k \neq i,j} p_{jk} \log \frac{p_{jk}}{\bar{p}_k} \geq 0,$$

with equality if and only if $p_{jk} = \bar{p}_k$ for all $k \neq i,j$. It can now be seen that

$$\mathrm{JS}(i,j) = 1 - \frac{1}{2}\left(\sum_{k \neq i,j} p_{ik} \log \frac{p_{ik}}{\bar{p}_k}\right) - \frac{1}{2}\left(\sum_{k \neq i,j} p_{jk} \log \frac{p_{jk}}{\bar{p}_k}\right) \leq 1$$

and, more specifically, that $\mathrm{JS}(i,j) = 1$ if and only if $p_{ik} = p_{jk}$ for all $k \neq i,j$. This proves the first part of the proposition.

Consider an author $k \neq i, j$. Obviously,

$$\log \frac{p_{ik}}{\bar{p}_k} = -\log \frac{\bar{p}_k}{p_{ik}} = -\log \frac{p_{ik} + p_{jk}}{2p_{ik}} = -\log \frac{1}{2}\left(1 + \frac{p_{jk}}{p_{ik}}\right) = 1 - \log\left(1 + \frac{p_{jk}}{p_{ik}}\right) \leq 1,$$

with equality if and only if $p_{jk} = 0$. Similarly,

$$\log \frac{p_{jk}}{\bar{p}_k} \leq 1,$$

with equality if and only if $p_{ik} = 0$. It can now be seen that

$$\mathrm{JS}(i, j) = 1 - \frac{1}{2}\left(\sum_{k \neq i,j} p_{ik} \log \frac{p_{ik}}{\bar{p}_k}\right) - \frac{1}{2}\left(\sum_{k \neq i,j} p_{jk} \log \frac{p_{jk}}{\bar{p}_k}\right) \geq 0$$

and, more specifically, that $\mathrm{JS}(i, j) = 0$ if and only if for each $k \neq i, j$ either $p_{ik} = 0$ or $p_{jk} = 0$. This proves the second part of the proposition. □

Proposition 3.3. $\mathrm{B}(i, j)$ defined in (3.3) has a value of at most 1, and its value is 1 if and only if the probability distributions given by the p_{ik}s and p_{jk}s are identical. $\mathrm{B}(i, j)$ has a value of at least 0, and its value is 0 if and only if the probability distributions given by the p_{ik}s and p_{jk}s are non-overlapping.

Proof. Consider an author $k \neq i, j$. Obviously,

$$\frac{1}{4}(p_{ik} + p_{jk})^2 - p_{ik}p_{jk} = \frac{1}{4}(p_{ik} - p_{jk})^2 \geq 0,$$

with equality if and only if $p_{ik} = p_{jk}$. Consequently,

$$p_{ik}p_{jk} \leq \frac{1}{4}(p_{ik} + p_{jk})^2,$$

and hence

$$\sqrt{p_{ik}p_{jk}} \leq \frac{1}{2}(p_{ik} + p_{jk}),$$

again with equality if and only if $p_{ik} = p_{jk}$. It now follows that

$$\mathrm{B}(i, j) = \sum_{k \neq i,j} \sqrt{p_{ik}p_{jk}} \leq \frac{1}{2} \sum_{k \neq i,j} (p_{ik} + p_{jk}) = 1$$

and, more specifically, that $\mathrm{B}(i,j) = 1$ if and only if $p_{ik} = p_{jk}$ for all $k \neq i,j$. This proves the first part of the proposition. The second part of the proposition is trivial. □

Chapter 4

How to Normalize Co-Occurrence Data? An Analysis of Some Well-Known Similarity Measures[*]

Abstract

In scientometric research, the use of co-occurrence data is very common. In many cases, a similarity measure is employed to normalize the data. However, there is no consensus among researchers on which similarity measure is most appropriate for normalization purposes. In this chapter, we theoretically analyze the properties of similarity measures for co-occurrence data, focusing in particular on four well-known measures: the association strength, the cosine, the inclusion index, and the Jaccard index. We also study the behavior of these measures empirically. Our analysis reveals that there exist two fundamentally different types of similarity measures, namely set-theoretic measures and probabilistic measures. The association strength is a probabilistic measure, while the cosine, the inclusion index, and the Jaccard index are set-theoretic measures. Both our theoretical and our empirical results indicate that co-occurrence data can best be normalized using a probabilistic measure. This provides strong support for the use of the association strength in scientometric research.

[*] This chapter is based on Van Eck and Waltman (2009). During the preparation of the final version of this thesis, we became aware of the work of Leicht, Holme, and Newman (2006; see also Newman, 2010, Section 7.12). Leicht et al. present some arguments that are similar to the ones presented in this chapter.

4.1 Introduction

The use of co-occurrence data is very common in scientometric research. Co-occurrence data can be used for a multitude of purposes. Co-citation data, for example, can be used to study relations among authors or journals, co-authorship data can be used to study scientific cooperation, and data on co-occurrences of words can be used to construct so-called co-word maps, which are maps that provide a visual representation of the structure of a scientific field. Usually, when co-occurrence data is used, a transformation is first applied to the data. The aim of such a transformation is to derive similarities from the data or, more specifically, to normalize the data. For example, when researchers study relations among authors based on co-citation data, they typically derive similarities from the data and then analyze these similarities using multivariate analysis techniques such as multidimensional scaling and hierarchical clustering (e.g., McCain, 1990; White & Griffith, 1981; White & McCain, 1998). Likewise, when researchers use co-authorship data to study scientific cooperation, they typically apply a normalization to the data and then base their analysis on the normalized data (e.g., Glänzel, 2001; Luukkonen, Persson, & Sivertsen, 1992; Luukkonen, Tijssen, Persson, & Sivertsen, 1993).

In this chapter, our focus is methodological. We study various measures for deriving similarities from co-occurrence data. Basically, there are two approaches that can be taken to derive similarities from co-occurrence data. We refer to these approaches as the direct and the indirect approach, but the approaches are also known as the local and the global approach (Ahlgren et al., 2003; Jarneving, 2008). Similarity measures that implement the direct approach are referred to as direct similarity measures in this chapter, while similarity measures that implement the indirect approach are referred to as indirect similarity measures.

The indirect approach to derive similarities from co-occurrence data relies on co-occurrence profiles. The co-occurrence profile of an object is a vector that contains the number of co-occurrences of the object with each other object. Indirect similarity measures determine the similarity between two objects by comparing the co-occurrence profiles of the objects. The indirect approach is mainly used for co-citation data (e.g., McCain, 1990, 1991; White & Griffith, 1981; White & McCain, 1998). From a theoretical point of view, the approach is quite well understood (Ahlgren et al., 2003; Van Eck & Waltman, 2008).

In this chapter, we focus most of our attention on the direct approach to derive similarities from co-occurrence data. Direct similarity measures determine the similarity between two objects by taking the number of co-occurrences of the objects and adjusting this number for the total number of occurrences or co-occurrences of each of the objects. Researchers use several different direct similarity measures. The cosine and the Jaccard index are especially popular, but other measures are also regularly used. However, relatively little is known about the theoretical properties of the various measures. Also, there is no consensus among researchers on which measure is most appropriate for a particular purpose. In this chapter, we theoretically analyze some well-known direct similarity measures and we compare their properties. We also study the behavior of the measures empirically. Usually, when a direct similarity measure is applied to co-occurrence data, the purpose is to normalize the data, that is, to correct the data for differences in the total number of occurrences or co-occurrences of objects. The main question that we try to answer in this chapter is therefore as follows: Which direct similarity measures are appropriate for normalizing co-occurrence data and which are not? Interestingly, despite their popularity, the cosine and the Jaccard index turn out not to be appropriate measures for normalization purposes. We argue that an appropriate measure for normalizing co-occurrence data is the association strength (Van Eck & Waltman, 2007a; Van Eck, Waltman, Van den Berg, & Kaymak, 2006a), also referred to as the proximity index (e.g., Peters & Van Raan, 1993b; Rip & Courtial, 1984) or the probabilistic affinity index (e.g., Zitt, Bassecoulard, & Okubo, 2000). Although this measure is somewhat less well-known, it turns out to have the right theoretical properties for normalizing co-occurrence data.

This chapter is organized as follows. We first provide an overview of the most popular direct similarity measures. We then analyze these measures theoretically. We also look for empirical relations among the measures. Finally, we answer the question which direct similarity measures are appropriate for normalizing co-occurrence data and which are not.

4.2 Overview of Direct Similarity Measures

In this section, we provide an overview of the most popular direct similarity measures. The overview is based on a survey of the scientometric literature.

We first introduce some mathematical notation. Let $\mathbf{O}$ denote an occurrence matrix of order $m \times n$. The columns of $\mathbf{O}$ correspond with the objects of which we want to analyze the co-occurrences. There are n such objects, denoted by $1, \ldots, n$. The objects can be, for example, authors (e.g., White & McCain, 1998), countries (e.g., Glänzel, 2001; Zitt et al., 2000), documents (e.g., Gmür, 2003; Klavans & Boyack, 2006b), journals (e.g., Boyack, Klavans, & Börner, 2005; Klavans & Boyack, 2006a), Web pages (e.g., Vaughan, 2006; Vaughan & You, 2006), or words (e.g., Kopcsa & Schiebel, 1998). The rows of $\mathbf{O}$ usually correspond with documents. m then denotes the number of documents on which the co-occurrence analysis is based. Sometimes the rows of $\mathbf{O}$ do not correspond with documents. In Web co-link analysis, for example, the rows of $\mathbf{O}$ correspond with Web pages (e.g., Vaughan, 2006; Vaughan & You, 2006). Throughout this chapter, however, we assume for simplicity that the rows of $\mathbf{O}$ always correspond with documents. Another assumption that we make is that $\mathbf{O}$ is a binary matrix, that is, each element of $\mathbf{O}$ equals either zero or one. Let o_{ki} denote the element in the kth row and ith column of $\mathbf{O}$. o_{ki} equals one if object i occurs in the document that corresponds with the kth row of $\mathbf{O}$, and it equals zero otherwise. Let $\mathbf{C}$ denote the co-occurrence matrix of the objects $1, \ldots, n$. $\mathbf{C}$ is a symmetric non-negative matrix of order $n \times n$. Let c_{ij} denote the element in the ith row and jth column of $\mathbf{C}$. For $i \neq j$, c_{ij} equals the number of co-occurrences of objects i and j. For $i = j$, c_{ij} equals the number of occurrences of object i. Clearly, for all i and j,

$$c_{ij} = \sum_{k=1}^{m} o_{ki} o_{kj}. \tag{4.1}$$

It follows from this that $\mathbf{C} = \mathbf{O}^{\mathrm{T}}\mathbf{O}$, where $\mathbf{O}^{\mathrm{T}}$ denotes the transpose of $\mathbf{O}$. Moreover, the assumption that $\mathbf{O}$ is a binary matrix implies that $\mathbf{C}$ is an integer matrix.

As we discussed in the introduction, there are two types of measures for determining similarities between objects based on co-occurrence data. We refer to these two types of measures as direct similarity measures and indirect similarity measures. Indirect similarity measures, also known as global similarity measures (Ahlgren et al., 2003; Jarneving, 2008), determine the similarity between two objects i and j by comparing the ith and the jth row (or column) of the co-occurrence matrix $\mathbf{C}$. The more similar the co-occurrence profiles in these two rows (or columns) of $\mathbf{C}$, the higher the similarity between i and j. Indirect similarity measures are especially popular for author co-citation analysis (e.g., McCain, 1990; White & Griffith, 1981; White & McCain, 1998) and journal co-citation analysis (e.g., McCain, 1991). We refer to Ahlgren et al. (2003) and Van Eck and Waltman (2008) for a detailed discussion of the properties of various indirect similarity measures. In this chapter, we focus most of our attention on direct similarity measures, also known as local similarity measures (Ahlgren et al., 2003; Jarneving, 2008). Direct similarity measures determine the similarity between two objects i and j by taking the number of co-occurrences of i and j and adjusting this number for the total number of occurrences or co-occurrences of i and the total number of occurrences or co-occurrences of j. We note that in some studies similarities between objects are determined by comparing columns of the occurrence matrix $\mathbf{O}$ (e.g., Leydesdorff & Vaughan, 2006; Schneider, Larsen, & Ingwersen, 2009). In most cases, this approach is mathematically equivalent to the use of a direct similarity measure.[1]

Let s_i denote either the total number of occurrences of object i or the total number of co-occurrences of object i. In the first case we have

$$s_i = c_{ii} = \sum_{k=1}^{m} o_{ki}, \tag{4.2}$$

while in the second case we have

$$s_i = \sum_{j=1, j \neq i}^{n} c_{ij}. \tag{4.3}$$

Both definitions are used in scientometric research (see also Leydesdorff, 2008), but the first definition seems to be more popular. We now provide a formal definition of a direct similarity measure.

[1] Leydesdorff and Vaughan (2006) and Schneider et al. (2009) use the Pearson correlation to compare columns of the occurrence matrix $\mathbf{O}$. As shown by Guilford (1973), applying the Pearson correlation to a binary occurrence matrix is mathematically equivalent to applying the so-called phi coefficient to the corresponding co-occurrence matrix.

Definition 4.1. A *direct similarity measure* is defined as a function $S(c_{ij}, s_i, s_j)$ that has the following three properties:

- The domain of $S(c_{ij}, s_i, s_j)$ equals

  $$D_S = \left\{ (c_{ij}, s_i, s_j) \in \mathbb{R}^3 \,|0 \leq c_{ij} \leq \min(s_i, s_j) \text{ and } s_i, s_j > 0 \right\}, \qquad (4.4)$$

  where $\mathbb{R}$ denotes the set of all real numbers.

- The range of $S(c_{ij}, s_i, s_j)$ is a subset of $\mathbb{R}$.

- $S(c_{ij}, s_i, s_j)$ is symmetric in s_i and s_j, that is $S(c_{ij}, s_i, s_j) = S(c_{ij}, s_j, s_i)$, for all $(c_{ij}, s_i, s_j) \in D_S$.

Based on this definition, a number of observations can be made. First, the definition does not require that c_{ij}, s_i, and s_j have integer values. Allowing for non-integer values of c_{ij}, s_i, and s_j simplifies the mathematical analysis of direct similarity measures. Second, even though most direct similarity measures take values between zero and one, the definition allows measures to have a different range. And third, because the definition requires direct similarity measures to be symmetric in s_i and s_j, it does not cover asymmetric similarity measures such as those discussed by (Egghe & Michel, 2002, 2003). As a final observation, we note that Definition 4.1 is quite general. More specific definitions for special classes of direct similarity measures will be provided later on in this chapter. We now define the notion of monotonic relatedness of direct similarity measures.

Definition 4.2. Two direct similarity measures $S_1(c_{ij}, s_i, s_j)$ and $S_2(c_{ij}, s_i, s_j)$ are said to be *monotonically related* if and only if

$$S_1(c_{ij}, s_i, s_j) < S_1(c'_{ij}, s'_i, s'_j) \Leftrightarrow S_2(c_{ij}, s_i, s_j) < S_2(c'_{ij}, s'_i, s'_j) \qquad (4.5)$$

for all $(c_{ij}, s_i, s_j), (c'_{ij}, s'_i, s'_j) \in D_S$.

Monotonic relatedness of direct similarity measures is important because certain multivariate analysis techniques that are frequently used in scientometric research are insensitive to monotonic transformations of similarities. This is for example the case for ordinal or non-metric multidimensional scaling (e.g., Borg & Groenen, 2005) and for single linkage and complete linkage hierarchical clustering (e.g., Anderberg, 1973).

Based on a survey of the literature, we have identified the most popular direct similarity measures in the field of scientometrics. These measures are defined as

$$S_{\mathrm{A}}(c_{ij}, s_i, s_j) = \frac{c_{ij}}{s_i s_j}, \tag{4.6}$$

$$S_{\mathrm{C}}(c_{ij}, s_i, s_j) = \frac{c_{ij}}{\sqrt{s_i s_j}}, \tag{4.7}$$

$$S_{\mathrm{I}}(c_{ij}, s_i, s_j) = \frac{c_{ij}}{\min(s_i, s_j)}, \tag{4.8}$$

$$S_{\mathrm{J}}(c_{ij}, s_i, s_j) = \frac{c_{ij}}{s_i + s_j - c_{ij}}. \tag{4.9}$$

We refer to these measures as, respectively, the association strength, the cosine, the inclusion index, and the Jaccard index. Assuming that c_{ij} is an integer, each of the measures takes values between zero and one. Moreover, it is not difficult to see that the measures satisfy

$$S_{\mathrm{A}}(c_{ij}, s_i, s_j) \leq S_{\mathrm{J}}(c_{ij}, s_i, s_j) \leq S_{\mathrm{C}}(c_{ij}, s_i, s_j) \leq S_{\mathrm{I}}(c_{ij}, s_i, s_j). \tag{4.10}$$

We now discuss each of the measures.

The association strength defined in (4.6) is used by Van Eck and Waltman (2007a) and Van Eck, Waltman, et al. (2006a).[2] Under various names, the measure is also used in a number of other studies. Hinze (1994), Leclerc and Gagné (1994), Peters and Van Raan (1993b), and Rip and Courtial (1984) refer to the measure as the proximity index, while Leydesdorff (2008) and Zitt et al. (2000) refer to it as the probabilistic affinity (or activity) index. The measure is also employed by Luukkonen et al. (1992, 1993), but in their work it does not have a name. The association strength is proportional to the ratio between on the one hand the observed number of co-occurrences of objects i and j and on the other hand the expected number of co-occurrences of objects i and

[2] The definition of the association strength used in these papers differs slightly from the definition provided in (4.6). However, since the two definitions are proportional to each other, the difference between them is not important. Throughout this section, direct similarity measures that are proportional to each other will simply be regarded as equivalent.

j under the assumption that occurrences of i and j are statistically independent. We will come back to this interpretation later on in this chapter. The association strength corresponds with the pseudo-cosine measure discussed by Jones and Furnas (1987) and is monotonically related to the (pointwise) mutual information measure used in the field of computational linguistics (e.g., Church & Hanks, 1990; Manning & Schütze, 1999). Measures equivalent to the association strength sometimes also appear outside the field of scientometrics (T. F. Cox & Cox, 2001; M. A. A. Cox & Cox, 2008; Hubálek, 1982).

The cosine defined in (4.7) equals the ratio between on the one hand the number of times that objects i and j are observed together and on the other hand the geometric mean of the number of times that object i is observed and the number of times that object j is observed. The measure can be interpreted as the cosine of the angle between the ith and the jth column of the occurrence matrix $\mathbf{O}$, where the columns of $\mathbf{O}$ are regarded as vectors in an m-dimensional space (e.g., Salton & McGill, 1983). The cosine seems to be the most popular direct similarity measure in the field of scientometrics. Frequently cited studies in which the measure is used include Braam, Moed, and Van Raan (1991b, 1991a), Klavans and Boyack (2006a), Leydesdorff (1989), Peters and Van Raan (1993a), Peters, Braam, and Van Raan (1995), Small (1994), Small and Sweeney (1985), and Small et al. (1985). The popularity of the cosine is largely due to the work of Salton in the field of information retrieval (e.g., Salton, 1963; Salton & McGill, 1983). The cosine is therefore sometimes referred to as Salton's measure (e.g., Glänzel, 2001; Glänzel, Schubert, & Czerwon, 1999; Luukkonen et al., 1993; Schubert & Braun, 1990) or as the Salton index (e.g., Morillo, Bordons, & Gómez, 2003). In some studies, a measure called the equivalence index is used (e.g., Callon, Courtial, & Laville, 1991; Kostoff, Eberhart, & Toothman, 1999; Law & Whittaker, 1992; Palmer, 1999). This measure equals the square of the cosine. Outside the fields of scientometrics and information retrieval, the cosine is also known as the Ochiai coefficient (e.g., T. F. Cox & Cox, 2001; M. A. A. Cox & Cox, 2008; Hubálek, 1982; Sokal & Sneath, 1963).

Examples of the use of the inclusion index defined in (4.8) can be found in the work of Kostoff, del Río, Humenik, García, and Ramírez (2001), McCain (1995), Peters and Van Raan (1993b), Rip and Courtial (1984), Tijssen (1992, 1993), and Tijssen and Van Raan (1989). We note that a measure somewhat different from the one defined in

(4.8) is sometimes also called the inclusion index (e.g., Braam et al., 1991b; Kostoff et al., 1999; Peters et al., 1995; Qin, 2000). In the field of information retrieval, the inclusion index is referred to as the overlap measure (e.g., Jones & Furnas, 1987; Rorvig, 1999; Salton & McGill, 1983). More in general, the inclusion index is sometimes called the Simpson coefficient (e.g., T. F. Cox & Cox, 2001; M. A. A. Cox & Cox, 2008; Hubálek, 1982).

The Jaccard index defined in (4.9) equals the ratio between on the one hand the number of times that objects i and j are observed together and on the other hand the number of times that at least one of the two objects is observed. Small uses the Jaccard index in his early work on co-citation analysis (e.g., Small, 1973, 1981; Small & Greenlee, 1980). Other work in which the Jaccard index is used includes Heimeriks, Hörlesberger, and Van den Besselaar (2003), Kopcsa and Schiebel (1998), Peters and Van Raan (1993b), Peters et al. (1995), Rip and Courtial (1984), Van Raan and Tijssen (1993), Vaughan (2006), and Vaughan and You (2006). As shown by Anderberg (1973), the Jaccard index is monotonically related to the Dice coefficient, which is a well-known measure in information retrieval (e.g., Jones & Furnas, 1987; Rorvig, 1999; Salton & McGill, 1983) and other fields (e.g., T. F. Cox & Cox, 2001; M. A. A. Cox & Cox, 2008; Hubálek, 1982; Sokal & Sneath, 1963).

We note that, in addition to the four direct similarity measures discussed above, many more direct similarity measures have been used in scientometric research. However, the above four measures are by far the most popular ones, and we therefore focus most of our attention on them in this chapter. The relations among various direct similarity measures are summarized in Table 4.1.

In the field of scientometrics, a number of studies have been performed in which different direct similarity measures are compared with each other. Boyack et al. (2005), Gmür (2003), Klavans and Boyack (2006a), Leydesdorff (2008), Luukkonen et al. (1993), and Peters and Van Raan (1993b) report results of empirical comparisons of different measures. Theoretical analyses of relations between different measures can be found in the work of Egghe (2009) and Hamers et al. (1989). Properties of various measures are also studied theoretically by Egghe and Rousseau (2006). An extensive discussion of the issue of comparing different measures is provided by Schneider and Borlund (2007a, 2007b). Other work that might be of interest has been done in the

Table 4.1: Relations among various direct similarity measures.

Measure	Alternative names	Monotonically related measures
association strength	probabilistic affinity index proximity index pseudo-cosine	(pointwise) mutual information
cosine	Ochiai coefficient Salton's index/measure	equivalence index
inclusion index	overlap measure Simpson coefficient	
Jaccard index		Dice coefficient

field of information retrieval. In the information retrieval literature, empirical comparisons of different direct similarity measures are discussed by Chung and Lee (2001) and Rorvig (1999) and a theoretical comparison is presented by Jones and Furnas (1987).[3] We further note that general overviews of a large number of direct similarity measures and their properties can be found in the statistical literature (Anderberg, 1973; T. F. Cox & Cox, 2001; M. A. A. Cox & Cox, 2008; Gower, 1985; Gower & Legendre, 1986) and also in the biological literature (Hubálek, 1982; Sokal & Sneath, 1963).

4.3 Set-Theoretic Similarity Measures

In this section and in the next one, we are concerned with two special classes of direct similarity measures. We discuss the class of set-theoretic similarity measures in this section and the class of probabilistic similarity measures in the next section. It turns out that there is a fundamental difference between the cosine, the inclusion index, and the Jaccard index on the one hand and the association strength on the other hand. The first three measures all belong to the class of set-theoretic similarity measures, while the last measure belongs to the class of probabilistic similarity measures. We assume from now on that s_i denotes the total number of occurrences of object i, that is, we assume that

[3] The results reported by Jones and Furnas are probably not very relevant to scientometric research. This is because Jones and Furnas focus on the effect of term weights on similarity measures. In scientometric research, there is no natural analogue to the term weights used in information retrieval. The reason for this is that the occurrence matrices used in scientometric research contain elements that are usually restricted to zero and one, while the document-term matrices used in information retrieval contain term weights that often do not have this restriction.

the definition of s_i in (4.2) is adopted. From a theoretical point of view, this definition is more convenient than the definition of s_i in (4.3). We note that proofs of the theoretical results that we present in this section and in the next one are provided in the appendix.

Each column of an occurrence matrix can be seen as a representation of a set, namely the set of all documents in which a certain object occurs (cf. Egghe & Rousseau, 2006). Consequently, a natural approach to determine the similarity between two objects i and j seems to be to determine the similarity between on the one hand the set of all documents in which i occurs and on the other hand the set of all documents in which j occurs. We refer to direct similarity measures that take this approach as set-theoretic similarity measures. In other words, set-theoretic similarity measures are direct similarity measures that are based on the notion of similarity between sets. In this section, we theoretically analyze the properties of set-theoretic similarity measures. We note that these properties are also studied theoretically by Baulieu (1989, 1997), Egghe and Michel (2002, 2003), Egghe and Rousseau (2006), and Janson and Vegelius (1981).

There are a number of properties of which we believe that it is natural to expect that any set-theoretic similarity measure $S(c_{ij}, s_i, s_j)$ has them. Three of these properties are given below.

Property 4.1. If $c_{ij} = 0$, then $S(c_{ij}, s_i, s_j)$ takes its minimum value.

Property 4.2. For all $\alpha > 0$, $S(\alpha c_{ij}, \alpha s_i, \alpha s_j) = S(c_{ij}, s_i, s_j)$.

Property 4.3. If $s'_i > s_i$ and $c_{ij} > 0$, then $S(c_{ij}, s'_i, s_j) < S(c_{ij}, s_i, s_j)$.

Property 4.1 is based on the idea that the similarity between two sets should be minimal if the sets are disjoint, that is, if they have no elements in common. Property 4.2 is based on the idea that the similarity between two sets should remain unchanged in the case of a proportional increase or decrease in both the number of elements of each of the sets and the number of elements of the intersection of the sets. Egghe and Rousseau (2006) refer to this idea as replication invariance. It underlies the notion of Lorenz similarity that is studied by Egghe and Rousseau. A similar idea is also used by Janson and Vegelius (1981), who call it homogeneity. Property 4.3 is based on the idea that the similarity between two sets should decrease if an element is added to one of the sets and this element does not belong to the other set. A similar idea is used by Baulieu (1989, 1997). It is not difficult to see that Properties 4.1, 4.2, and 4.3 are independent of each other, that is, none of the properties is implied by the others. We regard Properties 4.1, 4.2, and 4.3 as the characterizing properties of set-theoretic similarity measures. This is formally stated in the following definition.

Definition 4.3. A *set-theoretic similarity measure* is defined as a direct similarity measure $S(c_{ij}, s_i, s_j)$ that has Properties 4.1, 4.2, and 4.3.

This definition implies that the cosine defined in (4.7) and the Jaccard index defined in (4.9) are set-theoretic similarity measures. The association strength defined in (4.6) does not have Property 4.2 and is therefore not a set-theoretic similarity measure. The inclusion index defined in (4.8) is also not a set-theoretic similarity measure. This is because the inclusion index does not have Property 4.3. However, the inclusion index does have the following property, which is a weakened version of Property 4.3.

Property 4.4. If $s_i' > s_i$ and $c_{ij} > 0$, then $S(c_{ij}, s_i', s_j) = S(c_{ij}, s_i, s_j)$.

This property naturally leads to the following definition.

Definition 4.4. A *weak set-theoretic similarity measure* is defined as a direct similarity measure $S(c_{ij}, s_i, s_j)$ that has Properties 4.1, 4.2, and 4.4.

It follows from this definition that the inclusion index is a weak set-theoretic similarity measure. We note that our definition of a set-theoretic similarity measure seems to be more restrictive than the definition of a Lorenz similarity function that is provided by Egghe and Rousseau (2006). This is because a Lorenz similarity function need not have Properties 4.1 and 4.3.

In addition to Properties 4.1, 4.2, and 4.3, there are some other properties that we consider indispensable for any set-theoretic similarity measure $S(c_{ij}, s_i, s_j)$. Four of these properties are given below.

Property 4.5. If $S(c_{ij}, s_i, s_j)$ takes its minimum value, then $c_{ij} = 0$.

Property 4.6. If $c_{ij} = s_i = s_j$, then $S(c_{ij}, s_i, s_j)$ takes its maximum value.

Property 4.7. If $S(c_{ij}, s_i, s_j)$ takes its maximum value, then $c_{ij} = s_i = s_j$.

Property 4.8. For all $\alpha > 0$, if $c_{ij} < s_i$ or $c_{ij} < s_j$, then $S(c_{ij} + \alpha, s_i + \alpha, s_j + \alpha) > S(c_{ij}, s_i, s_j)$.

Properties 4.5, 4.6, and 4.7 are based on the idea that the similarity between two sets should be minimal only if the sets are disjoint and that it should be maximal if and only if the sets are equal. Property 4.8 is based on the idea that the similarity between two sets should increase if the same element is added to both sets. It turns out that Properties 4.5, 4.6, 4.7, and 4.8 are implied by Properties 4.1, 4.2, and 4.3. This is stated by the following proposition.

Proposition 4.1. All set-theoretic similarity measures $S(c_{ij}, s_i, s_j)$ have Properties 4.5, 4.6, 4.7, and 4.8.

We note that weak set-theoretic similarity measures need not have Properties 4.5, 4.7, and 4.8. They do have Property 4.6.

We now consider the following two properties.

Property 4.9. If$s'_i s'_j > s_i s_j$ and $c_{ij} > 0$, then $S(c_{ij}, s'_i, s'_j) < S(c_{ij}, s_i, s_j)$. If $s'_i s'_j = s_i s_j$, then $S(c_{ij}, s'_i, s'_j) = S(c_{ij}, s_i, s_j)$.

Property 4.10. If $s'_i + s'_j > s_i + s_j$ and $c_{ij} > 0$, then $S(c_{ij}, s'_i, s'_j) < S(c_{ij}, s_i, s_j)$. If $s'_i + s'_j = s_i + s_j$, then $S(c_{ij}, s'_i, s'_j) = S(c_{ij}, s_i, s_j)$.

It is easy to see that these properties both imply Property 4.3. Hence, Properties 4.9 and 4.10 are both stronger than Property 4.3. It can further be seen that the cosine has Property 4.9 and that the Jaccard index has Property 4.10. The following two propositions indicate the importance of Properties 4.9 and 4.10.

Proposition 4.2. All set-theoretic similarity measures $S(c_{ij}, s_i, s_j)$ that have Property 4.9 are monotonically related to the cosine defined in (4.7).

Proposition 4.3. All set-theoretic similarity measures $S(c_{ij}, s_i, s_j)$ that have Property 4.10 are monotonically related to the Jaccard index defined in (4.9).

It follows from Proposition 4.2 that Properties 4.1, 4.2, and 4.9 characterize the class of all set-theoretic similarity measures that are monotonically related to the cosine. Likewise, it follows from Proposition 4.3 that Properties 4.1, 4.2, and 4.10 characterize the class of all set-theoretic similarity measures that are monotonically related to the Jaccard index. We now apply a similar idea to the inclusion index. The inclusion index has the following property.

Property 4.11. If $\min(s'_i, s'_j) > \min(s_i, s_j)$ and $c_{ij} > 0$, then $S(c_{ij}, s'_i, s'_j) < S(c_{ij}, s_i, s_j)$. If $\min(s'_i, s'_j) = \min(s_i, s_j)$, then $S(c_{ij}, s'_i, s'_j) = S(c_{ij}, s_i, s_j)$.

This property implies Property 4.4. Together with Properties 4.1 and 4.2, Property 4.11 characterizes the class of all weak set-theoretic similarity measures that are monotonically related to the inclusion index. This is an immediate consequence of the following proposition.

Proposition 4.4. All weak set-theoretic similarity measures $S(c_{ij}, s_i, s_j)$ that have Property 4.11 are monotonically related to the inclusion index defined in (4.8).

In the above discussion, we have introduced a large number of properties that a direct similarity measure may or may not have. For convenience, in Table 4.2 we summarize for the association strength, the cosine, the inclusion index, and the Jaccard index which of these properties they have and which they do not have. We note that the last two properties in the table will be introduced in the next section.

In order to provide some additional insight into the relations among various (weak and non-weak) set-theoretic similarity measures, we now introduce what we call the generalized similarity index (for a similar idea, see Warrens, 2008).

Definition 4.5. The *generalized similarity index* is defined as a direct similarity measure that is given by

$$S_{\mathrm{G}}(c_{ij}, s_i, s_j; p) = \frac{2^{1/p} c_{ij}}{\left(s_i^p + s_j^p\right)^{1/p}}, \tag{4.11}$$

where p denotes a parameter that takes values in $\mathbb{R} \setminus \{0\}$.

For all values of the parameter p, the generalized similarity index takes values between zero and one. The index equals the ratio between on the one hand the number of times that objects i and j are observed together and on the other hand a power mean of the number of times that object i is observed and the number of times that object j is observed. (Power means, also known as generalized means or Hölder means, are a generalization of arithmetic, geometric, and harmonic means.) An interesting property of the generalized similarity index is that, for various values of p, the index reduces to a well-known (weak or non-weak) set-theoretic similarity measure. More specifically, it can be seen that

$$\lim_{p \to -\infty} S_{\mathrm{G}}(c_{ij}, s_i, s_j; p) = \frac{c_{ij}}{\min(s_i, s_j)}, \tag{4.12}$$

Table 4.2: Summary of the properties of a number of direct similarity measures. If a measure has a certain property, this is indicated using a × symbol.

	Property												
	4.1	4.2	4.3	4.4	4.5	4.6	4.7	4.8	4.9	4.10	4.11	4.12	4.13
Association strength	×		×	×	×		×		×			×	×
Cosine	×	×	×	×	×	×	×	×	×			×	
Inclusion index	×	×		×	×	×					×	×	
Jaccard index	×	×	×	×	×	×	×	×		×			

$$S_{\mathrm{G}}(c_{ij}, s_i, s_j; -1) = \frac{1}{2}\left(\frac{c_{ij}}{s_i} + \frac{c_{ij}}{s_j}\right), \tag{4.13}$$

$$\lim_{p \to 0} S_{\mathrm{G}}(c_{ij}, s_i, s_j; p) = \frac{c_{ij}}{\sqrt{s_i s_j}}, \tag{4.14}$$

$$S_{\mathrm{G}}(c_{ij}, s_i, s_j; 1) = \frac{2c_{ij}}{s_i + s_j}, \tag{4.15}$$

$$S_{\mathrm{G}}(c_{ij}, s_i, s_j; 2) = \frac{\sqrt{2}c_{ij}}{\sqrt{s_i^2 + s_j^2}}, \tag{4.16}$$

$$\lim_{p \to \infty} S_{\mathrm{G}}(c_{ij}, s_i, s_j; p) = \frac{c_{ij}}{\max(s_i, s_j)}, \tag{4.17}$$

where (4.12), (4.14), and (4.17) follow from the properties of power means as discussed by, for example, Hardy, Littlewood, and Pólya (1952). Equations (4.12) and (4.13) indicate that for $p \to -8$ the generalized similarity index equals the inclusion index and that for $p = -1$ it equals the so-called joint conditional probability measure that is used by McCain (1995). The latter measure is more generally known as one of the Kulczynski coefficients (e.g., T. F. Cox & Cox, 2001; M. A. A. Cox & Cox, 2008; Hubálek, 1982; Sokal & Sneath, 1963). It is easy to see that this measure is a set-theoretic similarity measure. Equations (4.14) and (4.15) indicate that for $p \to 0$ the generalized similarity index equals the cosine and that for $p = 1$ it equals the Dice coefficient. It follows from (4.9) and (4.15) that

$$S_{\mathrm{G}}(c_{ij}, s_i, s_j; 1) = \frac{2S_{\mathrm{J}}(c_{ij}, s_i, s_j)}{S_{\mathrm{J}}(c_{ij}, s_i, s_j) + 1}, \tag{4.18}$$

which implies that for $p = 1$ the generalized similarity index is monotonically related to the Jaccard index. Equations (4.16) and (4.17) indicate that for $p = 2$ and $p \to 8$ the generalized similarity index equals, respectively, the measures N and O_2 that are studied by Egghe and Michel (2002, 2003) and Egghe and Rousseau (2006). It is clear that N is a set-theoretic similarity measure and that O_2 is a weak set-theoretic similarity measure. Measures equivalent to (4.17) are also discussed by T. F. Cox and Cox (2001), M. A. A. Cox and Cox (2008) and Hubálek (1982).

The following proposition points out an important property of the generalized similarity index.

Proposition 4.5. For all finite values of the parameter p, the generalized similarity index defined in (4.11) is a set-theoretic similarity measure.

This proposition states that the generalized similarity index describes an entire class of set-theoretic similarity measures. Each member of this class corresponds with a particular value of p. Only in the limit case in which $p \to \pm 8$, the generalized similarity index is not a set-theoretic similarity measure. In this limit case, the generalized similarity index is a weak set-theoretic similarity measure.

4.4 Probabilistic Similarity Measures

In the previous section, we discussed the class of set-theoretic similarity measures. The cosine, the inclusion index, and the Jaccard index turned out to be (weak or non-weak) set-theoretic similarity measures. The association strength, however, turned out not to belong to the class of set-theoretic similarity measures. In this section, we discuss the class of probabilistic similarity measures. This is the class to which the association strength turns out to belong.

We are interested in direct similarity measures $S(c_{ij}, s_i, s_j)$ that have the following two properties.

Property 4.12. If $s_1 = s_2 = \cdots = s_n$, then $S(c_{ij}, s_i, s_j) = \alpha c_{ij}$ for all $i \neq j$ and for some $\alpha > 0$.

Property 4.13. For all $\alpha > 0$, $S(\alpha c_{ij}, \alpha s_i, s_j) = S(c_{ij}, s_i, s_j)$.

Property 4.12 requires that, if all objects occur equally frequently, the similarity between two objects is proportional to the number of co-occurrences of the objects. Property 4.13 requires that the similarity between two objects remains unchanged in the case of a proportional increase or decrease in on the one hand the number of co-occurrences of the objects and on the other hand the number of occurrences of one of the objects. (Notice the difference between this property and Property 4.2.) We regard Properties 4.12 and 4.13 as the characterizing properties of probabilistic similarity measures. This results in the following definition.

Definition 4.6. A *probabilistic similarity measure* is defined as a direct similarity measure $S(c_{ij}, s_i, s_j)$ that has Properties 4.12 and 4.13.

The cosine, the inclusion index, and the Jaccard index do not have Property 4.13 and therefore are not probabilistic similarity measures. The association strength, on the other hand, is a probabilistic similarity measure, since it has both Property 4.12 and Property 4.13. In this respect, the association strength is quite unique, as the following proposition indicates.

Proposition 4.6. All probabilistic similarity measures are proportional to the association strength defined in (4.6).

This proposition states that the class of probabilistic similarity measures consists only of the association strength and of measures that are proportional to the association strength. There are no other measures that belong to the class of probabilistic similarity measures. The following result is an immediate consequence of Proposition 4.6.

Corollary 4.7. *A direct similarity measure cannot be both a (weak or non-weak) set-theoretic similarity measure and a probabilistic similarity measure.*

This result makes clear that there is a fundamental difference between set-theoretic similarity measures and probabilistic similarity measures. In other words, there is a fundamental difference between measures such as the cosine, the inclusion index, and the Jaccard index on the one hand and the association strength on the other hand. We will come back to this difference later on in this chapter.

We now explain the rationale for Properties 4.12 and 4.13. To do so, we first discuss why direct similarity measures are applied to co-occurrence data. The number of co-occurrences of two objects can be seen as the result of two independent effects. We refer to these effects as the similarity effect and the size effect.[4] The similarity effect is the effect that, other things being equal, more similar objects have more co-occurrences. The size effect is the effect that, other things being equal, an object that occurs more frequently has more co-occurrences with other objects. If one is interested in the similarity between two objects, the number of co-occurrences of the objects is in general not an appropriate measure. This is because, due to the size effect, the number of co-occurrences is likely to give a distorted picture of the similarity between the objects (see also Waltman & Van Eck, 2007). Two frequently occurring objects, for example,

[4] The similarity effect and the size effect can be seen as analogous to what statisticians call, respectively, interaction effects and main effects.

may have a large number of co-occurrences and may therefore look very similar. However, it is quite well possible that the large number of co-occurrences of the objects is completely due to their high frequency of occurrence (i.e., the size effect) and has nothing to do with their similarity. Usually, when a direct similarity measure is applied to co-occurrence data, the aim is to correct the data for the size effect.

Based on the above discussion, the idea underlying Property 4.12 can be explained as follows. Property 4.12 is concerned with the behavior of a direct similarity measure in the special case in which all objects occur equally frequently. In this special case, the size effect is equally strong for all objects, which means that, unlike in the more general case, the number of co-occurrences of two objects is an appropriate measure of the similarity between the objects. Taking this into account, it is natural to expect that in the special case considered by Property 4.12 a direct similarity measure does not transform the co-occurrence frequencies of objects in any significant way. Property 4.12 implements this idea by requiring that, if all objects occur equally frequently, the similarity between two objects is proportional to the number of co-occurrences of the objects.

We now consider Property 4.13. The idea underlying this property can best be clarified by means of an example. Consider an arbitrary object i, and suppose that the total number of occurrences of i doubles. It can then be expected that the total number of co-occurrences of i also doubles, at least approximately. Suppose that the total number of co-occurrences of i indeed doubles and that the new co-occurrences of i are distributed over the other objects in the same way as the old co-occurrences of i. This simply means that the number of co-occurrences of i with each other object doubles. We believe that this increase in the number of occurrences and co-occurrences of i should not have any influence on the similarities between i and the other objects. This is because the number of occurrences of i and the number of co-occurrences of i with each other object have all increased proportionally, namely by a factor of two. Hence, relatively speaking, the frequency with which i co-occurs with each other object has not changed. This means that the increase in the number of co-occurrences of i with each other object is completely due to the size effect and has not been caused by the similarity effect. Taking this into account, it is natural to expect that the similarities between i and the other objects remain unchanged. Property 4.13 implements this idea. It does so not only for the case in which the number of occurrences and co-occurrences of an object doubles but more generally for any proportional increase or decrease in the number of occurrences and co-occurrences of an object. We note that the idea underlying Property 4.13 is not new. Ahlgren et al. (2003) and Van Eck and Waltman (2008) study properties of indirect similarity measures. A property that turns out to be particularly important is the so-called property of coordinate-wise scale invariance. Interestingly, this property relies on exactly the same idea as Property 4.13. Hence, direct similarity measures that have Property 4.13 and indirect similarity measures that have the property of coordinate-wise scale invariance are based on similar principles.

Finally, we discuss the probabilistic interpretation of probabilistic similarity measures (see also Leclerc & Gagné, 1994; Luukkonen et al., 1992, 1993; Zitt et al., 2000). Let p_i denote the probability that object i occurs in a randomly chosen document. It is clear that $p_i = s_i/m$. If two objects i and j occur independently of each other, the probability that they co-occur in a randomly chosen document equals $p_{ij} = p_i p_j$. The expected number of co-occurrences of i and j then equals $e_{ij} = mp_{ij} = mp_i p_j = s_i s_j/m$. A natural way to measure the similarity between i and j is to calculate the ratio between on the one hand the observed number of co-occurrences of i and j and on the other hand the expected number of co-occurrences of i and j under the assumption that i and j occur independently of each other (for a similar argument in a more general context, see de Solla Price, 1981). This results in a measure that equals c_{ij}/e_{ij}. This measure has a straightforward probabilistic interpretation. If $c_{ij}/e_{ij} > 1$, i and j co-occur more frequently than would be expected by chance. If, on the other hand, $c_{ij}/e_{ij} < 1$, i and j co-occur less frequently than would be expected by chance. It is easy to see that $c_{ij}/e_{ij} = mS_{\mathrm{A}}(c_{ij}, s_i, s_j)$. Hence, the measure c_{ij}/e_{ij} is proportional to the association strength and, consequently, belongs to the class of probabilistic similarity measures. Since probabilistic similarity measures are all proportional to each other (this follows from Proposition 4.6), they all have a similar probabilistic interpretation as the measure c_{ij}/e_{ij}.

4.5 Empirical Comparison

In the previous two sections, the differences between a number of well-known direct similarity measures were analyzed theoretically. It turned out that some measures have fundamentally different properties than others. An obvious question now is whether in practical applications there is much difference between the various measures. This is the question with which we are concerned in this section.

Leydesdorff (2008) reports the results of an empirical comparison of a number of direct and indirect similarity measures (for a theoretical explanation for some of the results, see Egghe, 2009). The measures are applied to a data set consisting of the co-citation frequencies of 24 authors, 12 from the field of information retrieval and 12 from the field of scientometrics.[5] It turns out that the direct similarity measures are strongly correlated with each other. The Spearman rank correlations between the association strength (referred to as the probabilistic affinity or activity index), the cosine, and the Jaccard index are all above 0.98. Hence, for the particular data set studied by Leydesdorff, there does not seem to be much difference between various direct similarity measures.

In this section, we examine whether the results reported by Leydesdorff hold more generally. To do so, we study three data sets, one consisting of co-citation frequencies of authors, one consisting of co-citation frequencies of journals, and one consisting of co-occurrence frequencies of terms. We refer to these data sets as, respectively, the author data set, the journal data set, and the term data set. The author data set consists of the co-citation frequencies of 100 authors in the field of information science in the period 1988–1995. The data set is studied extensively in a well-known paper by White and McCain (1998) (see also White, 2003b), and it is also used in one of our earlier papers (Van Eck & Waltman, 2008). The journal data set has not been studied before. The data set consists of the co-citation frequencies of 389 journals belonging to at least one of the following five subject categories of Thomson Reuters: *Business*, *Business-Finance*, *Economics*, *Management*, and *Operations Research & Management Science*. The co-citation frequencies of the journals were determined based on citations in articles published between 2005 and 2007 to articles published in 2005. The term data set consists of the co-occurrence frequencies of 332 terms in the field of computational intelligence in the period 1996–2000. Co-occurrences of terms were counted in abstracts of articles published in important journals and conference proceedings in the computational intelligence field. For a more detailed description of the term data set,

[5] The same data set is also studied by Ahlgren et al. (2003), Leydesdorff and Vaughan (2006), and Waltman and Van Eck (2007).

Table 4.3: Main characteristics of the author data set, the journal data set, and the term data set.

	Author data set	Journal data set	Term data set
# objects	100	389	332
# documents	5 463	24 106	6 235
# occurrences	7 768	32 697	26 211
# co-occurrences	22 520	13 378	60 640
% zeros in co-occurrence matrix	26%	93%	74%

we refer to an earlier paper (Van Eck & Waltman, 2007a). In Table 4.3, we summarize the main characteristics of the three data sets that we study.

