Jointly published by Akadémiai Kiadó, Budapest and Springer, Dordrecht

Scientometrics, Vol. 78, No. 1 (2009) 113–130
DOI: 10.1007/s11192-007-1961-z

Similarity measures for document mapping: A comparative study on the level of an individual scientist

CHRISTIAN STERNITZKE,[a,b] ISUMO BERGMANN[b]

[a] *Technische Universität Ilmenau, PATON – Landespatentzentrum Thüringen, PF 100 565, D-98684 Ilmenau, Germany*
[b] *Universität Bremen, Institut für Projektmanagement und Innovation (IPMI), Bremen, Germany*

This paper investigates the utility of the Inclusion Index, the Jaccard Index and the Cosine Index for calculating similarities of documents, as used for mapping science and technology. It is shown that, provided that the same content is searched across various documents, the Inclusion Index generally delivers more exact results, in particular when computing the degree of similarity based on citation data. In addition, various methodologies such as co-word analysis, Subject–Action–Object (SAO) structures, bibliographic coupling, co-citation analysis, and self-citation links are compared. We find that the two former ones tend to describe rather semantic similarities that differ from knowledge flows as expressed by the citation-based methodologies.

Introduction

Mapping of documents has been a discussion topic in scientometric research for a number of years (for a review, see e.g. [BOERNER & AL., 2003]). In general, the procedure follows a three-step process. First, (bibliographic) items are selected that serve as a basis for comparing documents. Here, a variety of methodologies exists: KESSLER [1963] suggested the use of the references contained in papers, whereas documents with the same references are regarded as very similar in nature. This approach is known as bibliographic coupling. In contrast, SMALL [1973] and MARSHAKOVA [1973] proposed not to use references (i.e. backward citations) but the citations a paper receives (i.e. so-called forward citations). This approach was named

Received December 6, 2007

Address for correspondence:
CHRISTIAN STERNITZKE
E-mail: cs@sternitzke.com

0138–9130/US $ 20.00

co-citation analysis. Another methodology uses words as items that are employed in, for instance, title and abstract to describe similarities between documents. This approach became known as co-word analysis (see e.g. [RIP & COURTIAL, 1984; CALLON & AL., 1991]). Similar approaches deploy advanced text-mining or artificial intelligence techniques, relying not solely on words but semantic structures of texts. Subject–Action–Object (SAO) structures extracted from full-text documents are an example (see [INVENTION MACHINE CORPORATION, NO DATE] and [TSOURIKOV & AL., 2000]).

In a second step, similarities are computed based on the above-mentioned items. Measures such as the Pearson correlation coefficient, Salton's Cosine formula, the Jaccard Index, or the Inclusion Index are possible (for discussions on the pros and cons of some of these measures see [HAMERS & AL., 1989; PETERS & AL., 1995; QIN, 2000; AHLGREN & AL., 2003]).

Finally, in the third step, the previously computed data is visualized by means of multivariate analyses such as cluster analysis or multidimensional scaling (MDS) (see e.g. [LEYDESDORFF, 1987]). A further but different approach is deploying graph-theoretical algorithms on citation links between documents, resulting in a citation network for the documents under consideration (for an example, see [CLARKSON, 2004; RAMLOGAN & AL., 2007]).

All three steps have an impact on the results of the analysis. We argue that the first step is the most important one because different items such as backward or forward citations, words, SAO structures, etc. represent different characteristics of similarities. In addition, these items are highly affected by data availability. Since most scientific articles cite other papers, data for bibliographic coupling should be available for the vast majority of scientific publications. Forward citations are, in contrast, highly skewed since only few papers receive many citations, and many papers receive few citations. So co-citation analysis is more difficult to conduct due to the inherently sparse availability of citation data. In addition, the amount of citations a paper receives depends on the future, whereas a reference list of backward citations is fixed. There exists also a bias for younger documents that have accumulated fewer citations than older documents. Co-word analysis is, when employing the Science Citation Index (SCI), limited in scope because particularly older records in the database do not contain abstracts. Furthermore, word lists are frequently cleaned by means of stopword lists (see e.g. [BLANCHARD, 2007]), and there are various ways in manipulating such lists. Approaches that employ semantic analyses, such as SAO structures, unfold their power when using full-texts of documents that are not provided by the SCI. Hence, semantic analysis cannot as easily be conducted as co-word analysis relying solely on titles and abstracts.