In order to examine how the association strength, the cosine, the inclusion index, and the Jaccard index are empirically related to each other, we analyzed each of the three data sets as follows. We first calculated for each combination of two objects the value of each of the four similarity measures. For each combination of two similarity measures, we then drew a scatter plot that shows how the values of the two measures are related to each other. The scatter plots obtained for the author data set and the term data set are shown in Figures 4.1 and 4.2, respectively. The scatter plots obtained for the journal data set look very similar to the ones obtained for the term data set and are therefore not shown. After drawing the scatter plots, we determined for each combination of two similarity measures how strongly the values of the measures are correlated with each other. We calculated both the Pearson correlation and the Spearman correlation. The Pearson correlation was used to measure the degree to which the values of two measures are linearly related, while the Spearman correlation was used to measure the degree to which the values of two measures are monotonically related. When calculating the Pearson and Spearman correlations between the values of two measures, we only took into account values above zero.[6] The correlations obtained for the three data sets are

[6] If two objects have zero co-occurrences, all four similarity measures have a value of zero. Co-occurrence matrices usually contain a large number of zeros (see Table 4.3). This leads to high correlations (close to one) between the values of the four similarity measures. We regard these high correlations as problematic because they do not properly reflect how the similarity measures are related to each other in the case of objects with a non-zero number of co-occurrences. To avoid the problem of the high correlations, we only took into account values above zero when calculating correlations between the values of the four similarity measures.

reported in Tables 4.4, 4.5, and 4.6. In each table, the values in the upper right part are Pearson correlations while the values in the lower left part are Spearman correlations.

The scatter plots in Figures 4.1 and 4.2 clearly show that in practical applications there can be substantial differences between different direct similarity measures. This is confirmed by the correlations in Tables 4.4, 4.5, and 4.6. These results differ from the ones reported by Leydesdorff (2008), who finds no substantial differences between different direct similarity measures. The difference between our results and the results of Leydesdorff is probably due to the unusual nature of the data set studied in Leydesdorff, in particular the small number of objects in the data set (24 authors) and the division of the objects into two strongly separated groups (the information retrieval researchers and the scientometricians). When looking in more detail at the scatter plots in Figures 4.1 and 4.2, it can be seen that the similarity measures that are strongest related to each other are the cosine and the Jaccard index. The same observation can be made in Tables 4.4, 4.5, and 4.6. The relatively strong relation between the cosine and the Jaccard index has been observed before and is discussed by Egghe (2009), Hamers et al. (1989), and Leydesdorff (2008). Apart from the relation between the cosine and the Jaccard index, the relations between the different similarity measures are quite weak. This is especially the case for the relations between the association strength and the other three measures. Consider, for example, how the association strength and the inclusion index are related to each other in the term data set. As can be seen in Figure 4.2, a low value of the association strength sometimes corresponds with a high value of the inclusion index and, the other way around, a low value of the inclusion index sometimes corresponds with a high value of the association strength. This clearly indicates that the relation between the two measures is rather weak, which is confirmed by the correlations in Table 4.6. It is further interesting to compare our empirical results with the theoretical results presented by Egghe (2009). Egghe mathematically studies relations between various (weak and non-weak) set-theoretic similarity measures under the simplifying assumption that the ratio of the number of occurrences of two objects is fixed. He proves that, under this assumption, there exist simple monotonic (often linear) relations between many measures. However, especially for the inclusion index, the scatter plots in Figures 4.1 and 4.2 do not show such relations. Our empirical results therefore seem to

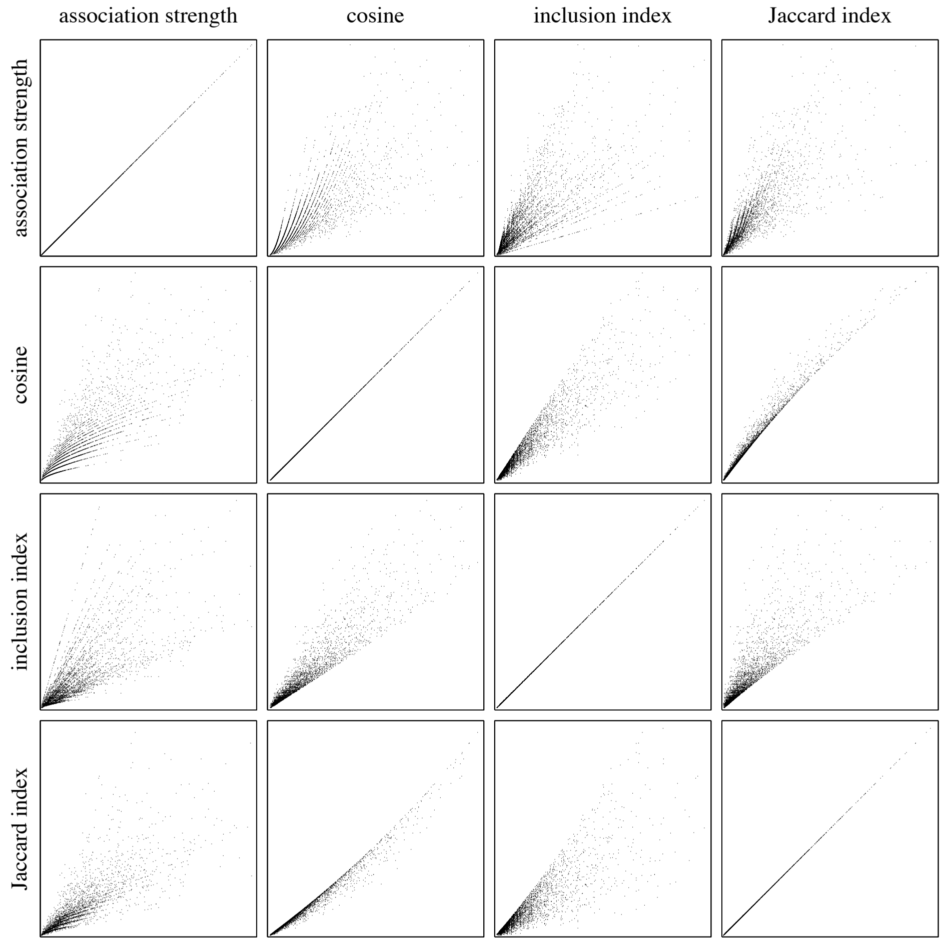


Figure 4.1: Scatter plots obtained for the author data set. In each plot, the lower left corner corresponds with the origin. The scales used for the different similarity measures are not the same.

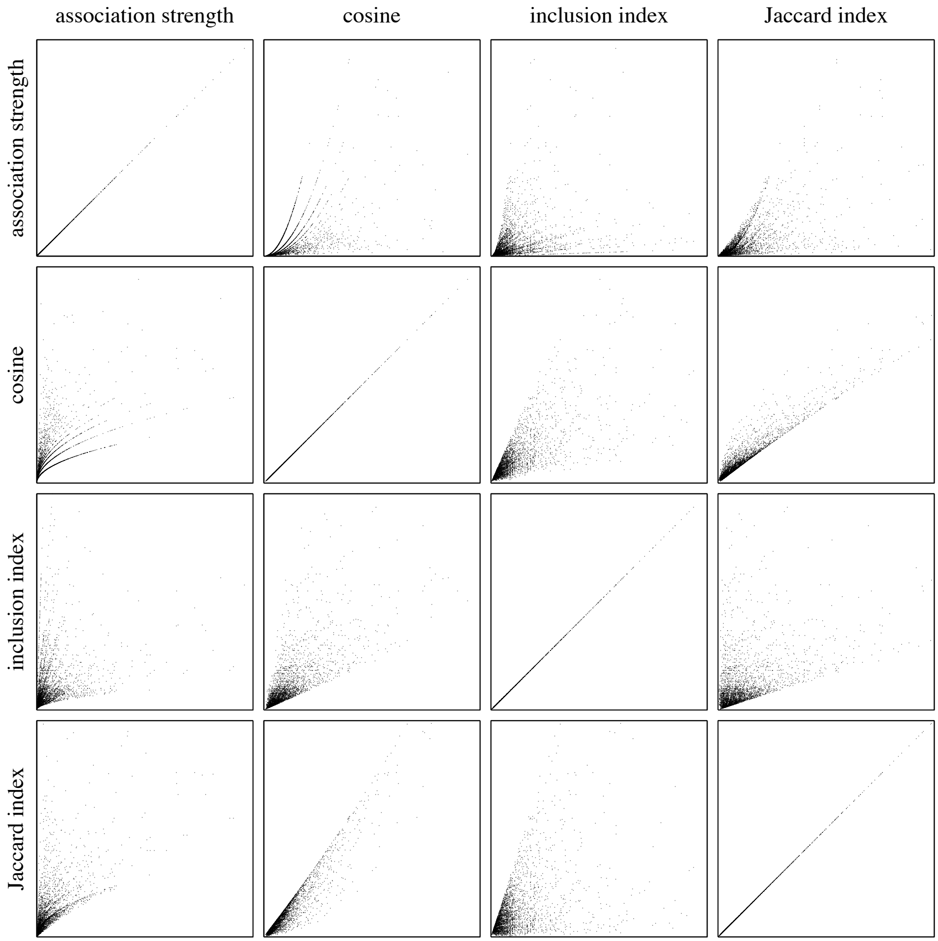


Figure 4.2: Scatter plots obtained for the term data set. In each plot, the lower left corner corresponds with the origin. The scales used for the different similarity measures are not the same.

Table 4.4: Correlations obtained for the author data set.

	Association strength	Cosine	Inclusion index	Jaccard index
Association strength		0.824	0.721	0.823
Cosine	0.913		0.929	0.987
Inclusion index	0.847	0.964		0.866
Jaccard index	0.920	0.994	0.931	

Table 4.5: Correlations obtained for the journal data set.

	Association strength	Cosine	Inclusion index	Jaccard index
Association strength		0.602	0.556	0.554
Cosine	0.892		0.800	0.971
Inclusion index	0.808	0.881		0.644
Jaccard index	0.832	0.952	0.708	

Table 4.6: Correlations obtained for the term data set.

	Association strength	Cosine	Inclusion index	Jaccard index
Association strength		0.653	0.347	0.688
Cosine	0.786		0.736	0.950
Inclusion index	0.562	0.799		0.511
Jaccard index	0.776	0.916	0.520	

indicate that the practical relevance of the theoretical results presented by Egghe might be somewhat limited.

The general conclusion that can be drawn from our empirical analysis is that there are quite significant differences between various direct similarity measures and, hence, that in practical applications it is important to use the measure that is most appropriate for one's purposes. In the next section, we discuss how an appropriate similarity measure can be chosen based on sound theoretical considerations. We focus in particular on the case in which a similarity measure is used for normalization purposes.

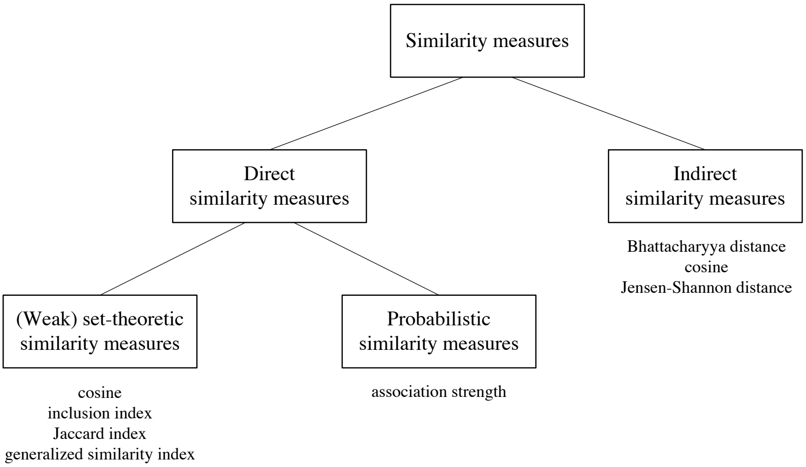

Similarity measures
Direct similarity measures
Indirect similarity measures
Bhattacharyya distance
cosine
Jensen-Shannon distance
(Weak) set-theoretic similarity measures
Probabilistic similarity measures
cosine
inclusion index
Jaccard index
generalized similarity index
association strength

Figure 4.3: Different types of similarity measures.

4.6 How To Normalize Co-Occurrence Data?

As we discussed in the previous sections, there are various ways in which similarities between objects can be determined based on co-occurrence data. The different types of similarity measures that can be used are shown in Figure 4.3. The first decision that one has to make is whether to use a direct or an indirect similarity measure. If one decides to use a direct similarity measure, one then has to decide whether to use a probabilistic or a set-theoretic similarity measure.

We first briefly discuss the use of indirect similarity measures. As pointed out by Schneider and Borlund (2007a), from a statistical perspective the use of an indirect similarity measure is a quite unconventional approach.[7] However, despite being unconventional, we do not believe that the approach has any fundamental statistical problems.[8] Appropriate indirect similarity measures include the Bhattacharyya distance, the

[7] A similar approach is sometimes taken in psychological research (e.g., Rosenberg & Jones, 1972; Rosenberg, Nelson, & Vivekananthan, 1968). In the psychological literature, there is some discussion about the advantages and disadvantages of this approach (Drasgow & Jones, 1979; Simmen, 1996; Van der Kloot & Van Herk, 1991).

[8] One of the issues that is sometimes raised is how the diagonal of a co-occurrence matrix should be treated. From a theoretical point of view, there are in our opinion two satisfactory solutions. One solution is to treat diagonal elements as missing values. The other solution is to set diagonal elements equal to the number of times objects occur at least twice in the same document (see also Ahlgren et al., 2003).

cosine,[9] and the Jensen-Shannon distance. These measures are known to have good theoretical properties (Van Eck & Waltman, 2008). A very popular indirect similarity measure, especially for author co-citation analysis (e.g., McCain, 1990; White & Griffith, 1981; White & McCain, 1998), is the Pearson correlation. However, this measure does not have good theoretical properties and should therefore not be used (Ahlgren et al., 2003; Van Eck & Waltman, 2008). The chi-squared distance, which is proposed as an indirect similarity measure by Ahlgren et al. (2003), also does not have all the theoretical properties that we believe an appropriate indirect similarity measure should have (Van Eck & Waltman, 2008). We note that theoretical studies of indirect similarity measures can also be found in the psychometric literature (e.g., Zegers & Ten Berge, 1985). In this literature, the cosine is referred to as Tucker's congruence coefficient.

The notions of direct and indirect similarity are fundamentally different. Direct and indirect similarity measures may therefore lead to significantly different results (e.g., Schneider et al., 2009). In general, we believe the notion of direct similarity to be closer to the intuitive idea of similarity. Consider two objects that do not co-occur at all but that have quite similar co-occurrence profiles. The direct similarity between the objects will be very low, while the indirect similarity between the objects will be quite high. However, a high similarity between two objects that do not co-occur can be rather counterintuitive, at least in certain contexts. For that reason, we believe that in general the notion of direct similarity is more natural than the notion of indirect similarity. We note, however, that indirect similarity measures may also have an advantage over direct similarity measures. Compared with direct similarity measures, indirect similarity measures are calculated based on a larger amount of data and most likely they therefore involve less statistical uncertainty.

In the rest of this section, we focus our attention on the use of direct similarity measures. Direct similarity measures determine the similarity between two objects by taking the number of co-occurrences of the objects and adjusting this number for the total number of occurrences of each of the objects. In scientometric research, when a direct similarity measure is applied to co-occurrence data, the aim usually is to normalize the data, that is, to correct the data for differences in the number of occurrences of

[9] There are two different similarity measures, a direct and an indirect one, that are both referred to as the cosine. Here we mean the cosine as discussed by, for example, Ahlgren et al. (2003) and Van Eck and Waltman (2008). This is a different measure than the one defined in (4.7).

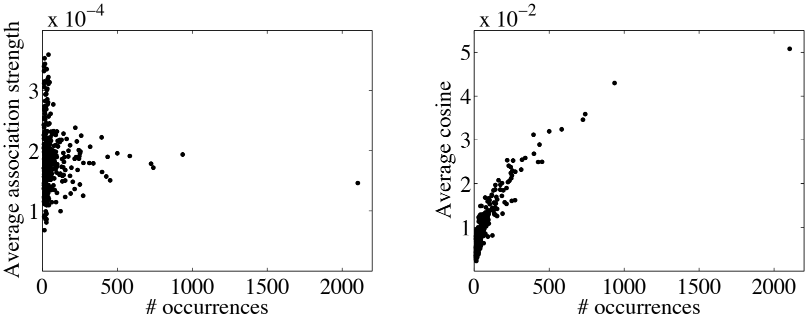


Figure 4.4: Relation between on the one hand the number of occurrences of a term and on the other hand the average similarity of a term with other terms. In the left panel, similarities are determined using the association strength. In the right panel, similarities are determined using the cosine.

objects. This brings us to the main question of this chapter: How should co-occurrence data be normalized? Or, in other words, which direct similarity measures are appropriate for normalizing co-occurrence data and which are not? We argue that co-occurrence data should always be normalized using a probabilistic similarity measure. Other direct similarity measures are not appropriate for normalization purposes. In particular, set-theoretic similarity measures should not be used to normalize co-occurrence data.

To see why probabilistic similarity measures have the right properties for normalizing co-occurrence data, recall that the number of co-occurrences of two objects can be seen as the result of two independent effects, the similarity effect and the size effect. As we discussed earlier in this chapter, probabilistic similarity measures correct for the size effect. This follows from Property 4.13. Set-theoretic similarity measures do not have this property, and they therefore do not properly correct for the size effect. As a consequence, set-theoretic similarity measures have, on average, higher values for objects that occur more frequently (see also Luukkonen et al., 1993; Zitt et al., 2000). The values of probabilistic similarity measures, on the other hand, do not depend on how frequently objects occur. This difference between set-theoretic and probabilistic similarity measures can easily be demonstrated empirically. In Figure 4.4, this is done for the term data set discussed in the previous section. (The author data set and the journal data set yield similar results.) The figure shows the relation between on the one hand the number of occurrences of a term and on the other hand the average similarity of a term with other terms. In the left panel of the figure, similarities are determined using a probabilistic similarity measure, namely the association strength. In this panel, there is no substantial correlation between the number of occurrences of a term and the average similarity of a term ($r = -0.069$, $\rho = -0.029$). This is very different in the right panel, in which similarities are determined using a set-theoretic similarity measure, namely the cosine. (The inclusion index and the Jaccard index yield similar results.) In the right panel, there is a strong positive correlation between the number of occurrences of a term and the average similarity of a term ($r = 0.839$, $\rho = 0.882$). Results such as those shown in the right panel clearly indicate that set-theoretic similarity measures do not properly correct for the size effect and, consequently, do not properly normalize co-occurrence data. It follows from this observation that one should be very careful with the interpretation of similarities that have been derived from co-occurrence data using a set-theoretic similarity measure (see also Luukkonen et al., 1993; Zitt et al., 2000). Moreover, when such similarities are analyzed using multivariate analysis techniques such as multidimensional scaling or hierarchical clustering, one should pay special attention to possible artifacts in the results of the analysis. When using multidimensional scaling, for example, it is our experience that frequently occurring objects tend to cluster together in the center of a solution.

To provide some additional insight why probabilistic similarity measures are more appropriate for normalization purposes than set-theoretic similarity measures, we now compare the main ideas underlying these two types of measures. Suppose that we are performing a co-word analysis and that we want to determine the similarity between two words, word i and word j. We consider two hypothetical scenarios, to which we refer as scenario 1 and scenario 2. The scenarios are summarized in Table 4.7, and they are illustrated graphically in the left and right panels of Figure 4.5. In each panel of the figure, the light gray rectangle represents the set of all documents used in the co-word analysis, the dark gray circle represents the set of all documents in which word i occurs, and the striped circle represents the set of all documents in which word j occurs. The area of a rectangle or circle is proportional to the number of documents in the corresponding set.

As can be seen in Table 4.7 and Figure 4.5, in scenario 1 words i and j both occur

Table 4.7: Summary of two hypothetical scenarios in a co-word analysis.

	Scenario 1	Scenario 2
m	1 000	1 000
s_i	300	20
s_j	300	20
c_{ij}	90	6
Association strength	0.001	0.015
Cosine	0.300	0.300
Inclusion index	0.300	0.300
Jaccard index	0.176	0.176

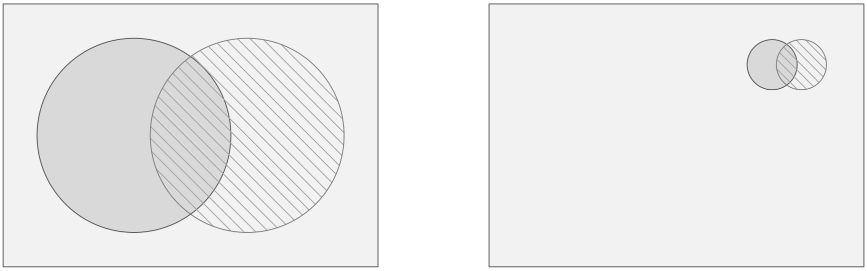


Figure 4.5: Graphical illustration of two hypothetical scenarios in a co-word analysis. Scenario 1 is shown in the left panel. Scenario 2 is shown in the right panel.

quite frequently, while in scenario 2 they both occur relatively infrequently. In both scenarios, however, the relative overlap of the set of documents in which word i occurs and the set of documents in which word j occurs is the same. That is, in both scenarios word i occurs in 30% of the documents in which word j occurs and, the other way around, word j occurs in 30% of the documents in which word i occurs. Because the relative overlap is the same, set-theoretic similarity measures, such as the cosine, the inclusion index, and the Jaccard index, yield the same similarity between words i and j in both scenarios (see Table 4.7). This is a consequence of Property 4.2. At first sight, it might seem a natural result to have the same similarity between words i and j in both scenarios. However, we argue that this result is far from natural, at least for normalization purposes.

We first consider scenario 1 in more detail. In this scenario, words i and j each occur in 30% of all documents. If there is no special relation between words i and j

and if, as a consequence, occurrences of the two words are statistically independent, one would expect the two words to co-occur in approximately $30\% \times 30\% = 9\%$ of all documents. As can be seen in Table 4.7, words i and j co-occur in exactly 9% of all documents. Hence, occurrences of words i and j seem to be statistically independent, at least approximately, and there seems to be no strong relation between the two words.

We now consider scenario 2. In this scenario, words i and j each occur in 2% of all documents. If occurrences of words i and j are statistically independent, one would expect the two words to co-occur in approximately 0.04% of all documents. However, words i and j co-occur in 0.6% of all documents, that is, they co-occur 15 times more frequently than would be expected under the assumption of statistical independence. Hence, there seems to be a quite strong relation between words i and j, definitely much stronger than in scenario 1.

It is clear that set-theoretic similarity measures yield results that do not properly reflect the difference between scenario 1 and scenario 2. This is because set-theoretic similarity measures are based on the idea of measuring the relative overlap of sets instead of the idea of measuring the deviation from statistical independence. Probabilistic similarity measures, such as the association strength, are based on the latter idea, and they therefore yield results that do properly reflect the difference between scenario 1 and scenario 2. As can be seen in Table 4.7, the association strength indicates that in scenario 2 the similarity between words i and j is 15 times higher than in scenario 1. This reflects that in scenario 2 the co-occurrence frequency of words i and j is 15 times higher than would be expected under the assumption of statistical independence while in scenario 1 the co-occurrence frequency of the two words equals the expected co-occurrence frequency under the independence assumption.

4.7 Conclusions

We have studied the application of direct similarity measures to co-occurrence data. Our survey of the scientometric literature has indicated that the most popular direct similarity measures are the association strength, the cosine, the inclusion index, and the Jaccard index. We have therefore focused most of our attention on these four measures. To make a well-considered decision which measure is most appropriate for one's purposes, we believe it to be indispensable to have a good theoretical understanding of the properties of the various measures. In this chapter, we have analyzed these properties in considerable detail. Our analysis has revealed that there are two fundamentally different types of direct similarity measures. On the one hand, there are set-theoretic similarity measures, which can be interpreted as measures of the relative overlap of two sets. On the other hand, there are probabilistic similarity measures, which can be interpreted as measures of the deviation of observed co-occurrence frequencies from expected co-occurrence frequencies under an independence assumption. The cosine, the inclusion index, and the Jaccard index are examples of set-theoretic similarity measures, while the association strength is an example of a probabilistic similarity measure. Set-theoretic and probabilistic similarity measures serve different purposes, and it therefore makes no sense to argue that one measure is always better than another. In scientometric research, however, similarity measures are usually used for normalization purposes, and we have argued that in that specific case probabilistic similarity measures are much more appropriate than set-theoretic ones. Consequently, for most applications of direct similarity measures in scientometric research, we advise against the use of set-theoretic similarity measures and we recommend the use of a probabilistic similarity measure.

In addition to our theoretical analysis, we have also performed an empirical analysis of the behavior of various direct similarity measures. The analysis has shown that in practical applications the differences between various direct similarity measures can be quite large. This indicates that the issue of choosing an appropriate similarity measure is not only of theoretical interest but also has a high practical relevance. Another empirical observation that we have made is that set-theoretic similarity measures yield systematically higher values for frequently occurring objects than for objects that occur only a limited number of times. This confirms our theoretical finding that set-theoretic similarity measures do not properly correct for size effects. Probabilistic similarity measures do not have this problem.

There is one final comment that we would like to make. Above, we have argued in favor of the use of probabilistic similarity measures in scientometric research. Since probabilistic similarity measures are all proportional to each other, it does not really matter which probabilistic similarity measure one uses. In this chapter, we have focused most of our attention on one particular probabilistic similarity measure, namely the association strength defined in (4.6). This measure shares with many other direct similarity measures the property that it takes values between zero and one. For practical purposes, however, it may be convenient not to use the measure in (4.6) directly but instead to multiply this measure by the number of documents m (e.g., Van Eck & Waltman, 2007a; Van Eck, Waltman, et al., 2006a). This results in a slight variant of the association strength. We have pointed out that this variant has the appealing property that it equals one if the observed co-occurrence frequency of two objects equals the co-occurrence frequency that would be expected under the assumption that occurrences of the objects are statistically independent. It takes a value above or below one if the observed co-occurrence frequency is, respectively, higher or lower than the expected co-occurrence frequency under the independence assumption.

4.A Appendix

In this appendix, we prove the theoretical results presented in the chapter.

Proof of Proposition 4.1. We prove each property separately.

(Property 4.5) This property follows from Property 4.3. Property 4.3 implies that, if $c_{ij} > 0$, $S(c_{ij}, s_i, s_j) > S(c_{ij}, s_i + 1, s_j)$. Hence, if $c_{ij} > 0$, $S(c_{ij}, s_i, s_j)$ cannot take its minimum value. This means that $S(c_{ij}, s_i, s_j)$ can take its minimum value only if $c_{ij} = 0$. This proves Property 4.5.

(Property 4.6) This property follows from Properties 4.1, 4.2, and 4.3. Suppose that $c_{ij} = s_i = s_j$. For all $(c'_{ij}, s'_i, s'_j) \in \mathcal{D}_S$ such that $c'_{ij} = 0$, Property 4.1 implies that $S(c'_{ij}, s'_i, s'_j) \leq S(c_{ij}, s_i, s_j)$. For all $(c'_{ij}, s'_i, s'_j) \in \mathcal{D}_S$ such that $c'_{ij} > 0$, Property 4.3 implies that $S(c'_{ij}, s'_i, s'_j) \leq S(c'_{ij}, c'_{ij}, c'_{ij})$ and Property 4.2 implies that $S(c'_{ij}, c'_{ij}, c'_{ij}) = S(c_{ij}, s_i, s_j)$. Hence, for all $(c'_{ij}, s'_i, s'_j) \in \mathcal{D}_S$, $S(c'_{ij}, s'_i, s'_j) \leq S(c_{ij}, s_i, s_j)$. This means that, if $c_{ij} = s_i = s_j$, $S(c_{ij}, s_i, s_j)$ takes its maximum value. This proves Property 4.6.

(Property 4.7) This property follows from Properties 4.1, 4.3, and 4.5. Properties 4.1 and 4.5 imply that, if $c_{ij} = 0$, $S(c_{ij}, s_i, s_j)$ cannot take its maximum value. Property 4.3 implies that, if $0 < c_{ij} < s_i$ or $0 < c_{ij} < s_j$, $S(c_{ij}, s_i, s_j) < S(c_{ij}, c_{ij}, c_{ij})$. Hence, if $0 < c_{ij} < s_i$ or $0 < c_{ij} < s_j$, $S(c_{ij}, s_i, s_j)$ cannot take its maximum value. It now follows that $S(c_{ij}, s_i, s_j)$ can take its maximum value only if $c_{ij} = s_i = s_j$. This proves Property 4.7.

(Property 4.8) This property follows from Properties 4.1, 4.2, 4.3, and 4.5. If $c_{ij} = 0$, the property follows trivially from Properties 4.1 and 4.5. We therefore focus on the case in which $c_{ij} > 0$. Suppose, without loss of generality, that $0 < c_{ij} < s_i$. Consider an arbitrary constant $\alpha > 0$, and let $\beta = (c_{ij} + \alpha)/c_{ij}$. Property 4.2 implies that $S(\beta c_{ij}, \beta s_i, \beta s_j) = S(c_{ij}, s_i, s_j)$. Moreover, because $\beta c_{ij} = c_{ij} + \alpha$, $\beta s_i > s_i + \alpha$, and $\beta s_j \geq s_j + \alpha$, Property 4.3 implies that $S(\beta c_{ij}, \beta s_i, \beta s_j) < S(c_{ij} + \alpha, s_i + \alpha, s_j + \alpha)$. It now follows that $S(c_{ij} + \alpha, s_i + \alpha, s_j + \alpha) > S(c_{ij}, s_i, s_j)$. This proves Property 4.8. □

Proof of Proposition 4.2. Let $S(c_{ij}, s_i, s_j)$ denote an arbitrary set-theoretic similarity measure that has Property 4.9. We start by showing that for all $(c_{ij}, s_i, s_j), (c'_{ij}, s'_i, s'_j) \in \mathcal{D}_S$ the properties of set-theoretic similarity measures together with Property 4.9 are sufficient to determine whether $S(c_{ij}, s_i, s_j)$ is greater than, less than, or equal to $S(c'_{ij}, s'_i, s'_j)$. Suppose first that $c_{ij}, c'_{ij} > 0$. Let $\alpha = c_{ij}/c'_{ij}$. Property 4.2 implies that $S(\alpha c'_{ij}, \alpha s'_i, \alpha s'_j) = S(c'_{ij}, s'_i, s'_j)$. Moreover, taking into account that $c_{ij} = \alpha c'_{ij}$, it can be seen that Property 4.9 determines whether $S(c_{ij}, s_i, s_j)$ is greater than, less than, or equal to $S(\alpha c'_{ij}, \alpha s'_i, \alpha s'_j)$. Hence, if $c_{ij}, c'_{ij} > 0$, Properties 4.2 and 4.9 are sufficient to determine whether $S(c_{ij}, s_i, s_j)$ is greater than, less than, or equal to $S(c'_{ij}, s'_i, s'_j)$. Suppose next that $c_{ij} = 0$ or $c'_{ij} = 0$. Property 4.1 implies that $S(c_{ij}, s_i, s_j) = S(c'_{ij}, s'_i, s'_j)$ if $c_{ij} = c'_{ij} = 0$. Furthermore, Properties 4.1 and 4.5 imply that $S(c_{ij}, s_i, s_j) > S(c'_{ij}, s'_i, s'_j)$ if $c_{ij} > c'_{ij} = 0$ and, conversely, that $S(c_{ij}, s_i, s_j) < S(c'_{ij}, s'_i, s'_j)$ if $c'_{ij} > c_{ij} = 0$. Hence, if $c_{ij} = 0$ or $c'_{ij} = 0$, Properties 4.1 and 4.5 are sufficient to determine whether $S(c_{ij}, s_i, s_j)$ is greater than, less than, or equal to $S(c'_{ij}, s'_i, s'_j)$. It now follows that for all $(c_{ij}, s_i, s_j), (c'_{ij}, s'_i, s'_j) \in \mathcal{D}_S$ the properties of set-theoretic similarity measures together with Property 4.9 are sufficient to determine whether $S(c_{ij}, s_i, s_j)$ is greater than, less than, or equal to $S(c'_{ij}, s'_i, s'_j)$. This implies that all set-theoretic similarity measures that have Property 4.9 are monotonically related to each other. One of these measures is the cosine defined in (4.7). Hence, all set-theoretic similarity measures that have Property 4.9 are monotonically related to the cosine. This completes the proof of the proposition. □

Proof of Proposition 4.3. The proof is analogous to the proof of Proposition 4.2 provided above. □

Proof of Proposition 4.4. Let $S(c_{ij}, s_i, s_j)$ denote an arbitrary weak set-theoretic similarity measure that has Property 4.11. Property 4.11 implies that, if $c_{ij} > 0$, $S(c_{ij}, s_i, s_j)$ $> S(c_{ij}, s_i + 1, s_j + 1)$. Hence, if $c_{ij} > 0$, $S(c_{ij}, s_i, s_j)$ cannot take its minimum value. This means that $S(c_{ij}, s_i, s_j)$ can take its minimum value only if $c_{ij} = 0$. In other words, $S(c_{ij}, s_i, s_j)$ has Property 4.5. This shows that all weak set-theoretic similarity measures $S(c_{ij}, s_i, s_j)$ that have Property 4.11 also have Property 4.5. The rest of the proof is now analogous to the proof of Proposition 4.2 provided above. □

Proof of Proposition 4.5. It is easy to see that for all finite values of the parameter p the generalized similarity index defined in (4.11) has Properties 4.1, 4.2, and 4.3. Hence, it follows from Definition 4.3 that for all finite values of the parameter p the generalized similarity index is a set-theoretic similarity measure. This completes the proof of the proposition. □

Proof of Proposition 4.6. Let $S(c_{ij}, s_i, s_j)$ denote an arbitrary probabilistic similarity measure. Furthermore, let $c'_{ij} = c_{ij}/(s_i s_j)$ for all $i \neq j$, and let $s'_i = 1$ for all i. It follows from Property 4.13 that $S(c_{ij}, s_i, s_j) = S(c'_{ij}, s'_i, s'_j)$ for all $i \neq j$, and it follows from Property 4.12 that $S(c'_{ij}, s'_i, s'_j) = \alpha c'_{ij}$ for all $i \neq j$ and for some $\alpha > 0$. Hence, for all $i \neq j$ and for some $\alpha > 0$, $S(c_{ij}, s_i, s_j) = S(c'_{ij}, s'_i, s'_j) = \alpha c'_{ij} = \alpha c_{ij}/(s_i s_j) = \alpha S_{\mathrm{A}}(c_{ij}, s_i, s_j)$. In other words, $S(c_{ij}, s_i, s_j)$ is proportional to the association strength defined in (4.6). This completes the proof of the proposition. □

Proof of Corollary 4.7. The association strength defined in (4.6) does not have Property 4.2 and is therefore not a (weak or non-weak) set-theoretic similarity measure. The same is true for all measures that are proportional to the association strength. Consequently, it follows from Proposition 4.6 that a probabilistic similarity measure cannot also be a (weak or non-weak) set-theoretic similarity measure. This completes the proof of the corollary. □

Chapter 5

A Comparison of Two Techniques for Bibliometric Mapping: Multidimensional Scaling and VOS[*]

Abstract

VOS is a new mapping technique that can serve as an alternative to the well-known technique of multidimensional scaling. We present an extensive comparison between the use of multidimensional scaling and the use of VOS for constructing bibliometric maps. In our theoretical analysis, we show the mathematical relation between the two techniques. In our empirical analysis, we use the techniques for constructing maps of authors, journals, and keywords. Two commonly used approaches to bibliometric mapping, both based on multidimensional scaling, turn out to produce maps that suffer from artifacts. Maps constructed using VOS turn out not to have this problem. We conclude that in general maps constructed using VOS provide a more satisfactory representation of a data set than maps constructed using well-known multidimensional scaling approaches.

[*] This chapter is based on Van Eck, Waltman, Dekker, and Van den Berg (2010).

5.1 Introduction

In the fields of bibliometrics and scientometrics, the idea of constructing science maps based on bibliographic data has intrigued researchers already for several decades. Many different types of maps have been studied. The various types of maps show relations among, for example, authors, documents, journals, or keywords, and they have usually been constructed based on citation, co-citation, or bibliographic coupling data or based on data on co-occurrences of keywords in documents. Quite some different techniques are available that can be used for constructing bibliometric maps. Without doubt, the most popular technique is the technique of multidimensional scaling (MDS).[1] MDS has been widely used for constructing maps of authors (e.g. McCain, 1990; White & Griffith, 1981; White & McCain, 1998), documents (e.g., Griffith et al., 1974; Small & Garfield, 1985; Small et al., 1985), journals (e.g., McCain, 1991), and keywords (e.g., Peters & Van Raan, 1993a, 1993b; Tijssen & Van Raan, 1989). Recently, a new mapping technique was introduced that is intended as an alternative to MDS (Van Eck & Waltman, 2007b). This new mapping technique is called VOS, which stands for *visualization of similarities*. VOS has been used for constructing bibliometric maps in a number of studies (Van Eck & Waltman, 2007a; Van Eck, Waltman, Noyons, & Buter, 2010; Van Eck, Waltman, et al., 2006a; Waaijer, Van Bochove, & Van Eck, 2010, 2011).

An extensive comparison between the use of MDS and the use of VOS for constructing bibliometric maps does not yet exist. In this chapter, we present such a comparison. We perform both a theoretical and an empirical analysis. In our theoretical analysis, we discuss the mathematics underlying MDS and VOS and we point out how the two techniques are mathematically related to each other. In our empirical analysis, we compare three approaches for constructing bibliometric maps. Two approaches rely on MDS, and the third approach relies on VOS. We use three data sets in our empirical analysis. One data set comprises co-citations of authors in the field of information science, another data set comprises co-citations of journals in the social sciences, and the third data set comprises co-occurrences of keywords in the field of operations research. Our empirical analysis indicates that maps constructed using either of the MDS approaches may

[1] Other techniques include the VxOrd technique (e.g., Boyack et al., 2005; Klavans & Boyack, 2006b), the graph drawing techniques of Kamada and Kawai (1989) and Fruchterman and Reingold (1991), and the pathfinder network technique (e.g., Schvaneveldt, 1990; Schvaneveldt et al., 1988; White, 2003b). For overviews of various techniques, we refer to Börner et al. (2003) and White and McCain (1997).

suffer from certain artifacts. Maps constructed using the VOS approach do not have this problem. Based on this observation, we conclude that in general maps constructed using the VOS approach provide a more satisfactory representation of the underlying data set than maps constructed using either of the MDS approaches.

This chapter is organized as follows. First, we discuss the use of MDS and VOS for constructing bibliometric maps and we study the mathematical relationship between the two techniques. Next, we present an empirical comparison of three approaches for constructing bibliometric maps, two approaches relying on MDS and one approach relying on VOS. Finally, we summarize the conclusions of our research.

5.2 Multidimensional Scaling

In this section, we discuss the way in which MDS is typically used for constructing bibliometric maps. For more detailed discussions of MDS, we refer to Borg and Groenen (2005) and T. F. Cox and Cox (2001). From now on, we assume that the construction of bibliometric maps is done based on co-occurrence data (which includes co-citation data and bibliographic coupling data as special cases). We use the following mathematical notation. There are n items to be mapped, which are denoted by $1, \ldots, n$. The items can be, for example, authors, documents, journals, or keywords. For $i \neq j$, the number of co-occurrences of items i and j is denoted by c_{ij} (where $c_{ij} = c_{ji}$). The total number of co-occurrences of item i is denoted by c_i. Hence, $c_i = \sum_{j \neq i} c_{ij}$.

Below, we first discuss the calculation of similarities between items, and we then discuss the technique of MDS.

5.2.1 Similarity Measures

MDS is usually not applied directly to co-occurrence frequencies. This is because in general co-occurrence frequencies do not properly reflect similarities between items (e.g., Waltman & Van Eck, 2007). To see this, suppose that journals A and B publish very similar articles. Suppose also that per year journal A publishes ten times as many articles as journal B. Other things being equal, one would expect journal A to receive about ten times as many citations as journal B and to have about ten times as many co-citations with other journals as journal B. It is clear that the fact that journal A has more co-citations with other journals than journal B does not indicate that journal A is more similar to other journals than journal B. It only indicates that journal A publishes more articles than journal B. Because of this, co-occurrence frequencies in general do not properly reflect similarities between items.

To determine similarities between items, co-occurrence frequencies are usually transformed using a similarity measure. Two types of similarity measures can be distinguished, namely direct and indirect similarity measures.[2] Direct similarity measures (Van Eck & Waltman, 2009; also known as local similarity measures, see Ahlgren et al., 2003) determine the similarity between two items by applying a normalization to the co-occurrence frequency of the items. The underlying idea is that co-occurrence frequencies can be interpreted as similarities only after one has corrected for the fact that for some items the total number of occurrences or co-occurrences may be much larger than for other items. Indirect similarity measures (also known as global similarity measures) determine the similarity between two items by comparing two vectors of co-occurrence frequencies. This is based on the idea that the similarity of two items should depend on the way in which each of the two items is related to all other items. The more two items have similar relations with other items, the more the two items should be considered similar. Most researchers interested in mapping authors or journals based on co-citation data rely on indirect similarity measures. Other researchers rely on direct similarity measures. However, direct and indirect similarity measures can both be applied to any type of co-occurrence data. There is, for example, no reason to confine the use of indirect similarity measures to author and journal co-citation data.

Various direct similarity measures are being used in the literature. Especially the cosine and the Jaccard index are very popular. In a recent study (Van Eck & Waltman, 2009), we extensively analyzed a number of well-known direct similarity measures. We argued that the most appropriate measure for normalizing co-occurrence frequencies is the so-called association strength (e.g., Van Eck & Waltman, 2007a; Van Eck, Waltman, et al., 2006a). This measure is also known as the proximity index (e.g., Peters & Van Raan, 1993b; Rip & Courtial, 1984) or as the probabilistic affinity index (e.g., Zitt

[2] Sometimes a distinction is made between similarity measures calculated based on a rectangular occurrence matrix and similarity measures calculated based on a square symmetric co-occurrence matrix (e.g., Schneider et al., 2009). It can be shown that this distinction is mathematically equivalent with our distinction between direct and indirect similarity measures (see also Van Eck & Waltman, 2009).

et al., 2000). The association strength of items i and j is given by

$$\mathrm{AS}_{ij} = \frac{c_{ij}}{c_i c_j}. \tag{5.1}$$

It can be shown that the association strength of items i and j is proportional to the ratio between on the one hand the observed number of co-occurrences of i and j and on the other hand the expected number of co-occurrences of i and j under the assumption that co-occurrences of i and j are statistically independent (Van Eck & Waltman, 2009).

For a long time, the Pearson correlation has been the most popular indirect similarity measure in the literature (e.g., McCain, 1990, 1991; White & Griffith, 1981; White & McCain, 1998). Nowadays, however, it is well known that the use of the Pearson correlation as an indirect similarity measure is not completely satisfactory (Ahlgren et al., 2003; Van Eck & Waltman, 2008). A more satisfactory indirect similarity measure is the well-known cosine.[3] The cosine of items i and j is given by

$$\mathrm{COS}_{ij} = \frac{\sum_{k \neq i,j} c_{ik} c_{jk}}{\sqrt{\sum_{k \neq i,j} c_{ik}^2 \sum_{k \neq i,j} c_{jk}^2}}. \tag{5.2}$$

For a discussion of some other indirect similarity measures, we refer to an earlier paper (Van Eck & Waltman, 2008).

5.2.2 The Technique of Multidimensional Scaling

After similarities between items have been calculated, a map is constructed by applying MDS to the similarities. The aim of MDS is to locate items in a low-dimensional space in such a way that the distance between any two items reflects the similarity or relatedness of the items as accurately as possible. The stronger the relation between two items, the smaller the distance between the items.

Let s_{ij} denote the similarity between items i and j given by some direct or indirect similarity measure. For each pair of items i and j, MDS requires as input a proximity p_{ij} (i.e., a similarity or dissimilarity) and, optionally, a weight w_{ij} ($w_{ij} \geq 0$). In the bibliometric mapping literature, the proximities p_{ij} are typically set equal to the similarities

[3] There are two different similarity measures, a direct and an indirect one, that are both referred to as the cosine. These two measures should not be confused with each other.

s_{ij}. The weights w_{ij} are typically not provided, in which case MDS uses $w_{ij} = 1$ for all i and j. To determine the locations of items in a map, MDS minimizes a so-called stress function. The most commonly used stress function is given by

$$\sigma(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \frac{\sum_{i<j} w_{ij} \left(f(p_{ij}) - \|\mathbf{x}_i - \mathbf{x}_j\|\right)^2}{\sum_{i<j} w_{ij} f(p_{ij})^2}, \tag{5.3}$$

where f denotes a transformation function for the proximities p_{ij} and $\mathbf{x}_i$ denotes the location of item i.[4] Typically, bibliometric maps have two dimensions and rely on the Euclidean distance measure. This means that $\mathbf{x}_i = (x_{i1}, x_{i2})$ and that

$$\|\mathbf{x}_i - \mathbf{x}_j\| = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2}. \tag{5.4}$$

As can be seen from Equation 5.3, MDS determines the locations of items in a map by minimizing the (weighted) sum of the squared differences between on the one hand the transformed proximities of items and on the other hand the distances between items in the map. For this idea to make sense, the transformation function f has to be increasing when the proximities p_{ij} are dissimilarities and decreasing when the proximities p_{ij} are similarities.

Depending on the transformation function f, different types of MDS can be distinguished. The three most important types of MDS are ratio MDS, interval MDS, and ordinal MDS. Ratio and interval MDS are also referred to as metric MDS, while ordinal MDS is also referred to as non-metric MDS. Ratio MDS treats the proximities p_{ij} as measurements on a ratio scale. Likewise, interval and ordinal MDS treat the proximities p_{ij} as measurements on, respectively, an interval and an ordinal scale.[5] In ratio MDS, f is a linear function without an intercept. In interval MDS, f can be any linear function, and in ordinal MDS, f can be any monotone function. We note that it makes no sense to use ratio MDS when the proximities p_{ij} are similarities. This is because f would then have to be a linearly decreasing function through the origin, which means

[4] The stress function in Equation 5.3 is referred to as the normalized raw stress function. Various alternative stress functions are discussed in the MDS literature (e.g., Borg & Groenen, 2005). In this chapter, however, we do not consider these alternative stress functions. The normalized raw stress function is used by most MDS programs, including the PROXSCAL program in SPSS. Some MDS programs, such as the ALSCAL program in SPSS, use a somewhat different stress function.

[5] For a discussion of the concepts of ratio scale, interval scale, and ordinal scale, see Stevens (1946).

that all transformed proximities would be negative or zero. In the bibliometric mapping literature, researchers often do not state which type of MDS they use. The proximities p_{ij} are typically set equal to the similarities s_{ij}, which means that ratio MDS cannot be used. There are a few well-known studies in which the use of ordinal MDS is reported (McCain, 1990; White & Griffith, 1981; White & McCain, 1998).

The stress function in Equation 5.3 can be minimized using an iterative algorithm. Various different algorithms are available. A popular algorithm is the SMACOF algorithm (e.g., Borg & Groenen, 2005). This algorithm relies on a technique known as iterative majorization. The SMACOF algorithm is used by the PROXSCAL program in SPSS.

5.3 VOS

In this section, we discuss the use of VOS for constructing bibliometric maps. The aim of VOS is the same as that of MDS. Hence, VOS aims to locate items in a low-dimensional space in such a way that the distance between any two items reflects the similarity or relatedness of the items as accurately as possible. As discussed below, VOS differs from MDS in the way in which it attempts to achieve this aim.

For each pair of items i and j, VOS requires as input a similarity s_{ij} ($s_{ij} \geq 0$). VOS treats the similarities s_{ij} as measurements on a ratio scale. The similarities s_{ij} are typically calculated using the association strength defined in Equation 5.1 (e.g., Van Eck & Waltman, 2007a; Van Eck, Waltman, et al., 2006a). VOS determines the locations of items in a map by minimizing

$$V(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \sum_{i<j} s_{ij} \left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2 \tag{5.5}$$

subject to

$$\frac{2}{n(n-1)} \sum_{i<j} \left\|\mathbf{x}_i - \mathbf{x}_j\right\| = 1. \tag{5.6}$$

Hence, the idea of VOS is to minimize a weighted sum of the squared distances between all pairs of items. The squared distance between a pair of items is weighed by the similarity between the items. To avoid trivial solutions in which all items have the same location, the constraint is imposed that the average distance between two items must be equal to one.

There are two computer programs in which the VOS mapping technique has been implemented. Both programs are freely available. A simple open source program is available at http://www.neesjanvaneck.nl/vos/, and a more advanced program called VOSviewer (Van Eck & Waltman, 2010) is available at http://www.vosviewer.com. The two programs both use a variant of the SMACOF algorithm mentioned above to perform the minimization of Equation 5.5 subject to Equation 5.6.

We note that the objective function in Equation 5.5 has an interesting property.[6] To show this property, we introduce the idea of the ideal location of an item. We define the ideal location of item i as

$$\mathbf{x}_i^* = \frac{\sum_{j \neq i} s_{ij} \mathbf{x}_j}{\sum_{j \neq i} s_{ij}}. \tag{5.7}$$

That is, the ideal location of item i is defined as a weighted average of the locations of all other items, where the location of an item is weighed by the item's similarity with item i. (Notice the analogy with the concept of center of gravity in physics.) The ideal location of an item seems to be the most natural location an item can have. Because of this, it seems desirable that items are located as close as possible to their ideal location. This is exactly what the objective function in Equation 5.5 seeks to achieve. To see this, suppose that the locations of all items except item i are fixed, and ignore the constraint in Equation 5.6. Minimization of the objective function can then be performed analytically and results in $\mathbf{x}_i$ being equal to $\mathbf{x}_i^*$ defined in Equation 5.7. Hence, if the locations of all items except item i are fixed and if the constraint is ignored, minimization of the objective function causes item i to be located exactly at its ideal location. Of course, items do not have fixed locations, and solutions are determined not only by the objective function but also by the constraint. For these reasons, items will in general not be located exactly at their ideal location. However, due to the objective function, items at least tend to be located close to their ideal location.

[6] Mapping techniques based on the objective function in Equation 5.5 have also been proposed by Belkin and Niyogi (2003) and by Davidson, Hendrickson, Johnson, Meyers, and Wylie (1998). However, the constraints used by these researchers are different from the constraint in Equation 5.6. In our experience, the constraint in Equation 5.6 yields much more satisfactory results than the alternative constraints used by other researchers.

5.4 Relationship Between Multidimensional Scaling and VOS

In this section, we study the mathematical relationship between MDS and VOS. We show that, under certain conditions, MDS and VOS are closely related.

As discussed above, when researchers use MDS for constructing bibliometric maps, they typically rely on ordinal or interval MDS. However, when MDS is applied to similarities calculated using the association strength defined in Equation 5.1, the use of ordinal or interval MDS is not completely satisfactory. This can be seen as follows. Suppose that items i and j have twice as many co-occurrences as items i and k. Suppose also that the total number of co-occurrences of item j equals the total number of co-occurrences of item k. Calculation of similarities using the association strength then yields $s_{ij} = 2s_{ik}$. Based on this, it seems natural to expect that in a map that perfectly represents the co-occurrences the distance between items i and j equals half the distance between items i and k. Of course, due to the inherent limitations of a low-dimensional Euclidean space, a map in which co-occurrences are perfectly represented usually cannot be constructed. However, ordinal and interval MDS do not even try to construct such a map. This is because in some sense the transformation function f has too much freedom in these types of MDS. In ordinal MDS, for example, f can be any monotonically decreasing function, which means that any map in which the distance between items i and j is smaller than the distance between items i and k may serve as a perfect representation of the equality $s_{ij} = 2s_{ik}$. Hence, ordinal MDS may be indifferent between, for example, a map in which the distance between items i and j equals exactly half the distance between items i and k and a map in which the distance between items i and j is just slightly smaller than the distance between items i and k.

We now propose an alternative way in which MDS can be applied to similarities calculated using the association strength (or to any other similarities that can be treated as measurements on a ratio scale). Our alternative approach does not have the above-mentioned disadvantage of ordinal and interval MDS. In our approach, we choose the transformation function f to be simply the identity function, which means that $f(p_{ij}) = p_{ij}$. Using this transformation function, it is easy to see that minimization of the stress function in Equation 5.3 is equivalent with minimization of

$$\hat{\sigma}(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \sum_{i<j} w_{ij} \left\| \mathbf{x}_i - \mathbf{x}_j \right\|^2 - 2 \sum_{i<j} w_{ij} p_{ij} \left\| \mathbf{x}_i - \mathbf{x}_j \right\|. \tag{5.8}$$

Equation 5.8 makes sense only if the proximities p_{ij} are dissimilarities. Because of this, we cannot set the proximities p_{ij} equal to the similarities s_{ij}. Instead, we first have to convert the similarities s_{ij} into dissimilarities d_{ij}. Converting similarities into dissimilarities can be done in many ways. We use the conversion given by $d_{ij} = 1/s_{ij}$. This conversion has the natural property that if in a perfect map the distance between one pair of items is twice as large as the distance between another pair of items, the similarity between the first pair of items is twice as low as the similarity between the second pair of items. Substitution of $p_{ij} = d_{ij} = 1/s_{ij}$ in Equation 5.8 yields

$$\hat{\sigma}(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \sum_{i<j} w_{ij} \left\| \mathbf{x}_i - \mathbf{x}_j \right\|^2 - 2 \sum_{i<j} w_{ij} \frac{1}{s_{ij}} \left\| \mathbf{x}_i - \mathbf{x}_j \right\|. \tag{5.9}$$

If two items i and j do not have any co-occurrences with each other, Equation 5.1 implies that $s_{ij} = 0$. This results in a division by zero in Equation 5.9. To circumvent this problem, we do not set the weights w_{ij} equal to one, but we instead define the weights w_{ij} as an increasing function of the similarities s_{ij}. More specifically, we define $w_{ij} = s_{ij}$.[7] Equation 5.9 then becomes

$$\hat{\sigma}(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \sum_{i<j} s_{ij} \left\| \mathbf{x}_i - \mathbf{x}_j \right\|^2 - 2 \sum_{i<j} \left\| \mathbf{x}_i - \mathbf{x}_j \right\|. \tag{5.10}$$

Interestingly, there turns out to be a close relationship between on the one hand the problem of minimizing Equation 5.10 and on the other hand the problem of minimizing Equation 5.5 subject to Equation 5.6. This is stated formally in the following proposition.

Proposition 5.1.

[7] Hence, w_{ij} increases linearly with s_{ij}. This is the most natural way to define w_{ij}. If w_{ij} increases slower than linearly with s_{ij}, the division by zero problem remains. If w_{ij} increases faster than linearly with s_{ij}, there is no penalty for locating two completely non-similar items close to each other in a map. We further note that $w_{ij} = s_{ij}$ is equivalent with $w_{ij} = 1/d_{ij}$. This is exactly how weights are chosen in the well-known Sammon mapping variant of MDS (Sammon, 1969).

(i) If $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ is a globally optimal solution to the problem of minimizing Equation 5.10, then there exists a positive real number c such that $c\mathbf{X}$ is a globally optimal solution to the problem of minimizing Equation 5.5 subject to Equation 5.6.

(ii) If $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ is a globally optimal solution to the problem of minimizing Equation 5.5 subject to Equation 5.6, then there exists a positive real number c such that $c\mathbf{X}$ is a globally optimal solution to the problem of minimizing Equation 5.10.

The proof of this proposition is provided in Appendix 5.A. It follows from the proposition that, under certain conditions, MDS and VOS are closely related. More specifically, the proposition indicates that VOS can be regarded as a kind of weighted MDS with proximities and weights chosen in a special way.

5.5 Empirical Comparison

We now present an empirical comparison of three approaches for constructing bibliometric maps. Two approaches rely on MDS, and the third approach relies on VOS. The two MDS approaches differ from each other in the similarity measure they use. One MDS approach uses a direct similarity measure, namely the association strength defined in Equation 5.1. The other MDS approach uses an indirect similarity measure, namely the cosine defined in Equation 5.2. From now on, we refer to the two MDS approaches as the MDS-AS approach and the MDS-COS approach. Like the MDS-AS approach, the VOS approach also uses the association strength similarity measure. Because VOS has been developed to be used specifically in combination with this similarity measure, we do not study the use of VOS in combination with other similarity measures.

Below, we first discuss the data sets that we use in our empirical comparison, and we then discuss the results of the comparison. We also briefly consider the phenomenon of circular maps.

5.5.1 Data Sets

We use three data sets in our empirical comparison. One data set comprises co-citations of authors in the field of information science, another data set comprises co-citations of journals in the social sciences, and the third data set comprises co-occurrences of keywords in the field of operations research. We refer to the data sets as, respectively, the authors data set, the journals data set, and the keywords data set. All three data sets were obtained from the Web of Science database. We have made the data sets available at http://www.neesjanvaneck.nl/comparison_mds_vos/.

The authors data set was collected as follows. We first delineated the field of information science. To do so, we selected the 36 journals that, based on co-citation data, are most closely related to the *Journal of the American Society for Information Science and Technology* (*JASIST*).[8] These journals and *JASIST* itself constituted our set of information science journals. This set of journals is shown in Table 5.1. Next, we selected all articles with at least 4 citations (excluding self citations) that were published in our set of information science journals between 1999 and 2008. We then counted for each author the number of selected articles.[9] All authors with at least 3 selected articles were included in the authors data set. There were 405 authors that satisfied this criterion. Finally, we counted the number of co-citations of each pair of authors in the authors data set. The co-citation frequency of two authors takes into account all articles published by the authors in our set of information science journals between 1999 and 2008.

To collect the journals data set, we first selected all journals in the Web of Science database that belong to at least one social science subject category. We then counted the number of co-citations of each pair of journals. We took into account all citations from articles published between 2004 and 2008 to articles published at most 10 years earlier. Finally, we included in the journals data set all journals with more than 25 co-citations. There were 2079 journals that satisfied this criterion.

The keywords data set has already been used in an earlier paper (Van Eck, Waltman, Noyons, & Buter, 2010). The data set includes 831 keywords that were automatically identified in the abstracts (and titles) of 7492 articles published in 15 operations research journals between 2001 and 2006. The co-occurrence frequency of two keywords was obtained by counting the number of abstracts in which the keywords both occur.

[8] The *Journal of the American Society for Information Science and Technology* and its predecessor, the *Journal of the American Society for Information Science*, were treated as a single journal.

[9] Author name disambiguation was performed using an algorithm that we have developed ourselves. A few corrections were made manually. Unlike in some other author co-citation studies, all authors of an article were taken into account, not just the first author.

Table 5.1: Set of journals used to delineate the field of information science.

ACM Transactions on Information Systems
Annual Review of Information Science and Technology
Aslib Proceedings
Bulletin of the Medical Library Association
College and Research Libraries
Computers and the Humanities
Electronic Library
Information Processing and Management
Information Research-An International Electronic Journal
Information Retrieval
Information Technology and Libraries
Interlending and Document Supply
Journal of Academic Librarianship
Journal of Documentation
Journal of Information Science
Journal of Librarianship and Information Science
Journal of Scholarly Publishing
Journal of the American Society for Information Science and Technology & Knowledge Organization
Law Library Journal
Learned Publishing
Library and Information Science Research
Library Collections Acquisitions and Technical Services
Library Journal
Library Quarterly
Library Resources and Technical Services
Library Trends
Libri
Online
Online Information Review
Portal-Libraries and the Academy
Proceedings of the ASIS Annual Meeting
Program-Electronic Library and Information Systems
Reference and User Services Quarterly
Research Evaluation
Scientometrics
Serials Review

Table 5.2: Stress values calculated using Equation 5.3 for the MDS-AS and MDS-COS approaches.

	MDS-AS	MDS-COS
Authors	0.12	0.04
Journals	0.14	0.05
Keywords	0.16	0.07

5.5.2 Results

For each of the three data sets that we consider, three maps were constructed, one using the MDS-AS approach, one using the MDS-COS approach, and one using the VOS approach. All maps are two-dimensional. MDS was run using the PROXSCAL program in SPSS. Both MDS approaches used ordinal MDS.[10] 100 random starts of the optimization algorithm were used in all three mapping approaches.[11] For the MDS approaches, stress values calculated using Equation 5.3 are reported in Table 5.2. The nine maps that were obtained are available online at http://www.neesjanvaneck.nl/comparison_mds_vos/, where they can be examined in detail using the VOSviewer software (Van Eck & Waltman, 2010). The global structure of each of the maps is shown in Figure 5.1. In this figure, circles are used to indicate the location of an item. The size of a circle reflects an item's total number of co-occurrences. In order to facilitate the interpretation of the maps, items were clustered using a clustering technique. We used the clustering technique proposed by (Waltman, Van Eck, & Noyons, 2010). Colors are used to indicate the cluster to which an item belongs.

To evaluate the maps shown in Figure 5.1, our criterion is the accuracy with which distances in a map reflect the similarity or relatedness of items. Sometimes other criteria are considered important as well, such as a roughly equal distribution of items in a map or a clearly visible separation between clusters of items. It is argued that maps satisfying such 'aesthetic' criteria are easier to interpret. Clearly, different criteria can be conflict-

[10] Ties in the data were kept tied. This is sometimes referred to as the secondary approach to ties (Borg & Groenen, 2005). The secondary approach to ties is the default setting in the PROXSCAL program.

[11] In the case of the MDS-AS approach, rather stringent convergence criteria were required for the optimization algorithm. Without such criteria, the algorithm was very sensitive to local optima. Due to the stringent convergence criteria, the application of the MDS-AS approach to the journals data set took more than two days of computing time on a standard desktop computer. For comparison, the application of the VOS approach to the same data set took less than ten minutes of computing time.

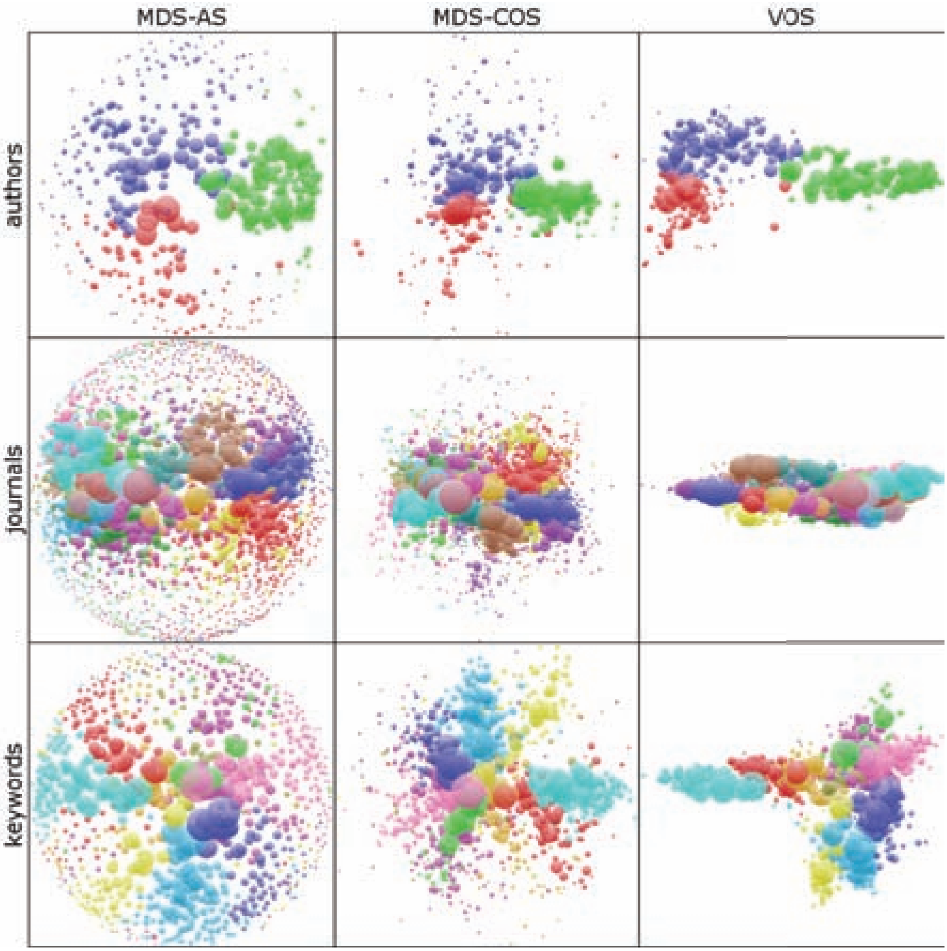


Figure 5.1: Global structure of nine maps. Each row corresponds with a data set. Each column corresponds with a mapping approach.

ing with each other. For example, having well-separated clusters of items may conflict with having distances that accurately reflect the similarity or relatedness of items. In this chapter, our choice is to focus exclusively on the latter criterion. This is consistent with the objective for which techniques such as MDS and VOS were originally developed. Other techniques, often referred to as graph-drawing techniques (e.g., Fruchterman & Reingold, 1991; Kamada & Kawai, 1989), were developed with a different objective in mind and give more weight to aesthetic criteria such as the ones mentioned above. However, these techniques, although valuable in their own right, are not the subject of study of this chapter.

As can be seen in Figure 5.1, the MDS-AS, MDS-COS, and VOS approaches produce quite different maps. Although all three approaches succeed to some extent in separating items belonging to different clusters, the global structure of the maps produced by the three approaches is very different. The MDS-AS approach produces maps with the shape of an almost perfect circle. The distribution of items within a circle is more or less uniform, in particular when the number of items is large, as in the case of the journals and keywords data sets. The maps produced by the MDS-COS approach also seem to have a tendency to be somewhat circular, but this effect is much weaker than in the case of the MDS-AS approach. A notable property of the maps produced by the two MDS approaches is that important items (i.e., items with a large number of co-occurrences) tend to be located toward the center of a map. This is especially clear in the case of the authors and keywords data sets. Many relatively unimportant items are scattered throughout the periphery of a map. In the maps produced by the VOS approach, no effects are visible similar to those observed in the case of the two MDS approaches. Hence, the VOS approach does not seem to have a tendency to produce circular maps. It also does not seem to locate important items toward the center of a map. Instead, the VOS approach seems to produce maps in which important and less important items are distributed fairly evenly over the central and peripheral areas.

We emphasize that the results shown in Figure 5.1 are quite robust. The results do not change much when interval MDS is used rather than ordinal MDS. Using MDS combined with direct similarity measures other than the association strength also does not have much effect on the results. Furthermore, the results shown in Figure 5.1 are relatively independent of the data sets that we use. We investigated numerous other data sets, and this yielded very similar results. However, the almost perfectly circular structure of maps produced by the MDS-AS approach was not observed in the case of data sets with only a relatively small number of items (e.g., less than 100 items). In the bibliometric mapping literature, a clear example of a circular map produced by MDS can be found in a study by Blatt (2009). Blatt used a data set of almost 5000 items. Most bibliometric mapping studies reported in the literature rely on data sets with a much smaller number of items. In such studies, MDS typically does not produce circular maps, although a tendency toward a circular structure sometimes seems visible.[12]

We now focus on one data set in more detail. This is the data set of authors in the field of information science. We note that somewhat similar data sets have also been analyzed in a paper by Persson (1994), in a well-known study by White and McCain (1998), and more recently in the work of Zhao and Strotmann (2008a, 2008b, 2008c) and C. Chen, Ibekwe-SanJuan, and Hou (2010). Maps of the authors data set constructed using the MDS-AS, MDS-COS, and VOS approaches are shown in Figures 5.2, 5.3, and 5.4, respectively. These are the same maps as the ones shown in the top row of Figure 5.1.

In various studies of the field of information science (e.g., Åström, 2007; White & McCain, 1998; Zhao & Strotmann, 2008a, 2008b, 2008c), it has been found that the field consists of two quite independent subfields. We adopt the terminology of Åström (2007) and refer to the subfields as information seeking and retrieval (ISR) and informetrics. Comparing the maps in Figures 5.2, 5.3, and 5.4, it can be observed that the separation of the subfields is clearly visible in the VOS map, somewhat less visible in the MDS-COS map, and least visible in the MDS-AS map.[13] In the VOS map, the right part represents the informetrics subfield (e.g., Egghe, Glänzel, and Van Raan) and the left part represents the ISR subfield (e.g., Baeza-Yates, Jansen, Robertson, Spink, Tenopir, and Wilson). There is only a relatively weak connection between the subfields. In the MDS-COS map, the middle right part represents the informetrics subfield and the rest of the map represents the ISR subfield. A striking property of the map is that the ISR subfield is rather scattered, with the most prominent authors (in terms of the number of co-citations) appearing in the center of the map and many somewhat less prominent authors appearing in the periphery. In the MDS-AS map, the middle right part represents the informetrics subfield and the rest of the map represents the ISR subfield. As noted earlier, the map has the shape of an almost perfect circle. The informetrics subfield is partly surrounded by the ISR subfield, with some empty space indicating the separation of the subfields. Prominent authors in the ISR subfield are located toward the center of

[12] We note that MDS is not the only mapping technique with a tendency to produce circular maps. See for example Boyack et al. (2005), Heimeriks et al. (2003), Klavans and Boyack (2006b), and Noll, Fröhlich, and Schiebel (2002).

[13] In the maps, the green cluster corresponds with the informetrics subfield and the blue and red clusters correspond with the ISR subfield.

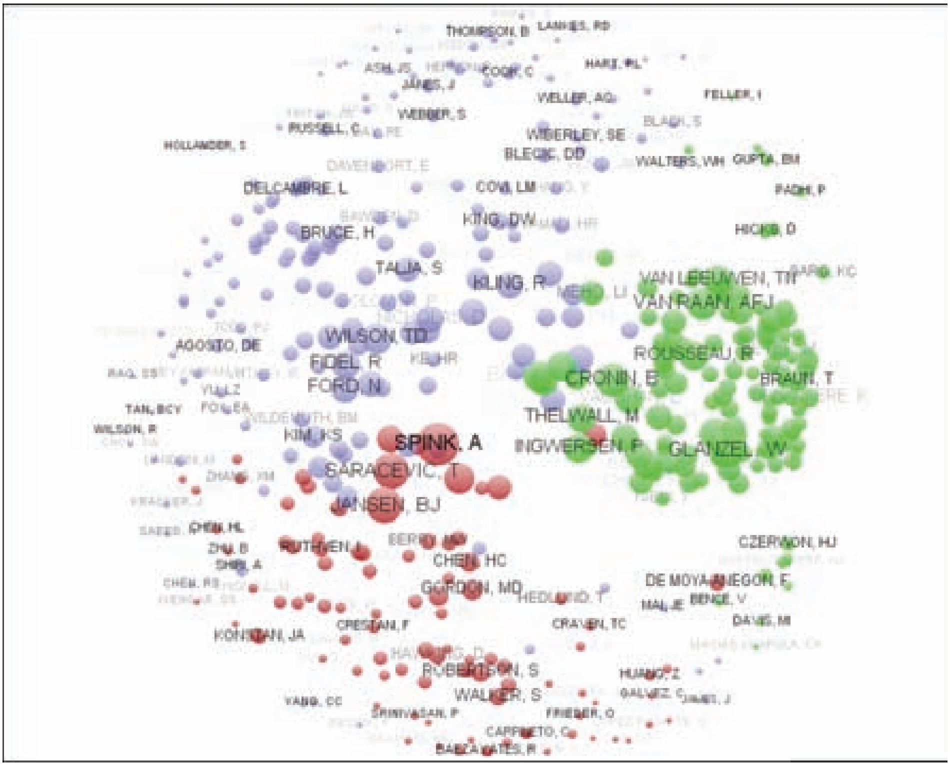


Figure 5.2: Map of the authors data set constructed using the MDS-AS approach.

the map. Less prominent authors tend to be located in the top and bottom parts of the map. This is quite similar to the MDS-COS map.

A distinction is sometimes made between "hard" and "soft" ISR research (e.g., Åström, 2007; Persson, 1994; White & McCain, 1998). Hard ISR research is system-oriented and is for example concerned with the development and the experimental evaluation of information retrieval algorithms. Soft ISR research, on the other hand, is user-oriented and studies for example users' information needs and information behavior. The distinction between hard and soft ISR research is visible in all three maps. In the VOS map, the lower left part represents hard ISR research (e.g., Baeza-Yates and Robertson) and the middle left and upper left parts represent soft ISR research (e.g., Jansen, Spink, Tenopir, and Wilson). In the MDS-COS and MDS-AS maps, the lower

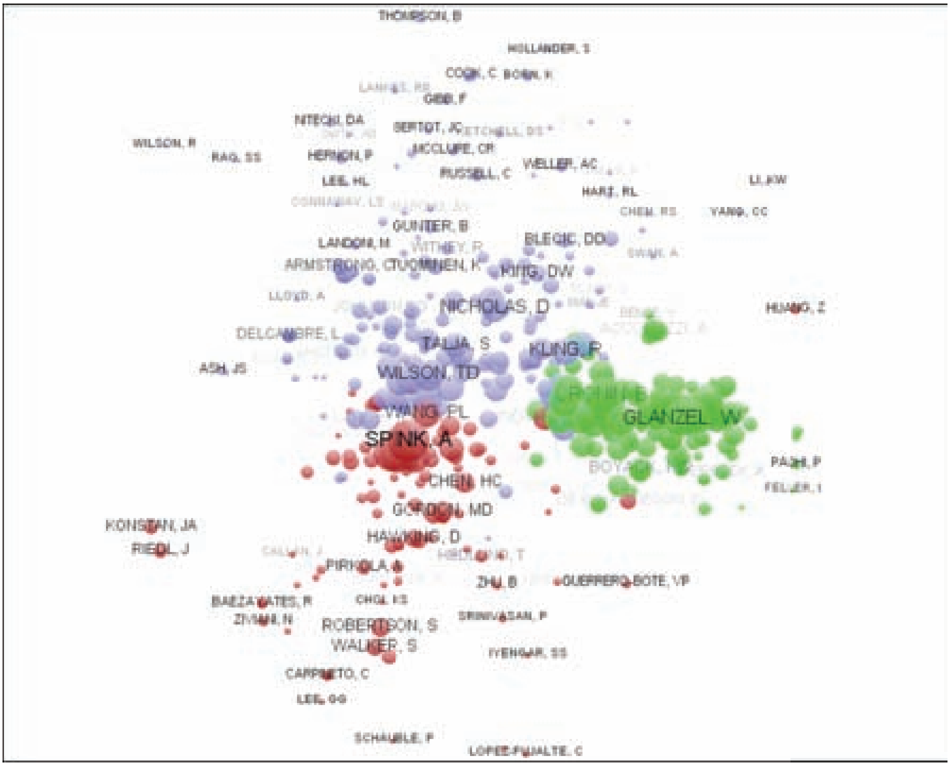


Figure 5.3: Map of the authors data set constructed using the MDS-COS approach.

part represents hard ISR research and the middle and upper parts represent soft ISR research. As can be seen from all three maps, there is much more soft ISR research than hard ISR research. This is similar to what was found by Åström (2007).

The above comparison of the three maps of the authors data set indicates that the MDS-AS, MDS-COS, and VOS approaches all three succeed reasonably well in locating similar authors close to each other. However, the comparison also makes clear that the MDS-AS and MDS-COS approaches suffer from serious artifacts. Both approaches have a tendency to locate the most prominent authors in the center of a map and less prominent authors in the periphery. Due to this tendency, the separation of subfields becomes more difficult to see. The MDS-AS approach also has a strong tendency to locate authors in a circular structure. This tendency further distorts the way in which

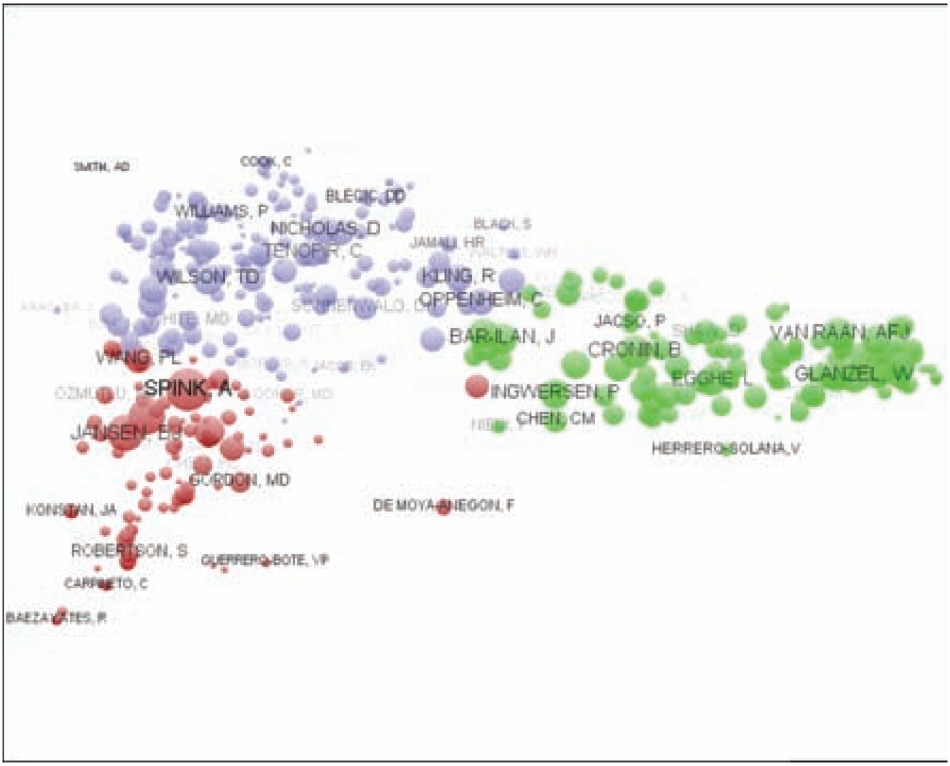


Figure 5.4: Map of the authors data set constructed using the VOS approach.

a field is represented. Unlike the two MDS approaches, the VOS approach does not seem to suffer from artifacts. That is, the VOS approach does not seem to impose any artificial structure on a map. Our findings based on the maps of the authors data set are confirmed when examining the maps of the journals and keywords data sets. A detailed discussion of the latter maps is beyond the scope of this chapter. We note, however, that an examination of these maps indicates the same artifacts of the MDS-AS and MDS-COS approaches as discussed above. The interested reader can verify this at http://www.neesjanvaneck.nl/comparison_mds_vos/.

The maps in Figures 5.2 and 5.3 indicate the consequences of the artifacts from which the MDS-AS and MDS-COS approaches suffer. In these maps, a number of prominent ISR authors (e.g., Spink, Wang, and Wilson) are located equally close or even closer to various informetrics authors than to some of their less prominent ISR colleagues. However, contrary to what the maps seem to suggest, there is in fact very little interaction between the prominent ISR authors and the informetrics authors. The relatively small distance between these two groups of authors therefore does not properly reflect the structure of the field of information science. The small distance is merely a technical artifact, caused by the tendency of the MDS-AS and MDS-COS approaches to locate important items in the center of a map. It follows from this observation that distances in maps constructed using the MDS approaches may not always give an accurate representation of the relatedness of items. Hence, in the case of the MDS approaches, the validity of the interpretation of a distance as an (inverse) measure of relatedness seems questionable. The VOS map in Figure 5.4 does properly reflect the large separation between the prominent ISR authors and the informetrics authors. In this map, the interpretation of a distance as a measure of relatedness therefore seems valid. We note that the journal and keyword maps available online provide similar examples of the consequences of the MDS artifacts.

5.5.3 Explanation for Circular Maps

Finally, let us consider the phenomenon of the circular maps produced by the MDS-AS approach in somewhat more detail. Although this phenomenon may seem puzzling at first sight, it actually has a quite straightforward explanation.[14] Co-occurrence data typically consists for a large part of zeros. For example, in the case of the authors, journals, and keywords data sets, respectively 73%, 75%, and 89% of all pairs of items have zero co-occurrences. It follows from Equation 5.1 that, when two items have a co-occurrence frequency of zero, their association strength equals zero as well. This means that in the MDS-AS approach MDS is typically applied to similarity data that consists largely of zeros. MDS attempts to determine the locations of items in a map in such a way that for each pair of items with a similarity of zero the distance between the items is the same. In the case of similarity data that consists largely of zeros, it is not possible to construct a low-dimensional map with exactly the same distance between each pair of items with a similarity of zero. MDS can only try to approximate such a map as closely as possible. Our empirical analysis indicates that the best possible

[14] For an explanation similar to ours, see Martín-Merino and Muñoz (2004).

approximation is a map with an almost perfectly circular structure. This is in fact not a very surprising finding, since it is well known in the MDS literature that MDS produces perfectly circular maps when all similarities between items are equal (Borg & Groenen, 2005; De Leeuw & Stoop, 1984; for a rigorous mathematical analysis, see Buja, Logan, Reeds, & Shepp, 1994). In our empirical analysis, not all similarities between items are equal but only a large proportion. The circular structure of our maps is therefore not perfect but almost perfect.

In our empirical analysis, the VOS approach is applied to the same similarity data as the MDS-AS approach. Hence, the VOS approach is also applied to similarity data that consists for a large part of zeros. This raises the question why, unlike the MDS-AS approach, the VOS approach does not produce circular maps. To answer this question, recall how MDS and VOS are related to each other. As discussed earlier, VOS can be regarded as a kind of weighted MDS with proximities and weights chosen in a special way. More precisely, in the case of VOS, the proximity of two items is set equal to the inverse of the similarity of the items. The weight of two items is set equal to the similarity of the items. From this point of view, one can say that the VOS approach distinguishes itself from the MDS-AS approach in that it does not give equal weight to all pairs of items. The VOS approach gives more weight to more similar pairs of items. It gives little weight to pairs of items with a low similarity. As mentioned above, similarity data is typically dominated by low values, in particular by zeros. These low values cause the MDS-AS approach to produce circular maps. In the case of the VOS approach, however, pairs of items with a low similarity receive little weight and therefore have little effect on a map. Because of this, the VOS approach does not produce circular maps.

5.6 Conclusions

VOS is a new mapping technique that is intended as an alternative to the well-known technique of MDS. We have presented an extensive comparison between the use of MDS and the use of VOS for constructing bibliometric maps. Our analysis has been partly theoretical and partly empirical. In our theoretical analysis, we have studied the mathematical relationship between MDS and VOS. We have shown that VOS can be regarded as a kind of weighted MDS with proximities and weights chosen in a special way. In our empirical analysis, we have compared three approaches for constructing bibliometric maps, two approaches relying on MDS and one approach relying on VOS. We have found that maps constructed using the VOS approach provide a more satisfactory representation of the underlying data set than maps constructed using either of the MDS approaches. The somewhat disappointing performance of the MDS approaches is due to two artifacts from which these approaches suffer. One artifact is the tendency to locate the most important items in the center of a map and less important items in the periphery. The other artifact is the tendency to locate items in a circular structure. Unlike the MDS approaches, the VOS approach does not seem to suffer from artifacts. It is worth emphasizing that our empirical findings are quite robust. We have made the same findings for three fairly different data sets. These data sets differ from each other in size (405, 831, or 2079 items), in type of item (authors, journals, or keywords), and in concept of similarity (co-citation in a reference list or co-occurrence in an abstract). We note, however, that in the case of small data sets (e.g., data sets of less than 100 items) the artifacts of the MDS approaches tend to be much less serious. Hence, the VOS approach yields improved results mainly in the case of medium and large data sets.

The interested reader who would like to try out the VOS approach to bibliometric mapping can easily do so using the VOSviewer software (Van Eck & Waltman, 2010) that is freely available at http://www.vosviewer.com. The software offers a graphical user interface that provides easy access to the VOS mapping technique. In addition, the software also comprehensively supports the visualization and interactive examination of bibliometric maps.

5.A Appendix

In this appendix, a proof of Proposition 5.1 is provided. The two parts of the proposition will be proven separately. Both parts will be proven by contradiction.

First consider part (i) of Proposition 5.1. Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ denote a globally optimal solution to the problem of minimizing Equation 5.10, and let $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ denote a globally optimal solution to the problem of minimizing Equation 5.5 subject to

Equation 5.6. Let c be given by

$$c = \frac{n(n-1)}{2\sum_{i<j} \|\mathbf{x}_i - \mathbf{x}_j\|}. \tag{5.11}$$

Furthermore, define $\mathbf{U} = c\mathbf{X}$ and $\mathbf{V} = \mathbf{Y}/c$. It follows from Equation 5.11 that $\mathbf{U}$ satisfies the constraint in Equation 5.6. Assume that $\mathbf{U}$ is not a globally optimal solution to the problem of minimizing Equation 5.5 subject to Equation 5.6. This assumption implies that

$$\sum_{i<j} s_{ij} \|\mathbf{u}_i - \mathbf{u}_j\|^2 > \sum_{i<j} s_{ij} \|\mathbf{y}_i - \mathbf{y}_j\|^2. \tag{5.12}$$

It then follows that

$$\sum_{i<j} s_{ij} \|\mathbf{x}_i - \mathbf{x}_j\|^2 > \sum_{i<j} s_{ij} \|\mathbf{v}_i - \mathbf{v}_j\|^2. \tag{5.13}$$

Extending both the left-hand side and the right-hand side of this inequality with an additional term, where the additional term in the left-hand side equals the additional term in the right-hand side, yields

$$\sum_{i<j} s_{ij} \|\mathbf{x}_i - \mathbf{x}_j\|^2 - 2\sum_{i<j} \|\mathbf{x}_i - \mathbf{x}_j\| > \sum_{i<j} s_{ij} \|\mathbf{v}_i - \mathbf{v}_j\|^2 - 2\sum_{i<j} \|\mathbf{v}_i - \mathbf{v}_j\|. \tag{5.14}$$

This inequality implies that $\mathbf{X}$ is not a globally optimal solution to the problem of minimizing Equation 5.10. However, this contradicts the way in which $\mathbf{X}$ was defined. Consequently, the assumption that $\mathbf{U}$ is not a globally optimal solution to the problem of minimizing Equation 5.5 subject to Equation 5.6 must be false. This proves part (i) of Proposition 5.1.

Now consider part (ii) of Proposition 5.1. This part will be proven in a similar way as part (i). Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ denote a globally optimal solution to the problem of minimizing Equation 5.5 subject to Equation 5.6, and let $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ denote a globally optimal solution to the problem of minimizing Equation 5.10. Let c be given by

$$c = \frac{2\sum_{i<j} \|\mathbf{y}_i - \mathbf{y}_j\|}{n(n-1)}. \tag{5.15}$$

Furthermore, define $\mathbf{U} = c\mathbf{X}$ and $\mathbf{V} = \mathbf{Y}/c$. It follows from Equation 5.15 that $\mathbf{V}$

satisfies the constraint in Equation 5.6. Assume that $\mathbf{U}$ is not a globally optimal solution to the problem of minimizing Equation 5.10. This assumption implies that

$$\sum_{i<j} s_{ij} \left\|\mathbf{u}_i - \mathbf{u}_j\right\|^2 - 2\sum_{i<j} \left\|\mathbf{u}_i - \mathbf{u}_j\right\| > \sum_{i<j} s_{ij} \left\|\mathbf{y}_i - \mathbf{y}_j\right\|^2 - 2\sum_{i<j} \left\|\mathbf{y}_i - \mathbf{y}_j\right\|. \quad (5.16)$$

In this inequality, the second term in the left-hand side equals the second term in the right-hand side. The inequality can therefore be simplified to

$$\sum_{i<j} s_{ij} \left\|\mathbf{u}_i - \mathbf{u}_j\right\|^2 > \sum_{i<j} s_{ij} \left\|\mathbf{y}_i - \mathbf{y}_j\right\|^2. \quad (5.17)$$

It then follows that

$$\sum_{i<j} s_{ij} \left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2 > \sum_{i<j} s_{ij} \left\|\mathbf{v}_i - \mathbf{v}_j\right\|^2. \quad (5.18)$$

This inequality implies that $\mathbf{X}$ is not a globally optimal solution to the problem of minimizing Equation 5.5 subject to Equation 5.6. However, this contradicts the way in which $\mathbf{X}$ was defined. Consequently, the assumption that $\mathbf{U}$ is not a globally optimal solution to the problem of minimizing Equation 5.10 must be false. This proves part (ii) of Proposition 5.1. The proof of the proposition is now complete.

Chapter 6

A Unified Approach to Mapping and Clustering of Bibliometric Networks[*]

Abstract

In the analysis of bibliometric networks, researchers often use mapping and clustering techniques in a combined fashion. Typically, however, mapping and clustering techniques that are used together rely on very different ideas and assumptions. We propose a unified approach to mapping and clustering of bibliometric networks. We show that the VOS mapping technique and a weighted and parameterized variant of modularity-based clustering can both be derived from the same underlying principle. We illustrate our proposed approach by producing a combined mapping and clustering of the most frequently cited publications that appeared in the field of information science in the period 1999-2008.

6.1 Introduction

In bibliometric and scientometric research, a lot of attention is paid to the analysis of networks of, for example, documents, keywords, authors, or journals. Mapping and clustering techniques are frequently used to study such networks. The aim of these techniques is to provide insight into the structure of a network. The techniques are used to address questions such as:

[*] This chapter is based on Waltman, Van Eck, and Noyons (2010).

- What are the main topics or the main research fields within a certain scientific domain?
- How do these topics or these fields relate to each other?
- How has a certain scientific domain developed over time?

To satisfactorily answer such questions, mapping and clustering techniques are often used in a combined fashion. Various different approaches are possible. One approach is to construct a map in which the individual nodes in a network are shown and to display a clustering of the nodes on top of the map, for example by marking off areas in the map that correspond with clusters (e.g., McCain, 1990; White & Griffith, 1981) or by coloring nodes based on the cluster to which they belong (e.g., Leydesdorff & Rafols, 2009; Van Eck, Waltman, Dekker, & Van den Berg, 2010). Another approach is to first cluster the nodes in a network and to then construct a map in which clusters of nodes are shown. This approach is for example taken in the work of Small and colleagues (e.g., Small et al., 1985) and in earlier work of our own institute (e.g., Noyons, Moed, & Van Raan, 1999). A third approach is to first construct a map in which the individual nodes in a network are shown and to then cluster the nodes based on their coordinates in the map (e.g., Boyack et al., 2005; Klavans & Boyack, 2006b).

In the bibliometric and scientometric literature, the most commonly used combination of a mapping and a clustering technique is the combination of multidimensional scaling and hierarchical clustering (for early examples, see McCain, 1990; Peters & Van Raan, 1993a; Small et al., 1985; White & Griffith, 1981). However, various alternatives to multidimensional scaling and hierarchical clustering have been introduced in the literature, especially in more recent work, and these alternatives are also often used in a combined fashion. A popular alternative to multidimensional scaling is the mapping technique of Kamada and Kawai (1989; see e.g. Leydesdorff & Rafols, 2009; Noyons & Calero-Medina, 2009), which is sometimes used together with the pathfinder network technique (Schvaneveldt et al., 1988; see e.g. C. Chen, 1999; de Moya-Anegón et al., 2007; White, 2003b). Two other alternatives to multidimensional scaling are the VxOrd mapping technique (e.g., Boyack et al., 2005; Klavans & Boyack, 2006b) and our own VOS mapping technique (e.g., Van Eck, Waltman, Dekker, & Van den Berg, 2010). Factor analysis, which has been used in a large number of studies (e.g., de Moya-

Anegón et al., 2007; Leydesdorff & Rafols, 2009; Zhao & Strotmann, 2008c), may be seen as a kind of clustering technique and, consequently, as an alternative to hierarchical clustering. Another alternative to hierarchical clustering is clustering based on the modularity function of Newman and Girvan (2004; see e.g. Wallace, Gingras, & Duhon, 2009; Zhang, Liu, Janssens, Liang, & Glänzel, 2010).

As we have discussed, mapping and clustering techniques have a similar objective, namely to provide insight into the structure of a network, and the two types of techniques are often used together in bibliometric and scientometric analyses. However, despite their close relatedness, mapping and clustering techniques have typically been developed separately from each other. This has resulted in techniques that have little in common. That is, mapping and clustering techniques are based on different ideas and rely on different assumptions. In our view, when a mapping and a clustering technique are used together in the same analysis, it is generally desirable that the techniques are based on similar principles as much as possible. This enhances the transparency of the analysis and helps to avoid unnecessary technical complexity. Moreover, by using techniques that rely on similar principles, inconsistencies between the results produced by the techniques can be avoided. In this chapter, we propose a unified approach to mapping and clustering of bibliometric networks. We show how a mapping and a clustering technique can both be derived from the same underlying principle. In doing so, we establish a relation between on the one hand the VOS mapping technique (Van Eck & Waltman, 2007b; Van Eck, Waltman, Dekker, & Van den Berg, 2010) and on the other hand clustering based on a weighted and parameterized variant of the well-known modularity function of Newman and Girvan (2004).

The chapter is organized as follows. We first present our proposal for a unified approach to mapping and clustering. We then discuss how the proposed approach is related to earlier work published in the physics literature. Next, we illustrate an application of the proposed approach by producing a combined mapping and clustering of frequently cited publications in the field of information science. Finally, we summarize the conclusions of our research. Some technical issues are elaborated in appendices.

6.2 Mapping and Clustering: A Unified Approach

Consider a network of n nodes. Suppose we want to create a mapping or a clustering of these nodes. c_{ij} denotes the number of links (e.g., co-occurrence links, co-citation links, or bibliographic coupling links) between nodes i and j ($c_{ij} = c_{ji} \geq 0$). s_{ij} denotes the association strength of nodes i and j (Van Eck & Waltman, 2009) and is given by

$$s_{ij} = \frac{2mc_{ij}}{c_i c_j}, \tag{6.1}$$

where c_i denotes the total number of links of node i and m denotes the total number of links in the network, that is,

$$c_i = \sum_{j \neq i} c_{ij} \qquad \text{and} \qquad m = \frac{1}{2} \sum_i c_i. \tag{6.2}$$

In the case of mapping, we need to find for each node i a vector $x_i \in \mathbb{R}^p$ that indicates the location of node i in a p-dimensional map (usually $p = 2$). In the case of clustering, we need to find for each node i a positive integer x_i that indicates the cluster to which node i belongs. Our unified approach to mapping and clustering is based on minimizing

$$V(x_1, \ldots, x_n) = \sum_{i<j} s_{ij} d_{ij}^2 - \sum_{i<j} d_{ij} \tag{6.3}$$

with respect to $x_1, \ldots, x_n$. d_{ij} denotes the distance between nodes i and j and is given by

$$d_{ij} = \|x_i - x_j\| = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2} \tag{6.4}$$

in the case of mapping and by

$$d_{ij} = \begin{cases} 0 & \text{if } x_i = x_j \\ 1/\gamma & \text{if } x_i \neq x_j \end{cases} \tag{6.5}$$

in the case of clustering. We refer to the parameter γ in (6.5) as the resolution parameter ($\gamma > 0$). The larger the value of this parameter, the larger the number of clusters that we obtain. Equation (6.3) can be interpreted in terms of attractive and repulsive forces between nodes. The first term in (6.3) represents an attractive force, and the second term represents a repulsive force. The higher the association strength of two nodes, the stronger the attractive force between the nodes. Since the strength of the repulsive force between two nodes does not depend on the association strength of the nodes, the overall effect of the two forces is that nodes with a high association strength are pulled towards each other while nodes with a low association strength are pushed away from each other.