As it was just briefly described, bibliometricians can choose among a variety of approaches to determine similarities between documents, but are these approaches alternatives to each other? We will try to answer this question and compare bibliographic coupling, co-citation analysis, co-word analysis, and SAO structures for a

set of publications originating from one prominent author in optoelectronics. Additionally, results from a citation network analysis will be compared.

Data and methodology

The dataset comprised 156 publications submitted to scientific journals between 1991 and 1999 by a prominent author in a new and emerging subfield in optoelectronics, found in the Web of Science (WoS). The focus on one author has several advantages: first, the author should have been aware of the same literature to cite in the papers in the course of time, leading to a high level of homogeneity in backward citations, which implies that bibliographic coupling should provide valuable results. Second, there should be a high overlap between the documents regarding the selection of words, grammatical terms, etc. Third, citation links between the documents are self-citations. Here, one can assume that the author cited all relevant self-created literature. Hence, a citation linkage should therefore be a strong indicator of similarity.[1] To enable a comparison between co-word analysis and SAO structures, only documents with at least ten different words were selected due to reasons explained below, leading in a reduction of the sample to 150 papers in total. Data on backward references was obtained from the Web of Science (WoS). Forward citations were elicited from SCISEARCH via STN International.[2]

Computations for bibliographic data were carried out in Microsoft Excel using the Add-on PATONanalist [BARTKOWSKI & AL., 2004; STERNITZKE & AL., 2007]. For the co-word analysis, words from titles and abstracts from the SCI were jointly investigated. The words were filtered by means of stopword lists to reduce noise, including the terms from RIJSBERGEN's [1979] list. The remaining words were treated with a Porter Stemmer [PORTER, 1980] to eliminate plural endings, etc. Finally, retained terms were standardized intellectually, searching for synonyms, etc. as it is recommended for such kind of analyses [JARNEVING, 2005].

The similarity measures deployed in this paper for bibliographic coupling, co-citation and co-word analysis are Salton's Cosine Index [SALTON & MACGILL, 1983] as already used for the same purpose recently by JARNEVING [2005], and the Inclusion Index. Another prominent index in this context is the Jaccard Index [JACCARD, 1901]. The index is calculated as the ratio of items (e.g. words, citation, etc.) being contained

[1] There is some noise in the data because in a few cases self-citations related to publications "in press", so no proper link between the two documents could be established.

[2] We used these two databases because we found for our dataset that in WoS slightly more than ten percent of the references are not linked properly, leading to omitted forward citations. The reasons appear to be that authors do not cite many papers correctly, in particular relating to page numbers starting with letters such as L (for letters) or R (for reviews).

in document i and j, normalized by the sum of the items in document i and j minus the nominator:

$$Jaccard\ \ Index\ = \frac{items_{ij}}{items_{i} + items_{j} - items_{ij}} \tag{1}$$

Salton's Cosine is computed as the ratio of items contained in document i and j, normalized by square root of the product of the items from document i and j:

$$Cosine\ Index = \frac{items_{ij}}{\sqrt{items_{i} \bullet items_{j}}} \tag{2}$$

The Inclusion Index takes into account the common items between two documents based on the minimum number of items from document i or j:

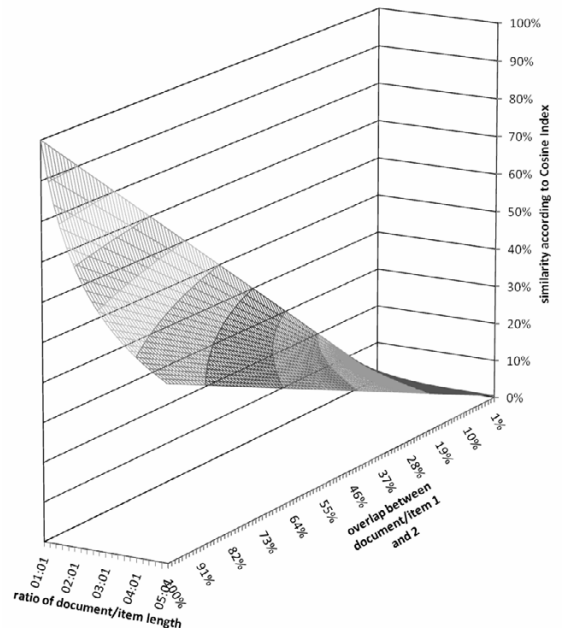
$$Inclusion\ Index = \frac{items_{ij}}{\min(items_{i}; items_{j})} \tag{3}$$