In the case of mapping, it has been shown that the above approach is equivalent to the VOS mapping technique (Van Eck & Waltman, 2007b; Van Eck, Waltman, Dekker, & Van den Berg, 2010), which is in turn closely related to the well-known technique of multidimensional scaling.

In the case of clustering, it can be shown (see Appendix 6.A) that minimizing (6.3) is equivalent to maximizing

$$\hat{V}(x_1, \ldots, x_n) = \frac{1}{2m} \sum_{i<j} \delta(x_i, x_j) w_{ij} \left(c_{ij} - \gamma \frac{c_i c_j}{2m} \right), \tag{6.6}$$

where $\delta(x_i, x_j)$ equals 1 if $x_i = x_j$ and 0 otherwise and where the weights w_{ij} are given by

$$w_{ij} = \frac{2m}{c_i c_j}. \tag{6.7}$$

Interestingly, if the resolution parameter γ and the weights w_{ij} are set equal to 1 in (6.6), then (6.6) reduces to the so-called modularity function introduced by Newman and Girvan (2004; see also Newman, 2004a). Clustering (also referred to as community detection) based on this modularity function (Newman, 2004b) is very popular among physicists and network scientists (for an extensive overview of the literature, see Fortunato, 2010). In bibliometric and scientometric research, modularity-based clustering has been used in a number of recent studies (P. Chen & Redner, 2010; Lambiotte & Panzarasa, 2009; Schubert & Soós, 2010; Takeda & Kajikawa, 2009; Wallace et al., 2009; Zhang et al., 2010). It follows from (6.6) and (6.7) that our proposed clustering technique can be seen as a kind of weighted variant of modularity-based clustering (see Appendix 6.B for a further discussion). However, unlike modularity-based clustering, our clustering technique has a resolution parameter γ. This parameter helps to deal with the resolution limit problem (Fortunato & Barthélemy, 2007) of modularity-based clustering. Due to this problem, modularity-based clustering may fail to identify small clusters. Using our clustering technique, small clusters can always be identified by choosing a sufficiently large value for the resolution parameter γ.

6.3 Related Work

Our unified approach to mapping and clustering is related to earlier work published in the physics literature. Here we summarize the most closely related work.

The above result showing how mapping and clustering can be performed in a unified and consistent way resembles to some extent a result derived by Noack (2009). Noack defined a parameterized objective function for a class of mapping techniques (referred to as force-directed layout techniques by Noack). This class of mapping techniques includes for example the well-known technique of Fruchterman and Reingold (1991). Noack showed that his parameterized objective function subsumes the modularity function of Newman and Girvan (2004). In this way, Noack established a relation between on the one hand a class of mapping techniques and on the other hand modularity-based clustering. Our result differs from the result of Noack in three ways. First, the result of Noack does not directly relate well-known mapping techniques such as the one of Fruchterman and Reingold to modularity-based clustering. Instead, Noack's result shows that the objective functions of some well-known mapping techniques and the modularity function of Newman and Girvan are special cases of the same parameterized function. Our result establishes a direct relation between a mapping technique that has been used in various applications, namely the VOS mapping technique, and a clustering technique. Second, the mapping and clustering techniques considered by Noack and the ones that we consider differ from each other by a weighing factor. This is the weighing factor given by (6.7). Third, the clustering technique considered by Noack is unparameterized, while our clustering technique has a resolution parameter γ.

A parameterized variant of the modularity function of Newman and Girvan (2004) was introduced by Reichardt and Bornholdt (2006; see also Heimo, Kumpula, Kaski, & Saramäki, 2008; Kumpula, Saramäki, Kaski, & Kertész, 2007). Clustering based on this generalized modularity function is closely related to our proposed clustering technique.

In fact, setting the weights w_{ij} equal to 1 in (6.6) essentially yields the function of Reichardt and Bornholdt.

6.4 Illustration of the Proposed Approach

We now illustrate an application of our unified approach to mapping and clustering. In Figure 6.1, we show a combined mapping and clustering of the 1242 most frequently cited publications that appeared in the field of information science in the period 1999–2008.[1] The mapping and the clustering were produced using our unified approach. This was done as follows. We first collected an initial set of publications. This set consisted of all Web of Science publications of the document types article and review published in 37 information science journals in the period 1999–2008 (for the list of journals, see Van Eck, Waltman, Dekker, & Van den Berg, 2010, Table 1). Publications without references were not included. We then extended the initial set of publications with all Web of Science publications in the period 1999–2008 cited by or referring to at least five publications in the initial set of publications. In this way, we ended up with a set of 9948 publications. For each publication in this set, we counted the number of citations from other publications in the set. We selected the 1242 publications with at least eight citations for further analysis. For these publications, we determined the number of co-citation links and the number of bibliographic coupling links. These two types of links were added together and served as input for both our mapping technique and our clustering technique.[2] In the case of our clustering technique, we tried out a number of different values for the resolution parameter γ. After some experimenting, we decided to set this parameter equal to 2. This turned out to yield a clustering with a satisfactory level of detail.

The combined mapping and clustering shown in Figure 6.1 provides an overview of the structure of the field of information science. The left part of the map represents what is sometimes referred to as the information seeking and retrieval (ISR) subfield (Åström,

[1] For other bibliometric studies of the field of information science at the level of individual publications, we refer to Åström (2007) and C. Chen et al. (2010).

[2] Our techniques for mapping and clustering both require solving an optimization problem. In the case of mapping, we minimized (6.3) using a majorization algorithm (similar to Borg & Groenen, 2005, Chapter 8). In the case of clustering, we maximized (6.8) using a top-down divisive algorithm combined with some local search heuristics.

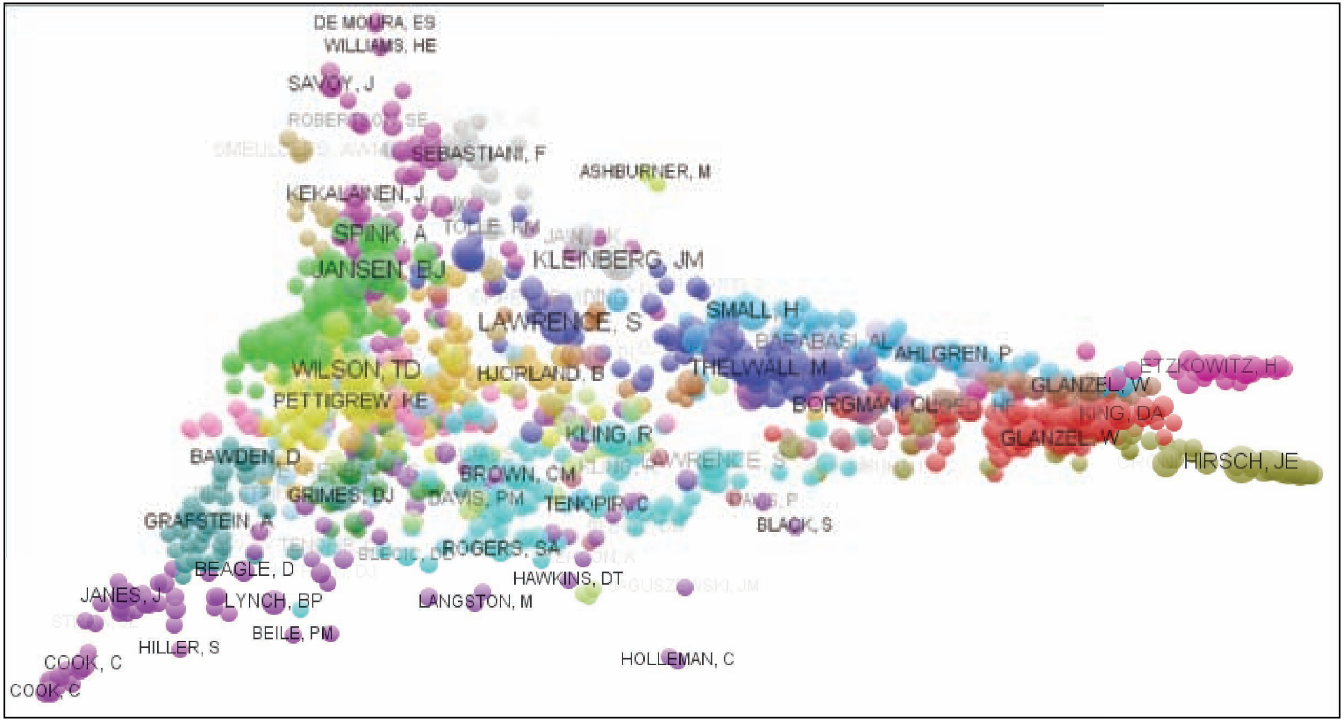


Figure 6.1: Combined mapping and clustering of the 1242 most frequently cited publications that appeared in the field of information science in the period 1999–2008. Publications are labeled with the name of the first author. Colors are used to indicate clusters.

Table 6.1: Summary of the contents of the eight informetrics clusters. The four authors with the largest number of publications in a cluster are listed as important authors in the second column. The color used to indicate a cluster in Figure 6.1 is shown in the fourth column.

No of pub.	Important authors	Main topics	Color
123	Rousseau, R.; Glänzel, W.; Moed, H.F.; Van Raan, A.F.J.	Citation analysis; research evaluation; general scientometric topics	🟥
101	Thelwall, M.; Vaughan, L.; Bar-Ilan, J.; Wilkinson, D.	Webometrics	🟦
73	Leydesdorff, L.; Chen, C.M.; White, H.D.; Small, H.	Mapping and visualization of science	🟦
53	Egghe, L.; Burrell, Q.L.; Daniel, H.D.; Glänzel, W.	*h*-index; citation distributions; Google Scholar	🟩
48	Glänzel, W.; Cronin, B.; Bozeman, B.; Shaw, D.	Scientific collaboration; co-authorship	🟫
46	Meyer, M.; Leydesdorff, L.; Tijssen, R.J.W.; Zimmermann, E.	Science and technology studies; patent analysis	🟪
26	Nisonger, T.E.; Cronin, B.; Shaw, D.; Wilson, C.S.	Studies of the library and information science field	🟥
14	Newman, M.E.J.; Barabasi, A.L.; Albert, R.; Jeong, H.	Complex networks; scientific collaboration networks	🟪

2007), and the right part of the map represents the informetrics subfield. The distinction between these two subfields is well known and has been observed in a number of studies. However, consistent with recent work by Åström (2007), the separation that we observe between the two subfields is less strong than in the influential study of White and McCain (1998). Within the ISR subfield, a further distinction can be made between "hard" (system-oriented) and "soft" (user-oriented) research (e.g., Åström, 2007). Hard ISR research is located in a relatively small area in the upper left part of our map, while soft ISR research is located in a much larger area in the middle and lower left part of the map.

The clustering shown in Figure 6.1 consists of 25 clusters. The distribution of the number of publications per cluster has a mean of 49.7 and a standard deviation of 31.5. There is one very small cluster consisting of just two publications. These two publications are concerned with the use of information science techniques to support biological research. The largest cluster consists of 123 publications. The publications in this cluster deal with citation analysis and some related bibliometric and scientometric topics. Out of the 25 clusters, eight clusters are used to cover the informetrics subfield. We have examined these clusters in more detail. A summary of the contents of the eight informetrics clusters is provided in Table 6.1.

The results presented above illustrate an application of our unified approach to mapping and clustering. Our approach seems to yield an accurate and detailed picture of the structure of the field of information science. The interested reader is invited to examine the results in more detail at http://www.ludowaltman.nl/unified_approach/. On this web page, the combined mapping and clustering shown in Figure 6.1 can be inspected using the VOSviewer software (Van Eck & Waltman, 2010). The clustering is also available in a spreadsheet file.

6.5 Conclusions

Mapping and clustering are complementary to each other. Mapping can be used to obtain a fairly detailed picture of the structure of a bibliometric network. For practical purposes, however, the picture will usually be restricted to just two dimensions. Hence, relations in more than two dimensions will usually not be visible. Clustering, on the other hand, does not suffer from dimensional restrictions. However, the price to be paid is that clustering works with binary rather than continuous dimensions. As a consequence, clustering tends to provide a rather coarse picture of the structure of a bibliometric network.[3]

Given the complementary nature of mapping and clustering and given the frequent combined use of mapping and clustering techniques, we believe that a unified approach to mapping and clustering can be highly valuable. A unified approach ensures that the mapping and clustering techniques on which one relies are based on similar ideas and similar assumptions. By taking a unified approach, inconsistencies between the results produced by mapping and clustering techniques can be avoided.

[3] In this chapter, we have been concerned with clustering techniques that require each node in a bibliometric network to be assigned to exactly one cluster. These are the most commonly used clustering techniques. We have not discussed clustering techniques that allow nodes to be assigned to multiple clusters (e.g., Fortunato, 2010, Section 11). The latter techniques provide a more detailed picture of the structure of a bibliometric network.

In this chapter, we have elaborated a proposal for a unified approach to mapping and clustering. Our proposal unifies the VOS mapping technique with a weighted and parameterized variant of modularity-based clustering. As discussed elsewhere (Van Eck & Waltman, 2007b; Van Eck, Waltman, Dekker, & Van den Berg, 2010), the VOS mapping technique is closely related to the well-known technique of multidimensional scaling, which has a long history in the statistical literature (for an extensive overview, see Borg & Groenen, 2005). Modularity-based clustering, on the other hand, is a recent result from the physics literature (Newman, 2004a, 2004b; Newman & Girvan, 2004). It follows from this that our proposed unified approach establishes a connection between on the one hand a long-lasting research stream in the field of statistics and on the other hand a much more recent research stream in the field of physics.

Our unified approach to mapping and clustering can be especially useful when multiple maps of the same domain are needed, each at a different level of detail. For example, when bibliometric mapping is used for science policy purposes, two maps may be needed. On the one hand a detailed map may be needed that can be carefully validated by experts in the domain of interest, and on the other hand a much more general map may be needed that can be provided to science politicians and research managers. The former map may show the individual nodes in a bibliometric network, while the latter map may show clusters of nodes. Expert validation, which is a crucial step in the use of bibliometric mapping for science policy purposes (Noyons, 1999), of course only makes sense when the map presented to domain experts shows essentially the same structure of the domain of interest as the map presented to science politicians. A unified approach to mapping and clustering helps to avoid discrepancies between maps constructed at different levels of detail. In that way, a unified approach facilitates the use of bibliometric mapping in a science policy context.

In the latest version of our freely available VOSviewer software (Van Eck & Waltman, 2010, see http://www.vosviewer.com), we have incorporated algorithms that implement our unified approach to mapping and clustering. Stand-alone algorithms implementing our unified approach are available at http://www.ludowaltman.nl/unified_approach/.

6.A Appendix I

In this appendix, we prove that in the case of clustering minimizing (6.3) is equivalent to maximizing (6.6) with weights w_{ij} given by (6.7). Using (6.1) and (6.5), it can be seen that (6.3) can be rewritten as

$$V(x_1, \ldots, x_n) = \frac{1}{\gamma} \sum_{i<j} (1 - \delta(x_i, x_j)) \left(\frac{1}{\gamma} \frac{2mc_{ij}}{c_i c_j} - 1 \right), \tag{6.8}$$

where $\delta(x_i, x_j)$ equals 1 if $x_i = x_j$ and 0 otherwise. Let us define

$$\hat{V}(x_1, \ldots, x_n) = -\frac{\gamma^2}{2m} V(x_1, \ldots, x_n) + \frac{1}{2m} \sum_{i<j} \left(\frac{2mc_{ij}}{c_i c_j} - \gamma \right). \tag{6.9}$$

Notice that (6.9) is obtained from (6.8) by multiplying with a constant and by adding a constant. The multiplicative constant is always negative. It follows from this that minimizing (6.8) is equivalent to maximizing (6.9). Substituting (6.8) into (6.9) yields

$$\hat{V}(x_1, \ldots, x_n) = \frac{1}{2m} \sum_{i<j} \delta(x_i, x_j) \left(\frac{2mc_{ij}}{c_i c_j} - \gamma \right). \tag{6.10}$$

We have now shown that minimizing (6.3) is equivalent to maximizing (6.10). Furthermore, (6.10) can be rewritten as (6.6) with weights w_{ij} given by (6.7). This completes the proof.

6.B Appendix II

Our proposed clustering technique can be seen as a weighted and parameterized variant of modularity-based clustering. Modularity-based clustering maximizes (6.6) with weights w_{ij} that are set equal to 1. Our clustering technique maximizes (6.6) with weights w_{ij} that are given by (6.7). In this appendix, we provide an illustration of the effect of the weights w_{ij} in (6.7).

Consider a network of $n = 31$ nodes. Let

$$c_{ij} = \begin{cases} 10 & \text{if } 1 \leq i \leq 10 \text{ and } 1 \leq j \leq 10 \text{ and } i \neq j \\ 100 & \text{if } 11 \leq i \leq 20 \text{ and } 11 \leq j \leq 20 \text{ and } i \neq j \\ 100 & \text{if } 21 \leq i \leq 30 \text{ and } 21 \leq j \leq 30 \text{ and } i \neq j \\ 20 & \text{if } (1 \leq i \leq 10 \text{ and } j = 31) \text{ or } (i = 31 \text{ and } 1 \leq j \leq 10) \\ 50 & \text{if } (11 \leq i \leq 20 \text{ and } j = 31) \text{ or } (i = 31 \text{ and } 11 \leq j \leq 20) \\ 0 & \text{otherwise.} \end{cases} \tag{6.11}$$

Our clustering technique (with the resolution parameter γ set equal to 1) and modularity-based clustering both identify three clusters. They both produce a cluster that contains nodes $1, \ldots, 10$, another cluster that contains nodes $11, \ldots, 20$, and a third cluster that contains nodes $21, \ldots, 30$. However, the two clustering techniques do not agree on the cluster to which node 31 should be assigned. Our clustering technique assigns node 31 to the same cluster as nodes $1, \ldots, 10$, while modularity-based clustering assigns node 31 to the same cluster as nodes $11, \ldots, 20$. The disagreement on the assignment of node 31 is due to the effect of the weights w_{ij} in (6.7). It follows from (6.7) that, compared with modularity-based clustering, our clustering technique gives less weight to nodes with a larger total number of links. Nodes $11, \ldots, 20$ have a much larger total number of links than nodes $1, \ldots, 10$, and compared with modularity-based clustering our clustering technique therefore gives less weight to nodes $11, \ldots, 20$ and more weight to nodes $1, \ldots, 10$. Node 31 is strongly associated both with nodes $1, \ldots, 10$ and with nodes $11, \ldots, 20$. However, due to the difference in weighting, our clustering technique assigns node 31 to the same cluster as nodes $1, \ldots, 10$ while modularity-based clustering assigns node 31 to the same cluster as nodes $11, \ldots, 20$.

Which of the two assignments of node 31 is to be preferred? The total number of links of nodes $11, \ldots, 20$ is almost an order of magnitude larger than the total number of links of nodes $1, \ldots, 10$, but the number of links between node 31 and nodes $11, \ldots, 20$ is only 2.5 times larger than the number of links between node 31 and nodes $1, \ldots, 10$. Hence, from a relative point of view, node 31 has more links with nodes $1, \ldots, 10$ than with nodes $11, \ldots, 20$. Based on this observation, assigning node 31 to the same cluster as nodes $1, \ldots, 10$ seems preferable to assigning node 31 to the same cluster as nodes $11, \ldots, 20$. Hence, we believe that, at least in this particular example, the results produced by our clustering technique are preferable to the results produced by modularity-based clustering.

Chapter 7

VOSviewer: A Computer Program for Bibliometric Mapping*

Abstract

We present VOSviewer, a freely available computer program that we have developed for constructing and viewing bibliometric maps. Unlike most computer programs that are used for bibliometric mapping, VOSviewer pays special attention to the graphical representation of bibliometric maps. The functionality of VOSviewer is especially useful for displaying large bibliometric maps in an easy-to-interpret way.

The chapter consists of three parts. In the first part, an overview of VOSviewer's functionality for displaying bibliometric maps is provided. In the second part, the technical implementation of specific parts of the program is discussed. Finally, in the third part, VOSviewer's ability to handle large maps is demonstrated by using the program to construct and display a co-citation map of 5000 major scientific journals.

7.1 Introduction

Bibliometric mapping is an important research topic in the field of bibliometrics (for an overview, see Börner et al., 2003). Two aspects of bibliometric mapping that can

*This chapter is based on Van Eck and Waltman (2010).

be distinguished are the construction of bibliometric maps and the graphical representation of such maps. In the bibliometric literature, most attention is paid to the construction of bibliometric maps. Researchers for example study the effect of different similarity measures (e.g., Ahlgren et al., 2003; Klavans & Boyack, 2006a; Van Eck & Waltman, 2009), and they experiment with different mapping techniques (e.g., Boyack et al., 2005; Van Eck & Waltman, 2007a; White, 2003b). The graphical representation of bibliometric maps receives considerably less attention. Although some researchers seriously study issues concerning graphical representation (e.g., C. Chen, 2003a, 2006a; Skupin, 2004), most papers published in the bibliometric literature rely on simple graphical representations provided by computer programs such as SPSS and Pajek. For small maps containing no more than, say, 100 items, simple graphical representations typically yield satisfactory results. However, there seems to be a trend towards larger maps (e.g., Boyack et al., 2005; Klavans & Boyack, 2006b; Leydesdorff, 2004; Van Eck, Waltman, Noyons, & Buter, 2010), and for such maps simple graphical representations are inadequate. The graphical representation of large bibliometric maps can be much enhanced by means of, for example, zoom functionality, special labeling algorithms, and density metaphors. This kind of functionality is not incorporated into the computer programs that are commonly used by bibliometric researchers. In this chapter, we therefore introduce a new computer program for bibliometric mapping. This program pays special attention to the graphical representation of bibliometric maps.

The computer program that we introduce is called VOSviewer. VOSviewer is a program that we have developed for constructing and viewing bibliometric maps. The program is freely available to the bibliometric research community (see http://www.vosviewer.com). VOSviewer can for example be used to construct maps of authors or journals based on co-citation data or to construct maps of keywords based on co-occurrence data. The program offers a viewer that allows bibliometric maps to be examined in full detail. VOSviewer can display a map in various different ways, each emphasizing a different aspect of the map. It has functionality for zooming, scrolling, and searching, which facilitates the detailed examination of a map. The viewing capabilities of VOSviewer are especially useful for maps containing at least a moderately large number of items (e.g., at least 100 items). Most computer programs that are used for bibliometric mapping do not display such maps in a satisfactory way.

To construct a map, VOSviewer uses the VOS mapping technique (Van Eck & Waltman, 2007b; Van Eck, Waltman, Dekker, & Van den Berg, 2010), where VOS stands for *visualization of similarities*. For earlier studies in which the VOS mapping technique was used, we refer to Van Eck, Waltman, et al. (2006a), Van Eck and Waltman (2007a), Van Eck, Waltman, Noyons, and Buter (2010), and Waaijer et al. (2010, 2011). VOSviewer can display maps constructed using any suitable mapping technique. Hence, the program can be employed not only for displaying maps constructed using the VOS mapping technique but also for displaying maps constructed using techniques such as multidimensional scaling. VOSviewer runs on a large number of hardware and operating system platforms and can be started directly from the internet.

In the remainder of this chapter, we first discuss for what type of bibliometric maps VOSviewer is intended to be used. We then provide an overview of VOSviewer's functionality for displaying bibliometric maps. We also elaborate on the technical implementation of specific parts of the program. Finally, to demonstrate VOSviewer's ability to handle large maps, we use the program to construct and display a co-citation map of 5000 major scientific journals.

7.2 Types of Bibliometric Maps

Two types of maps can be distinguished that are commonly used in bibliometric research.[1] We refer to these types of maps as distance-based maps and graph-based maps. Distance-based maps are maps in which the distance between two items reflects the strength of the relation between the items. A smaller distance generally indicates a stronger relation. In many cases, items are distributed quite unevenly in distance-based maps. On the one hand this makes it easy to identify clusters of related items, but on the other hand this sometimes makes it difficult to label all the items in a map without having labels that overlap each other. Graph-based maps are maps in which the distance between two items need not reflect the strength of the relation between the items. Instead, lines are drawn between items to indicate relations. Items are often distributed in a fairly uniform way in graph-based maps. This may have the advantage that there are

[1] We do not consider maps that are primarily intended for showing developments over time. Such maps are for example provided by the HistCite software of Eugene Garfield (e.g., Garfield, 2009).

Table 7.1: Some mapping techniques for constructing distance-based and graph-based maps.

Distance-based maps	Graph-based maps
Multidimensional scaling	Kamada-Kawai
VOS	Fruchterman-Reingold
VxOrd	Pathfinder networks
Kopcsa-Schiebel	

less problems with overlapping labels. In our opinion, a disadvantage of graph-based maps compared with distance-based maps is that it typically is more difficult to see the strength of the relation between two items. Clusters of related items may also be more difficult to detect.

In Table 7.1, we list some mapping techniques that are used in bibliometric research to construct distance-based and graph-based maps. For constructing distance-based maps, multidimensional scaling (e.g., Borg & Groenen, 2005) is by far the most popular technique in the field of bibliometrics. An alternative to multidimensional scaling is the VOS mapping technique (Van Eck & Waltman, 2007b; Van Eck, Waltman, Dekker, & Van den Berg, 2010). In general, this technique produces better structured maps than multidimensional scaling (Van Eck, Waltman, Dekker, & Van den Berg, 2008, 2010). Another technique for constructing distance-based maps is VxOrd (Davidson, Wylie, & Boyack, 2001; Klavans & Boyack, 2006b).[2] This technique is especially intended for constructing maps that contain very large numbers of items (more than 700,000 items in Klavans & Boyack, 2006b). A disadvantage of VxOrd is that a complete specification of how the technique works is not available. A fourth technique for constructing distance-based maps was proposed by Kopcsa and Schiebel (1998). This technique is implemented in a computer program called BibTechMon.

For constructing graph-based maps, researchers in the field of bibliometrics (e.g. de Moya-Anegón et al., 2007; Leydesdorff & Rafols, 2009; Vargas-Quesada & de Moya-Anegón, 2007; White, 2003b) usually use a mapping technique developed by Kamada and Kawai (1989). Sometimes an alternative technique proposed by Fruchterman and Reingold (1991) is used (e.g., Bollen et al., 2009; Leydesdorff, 2004). A popular com-

[2] A computer implementation of VxOrd is available at http://www.cs.sandia.gov/ smartin/software.html as part of the DrL toolbox.

puter program in which both techniques are implemented is Pajek (De Nooy et al., 2005). Some researchers (e.g., de Moya-Anegón et al., 2007; Vargas-Quesada & de Moya-Anegón, 2007; White, 2003b) combine the Kamada-Kawai technique with the technique of pathfinder networks (Schvaneveldt, 1990; Schvaneveldt et al., 1988). Two other computer programs that can be used to construct graph-based maps are CiteSpace (C. Chen, 2006a) and the Network Workbench Tool. Even more programs are available in the field of social network analysis (for an overview, see Huisman & Van Duijn, 2005).

Distance-based and graph-based maps both have advantages and disadvantages. In VOSviewer, we have chosen to support only distance-based maps. VOSviewer can be employed to view any two-dimensional distance-based map, regardless of the mapping technique that has been used to construct the map. One can employ VOSviewer to view multidimensional scaling maps constructed using statistical packages such as SAS, SPSS, and R, but one can also employ VOSviewer to view maps constructed using other, less common techniques. Because the VOS mapping technique shows a very good performance (Van Eck et al., 2008; Van Eck, Waltman, Dekker, & Van den Berg, 2010), this technique has been fully integrated into VOSviewer. This means that VOSviewer can be used not only to view VOS maps but also to construct them. Hence, no separate computer program is needed for constructing VOS maps.

7.3 Functionality of VOSviewer

In this section, we provide an overview of VOSviewer's functionality for displaying bibliometric maps.[3] We use a data set that consists of co-citation frequencies of journals belonging to at least one of the following five closely related subject categories of Thomson Reuters: *Business*, *Business-Finance*, *Economics*, *Management*, and *Operations Research & Management Science*. The co-citation frequencies of journals were determined based on citations in articles published between 2005 and 2007 to articles published in 2005. A journal was included in the data set only if it had at least 25 co-citations. There were 232 journals that satisfied this condition. Based on a clustering

[3] For a more extensive discussion of the functionality of VOSviewer, we refer to the VOSviewer manual, which is available at http://www.vosviewer.com.

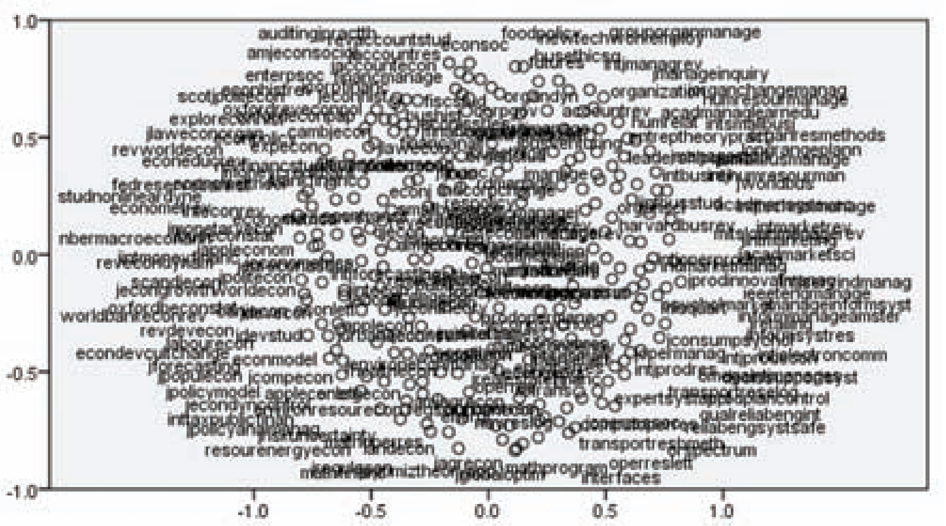


Figure 7.1: Map obtained using SPSS.

technique, the journals in the data set were divided into five clusters. The data set is available at http://www.vosviewer.com.

Two maps constructed based on the journal co-citation data set are shown in Figures 7.1 and 7.2. The figures were obtained using, respectively, SPSS and Pajek, which are both commonly used computer programs for bibliometric mapping. The map shown in Figure 7.1 is a distance-based map constructed using multidimensional scaling. The map shown in Figure 7.2 is a graph-based map constructed using the Kamada-Kawai technique (Kamada & Kawai, 1989). As can be seen, SPSS and Pajek both provide rather simple graphical representations of bibliometric maps. The programs both have serious problems with overlapping labels. Due to these problems, maps can be difficult to interpret, especially in the details. In the rest of this section, we demonstrate how VOSviewer overcomes the limitations of simple graphical representations provided by programs such as SPSS and Pajek.

A screenshot of the main window of VOSviewer is shown in Figure 7.3. Depending on the available data, VOSviewer can display a map in three or four different ways. The different ways of displaying a map are referred to as the label view, the density view, the cluster density view, and the scatter view. We now discuss each of these views:

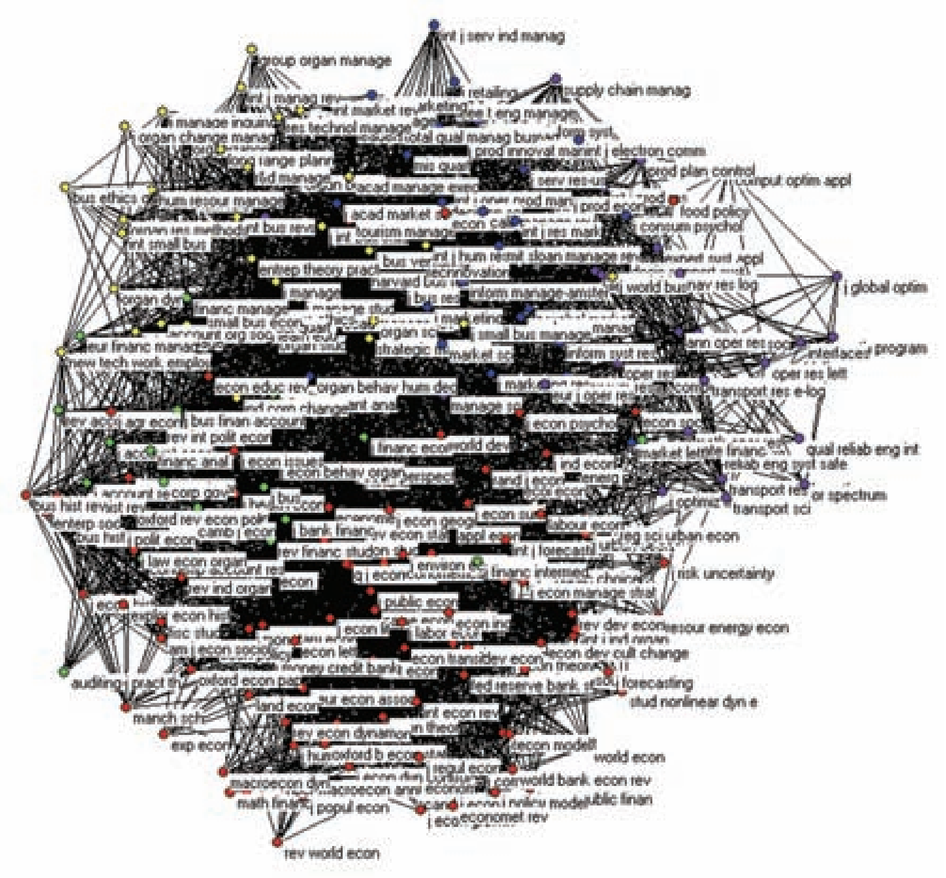


Figure 7.2: Map obtained using Pajek.

- *Label view*. In this view, items are indicated by a label and, by default, also by a circle. The more important an item, the larger its label and its circle. If colors have been assigned to items, each item's circle is displayed in the color of the item. By default, to avoid overlapping labels, only a subset of all labels is displayed. The label view is particularly useful for a detailed examination of a map.

  An example of the label view is shown in Figure 7.4. The map shown in this figure was constructed based on the journal co-citation data set discussed at the beginning of this section. Colors indicate the cluster to which a journal was assigned by

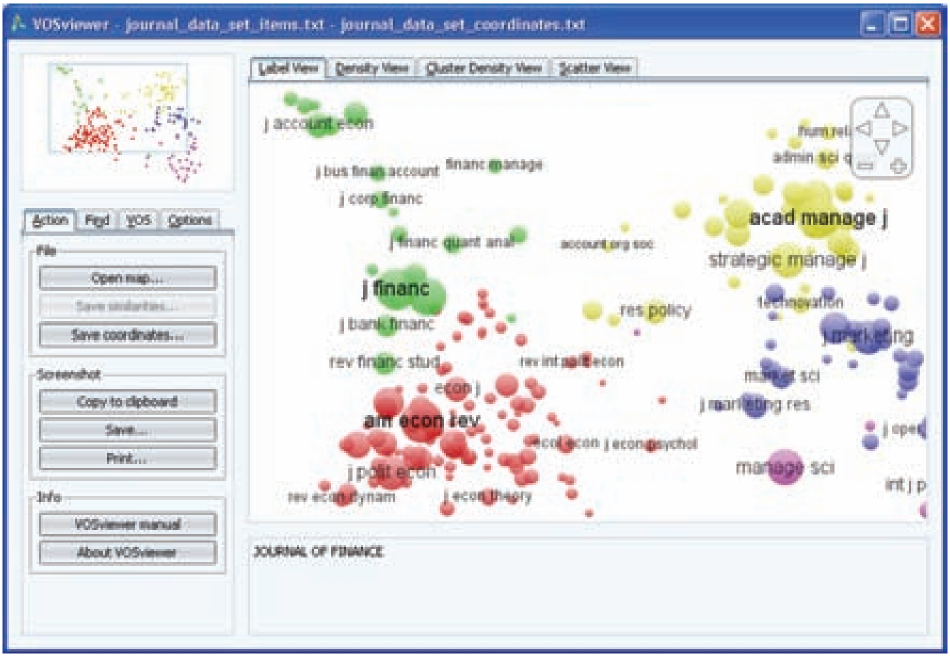


Figure 7.3: Screenshot of the main window of VOSviewer.

the clustering technique that we used. As can be seen, there is a strong agreement between the structure of the map and the clustering obtained using our clustering technique. The clustering also has a straightforward interpretation. The five clusters correspond with the following research fields: accounting/finance, economics, management, marketing, and operations research.[4] It is clear that the map shown in Figure 7.4 is much easier to interpret than the maps shown in Figures 7.1 and 7.2. This demonstrates one of the main advantages of VOSviewer over commonly used computer programs such as SPSS and Pajek.

- *Density view*. In this view, items are indicated by a label in a similar way as in the label view. Each point in a map has a color that depends on the density of items at

[4] Although this is not directly visible in Figure 7.4, we note that there is a large overlap in the map between the *Business* and *Management* subject categories of Thomson Reuters. This indicates an important difference between the clustering that we found and the clustering provided by the subject categories of Thomson Reuters.

that point. That is, the color of a point in a map depends on the number of items in the neighborhood of the point and on the importance of the neighboring items. The density view is particularly useful to get an overview of the general structure of a map and to draw attention to the most important areas in a map. We will discuss the technical implementation of the density view later on in this chapter.

An example of the density view is shown in Figure 7.5. The map shown in this figure is the same as the one shown in Figure 7.4. The density view immediately reveals the general structure of the map. Especially the economics and management areas turn out to be important. These areas are very dense, which indicates that overall the journals in these areas receive a lot of citations. It can also be seen that there is a clear separation between the fields of accounting, finance, and economics on the one hand and the fields of management, marketing, and operations research on the other hand. Like Figure 7.4, Figure 7.5 demonstrates VOSviewer's ability to provide easy-to-interpret graphical representations of bibliometric maps.

- *Cluster density view.* This view is available only if items have been assigned to clusters. The cluster density view is similar to the ordinary density view except that the density of items is displayed separately for each cluster of items. The cluster density view is particularly useful to get an overview of the assignment of items to clusters and of the way in which clusters of items are related to each other. We will discuss the technical implementation of the cluster density view later on in this chapter.

  Unfortunately, the cluster density view cannot be shown satisfactorily in black and white. We therefore do not show an example of the cluster density view.

- *Scatter view.* This view is a simple view in which items are indicated by a small circle and in which no labels are displayed. If colors have been assigned to items, each item's circle is displayed in the color of the item. The scatter view focuses solely on the general structure of a map and does not provide any detailed information.

In addition to the four views discussed above, another important feature of VOSviewer is its ability to handle large maps. VOSviewer can easily construct maps that contain

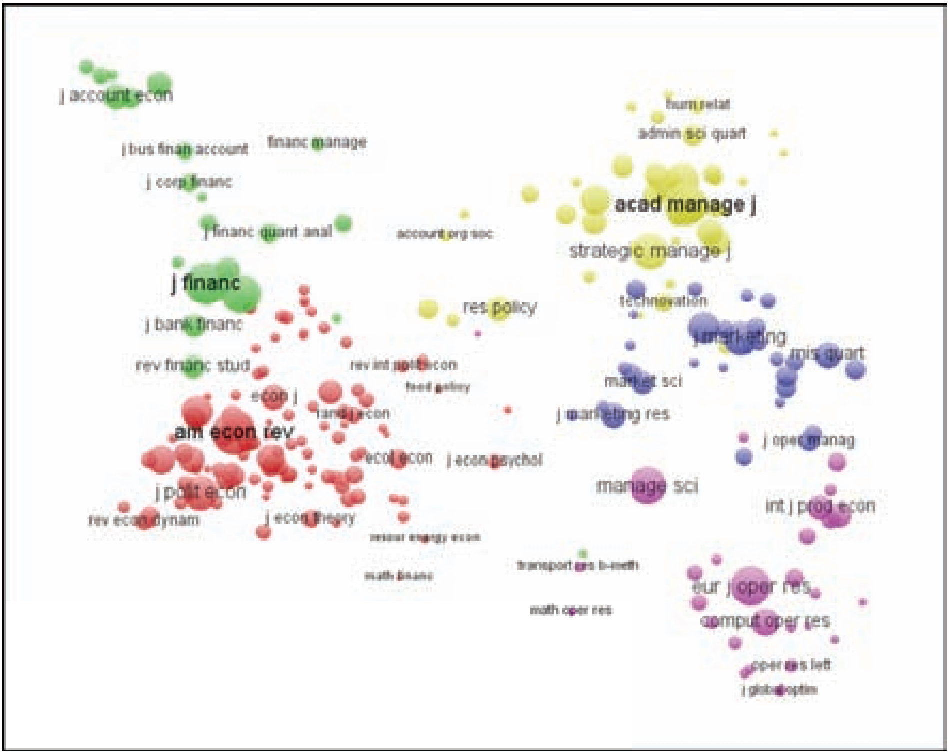


Figure 7.4: Screenshot of the label view.

several thousands of items, and it can display maps that contain more than 10,000 items. VOSviewer has functionality for zooming, scrolling, and searching, which facilitates the detailed examination of large maps. When displaying a map, VOSviewer uses a special algorithm to determine which labels can be displayed and which labels cannot be displayed without having labels that overlap each other. The further one zooms in on a specific area of a map, the more labels become visible. Later on in this chapter, we will demonstrate VOSviewer's ability to handle large maps by using the program to construct and display a co-citation map of 5000 major scientific journals. In the next two sections, however, we will first elaborate on the technical implementation of specific parts of VOSviewer.

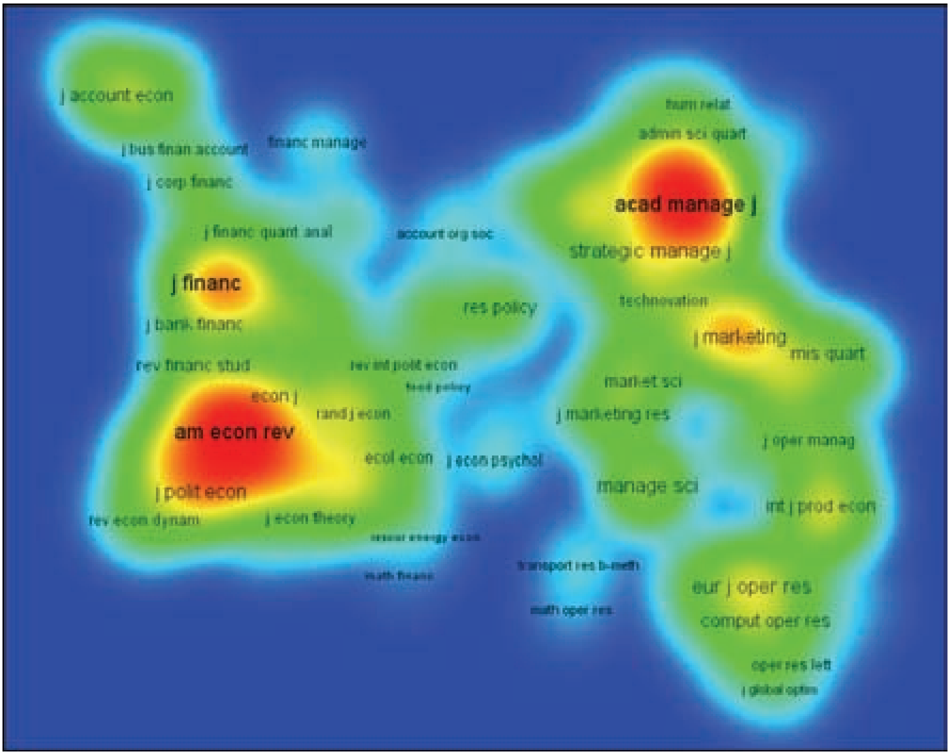


Figure 7.5: Screenshot of the density view.

7.4 Construction of a Map

VOSviewer constructs a map based on a co-occurrence matrix. The construction of a map is a process that consists of three steps. In the first step, a similarity matrix is calculated based on the co-occurrence matrix. In the second step, a map is constructed by applying the VOS mapping technique to the similarity matrix. And finally, in the third step, the map is translated, rotated, and reflected. We now discuss each of these steps in more detail.

7.4.1 Step 1: Similarity Matrix

The VOS mapping technique requires a similarity matrix as input. A similarity matrix can be obtained from a co-occurrence matrix by normalizing the latter matrix, that is, by correcting the matrix for differences in the total number of occurrences or co-occurrences of items. The most popular similarity measures for normalizing co-occurrence data are the cosine and the Jaccard index. VOSviewer, however, does not use one of these similarity measures. Instead, it uses a similarity measure known as the association strength (Van Eck & Waltman, 2007a; Van Eck, Waltman, et al., 2006a). This similarity measure is sometimes also referred to as the proximity index (e.g., Peters & Van Raan, 1993b; Rip & Courtial, 1984) or as the probabilistic affinity index (e.g., Zitt et al., 2000). Using the association strength, the similarity s_{ij} between two items i and j is calculated as

$$s_{ij} = \frac{c_{ij}}{w_i w_j}, \tag{7.1}$$

where c_{ij} denotes the number of co-occurrences of items i and j and where w_i and w_j denote either the total number of occurrences of items i and j or the total number of co-occurrences of these items. It can be shown that the similarity between items i and j calculated using (7.1) is proportional to the ratio between on the one hand the observed number of co-occurrences of items i and j and on the other hand the expected number of co-occurrences of items i and j under the assumption that occurrences of items i and j are statistically independent. We refer to Van Eck and Waltman (2009) for an extensive discussion of the advantages of the association strength over other similarity measures, such as the cosine and the Jaccard index.

7.4.2 Step 2: VOS Mapping Technique

We now discuss how the VOS mapping technique constructs a map based on the similarity matrix obtained in step 1. A more elaborate discussion of the VOS mapping technique, including an analysis of the relation between the VOS mapping technique and multidimensional scaling, is provided by Van Eck and Waltman (2007b) and Van Eck, Waltman, Dekker, and Van den Berg (2010). Some results of an empirical comparison between the VOS mapping technique and multidimensional scaling are reported by Van Eck et al. (2008); Van Eck, Waltman, Dekker, and Van den Berg (2010). A simple open source computer program that implements the VOS mapping technique is available at http://www.neesjanvaneck.nl/vos/.

Let n denote the number of items to be mapped. The VOS mapping technique constructs a two-dimensional map in which the items $1, \ldots, n$ are located in such a way that the distance between any pair of items i and j reflects their similarity s_{ij} as accurately as possible.[5] Items that have a high similarity should be located close to each other, while items that have a low similarity should be located far from each other. The idea of the VOS mapping technique is to minimize a weighted sum of the squared Euclidean distances between all pairs of items. The higher the similarity between two items, the higher the weight of their squared distance in the summation. To avoid trivial maps in which all items have the same location, the constraint is imposed that the average distance between two items must be equal to 1. In mathematical notation, the objective function to be minimized is given by

$$V(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \sum_{i<j} s_{ij} \left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2, \tag{7.2}$$

where the vector $\mathbf{x}_i = (x_{i1}, x_{i2})$ denotes the location of item i in a two-dimensional map and where $\|\cdot\|$ denotes the Euclidean norm. Minimization of the objective function is performed subject to the constraint

$$\frac{2}{n(n-1)} \sum_{i<j} \left\|\mathbf{x}_i - \mathbf{x}_j\right\| = 1. \tag{7.3}$$

The constrained optimization problem of minimizing (7.2) subject to (7.3) is solved numerically in two steps. The constrained optimization problem is first converted into an unconstrained optimization problem. The latter problem is then solved using a so-called majorization algorithm. The majorization algorithm used by VOSviewer is a variant of the SMACOF algorithm described in the multidimensional scaling literature (e.g., Borg & Groenen, 2005). To increase the likelihood of finding a globally optimal solution, the majorization algorithm can be run multiple times, each time using a different randomly generated initial solution.

[5] The VOS mapping technique can also be used to construct maps in more than two dimensions. However, VOSviewer does not support this. The VOS software available at http://www.neesjanvaneck.nl/vos/ does support the construction of maps in more than two dimensions.

7.4.3 Step 3: Translation, Rotation, and Reflection

The optimization problem discussed in step 2 does not have a unique globally optimal solution. This is because, if a solution is globally optimal, any translation, rotation, or reflection of the solution is also globally optimal (for a discussion of this issue in the multidimensional scaling context, see Borg & Groenen, 2005). It is of course important that VOSviewer produces consistent results. The same co-occurrence matrix should therefore always yield the same map (ignoring differences caused by local optima). To accomplish this, it is necessary to transform the solution obtained for the optimization problem discussed in step 2. VOSviewer applies the following three transformations to the solution:

- *Translation*. The solution is translated in such a way that it becomes centered at the origin.

- *Rotation*. The solution is rotated in such a way that the variance on the horizontal dimension is maximized. This transformation is known as principal component analysis.

- *Reflection*. If the median of $x_{11}, \ldots, x_{n1}$ is larger than 0, the solution is reflected in the vertical axis. If the median of $x_{12}, \ldots, x_{n2}$ is larger than 0, the solution is reflected in the horizontal axis.

These three transformations are sufficient to ensure that VOSviewer produces consistent results.

7.5 Density View and the Cluster Density View

In this section, we discuss the technical implementation of the density view and the cluster density view. Recall that in VOSviewer the cluster density view is available only if items have been assigned to clusters.

7.5.1 Density View

We first consider the density view (see also Van Eck & Waltman, 2007a). Similar ideas can be found in the work of, for example, Eilers and Goeman (2004) and Van Liere and De Leeuw (2003).

In the density view, the color of a point in a map is determined based on the item density of the point. Let $\bar{d}$ denote the average distance between two items, that is,

$$\bar{d} = \frac{2}{n(n-1)} \sum_{i<j} \|\mathbf{x}_i - \mathbf{x}_j\| . \tag{7.4}$$

The item density $D(\mathbf{x})$ of a point $\mathbf{x} = (x_1, x_2)$ is then defined as

$$D(\mathbf{x}) = \sum_{i=1}^{n} w_i K \left(\|\mathbf{x} - \mathbf{x}_i\| / \left(\bar{d}h\right)\right), \tag{7.5}$$

where $K : [0, \infty) \to [0, \infty)$ denotes a kernel function, $h > 0$ denotes a parameter called the kernel width,[6] and w_i denotes the weight of item i, that is, the total number of occurrences or co-occurrences of item i. The kernel function K must be non-increasing. VOSviewer uses a Gaussian kernel function given by

$$K(t) = \exp\left(-t^2\right). \tag{7.6}$$

It follows from (7.5) that the item density of a point in a map depends both on the number of neighboring items and on the weights of these items. The larger the number of neighboring items and the smaller the distances between these items and the point of interest, the higher the item density. Also, the higher the weights of the neighboring items, the higher the item density. We note that the calculation of item densities using (7.5) is similar to the estimation of a probability density function using the technique of kernel density estimation (e.g., Scott, 1992).

Item densities calculated using (7.5) are translated into colors using a color scheme. By default, VOSviewer uses a red-green-blue color scheme (see Figure 7.5). In this

[6] By default, VOSviewer uses $h = 0.125$. This generally seems to work fine. However, if necessary, the value of h can be changed.

color scheme, red corresponds with the highest item density and blue corresponds with the lowest item density.

Finally, we note that the above-described calculation of the color of a point in a map is performed only for a limited number of points. These points are located on a grid. The colors of points that do not lie on this grid are obtained through interpolation.

7.5.2 Cluster Density View

We now consider the cluster density view. In this view, the item density of a point in a map is calculated separately for each cluster. The item density of a point $\mathbf{x}$ for a cluster p, denoted by $D_p(\mathbf{x})$, is defined as

$$D_p(\mathbf{x}) = \sum_{i=1}^{n} I_p(i) w_i K \left(\|\mathbf{x} - \mathbf{x}_i\| \,/\, \left(\bar{d}h \right) \right), \tag{7.7}$$

where $I_p(i)$ denotes an indicator function that equals 1 if item i belongs to cluster p and that equals 0 otherwise. Like in the ordinary density view, K denotes the Gaussian kernel function given by (7.6).

After calculating item densities, the color of a point in a map is determined in two steps. Each cluster is associated with a color. In the first step, the colors of the clusters are mixed together. This is done by calculating a weighted average of the colors, where the weight of a color equals the item density for the corresponding cluster, as given by (7.7). In the second step, the color obtained in the first step is mixed with the (black or white) background color of the cluster density view. The proportion in which the two colors are mixed depends on the total item density of a point, (7.5). The lower the total item density of a point, the closer the color of the point is to the background color.

7.6 Large-Scale Application of VOSviewer

To demonstrate VOSviewer's ability to handle large maps, we use the program to construct and display a co-citation map of 5000 major scientific journals. For earlier studies in which journal maps of similar size were presented, we refer to Bollen et al. (2009), Boyack et al. (2005), and Leydesdorff (2004).

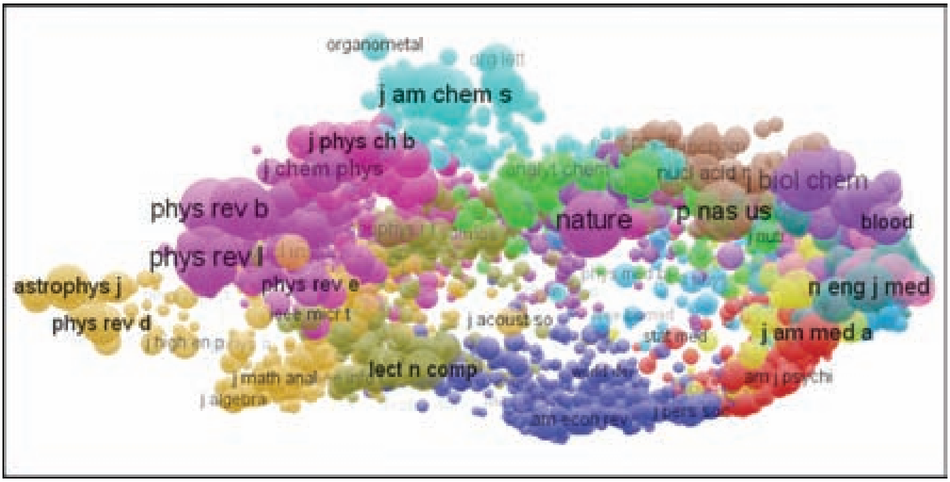


Figure 7.6: Co-citation map of 5000 major scientific journals (label view).

The journal co-citation map was constructed as follows. In the Web of Science database, we collected all citations from documents published in 2007 to documents published between 1997 and 2006. We only took into account documents of types article, note, and review. In total, we obtained about 17.5 million citations. It is well known that different scientific fields can have quite different citation practices. To correct for this, we source normalized all citations. By this we mean that if a document cites m other documents, we weighed each of the m citations by $1/m$ (cf. Small & Sweeney, 1985). 10,603 journals turned out to have been cited at least once. Out of these journals, we selected the 5000 journals with the largest number of source normalized citations. By multiplying the source-normalized citation matrix for these 5000 journals with its transpose, we obtained a source-normalized co-citation matrix. We used this matrix as input for VOSviewer. Based on the co-citation matrix, VOSviewer constructed the journal co-citation map that is shown in Figure 7.6. The interested reader may want to examine the map in full detail using VOSviewer. To do so, visit http://www.vosviewer.com/journalmap/.

Our journal co-citation map provides an overview of the structure of the scientific world. Clusters of related journals can be identified in the map, and these clusters can be linked to scientific fields. Clusters that are located close to each other in the map indicate closely related fields. As can be seen in Figure 7.6, the map has a more or less circular structure. The center of the map is relatively empty. At a global level, the interpretation of the map is fairly straightforward. The right part of the map covers the medical sciences. Moving counterclockwise from the medical sciences, the following major fields can be identified: life sciences, chemistry, physics, engineering, mathematics, computer science, social sciences, and psychology. Psychology is again closely related to the medical sciences, which completes the circular structure of the map.

There seems to be only one earlier study in which distance-based journal maps of similar size as our map were presented. This study was done by Boyack et al. (2005). Boyack et al. presented two kinds of journal maps, namely maps based on journal-to-journal citation data and maps based on journal co-citation data. Comparing the global structure of the maps of Boyack et al. with the global structure of our map, there turn out to be both some similarities and some differences. On the one hand, the way in which major scientific fields are located relative to each other is fairly similar in the maps of Boyack et al. and in our map. On the other hand, the general shape of the maps of Boyack et al. is quite different from the general shape of our map. In the maps of Boyack et al., clusters of journals are located more or less equally distributed within an almost perfect circle. This seems to be a structure that has been imposed by the VxOrd mapping technique used by Boyack et al. A disadvantage of this structure is that in the center of the maps of Boyack et al. different fields can be identified that do not really seem to have much in common. In our map constructed using VOSviewer, we cannot find any indications of a structure that has been imposed by the mapping technique. The general shape of our map seems to have been determined by the data rather than by the mapping technique that we used. A noticeable difference between our map and the maps of Boyack et al. is the relatively empty center of our map. Due to the relatively empty center, fields between which there are no strong relations are clearly separated from each other.

To show the importance of VOSviewer's viewing capabilities, we examine one particular area in our journal co-citation map in more detail. Suppose that we are interested in the interface between the sciences and the social sciences. As can be seen in Figure 7.6, an area where the sciences and the social sciences come together is between the fields of computer science (*Lecture Notes in Computer Science*) and economics (*Amer-*

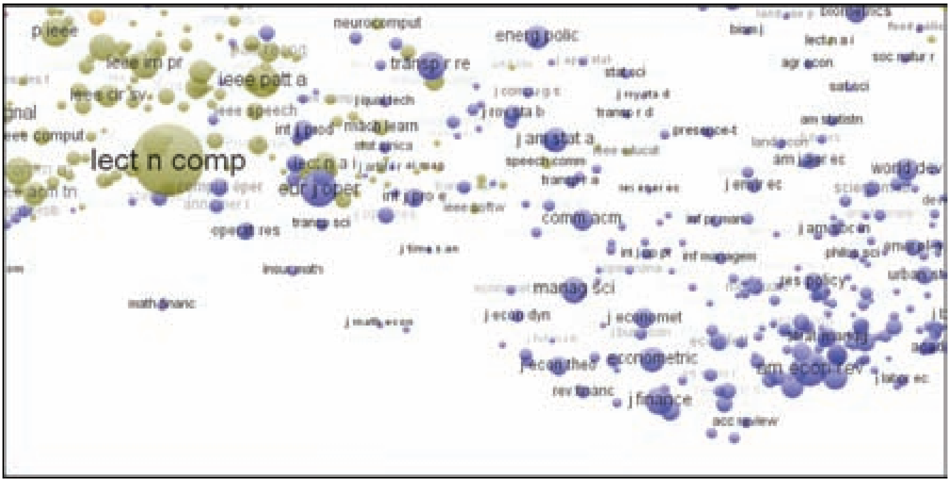


Figure 7.7: The area between the fields of computer science and economics.

ican Economic Review). However, Figure 7.6 does not provide any detailed insight into this area. We therefore use VOSviewer to zoom in on the area. This yields Figure 7.7. It is clear that Figure 7.7 shows much more detail than Figure 7.6. Unlike Figure 7.6, Figure 7.7 allows us to exactly identify the fields that are at the boundary between the sciences and the social sciences. These fields include artificial intelligence and machine learning (e.g., *Lecture Notes in Artificial Intelligence* and *Machine Learning*), operations research (e.g., *European Journal of Operational Research* and *Management Science*), statistics (e.g., *Journal of the American Statistical Association*), and transportation (e.g., *Transportation Research Record*).[7] Figure 7.7 illustrates the importance of VOSviewer's viewing capabilities. Without the zoom functionality of a computer program such as VOSviewer, only the global structure of a map can be inspected and detailed examinations of large maps such as our journal co-citation map are not possible.

7.7 Conclusion

In this chapter, we have presented VOSviewer, a freely available computer program for constructing and viewing bibliometric maps. Unlike programs such as SPSS and Pajek,

[7] Notice that *Scientometrics* is also visible in Figure 7.7 (in the right part of the figure).

which are commonly used for bibliometric mapping, VOSviewer pays special attention to the graphical representation of bibliometric maps. The functionality of VOSviewer is especially useful for displaying large bibliometric maps in an easy-to-interpret way.

VOSviewer has been used successfully in various projects carried out by the Centre for Science and Technology Studies. In future research on bibliometric mapping, we expect to rely heavily on VOSviewer. By making VOSviewer freely available to the bibliometric research community, we hope that others will benefit from it as well.

Chapter 8

Bibliometric Mapping of the Computational Intelligence Field[*]

Abstract

In this chapter, a bibliometric study of the computational intelligence field is presented. Bibliometric maps showing the associations between the main concepts in the field are provided for the periods 1996–2000 and 2001–2005. Both the current structure of the field and the evolution of the field over the last decade are analyzed. In addition, a number of emerging areas in the field are identified. It turns out that computational intelligence can best be seen as a field that is structured around four important types of problems, namely control problems, classification problems, regression problems, and optimization problems. Within the computational intelligence field, the neural networks and fuzzy systems subfields are fairly intertwined, whereas the evolutionary computation subfield has a relatively independent position.

8.1 Introduction

In this chapter, a bibliometric study of the field of computational intelligence (CI) is presented. The CI field is analyzed by means of bibliometric maps that show the associations between the main concepts in the field. The maps provide insight into the

[*] This chapter is based on Van Eck and Waltman (2007a).

structure of the CI field. More specifically, they visualize the division of the field into several subfields, and they indicate the relations between these subfields. By comparing bibliometric maps based on different periods of time, some insights are obtained into the evolution of the field over the last decade. The way in which the field has evolved is also studied through a quantitative analysis of the number of times researchers use specific concepts in their papers.

Bibliometric studies of the CI field are scarce. We are only aware of two studies in which the neural networks subfield is analyzed (Van Raan & Tijssen, 1993; Noyons & Van Raan, 1998). However, these studies are rather outdated, since they are based on data from the 1980s and the beginning of the 1990s. The present study is an extension of our earlier research (Van Eck, Waltman, Van den Berg, & Kaymak, 2006b; Van Eck, Waltman, et al., 2006a), in which we analyzed the CI field based on papers presented at the IEEE World Congress on Computational Intelligence in 2002 and 2006. In the present study, we use data from three major journals and three major conferences over the period 1996–2005. By considerably increasing the amount of data on which our analysis is based, we expect to improve the reliability of our results compared to our earlier research. In the present study, we also discuss a method for assessing the stability of a bibliometric map. In our opinion, the stability of bibliometric maps usually does not get sufficient attention in bibliometric studies. By taking into account the stability of a map, the reliability of a bibliometric analysis can be improved significantly. A third improvement over our earlier research is the refinement of our methodology for constructing so-called concept density maps. The refined methodology better visualizes the amount of attention researchers pay to the various research topics in a field of science.

Bibliometric maps can be constructed in many different ways. Overviews of various approaches to bibliometric mapping are provided by Börner et al. (2003) and by Noyons (2004). The closely related field of information visualization is covered by C. Chen (2006b). In this chapter, we are concerned with maps in which the distance between two objects indicates the strength of the association between the objects. Objects that are located close to each other are regarded as strongly associated, whereas objects that are located far from each other are regarded as weakly associated or as not associated at all. In the field of bibliometrics, a number of approaches have been proposed for constructing this type of map. Most of these approaches rely on the method of multidimensional scaling (Borg & Groenen, 2005). The most popular approach seems to be the one that is discussed by McCain (1990). A good example of the application of this approach is provided by White and McCain (1998). In this chapter, we use our own approach to constructing bibliometric maps. Rather than on multidimensional scaling, our approach relies on a closely related method called VOS, which is an abbreviation for *visualization of similarities*. In our experience, our approach to constructing bibliometric maps provides better results than the approaches that have been proposed in the bibliometric literature. The focus of this chapter, however, is not on the methodological aspect of our research. Although we do provide a detailed description of our approach to constructing bibliometric maps, we do not discuss the differences with and the advantages over alternative approaches.

The chapter is organized as follows. Our methodology for constructing bibliometric maps is discussed in Section 8.2. The bibliometric analysis of the CI field is presented in Section 8.3. Conclusions are drawn in Section 8.4.

8.2 Methodology

According to Börner et al. (2003), the process of constructing a bibliometric map can be divided into the following six steps: (1) collection of raw data, (2) selection of the type of item to analyze, (3) extraction of relevant information from the raw data, (4) calculation of similarities between items based on the extracted information, (5) positioning of items in a low-dimensional space based on the similarities, and (6) visualization of the low-dimensional space. We now discuss the way in which we implement each of these steps in this chapter. Our approach is summarized in Table 8.1.

The first step in the process of bibliometric mapping is the collection of raw data. In this chapter, the raw data consist of a corpus containing abstracts of papers from three major journals and three major conferences in the CI field.[1] The journals are the IEEE Transactions on Neural Networks, the IEEE Transactions on Fuzzy Systems, and the IEEE Transactions on Evolutionary Computation. The conferences are the International Joint Conference on Neural Networks, the IEEE International Conference on

[1] Actually, the corpus not only contains abstracts of papers, it also contains titles. Both abstracts and titles are used to construct bibliometric maps. However, for simplicity we will only refer to the abstracts in the rest of this chapter.

Table 8.1: Summary of our implementation of the process of bibliometric mapping.

Step of the mapping process	Implementation
(1) Collection of data	Abstracts of papers from journals and conferences in the CI field
(2) Selection of type of item	Concepts
(3) Extraction of information	Co-occurrence frequency (Subsection 8.2.1)
(4) Calculation of similarities	Association strength (Subsection 8.2.2)
(5) Positioning of items	VOS (Subsection 8.2.3)
(6) Visualization	Concept map (Subsection 8.2.4) Concept density map (Subsection 8.2.5)

Fuzzy Systems, and the IEEE Congress on Evolutionary Computation. Both the journals and the proceedings of the conferences are published by the IEEE Computational Intelligence Society. Two sets of data are collected, one containing abstracts from the period 1996–2000 and one containing abstracts from the period 2001–2005. In this way, separate bibliometric maps can be constructed for each of the two periods. The data are collected using two databases, IEEE Xplore and Elsevier Scopus. The latter database can be seen as an alternative to the well-known ISI Web of Science database. Compared to Web of Science, Scopus has the advantage that it also includes conference proceedings.

The second step in the process of bibliometric mapping is the selection of the type of item to analyze. According to Börner et al. (2003), journals, papers, authors, and descriptive terms or words are most commonly selected as the type of item to analyze. Each type of item provides a different visualization of a field of science and results in a different analysis. In the present study, we choose to analyze concepts.[2] A bibliometric map showing the associations between concepts in a scientific field is referred to as a concept map in this chapter. To avoid any possible confusion, we note that our con-

[2] According to the Merriam-Webster Online Dictionary, a concept is an abstract or generic idea generalized from particular instances. Concepts can be designated using terms. For example, the terms *neural network*, *fuzzy system*, and *genetic algorithm* designate three well-known concepts in the CI field. There may exist multiple terms designating the same concept. The terms *neural network* and *neural net*, for example, designate the same concept, and so do the terms *fuzzy system*, *fuzzy inference system*, and *fuzzy logic system*. Terms that designate the same concept are referred to as synonyms. In the case of synonyms, we have chosen a preferred term that we use to designate the corresponding concept in a consistent way throughout this chapter.

cept maps are very different from the concept maps originally introduced by Joseph D. Novak (Novak & Gowin, 1984).

The third step in the process of bibliometric mapping is the extraction of relevant information from the raw data collected in the first step. In this chapter, the relevant information consists of the co-occurrence frequencies of concepts. The co-occurrence frequency of two concepts is extracted from a corpus of abstracts by counting the number of abstracts in which the two concepts both occur. To identify the concepts that occur in an abstract, one needs a thesaurus of the scientific field with which one is concerned. Because a thesaurus of the CI field is not available to us, we construct one ourselves. The approach that we take to construct a thesaurus of the CI field is discussed in Subsection 8.2.1. We note that in the present study we do not use the same thesaurus as in our earlier research (Van Eck, Waltman, et al., 2006b, 2006a). This is because the present study covers a longer period of time and, as a consequence, the concepts of interest may differ from our earlier research.

The fourth step in the process of bibliometric mapping is the calculation of similarities between items based on the information extracted in the third step. In this chapter, similarities between items are calculated based on co-occurrence frequencies. In the bibliometric literature, two approaches can be distinguished for calculating similarities between items based on co-occurrence frequencies. One approach, which seems the most popular, is to use the Pearson correlation between the vectors of co-occurrence frequencies of two items as a measure of the items' similarity (McCain, 1990; White & McCain, 1998). The other approach is to normalize co-occurrence frequencies using, for example, the cosine measure, the inclusion index, or the Jaccard index (Peters & Van Raan, 1993b). In this chapter, we take the latter approach, since that approach is recommended in the statistical literature (Borg & Groenen, 2005). To normalize co-occurrence frequencies, we use a measure that we call association strength. A discussion of this measure is provided in Subsection 8.2.2.

The fifth step in the process of bibliometric mapping is the positioning of items in a low-dimensional space based on the similarities calculated in the fourth step. In this chapter, the low-dimensional space is referred to as a concept map and only two-dimensional concept maps are considered. In many studies (McCain, 1990; White & McCain, 1998; Peters & Van Raan, 1993b; Hinze, 1994), the fifth step in the process of bibliometric mapping is performed using the method of multidimensional scaling (Borg & Groenen, 2005). However, it is our experience that multidimensional scaling does not always provide satisfactory results when it is used for bibliometric mapping. More specifically, when a large proportion of the similarities equal zero, which occurs quite frequently in bibliometric mapping, multidimensional scaling always provides maps in which the items lie more or less equally distributed within a circle (in the case of a two-dimensional map). To avoid this problem, we use a method that is closely related to multidimensional scaling. The method, which is called VOS, is discussed in Subsection 8.2.3.

The sixth step in the process of bibliometric mapping is the visualization of the low-dimensional space that results from the fifth step. In our study, we use two different visualization approaches. We have implemented these approaches in two computer programs, which we call the concept map viewer and the concept density map viewer. The concept map viewer visualizes a concept map by displaying for each concept a label that indicates the location of the concept in the concept map. The concept density map viewer, on the other hand, displays labels only for a small number of frequently occurring concepts. In addition, this viewer uses colors to indicate the amount of attention researchers pay to the research topics located in the various areas of a concept map. The concept density map viewer is especially useful to get a quick overview of the division of a scientific field into several subfields and of the way in which subfields are related to each other. The visualizations provided by the concept map viewer and the concept density map viewer are discussed in more detail in Subsection 8.2.4 and 8.2.5, respectively.

An issue that, in our opinion, usually does not get sufficient attention in bibliometric studies is the stability of bibliometric maps. Taking into account the issue of stability can significantly improve the reliability of a bibliometric analysis. We discuss a method for assessing the stability of a bibliometric map in Subsection 8.2.6.

8.2.1 Thesaurus

To construct a thesaurus of the CI field, we make use of a term extraction tool that we have developed ourselves. The tool receives a corpus of abstracts as input. First, by using the MontyLingua software,[3] the tool assigns a part-of-speech category (like verb, noun, or adjective) to each word in the corpus. Then, based on the assigned part-of-speech categories, the tool selects words or sequences of words that are likely to be terms. This is accomplished using a regular expression similar to the one proposed by Justeson and Katz (1995). The output of the tool is a list of candidate terms sorted by frequency of occurrence in the corpus. We manually validate the list of candidate terms. For each candidate term, we decide whether the term is relevant to the CI field. Furthermore, when we consider a candidate term relevant, we identify its synonyms. Synonymy relations are important because terms that are synonymous designate the same concept. The identification of synonyms is also done manually. Using the above procedure, we obtain a simple thesaurus of the CI field consisting of the field's most important terms as well as the synonymy relations between these terms. This thesaurus allows us to identify the concepts that occur in an abstract.

8.2.2 Association Strength

To normalize co-occurrence frequencies of concepts, we use a measure that we call association strength. The aim of this measure is to normalize co-occurrence frequencies in such a way that concepts occurring in many abstracts and concepts occurring in only a few abstracts can be compared in a fair way. The association strength a_{ij} of the concepts i and j is defined as

$$a_{ij} = \frac{mc_{ij}}{c_{ii}c_{jj}} \qquad \text{for } i \neq j, \tag{8.1}$$

where c_{ij} denotes the number of abstracts in which the concepts i and j both occur, c_{ii} denotes the number of abstracts in which concept i occurs, and m denotes the total number of abstracts. The association strength of two concepts can be interpreted as the ratio between on the one hand the co-occurrence frequency of the concepts and on the other hand the expected co-occurrence frequency of the concepts obtained under the assumption that occurrences of the concepts are statistically independent (Van Eck, Waltman, et al., 2006b). To the best of our knowledge, there are, apart from our own research, only a few bibliometric studies in which the association strength measure is used (Peters & Van Raan, 1993b; Hinze, 1994; Rip & Courtial, 1984). In these studies, the measure

[3] See http://web.media.mit.edu/~hugo/montylingua/.

is referred to as the proximity index. In our opinion, however, the association strength measure is preferable over alternative measures for normalizing co-occurrence frequencies, like the cosine measure, the inclusion index, and the Jaccard index. This is because the alternative measures do not always make fair comparisons between concepts with a high frequency of occurrence and concepts with a low frequency of occurrence.

8.2.3 VOS

The positioning of concepts in a concept map based on their association strengths is accomplished using a method that we call VOS, which is an abbreviation for *visualization of similarities*. We now briefly introduce this method. A more elaborate discussion of VOS, including an analysis of the relationship between VOS and multidimensional scaling, is provided elsewhere (Van Eck & Waltman, 2007b).

Let there be n concepts. The aim of VOS is to provide a two-dimensional space in which the concepts $1, \ldots, n$ are located in such a way that the distance between any pair of concepts i and j reflects their association strength a_{ij} as accurately as possible. Concepts that have a high association strength should be located close to each other, whereas concepts that have a low association strength should be located far from each other. The idea of VOS is to minimize a weighted sum of the squared Euclidean distances between all pairs of concepts. The higher the association strength of two concepts, the higher the weight of their squared distance in the summation. To avoid solutions in which all concepts are located at the same coordinates, the constraint is imposed that the sum of all distances must equal some positive constant. In mathematical notation, the objective function to be minimized in VOS is given by

$$E(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \sum_{i<j} a_{ij} \|\mathbf{x}_i - \mathbf{x}_j\|^2, \tag{8.2}$$

where the vector $\mathbf{x}_i = (x_{i1}, x_{i2})$ denotes the location of concept i in a two-dimensional space and $\|\cdot\|$ denotes the Euclidean norm. Minimization of the objective function is performed subject to the constraint

$$\frac{1}{n(n-1)} \sum_{i<j} \|\mathbf{x}_i - \mathbf{x}_j\| = 1. \tag{8.3}$$

Note that the distances $\|\mathbf{x}_i - \mathbf{x}_j\|$ in the constraint are not squared. We numerically solve the constrained optimization problem of minimizing (8.2) subject to (8.3) in two steps. We first convert the constrained optimization problem into an unconstrained optimization problem. We then solve the latter problem using a majorization algorithm (Borg & Groenen, 2005). To reduce the effect of local minima, we run the majorization algorithm using ten random starts. A computer program that implements the majorization algorithm is available online.[4]

8.2.4 Concept Map Visualization

To visualize a concept map, we use a Java applet that we call the concept map viewer. The concept map viewer indicates the location of a concept in a concept map by displaying a label at that location. This label shows a term that designates the concept. The viewer has scroll, zoom, and search functionality to support a comprehensive examination of a concept map. In addition to visualizing the associations between concepts, the viewer also visualizes the importance of concepts and the distribution of the interest in concepts over the neural networks, fuzzy systems, and evolutionary computation subfields. The importance of a concept, measured by counting the number of abstracts in which the concept occurs, is indicated by the size of the label representing the concept. The distribution of the interest in a concept over the neural networks, fuzzy systems, and evolutionary computation subfields, measured by calculating for each subfield the proportion of the abstracts in which the concept occurs, is indicated by the color of the label representing the concept. A color consists of a red, green, and blue component, each of which has a value between 0 and 255. Consider the color of the label representing concept i. The red, green, and blue component of this color are given by

$$r\left(p_i^{\mathrm{FS}}, p_i^{\mathrm{NN}}, p_i^{\mathrm{EC}}\right) = \frac{p_i^{\mathrm{FS}}}{p_i^{\mathrm{FS}} + p_i^{\mathrm{NN}} + p_i^{\mathrm{EC}}} 180 + 75, \tag{8.4}$$

$$g\left(p_i^{\mathrm{FS}}, p_i^{\mathrm{NN}}, p_i^{\mathrm{EC}}\right) = \frac{p_i^{\mathrm{NN}}}{p_i^{\mathrm{FS}} + p_i^{\mathrm{NN}} + p_i^{\mathrm{EC}}} 180 + 75, \tag{8.5}$$

[4] See http://www.neesjanvaneck.nl/vos/.

and

$$b\left(p_i^{\mathrm{FS}}, p_i^{\mathrm{NN}}, p_i^{\mathrm{EC}}\right) = \frac{p_i^{\mathrm{EC}}}{p_i^{\mathrm{FS}} + p_i^{\mathrm{NN}} + p_i^{\mathrm{EC}}} 180 + 75, \tag{8.6}$$

respectively, where p_i^{FS} denotes the proportion of the abstracts from the IEEE Transactions on Fuzzy Systems and the IEEE International Conference on Fuzzy Systems in which concept i occurs, p_i^{NN} denotes the proportion of the abstracts from the IEEE Transactions on Neural Networks and the International Joint Conference on Neural Networks in which concept i occurs, and p_i^{EC} denotes the proportion of the abstracts from the IEEE Transactions on Evolutionary Computation and the IEEE Congress on Evolutionary Computation in which concept i occurs. Using (8.4), (8.5), and (8.6), the color of a label is not influenced by differences in the number of papers published in the neural networks, fuzzy systems, and evolutionary computation subfields.

8.2.5 Concept Density Map Visualization

A disadvantage of the concept map visualization discussed above is that labels of concepts usually overlap each other. This may obscure the overall structure of a concept map. Due to overlapping labels, it may for example be difficult to get a clear overview of the way in which a field of science is divided into subfields. To gain more insight into the overall structure of a concept map, we use a MATLAB program that we call the concept density map viewer. We refer to the maps shown by this viewer as concept density maps. Rather than displaying labels for all concepts, the concept density map viewer displays labels only for a small number of frequently occurring concepts. In addition, the viewer uses colors to indicate the amount of attention researchers pay to the research topics located in the various areas of a concept map. The amount of attention for a research topic is measured by counting the number of abstracts concerned with that topic. The idea of concept density maps has been introduced by Van Eck, Frasincar, and Van den Berg (2006). In this subsection, we present a refinement of their methodology for constructing concept density maps.

Concept density maps are based on the notion of concept density. The concept density at a specific location in a concept map depends both on the number of neighboring concepts and on the importance of these concepts. The higher the number of neighboring concepts and the smaller the distance between these concepts and the location under consideration, the higher the concept density. Also, the more important the neighboring concepts, as indicated by the number of abstracts in which they occur, the higher the concept density. The general idea of a concept density map is that the amount of attention researchers pay to a research topic located in a specific area of a concept map is indicated by the concept density in that area. In a concept density map, colors are used to display the concept density in the various areas of a concept map. In this way, areas with a high concept density can be easily identified. Such areas contain concepts that together receive a lot of attention from researchers. Most likely, the areas therefore point to important research topics.

We now discuss the construction of concept density maps. The concept density at a specific location in a concept map is calculated by first placing a so-called kernel function at each concept location and then taking a weighted average of the kernel functions. The weight of a kernel function is set equal to the number of abstracts in which the corresponding concept occurs. In mathematical notation, the concept density at location $\mathbf{x} = (x_1, x_2)$ is given by

$$D(\mathbf{x}) = \frac{1}{h^2 \sum_{i=1}^{n} c_{ii}} \sum_{i=1}^{n} c_{ii} K\left(\frac{x_1 - x_{i1}}{h}, \frac{x_2 - x_{i2}}{h}\right), \tag{8.7}$$

where K denotes a kernel function and h denotes a smoothing parameter. Recall further that c_{ii} denotes the number of abstracts in which concept i occurs and $\mathbf{x}_i = (x_{i1}, x_{i2})$ denotes the location of concept i in a concept map. The kernel function K must satisfy the conditions

$$\forall t_1, t_2, t_3, t_4 : t_1^2 + t_2^2 = t_3^2 + t_4^2 \Rightarrow K(t_1, t_2) = K(t_3, t_4), \tag{8.8}$$

$$\forall t_1, t_2, t_3, t_4 : t_1^2 + t_2^2 < t_3^2 + t_4^2 \Rightarrow K(t_1, t_2) \geq K(t_3, t_4), \tag{8.9}$$

and

$$\forall t_1, t_2 : K(t_1, t_2) \geq 0. \tag{8.10}$$

A kernel function satisfying these conditions is invariant to rotation. We require this property because concept maps are also invariant to rotation. In this chapter, we use the bivariate standard normal distribution for the kernel function K, which means that

$$K(t_1, t_2) = \frac{1}{2\pi} \exp\left(-\frac{t_1^2 + t_2^2}{2}\right). \tag{8.11}$$

The smoothness of the concept density function in (8.7) is determined by the smoothing parameter h. Choosing an appropriate value for h is essential. A too small value for h results in a concept density function that is too rough, whereas a too large value results in a concept density function that is too smooth. The coloring of a concept density map is based on concept densities calculated using (8.7). We use colors ranging from blue to red in our research. Blue areas in a concept density map have the lowest concept density and thus point to research topics that receive very little attention from researchers. Red areas, on the other hand, have the highest concept density and thus point to research topics that receive a lot of attention from researchers.

As a final remark, we note that the above approach to calculating concept densities is mathematically somewhat similar to the statistical technique of kernel density estimation. This technique is discussed by, for example, Scott (1992).

8.2.6 Stability

A bibliometric map can be considered stable if small changes in the underlying data produce only small changes in the map (De Leeuw & Meulman, 1986). Although the concept maps presented in this chapter are constructed using VOS, the stability of the maps can be analyzed in a similar way as in the case of maps constructed using multidimensional scaling methods. De Leeuw and Meulman (1986) propose to analyze the stability of multidimensional scaling maps by studying the effect of leaving out one object. Other approaches to stability analysis, proposed by Heiser and Meulman (1983a, 1983b) and Weinberg, Carroll, and Cohen (1984), investigate the effect of random sampling on multidimensional scaling maps. The latter approaches all rely on the statistical technique of bootstrapping.

Our analysis of the stability of our concept maps also focuses on the effect of random sampling. The approach that we take is quite similar to the one discussed by Heiser and Meulman (1983b). When constructing a concept map, the corpus of abstracts on which the map is based can be regarded as a sample, with each abstract representing an observation. The sample defines an empirical probability distribution over abstracts. A bootstrap sample is a sample that is drawn, with replacement, from this empirical probability distribution. A bootstrap sample has the same size as the original sample. In this chapter, 100 bootstrap samples are drawn in order to analyze the stability of a concept map. For each bootstrap sample, a concept map is constructed using the methodology discussed above. Since concept maps are invariant to rotation, reflection, translation, and dilation (i.e., stretching and shrinking), we cannot directly compare the concept maps obtained from the different bootstrap samples. Instead, we first use Procrustes rotation (Borg & Groenen, 2005) to match each concept map as closely as possible to the concept map obtained from the original sample. In this way, we end up with 100 concept maps that can be used to analyze the stability of individual concepts. For each concept, we thus have 100 locations, each obtained from a different bootstrap sample. To analyze the stability of a concept in a concept map, we draw an ellipse that covers most of the bootstrap locations of the concept. The ellipse is centered at the average of the bootstrap locations. The shape of the ellipse is based on the assumption of a bivariate normal sampling distribution and depends on the standard deviations and the correlation estimated using the bootstrap procedure. The size of the ellipse is determined in such a way that the ellipse covers exactly 90% of the bootstrap locations. In this way, an ellipse can be interpreted as an approximate 90% confidence region for the location of a concept.

8.3 Analysis

As stated before, our analysis is based on abstracts of papers from three major journals and three major conferences in the CI field. Furthermore, two time periods are considered in the analysis, 1996–2000 and 2001–2005. For each period, the number of abstracts that we obtained from the different journals and conference proceedings is reported in Table 8.2.[5] Based on the abstracts, we constructed a thesaurus of the CI field using the approach discussed in Subsection 8.2.1. We ended up with a thesaurus containing 376 concepts. However, when constructing concept maps of the CI field, we only

[5] Since the first issue of the IEEE Transactions on Evolutionary Computation appeared in 1997, abstracts from this journal were not available for the year 1996.

Table 8.2: Number of abstracts in the corpus.

Journal / conference proceedings	Number of abstracts	
	1996–2000	2001–2005
IEEE Trans. Neural Networks	701	682
IEEE Trans. Fuzzy Systems	272	360
IEEE Trans. Evolutionary Computation	89	203
Proc. Int. Joint Conf. Neural Networks	2761	2761
Proc. IEEE Int. Conf. Fuzzy Systems	1452	1148
Proc. IEEE Congr. Evolutionary Computation	960	1629
Total	6235	6783

included concepts that occurred in at least ten abstracts. This was done because we considered the amount of data on concepts occurring in less than ten abstracts too limited for a reliable analysis. In the periods 1996–2000 and 2001–2005, there were, respectively, 332 and 337 concepts that occurred in at least ten abstracts. For these concepts, we counted the co-occurrence frequencies. In both periods, 74% of the co-occurrence frequencies turned out to be equal to zero, which indicates that most combinations of concepts did not occur in any abstract at all. The concept maps that we constructed for the periods 2001–2005 and 1996–2000 are shown in Figures 8.1 and 8.4, respectively. The corresponding concept density maps are shown in Figures 8.2 and 8.5. Since the figures are printed in black and white, the coloring of the labels (see Subsection 8.2.4) is not visible in the concept maps. Similarly, in the concept density maps, colors indicating the density of concepts (see Subsection 8.2.5) are not visible. Instead, curves that indicate points of equal density are shown in the concept density maps. Concept maps and concept density maps with the correct coloring are available online.[6] We encourage the interested reader to have look at these maps, since they are much more insightful than maps printed in black and white. Moreover, we have also made available online our concept map viewer (see Subsection 8.2.4). Using this viewer, the concept maps in Figures 8.1 and 8.4 can be examined in much more detail. To provide some insight into the stability of our concept maps, approximate 90% confidence regions for a number of frequently occurring concepts in the periods 2001–2005 and 1996–2000 are shown

[6] See http://www.neesjanvaneck.nl/ijufks/.

in Figures 8.3 and 8.6, respectively. The confidence regions were calculated using the bootstrap approach discussed in Subsection 8.2.6.

8.3.1 Structure of the Computational Intelligence Field

To analyze the current structure of the CI field, we consider the maps for the period 2001–2005, which are shown in Figures 8.1, 8.2, and 8.3. Our initial expectation was to find three well-separated clusters of concepts, corresponding to the three well-known subfields of the CI field, that is, neural networks, fuzzy systems, and evolutionary computation. This is also what we found in our earlier research (Van Eck, Waltman, et al., 2006b, 2006a), in which we used a smaller data set and a smaller thesaurus than in the present study. However, somewhat to our surprise, there is no very clear correspondence between on the one hand the clusters that can be observed in our maps and on the other hand the three subfields of the CI field. The clusters can be seen most easily in the concept density map in Figure 8.2. The cluster in the right part of the map clearly corresponds to the evolutionary computation subfield, but the clusters in the left part of the map do not correspond one-to-one to the neural networks and fuzzy systems subfields. Instead, the clustering in the left part of the map seems to reflect different types of problems that are studied in the CI field. In the lower left part, there is a cluster for control problems. In the upper left part, there is a cluster for classification problems, that is, for problems involving the prediction of a class label. And in the center of the left part, there is a cluster for problems in which a continuous value has to be predicted. We will refer to the latter problems as regression problems. Moreover, the interpretation of clusters in terms of the type of problem with which they are concerned can also be applied to the cluster in the right part of the map. Since evolutionary computation primarily deals with optimization, this cluster can be seen as a cluster for optimization problems. So, following the above interpretation of the maps for the period 2001–2005, it turns out that, contrary to our expectation, the CI field is not structured around the three most important techniques studied in the field, that is, neural networks, fuzzy systems, and evolutionary computation. Instead, the field is structured around what seem to be the four main types of problems with which the field is concerned. These types of problems are control problems, classification problems, regression problems, and optimization problems.

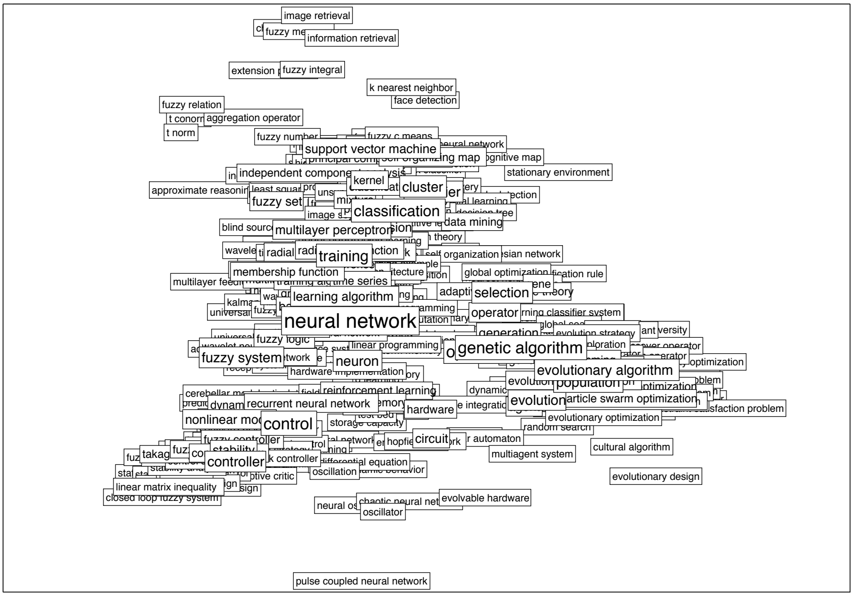


Figure 8.1: Concept map for the period 2001–2005.

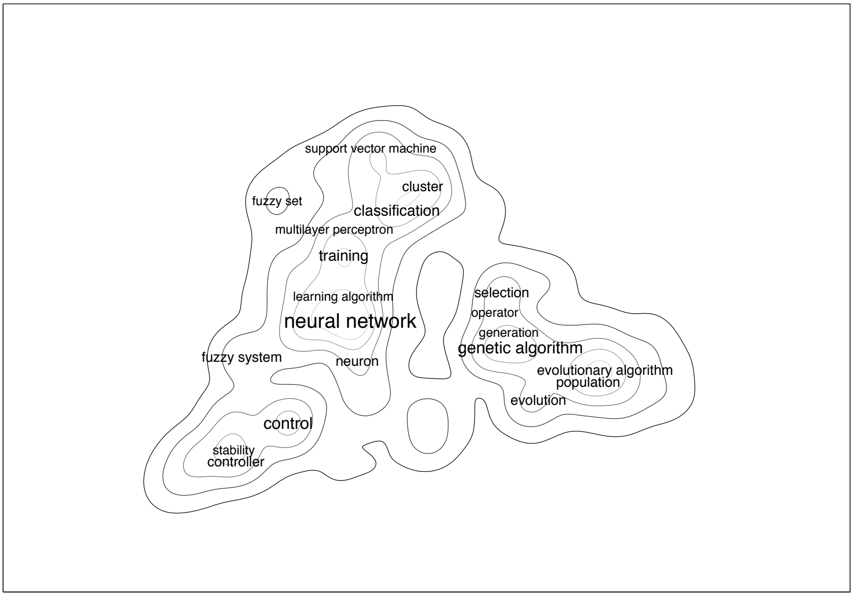


Figure 8.2: Concept density map for the period 2001–2005.

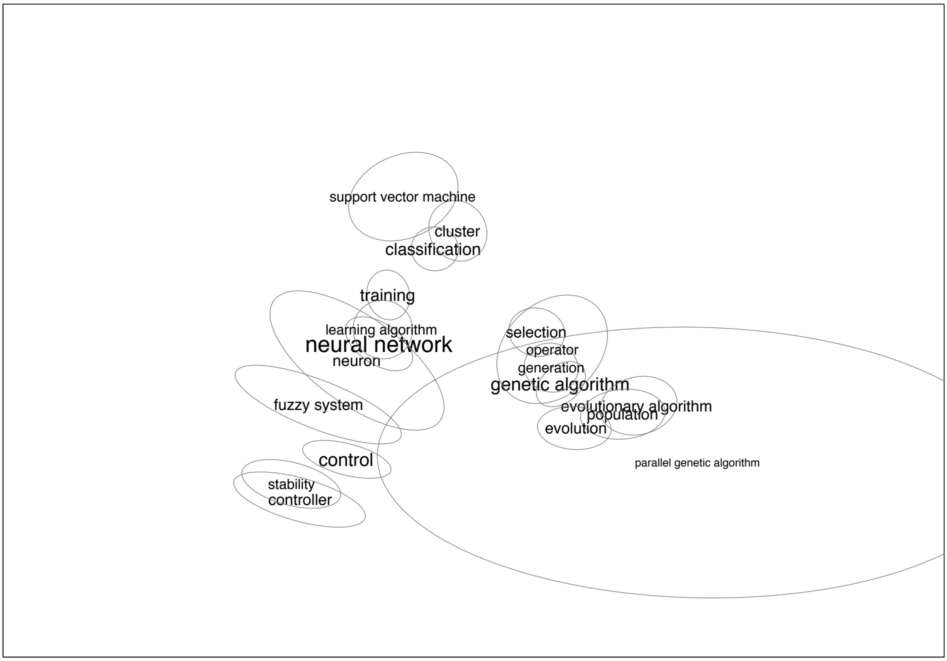


Figure 8.3: Approximate 90% confidence regions for a number of frequently occurring concepts in the period 2001–2005.

A closer examination of the concept map for the period 2001–2005, either using Figure 8.1 or using the concept map viewer available online, reveals that each of the three clusters in the left part of the map contains both concepts from the neural networks subfield and concepts from the fuzzy systems subfield. The control cluster is dominated by fuzzy systems concepts, but the cluster also contains some neural networks concepts, for example *recurrent neural network*, *neural network controller*, and *neural system*. Most concepts in the classification and regression clusters, on the other hand, belong to the neural networks subfield, but there are also a number of fuzzy systems concepts in these clusters. Some examples are *fuzzy c-means*, *fuzzy clustering*, and *fuzzy classifier* in the classification cluster and *membership function*, *fuzzy inference*, and *defuzzification* in the regression cluster. Together, all these examples clearly indicate that the clustering found in our maps does not coincide with the division of the CI field into the neural networks, fuzzy systems, and evolutionary computation subfields. More specifically, the neural networks and fuzzy systems subfields turn out to be fairly intertwined. The evolutionary computation subfield, on the other hand, has a relatively independent position within the CI field.