So the ratio of items does not play a role here. Hence, if the items from document i are fully contained in the much longer document j, the Inclusion Index will be 1.0. This index is, in particular, useful when searching for similar content in a variety of different documents since, in comparison to the Jaccard or Cosine Index, it is not biased by the number of items (e.g. the document length for co-word analysis) as the latter [HAMERS & AL., 1989; PETERS & AL., 1995; QIN, 2000]. Figure 1 illustrates this effect. Here, the ordinate provides the degree of similarity as computed by the Cosine or Jaccard Index. The long axis on the bottom layer represents the overlap between the items of document i and j, whereas the short axis in the bottom layer provides the ratio in item number (e.g. document length or length of the reference list if citations are counted) between document i and j. If, for example, all items from document i may be fully contained within document j (it will be a 100 percent overlap on the long axis on the bottom), but document j has five times more items than document i (i.e. a ratio of 5:1 on the short bottom axis), then it can easily be seen that the Cosine Index will become 45 percent, whereas the Jaccard Index yields a similarity degree of only 20 percent. The Inclusion Index, in comparison, would be 100 percent.

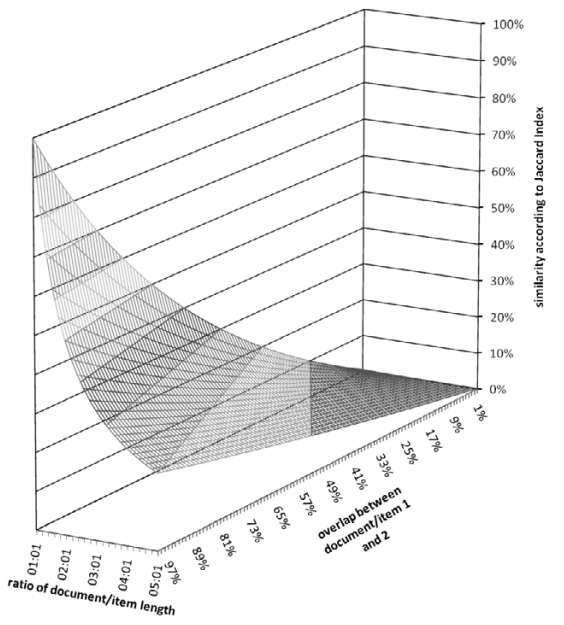
Data processing for SAO structures was conducted via the software Knowledgist from Invention Machine. The similarity measure used in this context takes into account the frequency of overlapping items occurring in both documents:

$$Sim_{ij} = \frac{SAO_{ij} + SAO_{ji}}{SAO_{i} + SAO_{j}} \tag{4}$$

A closer description of SAO structure processing can be found in MOEHRLE & AL. [2005], DREßLER [2006], or BERGMANN & AL. [2007]. Citation networks were visualized with UCINET and Netdraw [BORGATTI & AL., 1999] using a spring embedding algorithm [GOLBECK & MUTTON, 2006; KAMADA & KAWAI, 1989].



(a)



(b)

Figure 1. Simulation of overlap and document length on the Cosine (a) and Jaccard (b) Index

The similarity of the documents computed via bibliographic coupling, co-citation and SAO analysis was visualized by multidimensional scaling (Proxscal algorithm as contained in UCINET, drawings in Netdraw).