Based on the maps, some further observations on the structure of the CI field can be made. The concept density map in Figure 8.2 shows that the classification cluster and the regression cluster are only weakly separated from each other. The separation between other clusters is much stronger. One might even argue, based on the concept density map, that there is in fact one large cluster, which is concerned with both classification and regression problems. The weak separation between the classification cluster and the regression cluster seems to indicate that classification and regression problems are seen as fairly similar. This is probably due to the fact that important CI techniques like neural networks and fuzzy systems can be applied to both types of problems. Using the concept map, it can further be observed that within the classification cluster there is no clear separation between concepts related to classification (e.g., *classification*, *support vector machine*, and *neural network classifier*) on the one hand and concepts related to clustering (e.g., *cluster*, *fuzzy c-means*, and *fuzzy clustering*) on the other hand. Apparently, researchers do not see much difference between classification and clustering.

We now consider the map in Figure 8.3, which shows approximate 90% confidence regions for a number of frequently occurring concepts in the period 2001–2005. It can be seen that some concepts, like *neuron* and *fuzzy system*, are quite unstable. Other concepts, like *genetic algorithm* and *classification*, are much more stable. For comparison, the concept *parallel genetic algorithm*, which occurs in only ten abstracts, is also shown in the map. This concept is highly unstable, as indicated by its very large confidence region. Although concepts with confidence regions of this size are rather exceptional, it turns out that, on average, less frequently occurring concepts are also less stable. This is because the locations of these concepts in a concept map are calculated from a relatively small amount of data. The example of *parallel genetic algorithm* shows that one should be very careful when making detailed statements based on the location of a single concept, especially if the concept occurs in only a few abstracts. The above analysis of the structure of the CI field does not contain any very detailed statement, and it therefore does not depend too strongly on the exact locations of individual concepts.

In our opinion, a more detailed analysis may be possible, but such an analysis should be performed very carefully.

8.3.2 Evolution of the Computational Intelligence Field Over the Last Decade

To analyze the evolution of the CI field over the last decade, we first consider the differences in the number of occurrences of concepts in the periods 1996–2000 and 2001–2005. In Table 8.3, the concepts are listed that have the largest relative increase in their number of occurrences between the two periods. Only concepts occurring in at least 20 abstracts in the period 2001–2005 are shown. Similarly, the concepts with the largest relative decrease in their number of occurrences are listed in Table 8.4. This table only shows concepts that occur in at least 20 abstracts in the period 1996-2000. For each concept in Tables 8.3 and 8.4, the number of abstracts in which the concept occurs in the periods 1996-2000 and 2001–2005 is reported.

The data in Table 8.3 indicate a number of emerging areas in the CI field. Interestingly, most of these areas lie in the evolutionary computation subfield. The data reveal six emerging areas in this subfield. These areas are genetic regulatory networks, evolutionary multiobjective optimization, artificial immune systems, particle swarm optimization, ant colony optimization, and differential evolution. Furthermore, the interest of evolutionary computation researchers in the area of learning classifier systems has also increased considerably over the last years. As can be seen in Table 8.2, the recent developments in the evolutionary computation subfield have resulted in a large increase in the number of papers from this subfield. Another emerging area revealed by the data in Table 8.3 is support vector machines. Most abstracts containing the concept *support vector machine* belong to papers from the IEEE Transactions on Neural Networks or the International Joint Conference on Neural Networks. This shows that support vector machines research is usually seen as part of the neural networks subfield. Given the fairly large number of papers concerned with support vector machines, it is quite remarkable that the topic of support vector machines is not covered in two recent textbooks on CI (Engelbrecht, 2003; Konar, 2005). Apparently, there is no complete consensus within the CI community on the question whether support vector machines research belongs to the CI field at all. In the fuzzy systems subfield, research interest in the topic of fuzzy

Table 8.3: Concepts with the largest relative increase in their number of occurrences.

Concept	Number of occurrences	
	1996–2000	2001–2005
genetic regulatory network	0	26
NSGA-II	0	22
least squares support vector machine	1	27
artificial immune system	2	34
evolutionary multiobjective optimization	3	36
particle swarm optimization	10	113
pareto front	5	41
gaussian kernel	3	21
ant colony optimization	4	28
support vector machine	39	264
multiobjective evolutionary algorithm	11	70
learning classifier system	4	25
support vector	12	71
association rule	5	23
long term memory	5	21
pareto optimal solution	6	24
ant	14	51
immune system	10	34
kernel	54	173
multiobjective optimization	35	112
differential evolution	11	35
ant colony	8	25
gene	52	135
mutual information	19	49
image retrieval	11	27

association rules has increased significantly over the last decade. This is indicated by the concept *association rule* in Table 8.3.

Obviously, there must also be areas with a decreasing interest of CI researchers. These areas are indicated by the data in Table 8.4. In the neural networks subfield, interest in the area of feedforward neural networks has decreased considerably. The same is true for the area of fuzzy control in the fuzzy systems subfield. In the evolutionary

Table 8.4: Concepts with the largest relative decrease in their number of occurrences.

Concept	Number of occurrences	
	1996–2000	2001–2005
fuzzy constraint	21	4
constructive algorithm	28	8
cascade correlation	23	7
fuzzy logic control	48	15
multilayer feedforward neural network	44	16
control action	33	13
hidden unit	117	48
iris data	31	13
fuzzy number	63	27
evolutionary programming	90	39
fuzzy control system	73	32
feedforward neural network	184	82
sliding mode controller	20	9
universal approximator	31	14
fuzzy logic controller	128	58
defuzzification	44	20
knowledge base	78	37
PID controller	41	20
rule extraction	43	21
inverted pendulum	57	28
expert system	51	26
approximate reasoning	25	13
backpropagation	398	211
fuzzy controller design	22	12
output layer	42	23

computation subfield, the amount of research in the area of evolutionary programming has clearly decreased.

We now compare the maps for the period 1996–2000, shown in Figures 8.4, 8.5, and 8.6, to the maps for the period 2001–2005, shown in Figures 8.1, 8.2, and 8.3. The concept density map in Figure 8.5 reveals that in the period 1996–2000 the CI field was largely structured around the three most important techniques studied in the field, that is, neural networks, fuzzy systems, and evolutionary computation. The map clearly shows

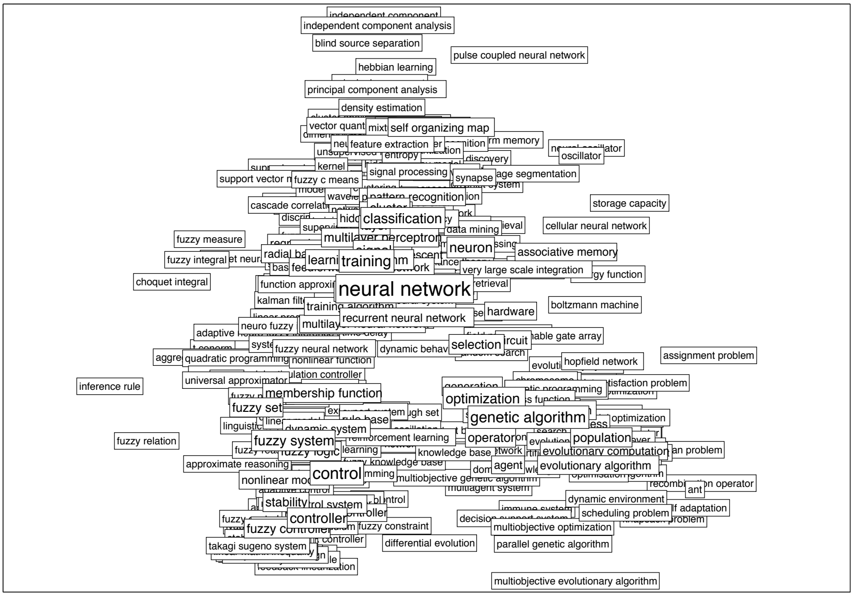


Figure 8.4: Concept map for the period 1996–2000.

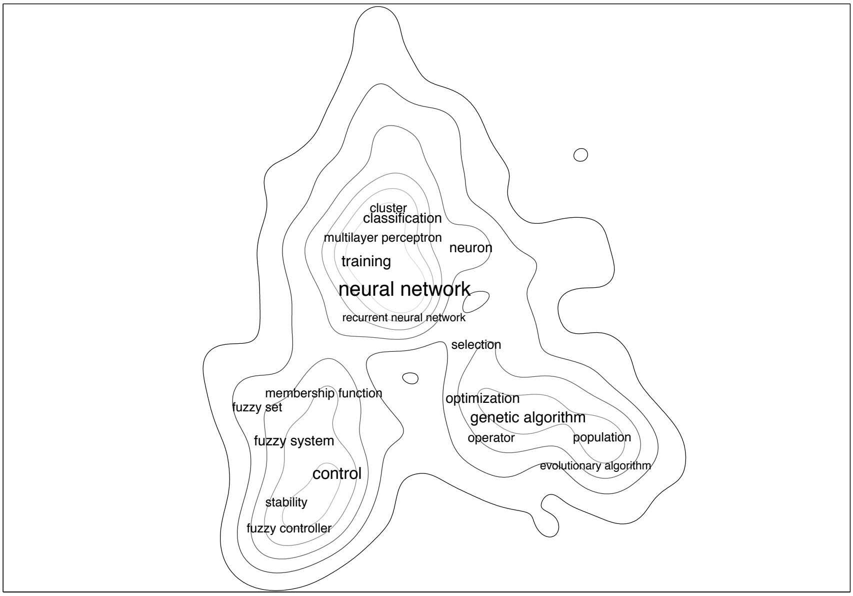


Figure 8.5: Concept density map for the period 1996–2000.

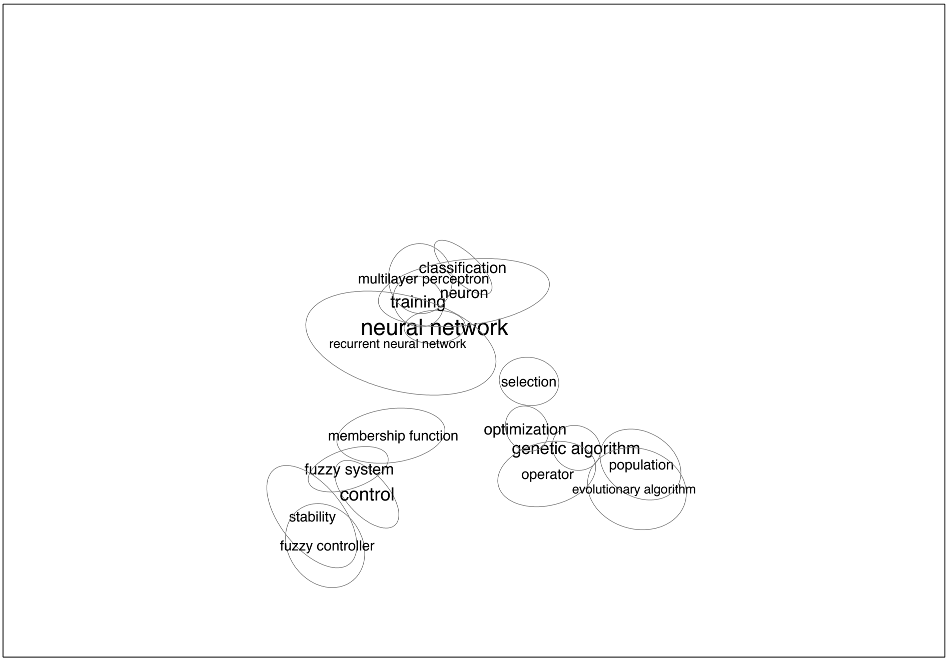


Figure 8.6: Approximate 90% confidence regions for a number of frequently occurring concepts in the period 1996–2000.

three clusters, each corresponding to one of the three techniques. The correspondence between the three clusters and the three techniques is not perfect. By examining the concept map for the period 1996–2000, either using Figure 8.4 or using the concept map viewer available online, it can be seen that some fuzzy systems concepts are located in the neural networks cluster. Most of these concepts have to do with classification (e.g., *fuzzy classifier* and *fuzzy classification*), clustering (e.g., *fuzzy clustering* and *fuzzy c-means*), or neuro-fuzzy systems (e.g., *fuzzy neural network* and *neuro-fuzzy inference system*). However, even though the correspondence between the three clusters and the three most important CI techniques is not perfect, it is clear that in the period 1996–2000 the CI field was much more structured around techniques than it was in the period 2001–2005. As discussed above, in the latter period the field was structured around four types of problems that each receive a lot of attention in the field.

Based on the concept density maps in Figures 8.2 and 8.5, some further observations on the evolution of the CI field can be made. One thing to note is that in the map for the period 1996–2000 concepts related to classification and concepts related to regression are located much closer to each other than in the map for the period 2001–2005. Apparently, nowadays research into classification problems on the one hand and into regression problems on the other hand is somewhat more separated than it was some years ago. Another observation is that concepts related to control and concepts related to neural networks have moved toward each other. This might be an indication that the application of neural network techniques to control problems has increased over the last decade.

8.4 Conclusions

In this chapter, we have presented a bibliometric study of the CI field. Based on our analysis, we can draw a number of conclusions. First of all, our initial expectation that the CI field is structured around the neural networks, fuzzy systems, and evolutionary computation subfields turns out to be too simplistic. As revealed by our bibliometric maps for the period 2001–2005, the CI field can best be seen as a field that is structured around four important types of problems, namely control problems, classification problems, regression problems, and optimization problems. Moreover, the neural networks and fuzzy systems subfields turn out to be fairly intertwined. Both subfields are concerned with control, classification, and regression problems. The evolutionary computation subfield mainly deals with optimization problems, and it therefore turns out to have a relatively independent position within the CI field. Interestingly, the intertwining of the neural networks and fuzzy systems subfields has increased considerably over the last decade. This can be seen by comparing the maps for the period 2001–2005 to the maps for the period 1996–2000. In the latter maps, the neural networks and fuzzy systems subfields are clearly separated from each other. Apparently, in the last decade there must have been some development in the CI field that has brought the neural networks and fuzzy systems subfields closer together. A possible explanation might be that more and more researchers recognize that in many cases neural network techniques and fuzzy system techniques are applied to rather similar problems, even though the techniques themselves are very different. As a consequence, more and more researchers become interested in comparing the two types of techniques, and they start combining them into hybrid systems. So, researchers focus less on one type of technique. Instead, they focus on the problem with which they are concerned, and they try to find the technique or the combination of techniques that solves the problem in the most satisfactory way.

Our analysis of the frequency with which researchers use specific concepts in their papers has revealed a number of emerging areas in the CI field. These areas are genetic regulatory networks, evolutionary multiobjective optimization, artificial immune systems, particle swarm optimization, ant colony optimization, differential evolution, and support vector machines. Interestingly, most of these areas lie in the evolutionary computation subfield, which suggests that this subfield has been particularly innovative over the last decade. We also note that it is not completely clear whether the area of support vector machines should be seen as part of the CI field at all. The interest of CI researchers in a number of more traditional research topics has decreased significantly over the last decade. These topics are feedforward neural networks, fuzzy control, and evolutionary programming.

Chapter 9

Summary and Future Research

9.1 Summary of the Thesis

Bibliometric mapping of science is concerned with quantitative methods for visually representing scientific literature based on bibliographic data. Bibliometric mapping has a rich history starting with the first pioneering efforts in the 1970s. During four decades of bibliometric mapping research, a large number of methods and techniques have been proposed and tested. Although this has not resulted in a single generally accepted methodological standard, it did result in a limited set of methods and techniques that are commonly used by a majority of the researchers.

In this thesis, a new methodology for bibliometric mapping has been presented. It has been argued that some commonly used methods and techniques for bibliometric mapping have important shortcomings. In particular, popular normalization methods, such as the cosine method and the Jaccard method, lack a solid mathematical justification, and popular multidimensional-scaling-based approaches for constructing bibliometric maps suffer from artifacts, especially when working with larger data sets. Also, the presentation of bibliometric maps is often done using very simple static pictures and without offering any possibility for interaction. The aim of the methodology introduced in this thesis is to provide improved methods and techniques for bibliometric mapping.

A general introduction into bibliometric mapping was provided in Chapter 1 of the thesis. An outline of the various steps of the bibliometric mapping process was also given in this chapter. In Chapters 2 to 8 of the thesis, seven separate studies were presented. The first six studies each focused on a specific step of the bibliometric mapping process. The seventh study was concerned with an application of bibliometric mapping. We will now summarize each of the studies.

In Chapter 2, a new technique for automatic term identification was introduced. This technique can be used to automatically select the terms to be shown in a term map. The technique looks at the way in which noun phrases are distributed over topics. The more the distribution of a noun phrase is biased towards a single topic, the more likely the noun phrase is to represent a relevant term in the domain of interest. The main conclusion that can be drawn from Chapter 2 is that for many purposes the proposed technique works sufficiently well, but that manual intervention remains necessary if a highly accurate selection of relevant terms is needed.

Chapters 3 and 4 were concerned with methods for normalizing relatedness scores of objects. These methods were referred to as similarity measures in these chapters. In Chapter 3, so-called indirect similarity measures were considered. In Chapter 4, the focus was on direct similarity measures. In both chapters, a strictly mathematical point of view was taken. More specifically, a number of properties were formulated that a reasonable similarity measure should satisfy, and it was derived which similarity measures indeed satisfy these properties and which do not. In Chapter 3, a number of indirect similarity measures were suggested that have satisfactory mathematical properties. In Chapter 4, a large family of direct similarity measures was considered, and it was concluded that within this family there is essentially only one measure, the so-called association strength measure, that has fully satisfactory mathematical properties. Other more commonly used measures, such as the cosine measure and the Jaccard measure, do not have fully satisfactory properties.

In Chapter 5, the VOS mapping technique was introduced. This technique can be seen as an alternative to the well-known technique of multidimensional scaling. The mathematical relation between the VOS mapping technique and multidimensional scaling was pointed out, and an empirical comparison was performed in which both techniques were used to construct a number of bibliometric maps. It was found that two commonly used multidimensional scaling approaches for constructing bibliometric maps suffer from artifacts. One artifact is the tendency to locate the most important objects in the center of a map and less important objects in the periphery. Another artifact is the tendency to locate objects in a circular structure. The VOS mapping technique turned out not to have these problems. Based on these observations, the conclusion was drawn that in general the VOS mapping technique produces more satisfactory bibliometric maps than the two commonly used multidimensional scaling approaches that were studied.

In Chapter 6, the VOS clustering technique was introduced. This technique can be used to cluster the objects in a bibliometric map. The technique can serve as an alternative to other clustering techniques, such as the commonly used technique of hierarchical clustering. It was shown in Chapter 6 that the VOS clustering technique can be derived from the same underlying mathematical principle as the VOS mapping technique. Because of this, the combination of the two VOS techniques provides a unified framework for mapping and clustering. The advantage of such a unified framework is that it will provide mapping and clustering results that are consistent with each other. In the literature, mapping and clustering techniques are often used together, but the techniques are typically based on different principles, which may lead to inconsistent results. It was also shown in Chapter 6 that the VOS clustering technique is closely related to modularity-based clustering, which is a popular clustering technique in the physics literature (Newman, 2004a, 2004b; Newman & Girvan, 2004). The unified mapping and clustering approach introduced in Chapter 6 was tested by constructing a map of highly cited publications in the field of information science.

Chapter 7 was concerned with the VOSviewer software for displaying and exploring bibliometric maps. The functionality of the software was presented, and the technical implementation of specific parts of the software was discussed. Also, an application was shown in which the software was used to construct and display a co-citation based map of 5000 major scientific journals.

Finally, in Chapter 8, an application of bibliometric mapping was presented. Bibliometric maps were constructed based on journal and conference publications in the field of computational intelligence. To study the evolution of the field over time, maps were produced for two time periods. Using the bibliometric maps, the main problems studied in the field of computational intelligence could be identified, and the position of the evolutionary computation, fuzzy systems, and neural networks subfields relative to each other could be analyzed.

9.2 Outlook and Directions for Future Research

The methods and techniques introduced in this thesis have been used in a number of scientific papers (Heersmink, Van den Hoven, Van Eck, & Van den Berg, 2011; Leydesdorff, Hammarfelt, & Akdag Salah, in press; Lu & Wolfram, 2010; Su & Lee, 2010; Tijssen, 2010; Waaijer et al., 2010, 2011; Waltman, Yan, & Van Eck, in press; Zuccala & Van Eck, 2011). More papers employing the methods and techniques introduced in this thesis are expected to appear in the near future. Especially the VOSviewer software is receiving more and more attention in the scientific community. The development of the VOSviewer software will continue, and it is hoped that the software will be of value to a large group of users, both inside and outside the field of bibliometrics, and also outside the academic world.

The bibliometric mapping methodology introduced in this thesis is also being used on a regular basis in commercial research projects conducted by the Centre for Science and Technology Studies of Leiden University. These projects are commissioned by governments, funding agencies, universities, and scientific publishers. In most cases, the projects have science policy or research management objectives. We expect the use of bibliometric mapping in a science policy and research management context to become more and more common. Because of this, the application of the methods and techniques introduced in this thesis for science policy and research management objectives may be an important topic for future research.

There are various other directions for future research. In particular, the methodology introduced in this thesis can be extended in a number of ways. Some possibilities in this direction are listed below:

- The technique for automatic term identification introduced in Chapter 2 requires the use of a clustering technique (i.e., probabilistic latent semantic analysis) for identifying topics. At the moment, we are investigating simpler techniques for automatic term identification that do not require the use of a clustering technique. Instead, these techniques identify terms directly based on their position in the network of co-occurrences of noun phrases. This approach is computationally much more efficient, and we also consider it conceptually more elegant. Our new approach to automatic term identification will be implemented in the next version of the VOSviewer software.

- It was found empirically that in some cases the association strength normalization method (see Chapter 4) does not yield a completely satisfactory normalization. An alternative, closely related normalization method is currently being tested.

- Another empirical observation is that, in the case of a map with two or more dimensions, the objective function of the VOS mapping technique (see Chapter 5) does not seem to have any non-global optima. Hence, optimization of the objective function seems easy, since there are no problems with local optima. This property of the objective function needs further mathematical investigation.

- In general, the VOS mapping technique produces well-structured maps. However, in the case of maps with lots of objects (see e.g. Section 7.6), the accuracy of the VOS mapping technique at the local level can be somewhat disappointing. Future research may be directed at improving the local accuracy of the VOS mapping technique in the case of maps with lots of objects. An interesting mapping technique that seems to yield accurate results both at the local and at the global level is the LinLog technique proposed by Noack (2007). A disadvantage of this technique is that it is based on an objective function that seems to be much more difficult to optimize than the objective function of the VOS mapping technique.

- The VOS clustering technique (see Chapter 6) produces non-overlapping clusters. This means that each object is assigned to exactly one cluster. In future research, variants of the VOS clustering technique may be developed that allow for overlapping clusters. In such variants, objects can be assigned to multiple clusters, resulting in a so-called fuzzy clustering of the objects. Another possibility is to develop variants of the VOS clustering technique that allow for hierarchically organized clusters.

- Bibliometric maps can be quite sensitive to noise in the underlying data. This noise can for example be a consequence of the relatively arbitrary decisions researchers make when choosing the references they cite or the terminology they use. To obtain some insight into the possible effect of noise on a bibliometric map, it would be desirable to have a quantitative measure of the sensitivity of a map to small changes in the underlying data. One possibility for calculating such a measure may be the use of a bootstrapping technique (in a somewhat similar way as in Chapter 8).

- Bibliometric mapping is often used for dynamic analyses, where the focus is on the changes that take place over time. Although the methods and techniques introduced in this thesis can be used for dynamic analyses (see Chapter 8), they have been developed primarily for static analyses. Static analyses, which focus on a single point in time, typically involve less difficulties than dynamic analyses. Future research may be aimed at developing a bibliometric mapping methodology that is intended specifically for dynamic analyses.

It is hoped that the above technical issues can be addressed in the near future.

Nederlandse Samenvatting (Summary in Dutch)

Dit proefschrift gaat over het maken van bibliometrische kaarten van de wetenschap. Bibliometrisch karteren houdt zich bezig met kwantitatieve methodes voor het visueel weergeven van wetenschappelijke literatuur op basis van bibliografische gegevens. Onderzoek op het gebied van bibliometrisch karteren heeft een rijke geschiedenis die teruggaat tot de jaren 70 van de vorige eeuw. Gedurende veertig jaar onderzoek zijn een groot aantal methodes en technieken geïntroduceerd en getest. Hoewel dit niet heeft geleid tot een algemeen geaccepteerde methodologische standaard, heeft het wel een beperkte verzameling van methodes en technieken opgeleverd die veelvuldig door onderzoekers worden gebruikt.

In dit proefschrift wordt een nieuwe methodologie voor bibliometrisch karteren gepresenteerd. Bepaalde veelgebruikte methodes en technieken voor bibliometrisch karteren hebben serieuze tekortkomingen. Populaire normalisatiemethodes, zoals de cosinus methode en de Jaccard methode, hebben bijvoorbeeld geen solide wiskundige onderbouwing. Populaire technieken voor het construeren van bibliometrische kaarten, gebaseerd op het idee van meerdimensionale schaling, hebben last van artefacten, in het bijzonder wanneer er met grote hoeveelheden gegevens wordt gewerkt. Verder worden voor de presentatie van bibliometrische kaarten vaak eenvoudige statische afbeeldingen gebruikt, zonder enige mogelijkheid voor interactie. Het doel van de methodologie die in dit proefschrift wordt geïntroduceerd is om verbeterde methodes en technieken voor bibliometrisch karteren te bieden.

Afgezien van een inleidend en een afsluitend hoofdstuk (hoofdstuk 1 en 9), bestaat dit proefschrift uit zeven hoofdstukken. Van deze zeven hoofdstukken hebben de eerste zes (hoofdstuk 2 t/m 7) een methodologisch karakter. Het zevende hoofdstuk (hoofdstuk 8) gaat over een toepassing. Hieronder worden de zeven hoofdstukken kort samengevat.

In hoofdstuk 2 wordt een nieuwe techniek voor het automatisch identificeren van termen geïntroduceerd. Deze techniek kan worden gebruikt om automatisch de termen te selecteren die in een termenkaart worden getoond. De techniek kijkt naar de verdeling van zelfstandignaamwoordgroepen over onderwerpen. Hoe meer de verdeling van een zelfstandignaamwoordgroep een afwijking heeft in de richting van één bepaald onderwerp, hoe waarschijnlijker het is dat deze zelfstandignaamwoordgroep een relevante term representeert. De belangrijkste conclusie van hoofdstuk 2 is dat de voorgestelde techniek voor veel doeleinden voldoende goed werkt, maar dat handmatige controle nodig blijft wanneer een hoge nauwkeurigheid van de termidentificatie vereist is.

Hoofdstuk 3 en 4 gaan over methodes voor het normaliseren van relatiesterktes van objecten. Deze methodes worden ook wel aangeduid als maten van gelijkenis. Hoofdstuk 3 gaat over zogeheten indirecte maten, terwijl hoofdstuk 4 over directe maten gaat. In beide hoofdstukken wordt een strikt wiskundige aanpak gehanteerd. Er worden eigenschappen geformuleerd die maten van gelijkenis redelijkerwijs zouden moeten hebben en er wordt gekeken welke maten deze eigenschappen inderdaad bezitten en welke niet. Hoofdstuk 3 levert een aantal suggesties op voor indirecte maten van gelijkenis met goede wiskundige eigenschappen. In hoofdstuk 4 wordt een grote verzameling van directe maten van gelijkenis beschouwd en wordt geconcludeerd dat er binnen deze verzameling in essentie slechts één maat is, de zogeheten associatiesterkte maat, die alle gewenste wiskundige eigenschappen bezit. Andere maten die veel vaker worden gebruikt, zoals de cosinus maat en de Jaccard maat, hebben niet alle gewenste eigenschappen.

In hoofdstuk 5 wordt de VOS karteringstechniek geïntroduceerd, waarbij de afkorting VOS staat voor ‘visualization of similarities’. De VOS karteringstechniek kan worden gezien als een alternatief voor de bekende meerdimensionale schaaltechniek. In hoofdstuk 5 wordt de wiskundige relatie tussen de twee technieken bestudeerd. Tevens wordt een empirische vergelijking uitgevoerd waarin beide technieken worden gebruikt om een aantal bibliometrische kaarten te construeren. Twee veelgebruikte benaderingen waarin meerdimensionale schaling wordt toegepast blijken last te hebben van artefacten. Een van de artefacten is de tendens om belangrijke objecten in het midden van een kaart te plaatsen en minder belangrijke objecten aan de rand. Een andere artefact is de tendens om objecten in een cirkelvormige structuur te plaatsen. De VOS karteringstechniek blijkt van deze problemen geen last te hebben. Op basis hiervan wordt geconcludeerd dat de VOS karteringstechniek over het algemeen betere bibliometrische kaarten oplevert dan de twee veelgebruikte meerdimensionale schaalbenaderingen.

In hoofdstuk 6 wordt de VOS clustertechniek geïntroduceerd. Deze techniek kan worden gebruikt om de objecten in een bibliometrische kaart te clusteren. De techniek kan dienen als een alternatief voor andere clustertechnieken, zoals de veelgebruikte hiërarchische technieken. In hoofdstuk 6 wordt aangetoond dat de VOS clustertechniek vanuit hetzelfde onderliggende wiskundige principe kan worden afgeleid als de VOS karteringstechniek. Hieruit volgt dat de combinatie van de twee VOS technieken tot een geünificeerde benadering voor karteren en clusteren leidt. Het voordeel van zo een geünificeerde benadering is dat het kaarten en clusters oplevert die onderling consistent zijn. In de literatuur worden karteringstechnieken en clustertechnieken vaak samen gebruikt, maar de technieken zijn gewoonlijk op verschillende principes gebaseerd, wat tot inconsistente resultaten kan leiden. In hoofdstuk 6 wordt ook aangetoond dat de VOS clustertechniek nauw verwant is aan clustertechnieken die gebaseerd zijn op zogeheten modulariteitsmaten. Dit type clustertechnieken is populair in the natuurkundige literatuur. Om de in hoofdstuk 6 voorgestelde geünificeerde benadering voor karteren en clusteren te testen wordt een bibliometrische kaart gemaakt van veelgeciteerde publicaties in de informatiewetenschappen.

Hoofdstuk 7 gaat over de VOSviewer software voor het weergeven en exploreren van bibliometrische kaarten. De functionaliteit van de software wordt besproken en er wordt nader ingegaan op de technische implementatie van specifieke onderdelen van de software. Tevens wordt een toepassing getoond waarin de software wordt gebruikt voor het construeren en weergeven van een op co-citaties gebaseerde kaart van 5000 grote wetenschappelijke tijdschriften.

Ten slotte is hoofdstuk 8 volledig gewijd aan een toepassing van bibliometrisch karteren. In deze toepassing worden bibliometrische kaarten geconstrueerd op basis van tijdschrift- en conferentiepublicaties in het vakgebied van de computationele intelligentie. Om de ontwikkeling van het vakgebied door de tijd heen te bekijken worden kaarten voor twee tijdsperiodes gemaakt. Op basis van de bibliometrische kaarten kunnen de belangrijkste problemen waar het vakgebied van de computationele intelligentie zich mee bezighoudt worden geïdentificeerd. Ook kan worden geanalyseerd hoe de drie voornaamste deelgebieden van dit vakgebied (evolutionair rekenen, fuzzy systemen en neurale netwerken) zich tot elkaar verhouden.

Bibliography

Ahlgren, P., Jarneving, B., & Rousseau, R. (2003). Requirements for a cocitation similarity measure, with special reference to Pearson's correlation coefficient. *Journal of the American Society for Information Science and Technology*, *54*(6), 550-560.

Anderberg, M. R. (1973). *Cluster analysis for applications*. Academic Press.

Åström, F. (2007). Changes in the LIS research front: Time-sliced cocitation analyses of LIS journal articles, 1990-2004. *Journal of the American Society for Information Science and Technology*, *58*(7), 947-957.

Baulieu, F. B. (1989). A classification of presence/absence based dissimilarity coefficients. *Journal of Classification*, *6*(1), 233-246.

Baulieu, F. B. (1997). Two variant axiom systems for presence/absence based dissimilarity coefficients. *Journal of Classification*, *14*(1), 159-170.

Belkin, M., & Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, *15*(6), 1373-1396.

Bensman, S. J. (2004). Pearson's r and author cocitation analysis: A commentary on the controversy. *Journal of the American Society for Information Science and Technology*, *55*(10), 935.

Bhattacharyya, A. (1943). On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society*, *35*, 99-109.

Blatt, E. M. (2009). Differentiating, describing, and visualizing scientific space: A novel approach to the analysis of published scientific abstracts. *Scientometrics*, *80*(2), 387-408.

Bollen, J., Van de Sompel, H., Hagberg, A., Bettencourt, L., Chute, R., Rodriguez, M. A., et al. (2009). Clickstream data yields high-resolution maps of science.

PLoS ONE, *4*(3), e4803.

Bookstein, A., & Swanson, D. R. (1974). Probabilistic models for automatic indexing. *Journal of the American Society for Information Science*, *25*(5), 312-318.

Borg, I., & Groenen, P. J. F. (2005). *Modern multidimensional scaling* (2nd ed.). Springer.

Börner, K. (2010). *Atlas of science: Visualizing what we know*. MIT Press.

Börner, K., Chen, C., & Boyack, K. W. (2003). Visualizing knowledge domains. *Annual Review of Information Science and Technology*, *37*(1), 179-255.

Börner, K., Palmer, F., Davis, J. M., Hardy, E., Uzzo, S. M., & Hook, B. J. (2009). Teaching children the structure of science. In *Proceedings of the SPIE conference on visualization and data analysis* (Vol. 7243, p. 724307). SPIE Press.

Bourigault, D. (1992). Surface grammatical analysis for the extraction of terminological noun phrases. In *Proceedings of the 14th conference on computational linguistics* (p. 977-981). Association for Computational Linguistics.

Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics*, *64*(3), 351-374.

Boyack, K. W., Wylie, B. N., & Davidson, G. S. (2002). Domain visualization using VxInsight for science and technology management. *Journal of the American Society for Information Science and Technology*, *53*(9), 764-774.

Braam, R. R., Moed, H. F., & Van Raan, A. F. J. (1991a). Mapping of science by combined co-citation and word analysis. II. Dynamical aspects. *Journal of the American Society for Information Science*, *42*(4), 252-266.

Braam, R. R., Moed, H. F., & Van Raan, A. F. J. (1991b). Mapping of science by combined co-citation and word analysis. I. Structural aspects. *Journal of the American Society for Information Science*, *42*(4), 233-251.

Buja, A., Logan, B. F., Reeds, J. A., & Shepp, L. A. (1994). Inequalities and positive-definite functions arising from a problem in multidimensional scaling. *Annals of Statistics*, *22*(1), 406-438.

Buter, R. K., & Noyons, E. C. M. (2001). Improving the functionality of interactive bibliometric science maps. *Scientometrics*, *51*(1), 55-68.

Cabré Castellví, M. T., Estopà Bagot, R., & Vivaldi Palatresi, J. (2001). Automatic term detection: A review of current systems. In D. Bourigault, C. Jacquemin, & M.-

C. L'Homme (Eds.), *Recent advances in computational terminology* (p. 53-87). John Benjamins.

Callon, M., Courtial, J.-P., & Laville, F. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry. *Scientometrics*, *22*(1), 155-205.

Callon, M., Courtial, J.-P., Turner, W. A., & Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. *Social Science Information*, *22*(2), 191-235.

Callon, M., Law, J., & Rip, A. (Eds.). (1986). *Mapping the dynamics of science and technology*. MacMillan Press.

Chen, C. (1999). Visualising semantic spaces and author co-citation networks in digital libraries. *Information Processing and Management*, *35*(3), 401-420.

Chen, C. (2003a). *Mapping scientific frontiers: The quest for knowledge visualization*. Springer.

Chen, C. (2003b). Visualizing scientific paradigms: An introduction. *Journal of the American Society for Information Science and Technology*, *54*(5), 392-393.

Chen, C. (2006a). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, *57*(3), 359-377.

Chen, C. (2006b). *Information visualization: Beyond the horizon* (2nd ed.). Springer.

Chen, C., Ibekwe-SanJuan, F., & Hou, J. (2010). The structure and dynamics of cocitation clusters: A multiple-perspective cocitation analysis. *Journal of the American Society for Information Science*, *61*(7), 1386-1409.

Chen, P., & Redner, S. (2010). Community structure of the physical review citation network. *Journal of Informetrics*, *4*(3), 278-290.

Chung, Y. M., & Lee, J. Y. (2001). A corpus-based approach to comparative evaluation of statistical term association measures. *Journal of the American Society for Information Science and Technology*, *52*(4), 283-296.

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*(1), 22-29.

Cox, M. A. A., & Cox, T. F. (2008). Multidimensional scaling. In C. Chen, W. Härdle, & A. Unwin (Eds.), *Handbook of data visualization* (p. 315-347). Springer.

Cox, T. F., & Cox, M. A. A. (2001). *Multidimensional scaling* (2nd ed.). Chapman & Hall/CRC.

Dagan, I., & Church, K. W. (1994). TERMIGHT: Identifying and translating technical terminology. In *Proceedings of the 4th conference on applied natural language processing* (p. 34-40). Association for Computational Lingustics.

Daille, B., Gaussier, É., & Langé, J.-M. (1994). Towards automatic extraction of monolingual and bilingual terminology. In *Proceedings of the 15th conference on computational linguistics* (p. 515-521). Association for Computational Linguistics.

Damerau, F. J. (1993). Generating and evaluating domain-oriented multi-word terms from texts. *Information Processing and Management*, *29*(4), 433-447.

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.

Davidson, G. S., Wylie, B. N., & Boyack, K. W. (2001). Cluster stability and the use of noise in interpretation of clustering. In *Proceedings of the IEEE symposium on information visualization 2001* (p. 23-30). IEEE Computer Society.

de Moya-Anegón, F., Herrero-Solana, V., & Jiménez-Contreras, E. (2006). A connectionist and multivariate approach to science maps: The SOM, clustering and MDS applied to library and information science research. *Journal of Information Science*, *32*(1), 63-77.

de Moya-Anegón, F., Vargas-Quesada, B., Chinchilla-Rodríguez, Z., Corera-Álvarez, E., Munoz-Fernández, F. J., & Herrero-Solana, V. (2007). Visualizing the marrow of science. *Journal of the American Society for Information Science and Technology*, *58*(14), 2167-2179.

de Solla Price, D. (1981). The analysis of scientometric matrices for policy implications. *Scientometrics*, *3*(1), 47-53.

De Bellis, N. (2009). *Bibliometrics and citation analysis: From the science citation index to cybermetrics*. Scarecrow Press.

De Leeuw, J., & Meulman, J. (1986). A special jackknife for multidimensional scaling. *Journal of Classification*, *3*, 97-112.

De Leeuw, J., & Stoop, I. (1984). Upper bounds for kruskal's stress. *Psychometrika*, *49*(3), 391-402.

De Nooy, W., Mrvar, A., & Batagelj, V. (2005). *Exploratory social network analysis with Pajek*. Cambridge University Press.

Ding, Y., Chowdhury, G. G., & Foo, S. (2001). Bibliometric cartography of information retrieval research by using co-word analysis. *Information Processing and Management*, *37*(6), 817-842.

Drasgow, F., & Jones, L. E. (1979). Multidimensional scaling of derived dissimilarities. *Multivariate Behavioral Research*, *14*(2), 227-244.

Drouin, P. (2003). Term extraction using non-technical corpora as a point of leverage. *Terminology*, *9*(1), 99-115.

Duncan, G. T., & Layard, M. W. J. (1973). A Monte-Carlo study of asymptotically robust tests for correlation coefficients. *Biometrika*, *60*(3), 551-558.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, *19*(1), 61-74.

Edgington, E. S. (1995). *Randomization tests* (3rd ed.). Marcel Dekker.

Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, *1*(1), 54-75.

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. Chapman & Hall.

Egghe, L. (2009). New relations between similarity measures for vectors based on vector norms. *Journal of the American Society for Information Science and Technology*, *60*(2), 232-239.

Egghe, L., & Michel, C. (2002). Strong similarity measures for ordered sets of documents in information retrieval. *Information Processing and Management*, *38*(6), 823-848.

Egghe, L., & Michel, C. (2003). Construction of weak and strong similarity measures for ordered sets of documents using fuzzy set techniques. *Information Processing and Management*, *39*(5), 771-807.

Egghe, L., & Rousseau, R. (2006). Classical retrieval and overlap measures satisfy the requirements for rankings based on a Lorenz curve. *Information Processing and Management*, *42*(1), 106-120.

Eilers, P. H. C., & Goeman, J. J. (2004). Enhancing scatterplots with smoothed densities. *Bioinformatics*, *20*(5), 623-628.

Engelbrecht, A. P. (2003). *Computational intelligence: An introduction*. John Wiley & Sons.

Eom, S. (2008). All author cocitation analysis and first author cocitation analysis: A comparative empirical investigation. *Journal of Informetrics*, *2*(1), 53-64.

Eto, H. (2000). Authorship and citation patterns in operational research journals in relation to competition and reform. *Scientometrics*, *47*(1), 25-42.

Eto, H. (2002). Authorship and citation patterns in Management Science in comparison with operational research. *Scientometrics*, *53*(3), 337-349.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, *486*(3–5), 75-174.

Fortunato, S., & Barthélemy, M. (2007). Resolution limit in community detection. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(1), 36-41.

Franklin, J. J., & Johnston, R. (1988). Co-citation bibliometric modeling as a tool for S&T and R&D management: Issues, applications, and developments. In A. F. J. Van Raan (Ed.), *Handbook of quantitative science and technology research* (p. 325-389). Elsevier Science Publishers.

Frantzi, K., Ananiadou, S., & Mima, H. (2000). Automatic recognition of multi-word terms: The C-value/NC-value method. *International Journal of Digital Libraries*, *3*(2), 117-132.

Fruchterman, T. M. J., & Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and Experience*, *21*(11), 1129-1164.

Garfield, E. (2009). From the science of science to Scientometrics visualizing the history of science with HistCite software. *Journal of Informetrics*, *3*(3), 173-179.

Garfield, E., Pudovkin, A. I., & Istomin, V. S. (2003). Why do we need algorithmic historiography? *Journal of the American Society for Information Science and Technology*, *54*(5), 400-412.

Glänzel, W. (2001). National characteristics in international scientific co-authorship relations. *Scientometrics*, *51*(1), 69-115.

Glänzel, W., Schubert, A., & Czerwon, H.-J. (1999). A bibliometric analysis of international scientific cooperation of the European Union (1985–1995). *Scientometrics*, *45*(2), 185-202.

Gmür, M. (2003). Co-citation analysis and the search for invisible colleges: A methodological evaluation. *Scientometrics*, *57*(1), 27-57.

Gower, J. C. (1985). Measures of similarity, dissimilarity, and distance. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences* (Vol. 5, p. 397-405). Wiley.

Gower, J. C., & Legendre, P. (1986). Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification*, *3*(1), 5-48.

Griffith, B. C., Small, H., Stonehill, J. A., & Dey, S. (1974). The structure of scientific literatures II: Toward a macro- and microstructure for science. *Science Studies*, *4*(4), 339-365.

Guilford, J. P. (1973). *Fundamental statistics in psychology and education* (5th ed.). McGraw-Hill.

Hamers, L., Hemeryck, Y., Herweyers, G., Janssen, M., Keters, H., Rousseau, R., et al. (1989). Similarity measures in scientometric research: The Jaccard index versus Salton's cosine formula. *Information Processing and Management*, *25*(3), 315-318.

Hardy, G. H., Littlewood, J. E., & Pólya, G. (1952). *Inequalities* (2nd ed.). Cambridge University Press.

Harter, S. P. (1975). A probabilistic approach to automatic keyword indexing. *Journal of the American Society for Information Science*, *26*(4), 197-206.

Healey, P., Rothman, H., & Hoch, P. K. (1986). An experiment in science mapping for research planning. *Research Policy*, *15*(5), 233-251.

Heersmink, R., Van den Hoven, J., Van Eck, N. J., & Van den Berg, J. (2011). Bibliometric mapping of computer and information ethics. *Ethics and Information Technology*, *13*(3), 241-249.

Heimeriks, G., Hörlesberger, M., & Van den Besselaar, P. (2003). Mapping communication and collaboration in heterogeneous research networks. *Scientometrics*, *58*(2), 391-413.

Heimo, T., Kumpula, J. M., Kaski, K., & Saramäki, J. (2008). Detecting modules in dense weighted networks with the Potts method. *Journal of Statistical Mechanics*, *8*, P08007.

Heiser, W. J., & Meulman, J. (1983a). Analyzing rectangular tables by joint and constrained multidimensional scaling. *Journal of Econometrics*, *22*, 139-167.

Heiser, W. J., & Meulman, J. (1983b). Constrained multidimensional scaling, including confirmation. *Applied Psychological Measurement*, *7*, 381-404.

Hinze, S. (1994). Bibliographical cartography of an emerging interdisciplinary discipline: The case of bioelectronics. *Scientometrics*, *29*(3), 353-376.

Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proceedings of the 15th conference on uncertainty in artificial intelligence* (p. 289-296). Morgan Kaufmann.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, *42*(1–2), 177-196.

Hood, W. W., & Wilson, C. S. (2001). The literature of bibliometrics, scientometrics, and informetrics. *Scientometrics*, *52*(2), 291-314.

Hubálek, Z. (1982). Coefficients of association and similarity, based on binary (presence-absence) data: An evaluation. *Biological Reviews*, *57*(4), 669-689.

Huisman, M., & Van Duijn, M. A. J. (2005). Software for social network analysis. In P. J. Carrington, J. Scott, & S. Wasserman (Eds.), *Models and methods in social network analysis* (p. 270-316). Cambridge University Press.

Jacquemin, C. (2001). *Spotting and discovering terms through natural language processing*. MIT Press.

Janson, S., & Vegelius, J. (1981). Measures of ecological association. *Oecologia*, *49*(3), 371-376.

Janssens, F., Glänzel, W., & De Moor, B. (2008). A hybrid mapping of information science. *Scientometrics*, *75*(3), 607-631.

Janssens, F., Leta, J., Glänzel, W., & De Moor, B. (2006). Towards mapping library and information science. *Information Processing and Management*, *42*(6), 1614-1642.

Jarneving, B. (2007). Bibliographic coupling and its application to research-front and other core documents. *Journal of Informetrics*, *1*(4), 287-307.

Jarneving, B. (2008). A variation of the calculation of the first author cocitation strength in author cocitation analysis. *Scientometrics*, *77*(3), 485-504.

Jones, W. P., & Furnas, G. W. (1987). Pictures of relevance: A geometric analysis of similarity measures. *Journal of the American Society for Information Science*, *38*(6), 420-442.

Justeson, J. S., & Katz, S. M. (1995). Technical terminology: Some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, *1*(1), 9-27.

Kageura, K., & Umino, B. (1996). Methods of automatic term recognition: A review. *Terminology*, *3*(2), 259-289.

Kamada, T., & Kawai, S. (1989). An algorithm for drawing general undirected graphs. *Information Processing Letters*, *31*(1), 7-15.

Kao, C. (2009). The authorship and country spread of Operation Research journals. *Scientometrics*, *78*(3), 397-407.

Kessler, M. M. (1963a). Bibliographic coupling between scientific papers. *American Documentation*, *14*(1), 10-25.

Kessler, M. M. (1963b). Bibliographic coupling extended in time: Ten case histories. *Information Storage and Retrieval*, *1*(4), 169-187.

Kim, W., & Wilbur, W. J. (2001). Corpus-based statistical screening for content-bearing terms. *Journal of the American Society for Information Science and Technology*, *52*(3), 247-259.

Klavans, R., & Boyack, K. W. (2006a). Identifying a better measure of relatedness for mapping science. *Journal of the American Society for Information Science and Technology*, *57*(2), 251-263.

Klavans, R., & Boyack, K. W. (2006b). Quantitative evaluation of large maps of science. *Scientometrics*, *68*(3), 475-499.

Klavans, R., & Boyack, K. W. (2009). Toward a consensus map of science. *Journal of the American Society for Information Science and Technology*, *60*(3), 455-476.

Konar, A. (2005). *Computational intelligence: Principles, techniques and applications*. Springer.

Kopcsa, A., & Schiebel, E. (1998). Science and technology mapping: A new iteration model for representing multidimensional relationships. *Journal of the American Society for Information Science*, *49*(1), 7-17.

Kostoff, R. N., & Block, J. A. (2005). Factor matrix text filtering and clustering. *Journal of the American Society for Information Science and Technology*, *56*(9), 946-968.

Kostoff, R. N., del Río, J. A., Humenik, J. A., García, E. O., & Ramírez, A. M. (2001).

Citation mining: Integrating text mining and bibliometrics for research user profiling. *Journal of the American Society for Information Science and Technology*, *52*(13), 1148-1156.

Kostoff, R. N., Eberhart, H. J., & Toothman, D. R. (1999). Hypersonic and supersonic flow roadmaps using bibliometrics and database tomography. *Journal of the American Society for Information Science*, *50*(5), 427-447.

Kowalski, C. J. (1972). On the effects of non-normality on the distribution of the sample product-moment correlation coefficient. *Applied Statistics*, *21*(1), 1-12.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, *22*(1), 79-86.

Kumpula, J. M., Saramäki, J., Kaski, K., & Kertész, J. (2007). Limited resolution in complex network community detection with Potts model approach. *The European Physical Journal B*, *56*(1), 41-45.

Lambiotte, R., & Panzarasa, P. (2009). Communities, knowledge creation, and information diffusion. *Journal of Informetrics*, *3*(3), 180-190.

Law, J., & Whittaker, J. (1992). Mapping acidification research: A test of the co-word method. *Scientometrics*, *23*(3), 417-461.

Leclerc, M., & Gagné, J. (1994). International scientific cooperation: The continentalization of science. *Scientometrics*, *31*(3), 261-292.

Leicht, E. A., Holme, P., & Newman, M. E. J. (2006). Vertex similarity in networks. *Physical Review E*, *73*(2), 026120.

Leopold, E., May, M., & Paaß, G. (2004). Data mining and text mining for science & technology research. In H. F. Moed, W. Glänzel, & U. Schmoch (Eds.), *Handbook of quantitative science and technology research* (p. 187-213). Kluwer Academic Publishers.

Leydesdorff, L. (1989). Words and co-words as indicators of intellectual organization. *Research Policy*, *18*(4), 209-223.

Leydesdorff, L. (2004). Clusters and maps of science journals based on bi-connected graphs in Journal Citation Reports. *Journal of Documentation*, *60*(4), 371-427.

Leydesdorff, L. (2005). Similarity measures, author cocitation analysis, and information theory. *Journal of the American Society for Information Science and Technology*, *56*(7), 769-772.

Leydesdorff, L. (2007). Should co-occurrence data be normalized? A rejoinder. *Journal of the American Society for Information Science and Technology*, *58*(14), 2411-2413.

Leydesdorff, L. (2008). On the normalization and visualization of author co-citation data: Salton's cosine versus the Jaccard index. *Journal of the American Society for Information Science and Technology*, *59*(1), 77-85.

Leydesdorff, L., Hammarfelt, B., & Akdag Salah, A. A. (in press). The structure of the Arts & Humanities Citation Index: A mapping on the basis of aggregated citations among 1,157 journals. *Journal of the American Society for Information Science and Technology*.

Leydesdorff, L., & Rafols, I. (2009). A global map of science based on the ISI subject categories. *Journal of the American Society for Information Science and Technology*, *60*(2), 348-362.

Leydesdorff, L., & Vaughan, L. (2006). Co-occurrence matrices and their applications in information science: Extending ACA to the Web environment. *Journal of the American Society for Information Science and Technology*, *57*(12), 1616-1628.

Leydesdorff, L., & Zaal, R. (1988). Co-words and citations relations between document sets and environments. In L. Egghe & R. Rousseau (Eds.), *Informetrics 87/88* (p. 105-119). Elsevier.

Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, *37*(1), 145-151.

Liu, Z. (2005). Visualizing the intellectual structure in urban studies: A journal co-citation analysis (1992-2002). *Scientometrics*, *62*(3), 385-402.

Lu, K., & Wolfram, D. (2010). Geographic characteristics of the growth of informetrics literature 1987-2008. *Journal of Informetrics*, *4*(4), 591-601.

Luukkonen, T., Persson, O., & Sivertsen, G. (1992). Understanding patterns of international scientific collaboration. *Science, Technology, and Human Values*, *17*(1), 101-126.

Luukkonen, T., Tijssen, R. J. W., Persson, O., & Sivertsen, G. (1993). The measurement of international scientific collaboration. *Scientometrics*, *28*(1), 15-36.

MacKay, D. J. C. (2003). *Information theory, inference, and learning algorithms*. Cambridge University Press.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.

Martín-Merino, M., & Muñoz, A. (2004). A new MDS algorithm for textual data analysis. In N. R. Pal, N. Kasabov, R. K. Mudi, S. Pal, & S. K. Parui (Eds.), *Proceedings of the 11th international conference on neural information processing* (Vol. 3316, p. 860-867). Springer.

Matsuo, Y., & Ishizuka, M. (2004). Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools*, *13*(1), 157-169.

Maynard, D., & Ananiadou, S. (2000). Identifying terms by their family and friends. In *Proceedings of the 18th conference on computational linguistics* (p. 530-536). Association for Computational Linguistics.

McCain, K. W. (1990). Mapping authors in intellectual space: A technical overview. *Journal of the American Society for Information Science*, *41*(6), 433-443.

McCain, K. W. (1991). Mapping economics through the journal literature: An experiment in journal cocitation analysis. *Journal of the American Society for Information Science*, *42*(4), 290-296.

McCain, K. W. (1995). The structure of biotechnology R & D. *Scientometrics*, *32*(2), 153-175.

McCain, K. W., Verner, J. M., Hislop, G. W., Evanco, W., & Cole, V. (2005). The use of bibliometric and knowledge elicitation techniques to map a knowledge domain: Software engineering in the 1990s. *Scientometrics*, *65*(1), 131-144.

Miguel, S., de Moya-Anegón, F., & Herrero-Solana, V. (2008). A new approach to institutional domain analysis: Multilevel research fronts structure. *Scientometrics*, *74*(3), 331-344.

Morillo, F., Bordons, M., & Gómez, I. (2003). Interdisciplinarity in science: A tentative typology of disciplines and research areas. *Journal of the American Society for Information Science and Technology*, *54*(13), 1237-1249.

Morris, S. A., & Van der Veer Martens, B. (2008). Mapping research specialties. *Annual Review of Information Science and Technology*, *42*(1), 213-295.

Newman, M. E. J. (2004a). Analysis of weighted networks. *Physical Review E*, *70*(5), 056131.

Newman, M. E. J. (2004b). Fast algorithm for detecting community structure in networks. *Physical Review E*, *69*(6), 066133.

Newman, M. E. J. (2010). *Networks: An introduction*. Oxford University Press.

Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, *69*(2), 026113.

Noack, A. (2007). Energy models for graph clustering. *Journal of Graph Algorithms and Applications*, *11*(2), 453-480.

Noack, A. (2009). Modularity clustering is force-directed layout. *Physical Review E*, *79*(2), 026102.

Noll, M., Fröhlich, D., & Schiebel, E. (2002). Knowledge maps of knowledge management tools - Information visualization with BibTechMon. In D. Karagiannis & U. Reimer (Eds.), *Proceedings of the 4th international conference on practical aspects of knowledge management* (Vol. 2569, p. 14-27). Springer.

Novak, J. D., & Gowin, D. B. (1984). *Learning how to learn*. Cambridge University Press.

Noyons, E. C. M. (1999). *Bibliometric mapping as a science policy and research management tool*. Unpublished doctoral dissertation, Leiden University.

Noyons, E. C. M. (2001). Bibliometric mapping of science in a policy context. *Scientometrics*, *50*(1), 83-98.

Noyons, E. C. M. (2004). Science maps within a science policy context. In H. F. Moed, W. Glänzel, & U. Schmoch (Eds.), *Handbook of quantitative science and technology research* (p. 237-255). Kluwer Academic Publishers.

Noyons, E. C. M., & Calero-Medina, C. (2009). Applying bibliometric mapping in a high level science policy context. *Scientometrics*, *79*(2), 261-275.

Noyons, E. C. M., Moed, H. F., & Van Raan, A. F. J. (1999). Integrating research performance analysis and science mapping. *Scientometrics*, *46*(3), 591-604.

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*(1), 68-81.

Palmer, C. L. (1999). Structures and strategies of interdisciplinary science. *Journal of the American Society for Information Science*, *50*(3), 242-253.

Pazienza, M. T., Pennacchiotti, M., & Zanzotto, F. M. (2005). Terminology extraction: An analysis of linguistic and statistical approaches. In S. Sirmakessis (Ed.), *Knowledge mining: Proceedings of the NEMIS 2004 final conference* (p. 255-279). Springer.

Pearson, E. S. (1931). The test of significance for the correlation coefficient. *Journal of the American Statistical Association*, *26*(174), 128-134.

Persson, O. (1994). The intellectual base and research fronts of JASIS 1986-1990. *Journal of the American Society for Information Science*, *45*(1), 31-38.

Peters, H. P. F., Braam, R. R., & Van Raan, A. F. J. (1995). Cognitive resemblance and citation relations in chemical engineering publications. *Journal of the American Society for Information Science*, *46*(1), 9-21.

Peters, H. P. F., & Van Raan, A. F. J. (1993a). Co-word-based science maps of chemical engineering. Part II: Representations by combined clustering and multidimensional scaling. *Research Policy*, *22*(1), 23-45.

Peters, H. P. F., & Van Raan, A. F. J. (1993b). Co-word-based science maps of chemical engineering. Part I: Representations by direct multidimensional scaling. *Research Policy*, *22*(1), 23-45.

Qin, J. (2000). Semantic similarities between a keyword database and a controlled vocabulary database: An investigation in the antibiotic resistance literature. *Journal of the American Society for Information Science*, *51*(3), 166-180.

Reichardt, J., & Bornholdt, S. (2006). Statistical mechanics of community detection. *Physical Review E*, *74*(1), 016110.

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

Rorvig, M. (1999). Images of similarity: A visual exploration of optimal similarity metrics and scaling properties of TREC topic-document sets. *Journal of the American Society for Information Science*, *50*(8), 639-651.

Rosenberg, S., & Jones, R. (1972). A method for investigating and representing a person's implicit theory of personality: Theodore Dreiser's view of people. *Journal of Personality and Social Psychology*, *22*(3), 372-386.

Rosenberg, S., Nelson, C., & Vivekananthan, P. S. (1968). A multidimensional approach to the structure of personality impressions. *Journal of Personality and Social Psychology*, *9*(4), 283-294.

Salton, G. (1963). Associative document retrieval techniques using bibliographic information. *Journal of the ACM*, *10*(4), 440-457.

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval.* McGraw-Hill.

Sammon, J. W. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, *C-18*(5), 401-409.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the international conference on new methods in language processing* (p. 44-49). University of Manchester Institute of Science and Technology.

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL SIGDAT workshop* (p. 47-50). University College.

Schneider, J. W. (2006). Concept symbols revisited: Naming clusters by parsing and filtering of noun phrases from citation contexts of concept symbols. *Scientometrics*, *68*(3), 573-593.

Schneider, J. W., & Borlund, P. (2007a). Matrix comparison, part 1: Motivation and important issues for measuring the resemblance between proximity measures or ordination results. *Journal of the American Society for Information Science and Technology*, *58*(11), 1586-1595.

Schneider, J. W., & Borlund, P. (2007b). Matrix comparison, part 2: Measuring the resemblance between proximity measures or ordination results by use of the Mantel and Procrustes statistics. *Journal of the American Society for Information Science and Technology*, *58*(11), 1596-1609.

Schneider, J. W., Larsen, B., & Ingwersen, P. (2009). A comparative study of first and all-author co-citation counting, and two different matrix generation approaches applied for author co-citation analyses. *Scientometrics*, *80*(1), 105-132.

Schubert, A., & Braun, T. (1990). International collaboration in the sciences, 1981–1985. *Scientometrics*, *19*(1–2), 3-10.

Schubert, A., & Soós, S. (2010). Mapping of science journals based on h-similarity. *Scientometrics*, *83*(2), 589-600.

Schvaneveldt, R. W. (Ed.). (1990). *Pathfinder associative networks: Studies in knowledge organization.* Ablex Publishing Corporation.

Schvaneveldt, R. W., Dearholt, D. W., & Durso, F. T. (1988). Graph theoretic foundations of pathfinder networks. *Computers and Mathematics with Applications*, *15*(4), 337-345.

Scott, D. W. (1992). *Multivariate density estimation: Theory, practice, and visualization*. John Wiley & Sons.

Shiffrin, R. M., & Börner, K. (2004). Mapping knowledge domains. *Proceedings of the National Academy of Sciences of the United States of America*, *101 (suppl. 1)*, 5183-5185.

Simmen, M. W. (1996). Multidimensional scaling of binary dissimilarities: Direct and derived approaches. *Multivariate Behavioral Research*, *31*(1), 47-67.

Skupin, A. (2004). The world of geography: Visualizing a knowledge domain with cartographic means. *Proceedings of the National Academy of Sciences of the United States of America*, *101 (suppl. 1)*, 5274-5278.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, *24*(4), 265-269.

Small, H. (1981). The relationship of information science to the social sciences: A co-citation analysis. *Information Processing and Management*, *17*(1), 39-50.

Small, H. (1994). A SCI-Map case study: Building a map of AIDS research. *Scientometrics*, *30*(1), 229-241.

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

Small, H. (1999). Visualizing science by citation mapping. *Journal of the American Society for Information Science*, *50*(9), 799-813.

Small, H. (2003). Paradigms, citations, and maps of science: A personal history. *Journal of the American Society for Information Science and Technology*, *54*(5), 394-399.

Small, H., & Garfield, E. (1985). The geography of science: Disciplinary and national mappings. *Journal of Information Science*, *11*(4), 147-159.

Small, H., & Greenlee, E. (1980). Citation context analysis of a co-citation cluster: Recombinant-DNA. *Scientometrics*, *2*(4), 277-301.

Small, H., & Griffith, B. C. (1974). The structure of scientific literatures I: Identifying and graphing specialties. *Science Studies*, *4*(1), 17-40.

Small, H., & Sweeney, E. (1985). Clustering the science citation index using co-citations. I. A comparison of methods. *Scientometrics*, *7*(3–6), 391-409.

Small, H., Sweeney, E., & Greenlee, E. (1985). Clustering the science citation index using co-citations. II. Mapping science. *Scientometrics*, *8*(5–6), 321-340.

Snedecor, G. W., & Cochran, W. G. (1989). *Statistical methods* (8th ed.). Iowa State University Press.

Sokal, R. R., & Sneath, P. H. A. (1963). *Principles of numerical taxonomy*. Freeman.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677-680.

Stout, W. F., Marden, J., & Travers, K. J. (2000). *Statistics: Making sense of data* (3rd ed.). Möbius Communications.

Su, H.-N., & Lee, P.-C. (2010). Mapping knowledge structure by keyword co-occurrence: A first look at journal papers in Technology Foresight. *Scientometrics*, *85*(1), 65-79.

Takeda, Y., & Kajikawa, Y. (2009). Optics: A bibliometric approach to detect emerging research domains and intellectual bases. *Scientometrics*, *78*(3), 543-558.

Tijssen, R. J. W. (1992). A quantitative assessment of interdisciplinary structures in science and technology: Co-classification analysis of energy research. *Research Policy*, *21*(1), 27-44.

Tijssen, R. J. W. (1993). A scientometric cognitive study of neural network research: Expert mental maps versus bibliometric maps. *Scientometrics*, *28*(1), 111-136.

Tijssen, R. J. W. (2010). Discarding the 'basic science/applied science' dichotomy: A knowledge utilization triangle classification system of research journals. *Journal of the American Society for Information Science and Technology*, *61*(9), 1842-1852.

Tijssen, R. J. W., & Van Raan, A. F. J. (1989). Mapping co-word structures: A comparison of multidimensional scaling and LEXIMAPPE. *Scientometrics*, *15*(3–4), 283-295.

Van der Kloot, W. A., & Van Herk, H. (1991). Multidimensional scaling of sorting data: A comparison of three procedures. *Multivariate Behavioral Research*, *26*(4), 563-581.

Van Eck, N. J., Frasincar, F., & Van den Berg, J. (2006). Visualizing concept associations using concept density maps. In *Proceedings of the 10th international conference on information visualisation* (p. 270-275). IEEE Computer Society.

Van Eck, N. J., & Waltman, L. (2007a). Bibliometric mapping of the computational intelligence field. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *15*(5), 625-645.

Van Eck, N. J., & Waltman, L. (2007b). VOS: A new method for visualizing similarities between objects. In H.-J. Lenz & R. Decker (Eds.), *Advances in data analysis: Proceedings of the 30th annual conference of the german classification society* (p. 299-306). Springer.

Van Eck, N. J., & Waltman, L. (2008). Appropriate similarity measures for author cocitation analysis. *Journal of the American Society for Information Science and Technology*, *59*(10), 1653-1661.

Van Eck, N. J., & Waltman, L. (2009). How to normalize cooccurrence data? An analysis of some well-known similarity measures. *Journal of the American Society for Information Science and Technology*, *60*(8), 1635-1651.

Van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2), 523-538.

Van Eck, N. J., Waltman, L., Dekker, R., & Van den Berg, J. (2008). *An experimental comparison of bibliometric mapping techniques.* (Paper presented at the 10th International Conference on Science and Technology Indicators, Vienna)

Van Eck, N. J., Waltman, L., Dekker, R., & Van den Berg, J. (2010). A comparison of two techniques for bibliometric mapping: Multidimensional scaling and VOS. *Journal of the American Society for Information Science and Technology*, *61*(12), 2405-2416.

Van Eck, N. J., Waltman, L., Noyons, E. C. M., & Buter, R. K. (2010). Automatic term identification for bibliometric mapping. *Scientometrics*, *82*(3), 581-596.

Van Eck, N. J., Waltman, L., Van den Berg, J., & Kaymak, U. (2006a). Visualizing the computational intelligence field. *IEEE Computational Intelligence Magazine*, *1*(4), 6-10.

Van Eck, N. J., Waltman, L., Van den Berg, J., & Kaymak, U. (2006b). Visualizing the WCCI 2006 knowledge domain. In *Proceedings of the 2006 IEEE international* *conference on fuzzy systems* (p. 7862-7869). IEEE Press.

Van Liere, R., & De Leeuw, W. (2003). GraphSplatting: Visualizing graphs as continuous fields. *IEEE Transactions on Visualization and Computer Graphics*, *9*(2), 206-212.

Van Raan, A. F. J., & Tijssen, R. J. W. (1993). The neural net of neural network research: An exercise in bibliometric mapping. *Scientometrics*, *26*(1), 169-192.

Vargas-Quesada, B., & de Moya-Anegón, F. (2007). *Visualizing the structure of science*. Springer.

Vaughan, L. (2006). Visualizing linguistic and cultural differences using Web co-link data. *Journal of the American Society for Information Science and Technology*, *57*(9), 1178-1193.

Vaughan, L., & You, J. (2006). Comparing business competition positions based on Web co-link data: The global market vs. the Chinese market. *Scientometrics*, *68*(3), 611-628.

Waaijer, C. J. F., Van Bochove, C. A., & Van Eck, N. J. (2010). Journal editorials give indication of driving science issues. *Nature*, *463*(7278), 157.

Waaijer, C. J. F., Van Bochove, C. A., & Van Eck, N. J. (2011). On the map: Nature and Science editorials. *Scientometrics*, *86*(1), 99-112.

Wallace, M. L., Gingras, Y., & Duhon, R. (2009). A new approach for detecting scientific specialties from raw cocitation networks. *Journal of the American Society for Information Science and Technology*, *60*(2), 240-246.

Waltman, L., & Van Eck, N. J. (2007). Some comments on the question whether co-occurrence data should be normalized. *Journal of the American Society for Information Science and Technology*, *58*(11), 1701-1703.

Waltman, L., Van Eck, N. J., & Noyons, E. C. M. (2010). A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics*, *4*(4), 629-635.

Waltman, L., Yan, E., & Van Eck, N. J. (in press). A recursive field-normalized bibliometric performance indicator: An application to the field of library and information science. *Scientometrics*.

Wang, X., McCallum, A., & Wei, X. (2007). Topical N-grams: Phrase and topic discovery, with an application to information retrieval. In *Proceedings of the 7th IEEE international conference on data mining* (p. 697-702). IEEE Computer

Society.

Warrens, M. J. (2008). *Similarity coefficients for binary data.* Unpublished doctoral dissertation, Leiden University.

Weinberg, S. L., Carroll, J. D., & Cohen, H. S. (1984). Confidence regions for INDSCAL using the jackknife and the bootstrap techniques. *Psychometrika*, *49*(4), 475-491.

White, H. D. (2003a). Author cocitation analysis and Pearson's *r*. *Journal of the American Society for Information Science and Technology*, *54*(13), 1250-1259.

White, H. D. (2003b). Pathfinder networks and author cocitation analysis: A remapping of paradigmatic information scientists. *Journal of the American Society for Information Science and Technology*, *54*(5), 423-434.

White, H. D. (2004). Replies and a correction. *Journal of the American Society for Information Science and Technology*, *55*(9), 843-844.

White, H. D., & Griffith, B. C. (1981). Author cocitation: A literature measure of intellectual structure. *Journal of the American Society for Information Science*, *32*(3), 163-171.

White, H. D., & McCain, K. W. (1997). Visualization of literatures. *Annual Review of Information Science and Technology*, *32*, 99-168.

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*(4), 327-355.

Zegers, F. E., & Ten Berge, J. M. F. (1985). A family of association coefficients for metric scales. *Psychometrika*, *50*(1), 17-24.

Zhang, L., Liu, X., Janssens, F., Liang, L., & Glänzel, W. (2010). Subject clustering analysis based on isi category classification. *Journal of Informetrics*, *4*(2), 185-193.

Zhao, D. (2006). Towards all-author co-citation analysis. *Information Processing and Management*, *42*(6), 1578-1591.

Zhao, D., & Strotmann, A. (2008a). Comparing all-author and first-author co-citation analyses of information science. *Journal of Informetrics*, *2*(3), 229-239.

Zhao, D., & Strotmann, A. (2008b). Evolution of research activities and intellectual influences in information science 1996-2005: Introducing author bibliographic-coupling analysis. *Journal of the American Society for Information Science and Technology*, *59*(13), 2070-2086.

Zhao, D., & Strotmann, A. (2008c). Information science during the first decade of the Web: An enriched author cocitation analysis. *Journal of the American Society for Information Science and Technology*, *59*(6), 916-937.

Zitt, M., Bassecoulard, E., & Okubo, Y. (2000). Shadows of the past in international cooperation: Collaboration profiles of the top five producers of science. *Scientometrics*, *47*(3), 627-657.

Zuccala, A. (2006). Author cocitation analysis is to intellectual structure as Web colink analysis is to . . . ? *Journal of the American Society for Information Science and Technology*, *57*(11), 1487-1502.

Zuccala, A., & Van Eck, N. J. (2011). Poverty research in a development policy context. *Development Policy Review*, *29*(3), 311-330.

Author Index

Curriculum Vitae

Nees Jan van Eck (1982) obtained his master's degree in Informatics & Economics with honors from Erasmus University Rotterdam in 2005. In the same year, he started his PhD research at the Econometric Institute of the Erasmus School of Economics. Nees Jan's PhD research is in the field of bibliometrics and scientometrics and focuses on bibliometric mapping of science. Results of his research have been presented at various international conferences and have been published both in bibliometrics/scientometrics journals (*Journal of Informetrics*, *Journal of the American Society for Information Science and Technology*, and *Scientometrics*) and in computer science journals (*IEEE Computational Intelligence Magazine* and *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*). As part of his PhD research, Nees Jan has developed a computer program for bibliometric mapping called VOSviewer (http://www.vosviewer.com).

Since 2009, Nees Jan has been working as a researcher at the Centre for Science and Technology Studies of Leiden University. He has continued his work on bibliometric mapping of science and on the development of the VOSviewer software, but he has also broadened his interests to other topics in the field of bibliometrics and scientometrics. In 2010, a short research note on one of his bibliometric mapping projects was published in *Nature*.

For more information, please visit http://www.neesjanvaneck.nl.

ERASMUS RESEARCH INSTITUTE OF MANAGEMENT (ERIM)

ERIM PH.D. SERIES
RESEARCH IN MANAGEMENT

ERIM Electronic Series Portal: http://hdl.handle.net/1765/1

Acciaro, M., *Bundling Strategies in Global Supply Chains*, Promotor: Prof.dr. H.E. Haralambides, EPS-2010-197-LIS, ISBN: 978-90-5892-240-3, http://hdl.handle.net/1765/19742

Agatz, N.A.H., *Demand Management in E-Fulfillment*, Promotor: Prof.dr.ir. J.A.E.E. van Nunen, EPS-2009-163-LIS, ISBN: 978-90-5892-200-7, http://hdl.handle.net/1765/15425

Alexiev, A., *Exploratory Innovation: The Role of Organizational and Top Management Team Social Capital*, Promotors: Prof.dr. F.A.J. van den Bosch & Prof.dr. H.W. Volberda, EPS-2010-208-STR, ISBN: 978-90-5892-249-6, http://hdl.handle.net/1765/20632

Althuizen, N.A.P., *Analogical Reasoning as a Decision Support Principle for Weakly Structured Marketing Problems*, Promotor: Prof.dr.ir. B. Wierenga, EPS-2006-095-MKT, ISBN: 90-5892-129-8, http://hdl.handle.net/1765/8190

Alvarez, H.L., *Distributed Collaborative Learning Communities Enabled by Information Communication Technology*, Promotor: Prof.dr. K. Kumar, EPS-2006-080-LIS, ISBN: 90-5892-112-3, http://hdl.handle.net/1765/7830

Appelman, J.H., *Governance of Global Interorganizational Tourism Networks: Changing Forms of Co-ordination between the Travel Agency and Aviation Sector*, Promotors: Prof.dr. F.M. Go & Prof.dr. B. Nooteboom, EPS-2004-036-MKT, ISBN: 90-5892-060-7, http://hdl.handle.net/1765/1199

Asperen, E. van, *Essays on Port, Container, and Bulk Chemical Logistics Optimization*, Promotor: Prof.dr.ir. R. Dekker, EPS-2009-181-LIS, ISBN: 978-90-5892-222-9, http://hdl.handle.net/1765/17626

Assem, M.J. van den, *Deal or No Deal? Decision Making under Risk in a Large-Stake TV Game Show and Related Experiments*, Promotor: Prof.dr. J. Spronk, EPS-2008-138-F&A, ISBN: 978-90-5892-173-4, http://hdl.handle.net/1765/13566

Baquero, G, *On Hedge Fund Performance, Capital Flows and Investor Psychology*, Promotor: Prof.dr.

M.J.C.M. Verbeek, EPS-2006-094-F&A, ISBN: 90-5892-131-X, http://hdl.handle.net/1765/8192

Benning, T.M., *A Consumer Perspective on Flexibility in Health Care: Priority Access Pricing and Customized Care*, Promotor: Prof.dr.ir. B.G.C. Dellaert, EPS-2011-241-MKT, ISBN: 978-90-5892-280-9, http://hdl.handle.net/1765/23670

Berens, G., *Corporate Branding: The Development of Corporate Associations and their Influence on Stakeholder Reactions*, Promotor: Prof.dr. C.B.M. van Riel, EPS-2004-039-ORG, ISBN: 90-5892-065-8, http://hdl.handle.net/1765/1273

Berghe, D.A.F. van den, *Working Across Borders: Multinational Enterprises and the Internationalization of Employment*, Promotors: Prof.dr. R.J.M. van Tulder & Prof.dr. E.J.J. Schenk, EPS-2003-029-ORG, ISBN: 90-5892-05-34, http://hdl.handle.net/1765/1041

Berghman, L.A., *Strategic Innovation Capacity: A Mixed Method Study on Deliberate Strategic Learning Mechanisms*, Promotor: Prof.dr. P. Mattyssens, EPS-2006-087-MKT, ISBN: 90-5892-120-4, http://hdl.handle.net/1765/7991

Bezemer, P.J., *Diffusion of Corporate Governance Beliefs: Board Independence and the Emergence of a Shareholder Value Orientation in the Netherlands*, Promotors: Prof.dr.ing. F.A.J. van den Bosch & Prof.dr. H.W. Volberda, EPS-2009-192-STR, ISBN: 978-90-5892-232-8, http://hdl.handle.net/1765/18458

Bijman, W.J.J., *Essays on Agricultural Co-operatives: Governance Structure in Fruit and Vegetable Chains*, Promotor: Prof.dr. G.W.J. Hendrikse, EPS-2002-015-ORG, ISBN: 90-5892-024-0, http://hdl.handle.net/1765/867

Binken, J.L.G., *System Markets: Indirect Network Effects in Action, or Inaction?*, Promotor: Prof.dr. S. Stremersch, EPS-2010-213-MKT, ISBN: 978-90-5892-260-1, http://hdl.handle.net/1765/21186

Blitz, D.C., *Benchmarking Benchmarks*, Promotors: Prof.dr. A.G.Z. Kemna & Prof.dr. W.F.C. Verschoor, EPS-2011-225-F&A, ISBN: 978-90-5892-268-7, http://hdl.handle.net/1765/22624

Bispo, A., *Labour Market Segmentation: An investigation into the Dutch hospitality industry*, Promotors: Prof.dr. G.H.M. Evers & Prof.dr. A.R. Thurik, EPS-2007-108-ORG, ISBN: 90-5892-136-9, http://hdl.handle.net/1765/10283

Blindenbach-Driessen, F., *Innovation Management in Project-Based Firms*, Promotor: Prof.dr. S.L. van de Velde, EPS-2006-082-LIS, ISBN: 90-5892-110-7, http://hdl.handle.net/1765/7828

Boer, C.A., *Distributed Simulation in Industry*, Promotors: Prof.dr. A. de Bruin & Prof.dr.ir. A. Verbraeck, EPS-2005-065-LIS, ISBN: 90-5892-093-3, http://hdl.handle.net/1765/6925

Boer, N.I., *Knowledge Sharing within Organizations: A situated and Relational Perspective*, Promotor: Prof.dr. K. Kumar, EPS-2005-060-LIS, ISBN: 90-5892-086-0, http://hdl.handle.net/1765/6770

Boer-Sorbán, K., *Agent-Based Simulation of Financial Markets: A modular, Continuous-Time Approach*, Promotor: Prof.dr. A. de Bruin, EPS-2008-119-LIS, ISBN: 90-5892-155-0, http://hdl.handle.net/1765/10870

Boon, C.T., *HRM and Fit: Survival of the Fittest!?*, Promotors: Prof.dr. J. Paauwe & Prof.dr. D.N. den Hartog, EPS-2008-129-ORG, ISBN: 978-90-5892-162-8, http://hdl.handle.net/1765/12606

Borst, W.A.M., *Understanding Crowdsourcing: Effects of Motivation and Rewards on Participation and Performance in Voluntary Online Activities*, Promotors: Prof.dr.ir. J.C.M. van den Ende & Prof.dr.ir. H.W.G.M. van Heck, EPS-2010-221-LIS, ISBN: 978-90-5892-262-5, http://hdl.handle.net/1765/ 21914

Braun, E., *City Marketing: Towards an Integrated Approach*, Promotor: Prof.dr. L. van den Berg, EPS-2008-142-MKT, ISBN: 978-90-5892-180-2, http://hdl.handle.net/1765/13694

Brito, M.P. de, *Managing Reverse Logistics or Reversing Logistics Management?*, Promotors: Prof.dr.ir. R. Dekker & Prof.dr. M. B. M. de Koster, EPS-2004-035-LIS, ISBN: 90-5892-058-5, http://hdl.handle.net/1765/1132

Brohm, R., *Polycentric Order in Organizations: A Dialogue between Michael Polanyi and IT-Consultants on Knowledge, Morality, and Organization*, Promotors: Prof.dr. G. W. J. Hendrikse & Prof.dr. H. K. Letiche, EPS-2005-063-ORG, ISBN: 90-5892-095-X, http://hdl.handle.net/1765/6911

Brumme, W.-H., *Manufacturing Capability Switching in the High-Tech Electronics Technology Life Cycle*, Promotors: Prof.dr.ir. J.A.E.E. van Nunen & Prof.dr.ir. L.N. Van Wassenhove, EPS-2008-126-LIS, ISBN: 978-90-5892-150-5, http://hdl.handle.net/1765/12103

Budiono, D.P., *The Analysis of Mutual Fund Performance: Evidence from U.S. Equity Mutual Funds*, Promotor: Prof.dr. M.J.C.M. Verbeek, EPS-2010-185-F&A, ISBN: 978-90-5892-224-3, http://hdl.handle.net/1765/18126

Burgers, J.H., *Managing Corporate Venturing: Multilevel Studies on Project Autonomy, Integration, Knowledge Relatedness, and Phases in the New Business Development Process*, Promotors: Prof.dr.ir.

F.A.J. Van den Bosch & Prof.dr. H.W. Volberda, EPS-2008-136-STR, ISBN: 978-90-5892-174-1, http://hdl.handle.net/1765/13484

Camacho, N.M., *Health and Marketing; Essays on Physician and Patient Decision-making*, Promotor: Prof.dr. S. Stremersch, EPS-2011-237-MKT, ISBN: 978-90-5892-284-7, http://hdl.handle.net/1765/23604

Campbell, R.A.J., *Rethinking Risk in International Financial Markets*, Promotor: Prof.dr. C.G. Koedijk, EPS-2001-005-F&A, ISBN: 90-5892-008-9, http://hdl.handle.net/1765/306

Carvalho de Mesquita Ferreira, L., *Attention Mosaics: Studies of Organizational Attention*, Promotors: Prof.dr. P.M.A.R. Heugens & Prof.dr. J. van Oosterhout, EPS-2010-205-ORG, ISBN: 978-90-5892-242-7, http://hdl.handle.net/1765/19882

Chen, C.-M., *Evaluation and Design of Supply Chain Operations Using DEA*, Promotor: Prof.dr. J.A.E.E. van Nunen, EPS-2009-172-LIS, ISBN: 978-90-5892-209-0, http://hdl.handle.net/1765/16181

Chen, H., *Individual Mobile Communication Services and Tariffs*, Promotor: Prof.dr. L.F.J.M. Pau, EPS-2008-123-LIS, ISBN: 90-5892-158-1, http://hdl.handle.net/1765/11141

Chen, Y., *Labour Flexibility in China's Companies: An Empirical Study*, Promotors: Prof.dr. A. Buitendam & Prof.dr. B. Krug, EPS-2001-006-ORG, ISBN: 90-5892-012-7, http://hdl.handle.net/1765/307

Damen, F.J.A., *Taking the Lead: The Role of Affect in Leadership Effectiveness*, Promotor: Prof.dr. D.L. van Knippenberg, EPS-2007-107-ORG, http://hdl.handle.net/1765/10282

Daniševská, P., *Empirical Studies on Financial Intermediation and Corporate Policies*, Promotor: Prof.dr. C.G. Koedijk, EPS-2004-044-F&A, ISBN: 90-5892-070-4, http://hdl.handle.net/1765/1518

Defilippi Angeldonis, E.F., *Access Regulation for Naturally Monopolistic Port Terminals: Lessons from Regulated Network Industries*, Promotor: Prof.dr. H.E. Haralambides, EPS-2010-204-LIS, ISBN: 978-90-5892-245-8, http://hdl.handle.net/1765/19881

Delporte-Vermeiren, D.J.E., *Improving the Flexibility and Profitability of ICT-enabled Business Networks: An Assessment Method and Tool*, Promotors: Prof. mr. dr. P.H.M. Vervest & Prof.dr.ir. H.W.G.M. van Heck, EPS-2003-020-LIS, ISBN: 90-5892-040-2, http://hdl.handle.net/1765/359

Derwall, J.M.M., *The Economic Virtues of SRI and CSR*, Promotor: Prof.dr. C.G. Koedijk, EPS-2007-101-F&A, ISBN: 90-5892-132-8, http://hdl.handle.net/1765/8986

Desmet, P.T.M., *In Money we Trust? Trust Repair and the Psychology of Financial Compensations*, Promotors: Prof.dr. D. De Cremer & Prof.dr. E. van Dijk, EPS-2011-232-ORG, ISBN: 978-90-5892-274-8, http://hdl.handle.net/1765/23268

Diepen, M. van, *Dynamics and Competition in Charitable Giving*, Promotor: Prof.dr. Ph.H.B.F. Franses, EPS-2009-159-MKT, ISBN: 978-90-5892-188-8, http://hdl.handle.net/1765/14526

Dietvorst, R.C., *Neural Mechanisms Underlying Social Intelligence and Their Relationship with the Performance of Sales Managers*, Promotor: Prof.dr. W.J.M.I. Verbeke, EPS-2010-215-MKT, ISBN: 978-90-5892-257-1, http://hdl.handle.net/1765/21188

Dietz, H.M.S., *Managing (Sales)People towards Perfomance: HR Strategy, Leadership & Teamwork*, Promotor: Prof.dr. G.W.J. Hendrikse, EPS-2009-168-ORG, ISBN: 978-90-5892-210-6, http://hdl.handle.net/1765/16081

Dijksterhuis, M., *Organizational Dynamics of Cognition and Action in the Changing Dutch and US Banking Industries*, Promotors: Prof.dr.ir. F.A.J. van den Bosch & Prof.dr. H.W. Volberda, EPS-2003-026-STR, ISBN: 90-5892-048-8, http://hdl.handle.net/1765/1037

Eijk, A.R. van der, *Behind Networks: Knowledge Transfer, Favor Exchange and Performance*, Promotors: Prof.dr. S.L. van de Velde & Prof.dr.drs. W.A. Dolfsma, EPS-2009-161-LIS, ISBN: 978-90-5892-190-1, http://hdl.handle.net/1765/14613

Elstak, M.N., *Flipping the Identity Coin: The Comparative Effect of Perceived, Projected and Desired Organizational Identity on Organizational Identification and Desired Behavior*, Promotor: Prof.dr. C.B.M. van Riel, EPS-2008-117-ORG, ISBN: 90-5892-148-2, http://hdl.handle.net/1765/10723

Erken, H.P.G., *Productivity, R&D and Entrepreneurship*, Promotor: Prof.dr. A.R. Thurik, EPS-2008-147-ORG, ISBN: 978-90-5892-179-6, http://hdl.handle.net/1765/14004

Essen, M. van, *An Institution-Based View of Ownership*, Promotos: Prof.dr. J. van Oosterhout & Prof.dr. G.M.H. Mertens, EPS-2011-226-ORG, ISBN: 978-90-5892-269-4, http://hdl.handle.net/1765/22643

Fenema, P.C. van, *Coordination and Control of Globally Distributed Software Projects*, Promotor: Prof.dr. K. Kumar, EPS-2002-019-LIS, ISBN: 90-5892-030-5, http://hdl.handle.net/1765/360

Feng, L., *Motivation, Coordination and Cognition in Cooperatives*, Promotor: Prof.dr. G.W.J.

Hendrikse, EPS-2010-220-ORG, ISBN: 90-5892-261-8, http://hdl.handle.net/1765/21680

Fleischmann, M., *Quantitative Models for Reverse Logistics*, Promotors: Prof.dr.ir. J.A.E.E. van Nunen & Prof.dr.ir. R. Dekker, EPS-2000-002-LIS, ISBN: 35-4041-711-7, http://hdl.handle.net/1765/1044

Flier, B., *Strategic Renewal of European Financial Incumbents: Coevolution of Environmental Selection, Institutional Effects, and Managerial Intentionality*, Promotors: Prof.dr.ir. F.A.J. van den Bosch & Prof.dr. H.W. Volberda, EPS-2003-033-STR, ISBN: 90-5892-055-0, http://hdl.handle.net/1765/1071

Fok, D., *Advanced Econometric Marketing Models*, Promotor: Prof.dr. Ph.H.B.F. Franses, EPS-2003-027-MKT, ISBN: 90-5892-049-6, http://hdl.handle.net/1765/1035

Ganzaroli, A., *Creating Trust between Local and Global Systems*, Promotors: Prof.dr. K. Kumar & Prof.dr. R.M. Lee, EPS-2002-018-LIS, ISBN: 90-5892-031-3, http://hdl.handle.net/1765/361

Gertsen, H.F.M., *Riding a Tiger without Being Eaten: How Companies and Analysts Tame Financial Restatements and Influence Corporate Reputation*, Promotor: Prof.dr. C.B.M. van Riel, EPS-2009-171-ORG, ISBN: 90-5892-214-4, http://hdl.handle.net/1765/16098

Gilsing, V.A., *Exploration, Exploitation and Co-evolution in Innovation Networks*, Promotors: Prof.dr. B. Nooteboom & Prof.dr. J.P.M. Groenewegen, EPS-2003-032-ORG, ISBN: 90-5892-054-2, http://hdl.handle.net/1765/1040

Gijsbers, G.W., *Agricultural Innovation in Asia: Drivers, Paradigms and Performance*, Promotor: Prof.dr. R.J.M. van Tulder, EPS-2009-156-ORG, ISBN: 978-90-5892-191-8, http://hdl.handle.net/1765/14524

Gong, Y., *Stochastic Modelling and Analysis of Warehouse Operations*, Promotors: Prof.dr. M.B.M. de Koster & Prof.dr. S.L. van de Velde, EPS-2009-180-LIS, ISBN: 978-90-5892-219-9, http://hdl.handle.net/1765/16724

Govers, R., *Virtual Tourism Destination Image: Glocal Identities Constructed, Perceived and Experienced*, Promotors: Prof.dr. F.M. Go & Prof.dr. K. Kumar, EPS-2005-069-MKT, ISBN: 90-5892-107-7, http://hdl.handle.net/1765/6981

Graaf, G. de, *Tractable Morality: Customer Discourses of Bankers, Veterinarians and Charity Workers*, Promotors: Prof.dr. F. Leijnse & Prof.dr. T. van Willigenburg, EPS-2003-031-ORG, ISBN: 90-5892-051-8, http://hdl.handle.net/1765/1038

Greeven, M.J., *Innovation in an Uncertain Institutional Environment: Private Software Entrepreneurs in Hangzhou, China*, Promotor: Prof.dr. B. Krug, EPS-2009-164-ORG, ISBN: 978-90-5892-202-1, http://hdl.handle.net/1765/15426

Groot, E.A. de, *Essays on Economic Cycles*, Promotors: Prof.dr. Ph.H.B.F. Franses & Prof.dr. H.R. Commandeur, EPS-2006-091-MKT, ISBN: 90-5892-123-9, http://hdl.handle.net/1765/8216

Guenster, N.K., *Investment Strategies Based on Social Responsibility and Bubbles*, Promotor: Prof.dr. C.G. Koedijk, EPS-2008-175-F&A, ISBN: 978-90-5892-206-9, http://hdl.handle.net/1765/1

Gutkowska, A.B., *Essays on the Dynamic Portfolio Choice*, Promotor: Prof.dr. A.C.F. Vorst, EPS-2006-085-F&A, ISBN: 90-5892-118-2, http://hdl.handle.net/1765/7994

Hagemeijer, R.E., *The Unmasking of the Other*, Promotors: Prof.dr. S.J. Magala & Prof.dr. H.K. Letiche, EPS-2005-068-ORG, ISBN: 90-5892-097-6, http://hdl.handle.net/1765/6963

Hakimi, N.A, *Leader Empowering Behaviour: The Leader's Perspective: Understanding the Motivation behind Leader Empowering Behaviour*, Promotor: Prof.dr. D.L. van Knippenberg, EPS-2010-184-ORG, http://hdl.handle.net/1765/17701

Halderen, M.D. van, *Organizational Identity Expressiveness and Perception Management: Principles for Expressing the Organizational Identity in Order to Manage the Perceptions and Behavioral Reactions of External Stakeholders*, Promotor: Prof.dr. S.B.M. van Riel, EPS-2008-122-ORG, ISBN: 90-5892-153-6, http://hdl.handle.net/1765/10872

Hartigh, E. den, *Increasing Returns and Firm Performance: An Empirical Study*, Promotor: Prof.dr. H.R. Commandeur, EPS-2005-067-STR, ISBN: 90-5892-098-4, http://hdl.handle.net/1765/6939

Hensmans, M., *A Republican Settlement Theory of the Firm: Applied to Retail Banks in England and the Netherlands (1830-2007)*, Promotors: Prof.dr. A. Jolink& Prof.dr. S.J. Magala, EPS-2010-193-ORG, ISBN 90-5892-235-9, http://hdl.handle.net/1765/19494

Hermans. J.M., *ICT in Information Services; Use and Deployment of the Dutch Securities Trade, 1860-1970*, Promotor: Prof.dr. drs. F.H.A. Janszen, EPS-2004-046-ORG, ISBN 90-5892-072-0, http://hdl.handle.net/1765/1793

Hernandez Mireles, C., *Marketing Modeling for New Products*, Promotor: Prof.dr. P.H. Franses, EPS-2010-202-MKT, ISBN 90-5892-237-3, http://hdl.handle.net/1765/19878

Hessels, S.J.A., *International Entrepreneurship: Value Creation Across National Borders*, Promotor: Prof.dr. A.R. Thurik, EPS-2008-144-ORG, ISBN: 978-90-5892-181-9, http://hdl.handle.net/1765/13942

Heugens, P.P.M.A.R., *Strategic Issues Management: Implications for Corporate Performance*, Promotors: Prof.dr.ir. F.A.J. van den Bosch & Prof.dr. C.B.M. van Riel, EPS-2001-007-STR, ISBN: 90-5892-009-9, http://hdl.handle.net/1765/358

Heuvel, W. van den, *The Economic Lot-Sizing Problem: New Results and Extensions*, Promotor: Prof.dr. A.P.L. Wagelmans, EPS-2006-093-LIS, ISBN: 90-5892-124-7, http://hdl.handle.net/1765/1805

Hoedemaekers, C.M.W., *Performance, Pinned down: A Lacanian Analysis of Subjectivity at Work*, Promotors: Prof.dr. S. Magala & Prof.dr. D.H. den Hartog, EPS-2008-121-ORG, ISBN: 90-5892-156-7, http://hdl.handle.net/1765/10871

Hooghiemstra, R., *The Construction of Reality: Cultural Differences in Self-serving Behaviour in Accounting Narratives*, Promotors: Prof.dr. L.G. van der Tas RA & Prof.dr. A.Th.H. Pruyn, EPS-2003-025-F&A, ISBN: 90-5892-047-X, http://hdl.handle.net/1765/871