For illustrative purposes, additional information was integrated into the visualizations, such as document age, and information on the content of the papers. Here, all papers were clustered manually based on title and abstract. Four different "classes" were chosen depending on the following terms in title or abstract: i) "light emitting diodes" (LEDs); ii) "laser diodes" (LDs); iii) both LDs *and* LEDs; and iv) characterizations of thin films and quantum well structures, including film growth. The inherent nature of titles and abstracts is to describe the major contents of a paper. Nevertheless, important aspects can also be described in the full-text of the papers, so this measure is not free from errors.

Results and discussion

First, the theoretical considerations regarding the efficiency of the Cosine/ Jaccard Index and the number of items per document are investigated empirically. Second, the threshold level of items is discussed that need to be included into an analysis in order to yield useable results. Third, the dataset visualizations by MDS is shown, and fourth, the results from a factor analysis on the different methodologies are discussed.

Jaccard versus Cosine Index

As we have already mentioned, a central issue when discussing the differences between similarity indices is the difference in items between the documents. Since the Inclusion Index is not affected by this phenomenon, we only discuss the effect of different item numbers for the Cosine Index and the Jaccard Index.

For the dataset under consideration, we tested the relevancy of the effect presented in Figure 1, i.e. the impact of the number of different items – namely co-words, backward references, and forward citations – on the two similarity indices. Only documents with at least ten co-words and three citations (as suggested by [SHARABCHIEV, 1989]) were taken into account. As a consequence, the impact of randomly involved words or citations on the results is limited. This limitation resulted in subsets of the 156 papers that were finally analyzed: 150 documents were included into the dataset for comparing co-words, 149 into the one for backward references, and 136 into the set for forward references.

Results can be found in Table 1. The first column therein provides the item-to-item ratio, ranging from smaller than 1.5:1 to larger than 10:1. This measure describes the ratio of the larger item list to the smaller one when comparing two documents.

Table 1. Item-to-item ratios

item/item ratio	Maximum of Cosine Index	Maximum of Jaccard Index	Co-word	Backward references	Forward citations
			Percentage [cumulative]	Percentage [cumulative]	Percentage [cumulative]
1.5:1	82%	67%	70%	39%	15%
2:1	71%	50%	88%	59%	26%
4:1	50%	25%	99%	87%	50%
6:1	41%	17%	100%	95%	61%
8:1	35%	13%	100%	98%	69%
10:1	32%	10%	100%	99%	74%
>10:1	<32%	<10%	100%	100%	100%

* Threshold level for inclusion: Co-words: 10, backward and forward references: 3, as suggested by SHARABCHIEV (1989)

The two following columns present the similarity degree as computed by the Cosine and Jaccard Index for the case the items to be compared overlap to 100 percent. For an item-to-item ratio of 1.5:1, the Cosine Index would yield a similarity degree of 82 percent, the Jaccard Index 67 percent, while these numbers drop in the case of a 10:1 ratio to 32 percent and 10 percent, respectively.

The distributions of the item-to-item ratios for co-words and citations indicate that for co-word analysis the effect is less severe than for citation data. With the threshold level given, only 12 percent of all documents have an item-to-item ratio larger 2:1, meaning that Cosine and Jaccard Index are lower than 71 and 50 percent, respectively (see Table 1). For backward references, this number increases from 12 to 41 percent, and for forward citations to even 74 percent. Therefore, one obtains a severe bias when using these two similarity indices for citation data. The Inclusion Index, however, would not be affected by these distributions.

Item threshold level

Computing similarities between documents that only have very few items increases the weight of every item substantially. For instance, when using the Inclusion Index for co-citation analysis, there are two documents i and j. Document i had received ten citations, document j only one. If both would have been cited by the same subsequent paper, then they would show a similarity degree of 100 percent. This seems to be somewhat odd. So another important issue in co-word analysis, bibliographic coupling, and co-citation analysis is the exclusion of documents with too few items to minimize random effects. As we have done in the previous section, the solution is to set threshold levels and define a minimum number of items a document needs to possess in order to be included into the analysis.

Could it be that the data we presented in the previous section suffers from inefficiently chosen threshold levels? Inefficiently in this case implies that the level was either set too high, meaning that too many documents drop out, or that it was set too low, with the result that single items receive a very high weight and can bias the analysis. In general, this is a classical precision-recall dilemma in information retrieval.