Hu, Y., *Essays on the Governance of Agricultural Products: Cooperatives and Contract Farming*, Promotors: Prof.dr. G.W.J. Hendrkse & Prof.dr. B. Krug, EPS-2007-113-ORG, ISBN: 90-5892-145-1, http://hdl.handle.net/1765/10535

Huang, X., *An Analysis of Occupational Pension Provision: From Evaluation to Redesigh*, Promotors: Prof.dr. M.J.C.M. Verbeek & Prof.dr. R.J. Mahieu, EPS-2010-196-F&A, ISBN: 90-5892-239-7, http://hdl.handle.net/1765/19674

Huij, J.J., *New Insights into Mutual Funds: Performance and Family Strategies*, Promotor: Prof.dr. M.C.J.M. Verbeek, EPS-2007-099-F&A, ISBN: 90-5892-134-4, http://hdl.handle.net/1765/9398

Huurman, C.I., *Dealing with Electricity Prices*, Promotor: Prof.dr. C.D. Koedijk, EPS-2007-098-F&A, ISBN: 90-5892-130-1, http://hdl.handle.net/1765/9399

Iastrebova, K, *Manager's Information Overload: The Impact of Coping Strategies on Decision-Making Performance*, Promotor: Prof.dr. H.G. van Dissel, EPS-2006-077-LIS, ISBN: 90-5892-111-5, http://hdl.handle.net/1765/7329

Iwaarden, J.D. van, *Changing Quality Controls: The Effects of Increasing Product Variety and Shortening Product Life Cycles,* Promotors: Prof.dr. B.G. Dale & Prof.dr. A.R.T. Williams, EPS-2006-084-ORG, ISBN: 90-5892-117-4, http://hdl.handle.net/1765/7992

Jalil, M.N., *Customer Information Driven After Sales Service Management: Lessons from Spare Parts Logistics*, Promotor: Prof.dr. L.G. Kroon, EPS-2011-222-LIS, ISBN: 978-90-5892-264-9, http://hdl.handle.net/1765/22156

Jansen, J.J.P., *Ambidextrous Organizations*, Promotors: Prof.dr.ir. F.A.J. Van den Bosch & Prof.dr. H.W. Volberda, EPS-2005-055-STR, ISBN: 90-5892-081-X, http://hdl.handle.net/1765/6774

Jaspers, F.P.H., *Organizing Systemic Innovation*, Promotor: Prof.dr.ir. J.C.M. van den Ende, EPS-2009-160-ORG, ISBN: 978-90-5892-197-), http://hdl.handle.net/1765/14974

Jennen, M.G.J., *Empirical Essays on Office Market Dynamics*, Promotors: Prof.dr. C.G. Koedijk & Prof.dr. D. Brounen, EPS-2008-140-F&A, ISBN: 978-90-5892-176-5, http://hdl.handle.net/1765/13692

Jiang, T., *Capital Structure Determinants and Governance Structure Variety in Franchising*, Promotors: Prof.dr. G. Hendrikse & Prof.dr. A. de Jong, EPS-2009-158-F&A, ISBN: 978-90-5892-199-4, http://hdl.handle.net/1765/14975

Jiao, T., *Essays in Financial Accounting*, Promotor: Prof.dr. G.M.H. Mertens, EPS-2009-176-F&A, ISBN: 978-90-5892-211-3, http://hdl.handle.net/1765/16097

Jong, C. de, *Dealing with Derivatives: Studies on the Role, Informational Content and Pricing of Financial Derivatives*, Promotor: Prof.dr. C.G. Koedijk, EPS-2003-023-F&A, ISBN: 90-5892-043-7, http://hdl.handle.net/1765/1043

Kaa, G. van, *Standard Battles for Complex Systems: Empirical Research on the Home Network*, Promotors: Prof.dr.ir. J. van den Ende & Prof.dr.ir. H.W.G.M. van Heck, EPS-2009-166-ORG, ISBN: 978-90-5892-205-2, http://hdl.handle.net/1765/16011

Kagie, M., *Advances in Online Shopping Interfaces: Product Catalog Maps and Recommender Systems,* Promotor: Prof.dr. P.J.F. Groenen, EPS-2010-195-MKT, ISBN: 978-90-5892-233-5, http://hdl.handle.net/1765/19532

Kappe, E.R., *The Effectiveness of Pharmaceutical Marketing*, Promotor: Prof.dr. S. Stremersch, EPS-2011-239-MKT, ISBN: 978-90-5892-283-0, http://hdl.handle.net/1765/23610

Karreman, B., *Financial Services and Emerging Markets*, Promotors: Prof.dr. G.A. van der Knaap & Prof.dr. H.P.G. Pennings, EPS-2011-223-ORG, ISBN: 978-90-5892-266-3, http://hdl.handle.net/1765/22280

Keizer, A.B., *The Changing Logic of Japanese Employment Practices: A Firm-Level Analysis of Four Industries*, Promotors: Prof.dr. J.A. Stam & Prof.dr. J.P.M. Groenewegen, EPS-2005-057-ORG, ISBN: 90-5892-087-9, http://hdl.handle.net/1765/6667

Kijkuit, R.C., *Social Networks in the Front End: The Organizational Life of an Idea,* Promotor: Prof.dr. B. Nooteboom, EPS-2007-104-ORG, ISBN: 90-5892-137-6, http://hdl.handle.net/1765/10074

Kippers, J., *Empirical Studies on Cash Payments,* Promotor: Prof.dr. Ph.H.B.F. Franses, EPS-2004-043-F&A, ISBN: 90-5892-069-0, http://hdl.handle.net/1765/1520

Klein, M.H., *Poverty Alleviation through Sustainable Strategic Business Models: Essays on Poverty Alleviation as a Business Strategy,* Promotor: Prof.dr. H.R. Commandeur, EPS-2008-135-STR, ISBN: 978-90-5892-168-0, http://hdl.handle.net/1765/13482

Knapp, S., *The Econometrics of Maritime Safety: Recommendations to Enhance Safety at Sea*, Promotor: Prof.dr. Ph.H.B.F. Franses, EPS-2007-096-ORG, ISBN: 90-5892-127-1, http://hdl.handle.net/1765/7913

Kole, E., *On Crises, Crashes and Comovements*, Promotors: Prof.dr. C.G. Koedijk & Prof.dr. M.J.C.M. Verbeek, EPS-2006-083-F&A, ISBN: 90-5892-114-X, http://hdl.handle.net/1765/7829

Kooij-de Bode, J.M., *Distributed Information and Group Decision-Making: Effects of Diversity and Affect*, Promotor: Prof.dr. D.L. van Knippenberg, EPS-2007-115-ORG, http://hdl.handle.net/1765/10722

Koppius, O.R., *Information Architecture and Electronic Market Performance*, Promotors: Prof.dr. P.H.M. Vervest & Prof.dr.ir. H.W.G.M. van Heck, EPS-2002-013-LIS, ISBN: 90-5892-023-2, http://hdl.handle.net/1765/921

Kotlarsky, J., *Management of Globally Distributed Component-Based Software Development Projects*, Promotor: Prof.dr. K. Kumar, EPS-2005-059-LIS, ISBN: 90-5892-088-7, http://hdl.handle.net/1765/6772

Krauth, E.I., *Real-Time Planning Support: A Task-Technology Fit Perspective*, Promotors: Prof.dr. S.L. van de Velde & Prof.dr. J. van Hillegersberg, EPS-2008-155-LIS, ISBN: 978-90-5892-193-2, http://hdl.handle.net/1765/14521

Kuilman, J., *The Re-Emergence of Foreign Banks in Shanghai: An Ecological Analysis*, Promotor: Prof.dr. B. Krug, EPS-2005-066-ORG, ISBN: 90-5892-096-8, http://hdl.handle.net/1765/6926

Kwee, Z., *Investigating Three Key Principles of Sustained Strategic Renewal: A Longitudinal Study of Long-Lived Firms*, Promotors: Prof.dr.ir. F.A.J. Van den Bosch & Prof.dr. H.W. Volberda, EPS-2009-174-STR, ISBN: 90-5892-212-0, http://hdl.handle.net/1765/16207

Lam, K.Y., *Reliability and Rankings*, Promotor: Prof.dr. P.H.B.F. Franses, EPS-2011-230-MKT, ISBN: 978-90-5892-272-4, http://hdl.handle.net/1765/22977

Langen, P.W. de, *The Performance of Seaport Clusters: A Framework to Analyze Cluster Performance and an Application to the Seaport Clusters of Durban, Rotterdam and the Lower Mississippi*, Promotors: Prof.dr. B. Nooteboom & Prof. drs. H.W.H. Welters, EPS-2004-034-LIS, ISBN: 90-5892-056-9, http://hdl.handle.net/1765/1133

Langhe, B. de, *Contingencies: Learning Numerical and Emotional Associations in an Uncertain World*, Promotors: Prof.dr.ir. B. Wierenga & Prof.dr. S.M.J. van Osselaer, EPS-2011-236-MKT, ISBN: 978-90-5892-278-6, http://hdl.handle.net/1765/1

Larco Martinelli, J.A., *Incorporating Worker-Specific Factors in Operations Management Models*, Promotors: Prof.dr.ir. J. Dul & Prof.dr. M.B.M. de Koster, EPS-2010-217-LIS, ISBN: 90-5892-263-2, http://hdl.handle.net/1765/21527

Le Anh, T., *Intelligent Control of Vehicle-Based Internal Transport Systems*, Promotors: Prof.dr. M.B.M. de Koster & Prof.dr.ir. R. Dekker, EPS-2005-051-LIS, ISBN: 90-5892-079-8, http://hdl.handle.net/1765/6554

Le-Duc, T., *Design and Control of Efficient Order Picking Processes*, Promotor: Prof.dr. M.B.M. de Koster, EPS-2005-064-LIS, ISBN: 90-5892-094-1, http://hdl.handle.net/1765/6910

Leeuwen, E.P. van, *Recovered-Resource Dependent Industries and the Strategic Renewal of Incumbent Firm: A Multi-Level Study of Recovered Resource Dependence Management and Strategic Renewal in the European Paper and Board Industry*, Promotors: Prof.dr.ir. F.A.J. Van den Bosch & Prof.dr. H.W. Volberda, EPS-2007-109-STR, ISBN: 90-5892-140-6, http://hdl.handle.net/1765/10183

Lentink, R.M., *Algorithmic Decision Support for Shunt Planning*, Promotors: Prof.dr. L.G. Kroon & Prof.dr.ir. J.A.E.E. van Nunen, EPS-2006-073-LIS, ISBN: 90-5892-104-2, http://hdl.handle.net/1765/7328

Li, T., *Informedness and Customer-Centric Revenue Management*, Promotors: Prof.dr. P.H.M. Vervest & Prof.dr.ir. H.W.G.M. van Heck, EPS-2009-146-LIS, ISBN: 978-90-5892-195-6,

http://hdl.handle.net/1765/14525

Liang, G., *New Competition: Foreign Direct Investment and Industrial Development in China*, Promotor: Prof.dr. R.J.M. van Tulder, EPS-2004-047-ORG, ISBN: 90-5892-073-9, http://hdl.handle.net/1765/1795

Liere, D.W. van, *Network Horizon and the Dynamics of Network Positions: A Multi-Method Multi-Level Longitudinal Study of Interfirm Networks*, Promotor: Prof.dr. P.H.M. Vervest, EPS-2007-105-LIS, ISBN: 90-5892-139-0, http://hdl.handle.net/1765/10181

Loef, J., *Incongruity between Ads and Consumer Expectations of Advertising*, Promotors: Prof.dr. W.F. van Raaij & Prof.dr. G. Antonides, EPS-2002-017-MKT, ISBN: 90-5892-028-3, http://hdl.handle.net/1765/869

Londoño, M. del Pilar, *Institutional Arrangements that Affect Free Trade Agreements: Economic Rationality Versus Interest Groups*, Promotors: Prof.dr. H.E. Haralambides & Prof.dr. J.F. Francois, EPS-2006-078-LIS, ISBN: 90-5892-108-5, http://hdl.handle.net/1765/7578

Lovric, M., *Behavioral Finance and Agent-Based Artificial Markets*, Promotors: Prof.dr. J. Spronk & Prof.dr.ir. U. Kaymak, EPS-2011-229-F&A, ISBN: 978-90-5892-258-8, http://hdl.handle.net/1765/22814

Maas, A.A., van der, *Strategy Implementation in a Small Island Context: An Integrative Framework*, Promotor: Prof.dr. H.G. van Dissel, EPS-2008-127-LIS, ISBN: 978-90-5892-160-4, http://hdl.handle.net/1765/12278

Maas, K.E.G., *Corporate Social Performance: From Output Measurement to Impact Measurement*, Promotor: Prof.dr. H.R. Commandeur, EPS-2009-182-STR, ISBN: 978-90-5892-225-0, http://hdl.handle.net/1765/17627

Maeseneire, W., de, *Essays on Firm Valuation and Value Appropriation*, Promotor: Prof.dr. J.T.J. Smit, EPS-2005-053-F&A, ISBN: 90-5892-082-8, http://hdl.handle.net/1765/6768

Mandele, L.M., van der, *Leadership and the Inflection Point: A Longitudinal Perspective*, Promotors: Prof.dr. H.W. Volberda & Prof.dr. H.R. Commandeur, EPS-2004-042-STR, ISBN: 90-5892-067-4, http://hdl.handle.net/1765/1302

Markwat, T.D., *Extreme Dependence in Asset Markets Around the Globe*, Promotor: Prof.dr. D.J.C. van Dijk, EPS-2011-227-F&A, ISBN: 978-90-5892-270-0, http://hdl.handle.net/1765/22744

Meer, J.R. van der, *Operational Control of Internal Transport*, Promotors: Prof.dr. M.B.M. de Koster & Prof.dr.ir. R. Dekker, EPS-2000-001-LIS, ISBN: 90-5892-004-6, http://hdl.handle.net/1765/859

Mentink, A., *Essays on Corporate Bonds*, Promotor: Prof.dr. A.C.F. Vorst, EPS-2005-070-F&A, ISBN: 90-5892-100-X, http://hdl.handle.net/1765/7121

Meuer, J., *Configurations of Inter-Firm Relations in Management Innovation: A Study in China's Biopharmaceutical Industry*, Promotor: Prof.dr. B. Krug, EPS-2011-228-ORG, ISBN: 978-90-5892-271-1, http://hdl.handle.net/1765/22745

Meyer, R.J.H., *Mapping the Mind of the Strategist: A Quantitative Methodology for Measuring the Strategic Beliefs of Executives*, Promotor: Prof.dr. R.J.M. van Tulder, EPS-2007-106-ORG, ISBN: 978-90-5892-141-3, http://hdl.handle.net/1765/10182

Miltenburg, P.R., *Effects of Modular Sourcing on Manufacturing Flexibility in the Automotive Industry: A Study among German OEMs*, Promotors: Prof.dr. J. Paauwe & Prof.dr. H.R. Commandeur, EPS-2003-030-ORG, ISBN: 90-5892-052-6, http://hdl.handle.net/1765/1039

Moerman, G.A., *Empirical Studies on Asset Pricing and Banking in the Euro Area*, Promotor: Prof.dr. C.G. Koedijk, EPS-2005-058-F&A, ISBN: 90-5892-090-9, http://hdl.handle.net/1765/6666

Moitra, D., *Globalization of R&D: Leveraging Offshoring for Innovative Capability and Organizational Flexibility*, Promotor: Prof.dr. K. Kumar, EPS-2008-150-LIS, ISBN: 978-90-5892-184-0, http://hdl.handle.net/1765/14081

Mol, M.M., *Outsourcing, Supplier-relations and Internationalisation: Global Source Strategy as a Chinese Puzzle*, Promotor: Prof.dr. R.J.M. van Tulder, EPS-2001-010-ORG, ISBN: 90-5892-014-3, http://hdl.handle.net/1765/355

Mom, T.J.M., *Managers' Exploration and Exploitation Activities: The Influence of Organizational Factors and Knowledge Inflows*, Promotors: Prof.dr.ir. F.A.J. Van den Bosch & Prof.dr. H.W. Volberda, EPS-2006-079-STR, ISBN: 90-5892-116-6, http://hdl.handle.net/1765

Moonen, J.M., *Multi-Agent Systems for Transportation Planning and Coordination*, Promotors: Prof.dr. J. van Hillegersberg & Prof.dr. S.L. van de Velde, EPS-2009-177-LIS, ISBN: 978-90-5892-216-8, http://hdl.handle.net/1765/16208

Mulder, A., *Government Dilemmas in the Private Provision of Public Goods*, Promotor: Prof.dr. R.J.M. van Tulder, EPS-2004-045-ORG, ISBN: 90-5892-071-2, http://hdl.handle.net/1765/1790

Muller, A.R., *The Rise of Regionalism: Core Company Strategies Under The Second Wave of Integration*, Promotor: Prof.dr. R.J.M. van Tulder, EPS-2004-038-ORG, ISBN: 90-5892-062-3, http://hdl.handle.net/1765/1272

Nalbantov G.I., *Essays on Some Recent Penalization Methods with Applications in Finance and Marketing*, Promotor: Prof. dr P.J.F. Groenen, EPS-2008-132-F&A, ISBN: 978-90-5892-166-6, http://hdl.handle.net/1765/13319

Nederveen Pieterse, A., *Goal Orientation in Teams: The Role of Diversity*, Promotor: Prof.dr. D.L. van Knippenberg, EPS-2009-162-ORG, http://hdl.handle.net/1765/15240

Nguyen, T.T., *Capital Structure, Strategic Competition, and Governance*, Promotor: Prof.dr. A. de Jong, EPS-2008-148-F&A, ISBN: 90-5892-178-9, http://hdl.handle.net/1765/14005

Nielsen, L.K., *Rolling Stock Rescheduling in Passenger Railways: Applications in Short-term Planning and in Disruption Management*, Promotor: Prof.dr. L.G. Kroon, EPS-2011-224-LIS, ISBN: 978-90-5892-267-0, http://hdl.handle.net/1765/22444

Niesten, E.M.M.I., *Regulation, Governance and Adaptation: Governance Transformations in the Dutch and French Liberalizing Electricity Industries*, Promotors: Prof.dr. A. Jolink & Prof.dr. J.P.M. Groenewegen, EPS-2009-170-ORG, ISBN: 978-90-5892-208-3, http://hdl.handle.net/1765/16096

Nieuwenboer, N.A. den, *Seeing the Shadow of the Self*, Promotor: Prof.dr. S.P. Kaptein, EPS-2008-151-ORG, ISBN: 978-90-5892-182-6, http://hdl.handle.net/1765/14223

Nijdam, M.H., *Leader Firms: The Value of Companies for the Competitiveness of the Rotterdam Seaport Cluster*, Promotor: Prof.dr. R.J.M. van Tulder, EPS-2010-216-ORG, ISBN: 978-90-5892-256-4, http://hdl.handle.net/1765/21405

Ning, H., *Hierarchical Portfolio Management: Theory and Applications*, Promotor: Prof.dr. J. Spronk, EPS-2007-118-F&A, ISBN: 90-5892-152-9, http://hdl.handle.net/1765/10868

Noeverman, J., *Management Control Systems, Evaluative Style, and Behaviour: Exploring the Concept and Behavioural Consequences of Evaluative Style*, Promotors: Prof.dr. E.G.J. Vosselman & Prof.dr. A.R.T. Williams, EPS-2007-120-F&A, ISBN: 90-5892-151-2, http://hdl.handle.net/1765/10869

Noordegraaf-Eelens, L.H.J., *Contested Communication: A Critical Analysis of Central Bank Speech*, Promotor: Prof.dr. Ph.H.B.F. Franses, EPS-2010-209-MKT, ISBN: 978-90-5892-254-0,

http://hdl.handle.net/1765/21061

Nuijten, I., *Servant Leadership: Paradox or Diamond in the Rough? A Multidimensional Measure and Empirical Evidence*, Promotor: Prof.dr. D.L. van Knippenberg, EPS-2009-183-ORG, http://hdl.handle.net/1765/21405

Oosterhout, J., van, *The Quest for Legitimacy: On Authority and Responsibility in Governance*, Promotors: Prof.dr. T. van Willigenburg & Prof.mr. H.R. van Gunsteren, EPS-2002-012-ORG, ISBN: 90-5892-022-4, http://hdl.handle.net/1765/362

Oosterhout, M., van, *Business Agility and Information Technology in Service Organizations*, Promotor: Prof,dr.ir. H.W.G.M. van Heck, EPS-2010-198-LIS, ISBN: 90-5092-236-6, http://hdl.handle.net/1765/19805

Oostrum, J.M., van, *Applying Mathematical Models to Surgical Patient Planning*, Promotor: Prof.dr. A.P.M. Wagelmans, EPS-2009-179-LIS, ISBN: 978-90-5892-217-5, http://hdl.handle.net/1765/16728

Osadchiy, S.E., *The Dynamics of Formal Organization: Essays on Bureaucracy and Formal Rules*, Promotor: Prof.dr. P.P.M.A.R. Heugens, EPS-2011-231-ORG, ISBN: 978-90-5892-273-1, http://hdl.handle.net/1765/23250

Otgaar, A.H.J., *Industrial Tourism: Where the Public Meets the Private*, Promotor: Prof.dr. L. van den Berg, EPS-2010-219-ORG, ISBN: 90-5892-259-5, http://hdl.handle.net/1765/21585

Ozdemir, M.N., *Project-level Governance, Monetary Incentives and Performance in Strategic R&D Alliances*, Promoror: Prof.dr.ir. J.C.M. van den Ende, EPS-2011-235-LIS, ISBN: 978-90-5892-282-3, http://hdl.handle.net/1765/23550

Paape, L., *Corporate Governance: The Impact on the Role, Position, and Scope of Services of the Internal Audit Function*, Promotors: Prof.dr. G.J. van der Pijl & Prof.dr. H. Commandeur, EPS-2007-111-MKT, ISBN: 90-5892-143-7, http://hdl.handle.net/1765/10417

Pak, K., *Revenue Management: New Features and Models*, Promotor: Prof.dr.ir. R. Dekker, EPS-2005-061-LIS, ISBN: 90-5892-092-5, http://hdl.handle.net/1765/362/6771

Pattikawa, L.H, *Innovation in the Pharmaceutical Industry: Evidence from Drug Introduction in the U.S.*, Promotors: Prof.dr. H.R.Commandeur, EPS-2007-102-MKT, ISBN: 90-5892-135-2, http://hdl.handle.net/1765/9626

Peeters, L.W.P., *Cyclic Railway Timetable Optimization*, Promotors: Prof.dr. L.G. Kroon & Prof.dr.ir. J.A.E.E. van Nunen, EPS-2003-022-LIS, ISBN: 90-5892-042-9, http://hdl.handle.net/1765/429

Pietersz, R., *Pricing Models for Bermudan-style Interest Rate Derivatives*, Promotors: Prof.dr. A.A.J. Pelsser & Prof.dr. A.C.F. Vorst, EPS-2005-071-F&A, ISBN: 90-5892-099-2, http://hdl.handle.net/1765/7122

Pince, C., *Advances in Inventory Management: Dynamic Models*, Promotor: Prof.dr.ir. R. Dekker, EPS-2010-199-LIS, ISBN: 978-90-5892-243-4, http://hdl.handle.net/1765/19867

Poel, A.M. van der, *Empirical Essays in Corporate Finance and Financial Reporting*, Promotors: Prof.dr. A. de Jong & Prof.dr. G.M.H. Mertens, EPS-2007-133-F&A, ISBN: 978-90-5892-165-9, http://hdl.handle.net/1765/13320

Popova, V., *Knowledge Discovery and Monotonicity*, Promotor: Prof.dr. A. de Bruin, EPS-2004-037-LIS, ISBN: 90-5892-061-5, http://hdl.handle.net/1765/1201

Potthoff, D., *Railway Crew Rescheduling: Novel Approaches and Extensions*, Promotors: Prof.dr. A.P.M. Wagelmans & Prof.dr. L.G. Kroon, EPS-2010-210-LIS, ISBN: 90-5892-250-2, http://hdl.handle.net/1765/21084

Pouchkarev, I., *Performance Evaluation of Constrained Portfolios*, Promotors: Prof.dr. J. Spronk & Dr. W.G.P.M. Hallerbach, EPS-2005-052-F&A, ISBN: 90-5892-083-6, http://hdl.handle.net/1765/6731

Prins, R., *Modeling Consumer Adoption and Usage of Value-Added Mobile Services*, Promotors: Prof.dr. Ph.H.B.F. Franses & Prof.dr. P.C. Verhoef, EPS-2008-128-MKT, ISBN: 978/90-5892-161-1, http://hdl.handle.net/1765/12461

Puvanasvari Ratnasingam, P., *Interorganizational Trust in Business to Business E-Commerce*, Promotors: Prof.dr. K. Kumar & Prof.dr. H.G. van Dissel, EPS-2001-009-LIS, ISBN: 90-5892-017-8, http://hdl.handle.net/1765/356

Quak, H.J., *Sustainability of Urban Freight Transport: Retail Distribution and Local Regulation in Cities*, Promotor: Prof.dr. M.B.M. de Koster, EPS-2008-124-LIS, ISBN: 978-90-5892-154-3, http://hdl.handle.net/1765/11990

Quariguasi Frota Neto, J., *Eco-efficient Supply Chains for Electrical and Electronic Products*, Promotors: Prof.dr.ir. J.A.E.E. van Nunen & Prof.dr.ir. H.W.G.M. van Heck, EPS-2008-152-LIS, ISBN: 978-90-5892-192-5, http://hdl.handle.net/1765/14785

Radkevitch, U.L, *Online Reverse Auction for Procurement of Services*, Promotor: Prof.dr.ir. H.W.G.M. van Heck, EPS-2008-137-LIS, ISBN: 978-90-5892-171-0, http://hdl.handle.net/1765/13497

Rijsenbilt, J.A., *CEO Narcissism; Measurement and Impact*, Promotors: Prof.dr. A.G.Z. Kemna & Prof.dr. H.R. Commandeur, EPS-2011-238-STR, ISBN: 978-90-5892-281-6, http://hdl.handle.net/1765/23554

Rinsum, M. van, *Performance Measurement and Managerial Time Orientation*, Promotor: Prof.dr. F.G.H. Hartmann, EPS-2006-088-F&A, ISBN: 90-5892-121-2, http://hdl.handle.net/1765/7993

Roelofsen, E.M., *The Role of Analyst Conference Calls in Capital Markets*, Promotors: Prof.dr. G.M.H. Mertens & Prof.dr. L.G. van der Tas RA, EPS-2010-190-F&A, ISBN: 978-90-5892-228-1, http://hdl.handle.net/1765/18013

Romero Morales, D., *Optimization Problems in Supply Chain Management*, Promotors: Prof.dr.ir. J.A.E.E. van Nunen & Dr. H.E. Romeijn, EPS-2000-003-LIS, ISBN: 90-9014078-6, http://hdl.handle.net/1765/865

Roodbergen, K.J., *Layout and Routing Methods for Warehouses*, Promotors: Prof.dr. M.B.M. de Koster & Prof.dr.ir. J.A.E.E. van Nunen, EPS-2001-004-LIS, ISBN: 90-5892-005-4, http://hdl.handle.net/1765/861

Rook, L., *Imitation in Creative Task Performance*, Promotor: Prof.dr. D.L. van Knippenberg, EPS-2008-125-ORG, http://hdl.handle.net/1765/11555

Rosmalen, J. van, *Segmentation and Dimension Reduction: Exploratory and Model-Based Approaches*, Promotor: Prof.dr. P.J.F. Groenen, EPS-2009-165-MKT, ISBN: 978-90-5892-201-4, http://hdl.handle.net/1765/15536

Roza, M.W., *The Relationship between Offshoring Strategies and Firm Performance: Impact of Innovation, Absorptive Capacity and Firm Size*, Promotors: Prof.dr. H.W. Volberda & Prof.dr.ing. F.A.J. van den Bosch, EPS-2011-214-STR, ISBN: 978-90-5892-265-6, http://hdl.handle.net/1765/22155

Rus, D., *The Dark Side of Leadership: Exploring the Psychology of Leader Self-serving Behavior*, Promotor: Prof.dr. D.L. van Knippenberg, EPS-2009-178-ORG, http://hdl.handle.net/1765/16726

Samii, R., *Leveraging Logistics Partnerships: Lessons from Humanitarian Organizations,* Promotors: Prof.dr.ir. J.A.E.E. van Nunen & Prof.dr.ir. L.N. Van Wassenhove, EPS-2008-153-LIS, ISBN:

978-90-5892-186-4, http://hdl.handle.net/1765/14519

Schaik, D. van, *M&A in Japan: An Analysis of Merger Waves and Hostile Takeovers,* Promotors: Prof.dr. J. Spronk & Prof.dr. J.P.M. Groenewegen, EPS-2008-141-F&A, ISBN: 978-90-5892-169-7, http://hdl.handle.net/1765/13693

Schauten, M.B.J., *Valuation, Capital Structure Decisions and the Cost of Capital,* Promotors: Prof.dr. J. Spronk & Prof.dr. D. van Dijk, EPS-2008-134-F&A, ISBN: 978-90-5892-172-7, http://hdl.handle.net/1765/13480

Schellekens, G.A.C., *Language Abstraction in Word of Mouth,* Promotor: Prof.dr.ir. A. Smidts, EPS-2010-218-MKT, ISBN: 978-90-5892-252-6, http://hdl.handle.net/1765/21580

Schramade, W.L.J., *Corporate Bonds Issuers,* Promotor: Prof.dr. A. De Jong, EPS-2006-092-F&A, ISBN: 90-5892-125-5, http://hdl.handle.net/1765/8191

Schweizer, T.S., *An Individual Psychology of Novelty-Seeking, Creativity and Innovation,* Promotor: Prof.dr. R.J.M. van Tulder, EPS-2004-048-ORG, ISBN: 90-5892-077-1, http://hdl.handle.net/1765/1818

Six, F.E., *Trust and Trouble: Building Interpersonal Trust Within Organizations*, Promotors: Prof.dr. B. Nooteboom & Prof.dr. A.M. Sorge, EPS-2004-040-ORG, ISBN: 90-5892-064-X, http://hdl.handle.net/1765/1271

Slager, A.M.H., *Banking across Borders*, Promotors: Prof.dr. R.J.M. van Tulder & Prof.dr. D.M.N. van Wensveen, EPS-2004-041-ORG, ISBN: 90-5892-066–6, http://hdl.handle.net/1765/1301

Sloot, L., *Understanding Consumer Reactions to Assortment Unavailability*, Promotors: Prof.dr. H.R. Commandeur, Prof.dr. E. Peelen & Prof.dr. P.C. Verhoef, EPS-2006-074-MKT, ISBN: 90-5892-102-6, http://hdl.handle.net/1765/7438

Smit, W., *Market Information Sharing in Channel Relationships: Its Nature, Antecedents and Consequences*, Promotors: Prof.dr.ir. G.H. van Bruggen & Prof.dr.ir. B. Wierenga, EPS-2006-076-MKT, ISBN: 90-5892-106-9, http://hdl.handle.net/1765/7327

Sonnenberg, M., *The Signalling Effect of HRM on Psychological Contracts of Employees: A Multi-level Perspective*, Promotor: Prof.dr. J. Paauwe, EPS-2006-086-ORG, ISBN: 90-5892-119-0, http://hdl.handle.net/1765/7995

Sotgiu, F., *Not All Promotions are Made Equal: From the Effects of a Price War to Cross-chain*

Cannibalization, Promotors: Prof.dr. M.G. Dekimpe & Prof.dr.ir. B. Wierenga, EPS-2010-203-MKT, ISBN: 978-90-5892-238-0, http://hdl.handle.net/1765/19714

Speklé, R.F., *Beyond Generics: A closer Look at Hybrid and Hierarchical Governance*, Promotor: Prof.dr. M.A. van Hoepen RA, EPS-2001-008-F&A, ISBN: 90-5892-011-9, http://hdl.handle.net/1765/357

Srour, F.J., *Dissecting Drayage: An Examination of Structure, Information, and Control in Drayage Operations*, Promotor: Prof.dr. S.L. van de Velde, EPS-2010-186-LIS, http://hdl.handle.net/1765/18231

Stam, D.A., *Managing Dreams and Ambitions: A Psychological Analysis of Vision Communication*, Promotor: Prof.dr. D.L. van Knippenberg, EPS-2008-149-ORG, http://hdl.handle.net/1765/14080

Stienstra, M., *Strategic Renewal in Regulatory Environments: How Inter- and Intra-organisational Institutional Forces Influence European Energy Incumbent Firms,* Promotors: Prof.dr.ir. F.A.J. Van den Bosch & Prof.dr. H.W. Volberda, EPS-2008-145-STR, ISBN: 978-90-5892-184-0, http://hdl.handle.net/1765/13943

Sweldens, S.T.L.R., *Evaluative Conditioning 2.0: Direct versus Associative Transfer of Affect to Brands,* Promotor: Prof.dr. S.M.J. van Osselaer, EPS-2009-167-MKT, ISBN: 978-90-5892-204-5, http://hdl.handle.net/1765/16012

Szkudlarek, B.A., *Spinning the Web of Reentry: [Re]connecting reentry training theory and practice*, Promotor: Prof.dr. S.J. Magala, EPS-2008-143-ORG, ISBN: 978-90-5892-177-2, http://hdl.handle.net/1765/13695

Tempelaar, M.P., *Organizing for Ambidexterity: Studies on the Pursuit of Exploration and Exploitation through Differentiation, Integration, Contextual and Individual Attributes*, Promotors: Prof.dr.ing. F.A.J. van den Bosch & Prof.dr. H.W. Volberda, EPS-2010-191-STR, ISBN: 978-90-5892-231-1, http://hdl.handle.net/1765/18457

Teunter, L.H., *Analysis of Sales Promotion Effects on Household Purchase Behavior*, Promotors: Prof.dr.ir. B. Wierenga & Prof.dr. T. Kloek, EPS-2002-016-MKT, ISBN: 90-5892-029-1, http://hdl.handle.net/1765/868

Tims, B., *Empirical Studies on Exchange Rate Puzzles and Volatility*, Promotor: Prof.dr. C.G. Koedijk, EPS-2006-089-F&A, ISBN: 90-5892-113-1, http://hdl.handle.net/1765/8066

Tiwari, V., *Transition Process and Performance in IT Outsourcing: Evidence from a Field Study and*

Laboratory Experiments, Promotors: Prof.dr.ir. H.W.G.M. van Heck & Prof.dr. P.H.M. Vervest, EPS-2010-201-LIS, ISBN: 978-90-5892-241-0, http://hdl.handle.net/1765/19868

Tröster, C., *Nationality Heterogeneity and Interpersonal Relationships at Work*, Promotor: Prof.dr. D.L. van Knippenberg, EPS-2011-233-ORG, http://hdl.handle.net/1765/23298

Tuk, M.A., *Is Friendship Silent When Money Talks? How Consumers Respond to Word-of-Mouth Marketing*, Promotors: Prof.dr.ir. A. Smidts & Prof.dr. D.H.J. Wigboldus, EPS-2008-130-MKT, ISBN: 978-90-5892-164-2, http://hdl.handle.net/1765/12702

Tzioti, S., *Let Me Give You a Piece of Advice: Empirical Papers about Advice Taking in Marketing*, Promotors: Prof.dr. S.M.J. van Osselaer & Prof.dr.ir. B. Wierenga, EPS-2010-211-MKT, ISBN: 978-90-5892-251-9, hdl.handle.net/1765/21149

Vaccaro, I.G., *Management Innovation: Studies on the Role of Internal Change Agents*, Promotors: Prof.dr. F.A.J. van den Bosch & Prof.dr. H.W. Volberda, EPS-2010-212-STR, ISBN: 978-90-5892-253-3, hdl.handle.net/1765/21150

Valck, K. de, *Virtual Communities of Consumption: Networks of Consumer Knowledge and Companionship*, Promotors: Prof.dr.ir. G.H. van Bruggen & Prof.dr.ir. B. Wierenga, EPS-2005-050-stocktickerMKT, ISBN: 90-5892-078-X, http://hdl.handle.net/1765/6663

Valk, W. van der, *Buyer-Seller Interaction Patterns During Ongoing Service Exchange*, Promotors: Prof.dr. J.Y.F. Wynstra & Prof.dr.ir. B. Axelsson, EPS-2007-116-MKT, ISBN: 90-5892-146-8, http://hdl.handle.net/1765/10856

Verheijen, H.J.J., *Vendor-Buyer Coordination in Supply Chains*, Promotor: Prof.dr.ir. J.A.E.E. van Nunen, EPS-2010-194-LIS, ISBN: 90-5892-234-2, http://hdl.handle.net/1765/19594

Verheul, I., *Is There a (Fe)male Approach? Understanding Gender Differences in Entrepreneurship*, Promotor: Prof.dr. A.R. Thurik, EPS-2005-054-ORG, ISBN: 90-5892-080-1, http://hdl.handle.net/1765/2005

Verwijmeren, P., *Empirical Essays on Debt, Equity, and Convertible Securities*, Promotors: Prof.dr. A. de Jong & Prof.dr. M.J.C.M. Verbeek, EPS-2009-154-F&A, ISBN: 978-90-5892-187-1, http://hdl.handle.net/1765/14312

Vis, I.F.A., *Planning and Control Concepts for Material Handling Systems*, Promotors: Prof.dr. M.B.M. de Koster & Prof.dr.ir. R. Dekker, EPS-2002-014-LIS, ISBN: 90-5892-021-6,

http://hdl.handle.net/1765/866

Vlaar, P.W.L., *Making Sense of Formalization in Interorganizational Relationships: Beyond Coordination and Control*, Promotors: Prof.dr.ir. F.A.J. Van den Bosch & Prof.dr. H.W. Volberda, EPS-2006-075-STR, ISBN 90-5892-103-4, http://hdl.handle.net/1765/7326

Vliet, P. van, *Downside Risk and Empirical Asset Pricing*, Promotor: Prof.dr. G.T. Post, EPS-2004-049-F&A, ISBN: 90-5892-07-55, http://hdl.handle.net/1765/1819

Vlist, P. van der, *Synchronizing the Retail Supply Chain*, Promotors: Prof.dr.ir. J.A.E.E. van Nunen & Prof.dr. A.G. de Kok, EPS-2007-110-LIS, ISBN: 90-5892-142-0, http://hdl.handle.net/1765/10418

Vries-van Ketel E. de, *How Assortment Variety Affects Assortment Attractiveness:A Consumer Perspective*, Promotors: Prof.dr. G.H. van Bruggen & Prof.dr.ir. A. Smidts, EPS-2006-072-MKT, ISBN: 90-5892-101-8, http://hdl.handle.net/1765/7193

Vromans, M.J.C.M., *Reliability of Railway Systems*, Promotors: Prof.dr. L.G. Kroon, Prof.dr.ir. R. Dekker & Prof.dr.ir. J.A.E.E. van Nunen, EPS-2005-062-LIS, ISBN: 90-5892-089-5, http://hdl.handle.net/1765/6773

Vroomen, B.L.K., *The Effects of the Internet, Recommendation Quality and Decision Strategies on Consumer Choice*, Promotor: Prof.dr. Ph.H.B.F. Franses, EPS-2006-090-MKT, ISBN: 90-5892-122-0, http://hdl.handle.net/1765/8067

Waal, T. de, *Processing of Erroneous and Unsafe Data*, Promotor: Prof.dr.ir. R. Dekker, EPS-2003-024-LIS, ISBN: 90-5892-045-3, http://hdl.handle.net/1765/870

Waard, E.J. de, *Engaging Environmental Turbulence: Organizational Determinants for Repetitive Quick and Adequate Responses*, Promotors: Prof.dr. H.W. Volberda & Prof.dr. J. Soeters, EPS-2010-189-STR, ISBN: 978-90-5892-229-8, http://hdl.handle.net/1765/18012

Wall, R.S., *Netscape: Cities and Global Corporate Networks*, Promotor: Prof.dr. G.A. van der Knaap, EPS-2009-169-ORG, ISBN: 978-90-5892-207-6, http://hdl.handle.net/1765/16013

Wang, Y., *Information Content of Mutual Fund Portfolio Disclosure*, Promotor: Prof.dr. M.J.C.M. Verbeek, EPS-2011-242-F&A, ISBN: 978-90-5892-285-4, http://hdl.handle.net/1765/1

Watkins Fassler, K., *Macroeconomic Crisis and Firm Performance*, Promotors: Prof.dr. J. Spronk & Prof.dr. D.J. van Dijk, EPS-2007-103-F&A, ISBN: 90-5892-138-3, http://hdl.handle.net/1765/10065

Weerdt, N.P. van der, *Organizational Flexibility for Hypercompetitive Markets: Empirical Evidence of the Composition and Context Specificity of Dynamic Capabilities and Organization Design Parameters*, Promotor: Prof.dr. H.W. Volberda, EPS-2009-173-STR, ISBN: 978-90-5892-215-1, http://hdl.handle.net/1765/16182

Wennekers, A.R.M., *Entrepreneurship at Country Level: Economic and Non-Economic Determinants*, Promotor: Prof.dr. A.R. Thurik, EPS-2006-81-ORG, ISBN: 90-5892-115-8, http://hdl.handle.net/1765/7982

Wielemaker, M.W., *Managing Initiatives: A Synthesis of the Conditioning and Knowledge-Creating View*, Promotors: Prof.dr. H.W. Volberda & Prof.dr. C.W.F. Baden-Fuller, EPS-2003-28-STR, ISBN: 90-5892-050-X, http://hdl.handle.net/1765/1042

Wijk, R.A.J.L. van, *Organizing Knowledge in Internal Networks: A Multilevel Study*, Promotor: Prof.dr.ir. F.A.J. van den Bosch, EPS-2003-021-STR, ISBN: 90-5892-039-9, http://hdl.handle.net/1765/347

Wolters, M.J.J., *The Business of Modularity and the Modularity of Business*, Promotors: Prof. mr. dr. P.H.M. Vervest & Prof.dr.ir. H.W.G.M. van Heck, EPS-2002-011-LIS, ISBN: 90-5892-020-8, http://hdl.handle.net/1765/920

Wubben, M.J.J., *Social Functions of Emotions in Social Dilemmas*, Promotors: Prof.dr. D. De Cremer & Prof.dr. E. van Dijk, EPS-2009-187-ORG, http://hdl.handle.net/1765/18228

Xu, Y., *Empirical Essays on the Stock Returns, Risk Management, and Liquidity Creation of Banks*, Promotor: Prof.dr. M.J.C.M. Verbeek, EPS-2010-188-F&A, http://hdl.handle.net/1765/18125

Yang, J., *Towards the Restructuring and Co-ordination Mechanisms for the Architecture of Chinese Transport Logistics*, Promotor: Prof.dr. H.E. Harlambides, EPS-2009-157-LIS, ISBN: 978-90-5892-198-7, http://hdl.handle.net/1765/14527

Yu, M., *Enhancing Warehouse Perfromance by Efficient Order Picking*, Promotor: Prof.dr. M.B.M. de Koster, EPS-2008-139-LIS, ISBN: 978-90-5892-167-3, http://hdl.handle.net/1765/13691

Zhang, X., *Strategizing of Foreign Firms in China: An Institution-based Perspective*, Promotor: Prof.dr. B. Krug, EPS-2007-114-ORG, ISBN: 90-5892-147-5, http://hdl.handle.net/1765/10721

Zhang, X., *Scheduling with Time Lags*, Promotor: Prof.dr. S.L. van de Velde, EPS-2010-206-LIS, ISBN:

978-90-5892-244-1, http://hdl.handle.net/1765/19928

Zhou, H., *Knowledge, Entrepreneurship and Performance: Evidence from Country-level and Firm-level Studies*, Promotors: Prof.dr. A.R. Thurik & Prof.dr. L.M. Uhlaner, EPS-2010-207-ORG, ISBN: 90-5892-248-9, http://hdl.handle.net/1765/20634

Zhu, Z., *Essays on China's Tax System*, Promotors: Prof.dr. B. Krug & Prof.dr. G.W.J. Hendrikse, EPS-2007-112-ORG, ISBN: 90-5892-144-4, http://hdl.handle.net/1765/10502

Zwan, P.W. van der, *The Entrepreneurial Process: An International Analysis of Entry and Exit*, EPS-2011-234-ORG, ISBN: 978-90-5892-277-9, http://hdl.handle.net/1765/23422

Zwart, G.J. de, *Empirical Studies on Financial Markets: Private Equity, Corporate Bonds and Emerging Markets*, Promotors: Prof.dr. M.J.C.M. Verbeek & Prof.dr. D.J.C. van Dijk, EPS-2008-131-F&A, ISBN: 978-90-5892-163-5, http://hdl.handle.net/1765/12703

Erasmus Research Institute of Management - ERiM

METHODOLOGICAL ADVANCES IN BIBLIOMETRIC MAPPING OF SCIENCE

Bibliometric mapping of science is concerned with quantitative methods for visually representing scientific literature based on bibliographic data. Since the first pioneering efforts in the 1970s, a large number of methods and techniques for bibliometric mapping have been proposed and tested. Although this has not resulted in a single generally accepted methodological standard, it did result in a limited set of commonly used methods and techniques.

In this thesis, a new methodology for bibliometric mapping is presented. It is argued that some well-known methods and techniques for bibliometric mapping have serious shortcomings. For instance, the mathematical justification of a number of commonly used normalization methods is criticized, and popular multidimensional-scaling-based approaches for constructing bibliometric maps are shown to suffer from artifacts, especially when working with larger data sets.

The methodology introduced in this thesis aims to provide improved methods and techniques for bibliometric mapping. The thesis contains an extensive mathematical analysis of normalization methods, indicating that the so-called association strength measure has the most satisfactory mathematical properties. The thesis also introduces the VOS technique for constructing bibliometric maps, where VOS stands for *visualization of similarities*. Compared with well-known multidimensional-scaling-based approaches, the VOS technique is shown to produce more satisfactory maps. In addition to the VOS mapping technique, the thesis also presents the VOS clustering technique. Together, these two techniques provide a unified framework for mapping and clustering. Finally, the VOSviewer software for constructing, displaying, and exploring bibliometric maps is introduced.

Design & layout: B&T Ontwerp en advies (www.b-en-t.nl)   Print: Haveka (www.haveka.nl)

ERiM

The Erasmus Research Institute of Management (ERIM) is the Research School (Onderzoekschool) in the field of management of the Erasmus University Rotterdam. The founding participants of ERIM are the Rotterdam School of Management (RSM), and the Erasmus School of Economics (ESE). ERIM was founded in 1999 and is officially accredited by the Royal Netherlands Academy of Arts and Sciences (KNAW). The research undertaken by ERIM is focused on the management of the firm in its environment, its intra- and interfirm relations, and its business processes in their interdependent connections.

The objective of ERIM is to carry out first rate research in management, and to offer an advanced doctoral programme in Research in Management. Within ERIM, over three hundred senior researchers and PhD candidates are active in the different research programmes. From a variety of academic backgrounds and expertises, the ERIM community is united in striving for excellence and working at the forefront of creating new business knowledge.

ERIM PhD Series
Research in Management

Erasmus Research Institute of Management - ERiM
Rotterdam School of Management (RSM)
Erasmus School of Economics (ESE)
Erasmus University Rotterdam (EUR)
P.O. Box 1738, 3000 DR Rotterdam,
The Netherlands

Tel.	+31 10 408 11 82
Fax	+31 10 408 96 40
E-mail	info@erim.eur.nl
Internet	www.erim.eur.nl