In order to shed more light on this phenomenon, we provided the statistics on the occurrence of documents with a certain number of items in Table 2. It can be seen that the occurrence of words follows a distribution similar to the normal distribution, while citation data is, as expected, rather skewed, with forward citations having a longer tail. Increasing the threshold level for citation data from e.g. three to ten would reduce the item-to-item ratio to a certain degree because the minimum number of citations being used for computing the item-to-item ratio triples, but at the same time the number of documents being included into the analysis would drop substantially. So setting the threshold level at 10 words and three citations assures that the majority of all papers can be integrated into the similarity analyses, leading to a high recall at the cost of some precision.

Table 2. Occurrence of items in documents: Co-words, backward references, and forward references

Items of document	Co-word		Backward references		Forward citations	
	Occurrence	Percentage [cumulative]	Occurrence*	Percentage [cumulative]	Occurrence*	Percentage [cumulative]
<10	6	4%	35	22%	40	26%
11–20	4	6%	77	72%	18	37%
21–30	7	11%	29	90%	11	44%
31–40	39	36%	6	94%	13	53%
41–50	46	65%	7	99%	5	56%
51–70	44	94%	2	100%	13	64%
71–100	7	98%	0	100%	15	74%
>100	3	100%	0	100%	41	100%

Visualization of the results

Even though we argued in the previous two sections that the Cosine Index yields problematic results, we computed the similarities between the documents based on co-words, bibliographic coupling, and co-citations with both the Cosine and Inclusion Index. SAO structures and citation networks were created as described in the Data and methodology section. The visualizations are presented in Figures 2–6. Here, older papers appear larger. The shape of the dots refers to the classes of the articles: squares represent characterizations; diamonds represent LED-related papers, triangles lasers, and circles both LEDs and lasers.

Both graphs showing bibliographic coupling (Figure 2) demonstrate similarities in a kind of cluster on the left side. This cluster includes all four classes of different document contents such as characterization of materials including layer growth, LEDs and/or laser diodes. Most of them are relatively old and are regarded as basis papers within the industry. This could mean that during later stages in the development simply more literature was available and could be cited than in the beginning. Towards the right side of the graph, clustering becomes somewhat dispersed and unclear, even though minor subclusters can be identified.

Co-citation analysis reveals a rather dispersed, random-like landscape of document similarities. Here, visual inspection favors the Inclusion Index which creates more loosely coupled clusters according to the four groups of document contents. There is hardly any similarity among the visualizations of bibliographic coupling and co-citations. As already mentioned, not all documents in fact received citations. These are excluded from the visualization. As was illustrated earlier, bias is also included when taking into account only few citations, as it was done for both bibliographic coupling and co-citations, because a citation can occur by random. So documents cited only once or twice, compared with documents with a much longer citation list, would tend to show a high degree of similarity.

The co-word analysis in Figure 4 provides a totally different picture than the previous graphs. However, the differences between the Inclusion and Cosine Index are not very large, which is rooted in the Gaussian shape-like distribution of co-words as described in Table 2. Even though somewhat biased when taking the discrimination into the four classes of documents into account because LEDs, laser diodes and types of characterization play an important role in title and abstract, the documents are grouped into several areas representing the different document classes. Here again, the Inclusion Index seems to discriminate better between classes than the Cosine Index.

Figure 5 highlights the results of the SAO analysis. Documents are distributed similar to the co-citation analysis, even though some clustering of the documents can be recognized. Comparing Figures 4 and 5, it appears that co-word analysis and SAO structures hardly describe the same. The (self) citation link analysis in Figure 6 reveals several clusters describing also document classes. The second cluster on the right side, for example, relates, according to title and abstract, to light emission in quantum wells, i.e. compound semiconductor (multilayer) films with dimensions in the nanoscale used for bright LEDs and lasers. Taking the different classes into account, the picture seems to be logic: having started with research on film growth and characterization, these developments were first used to create LEDs and then, laser diodes.

In conclusion, some methodologies appear to describe the content of the papers as several clusters, other approaches rather yield a random-like structure.

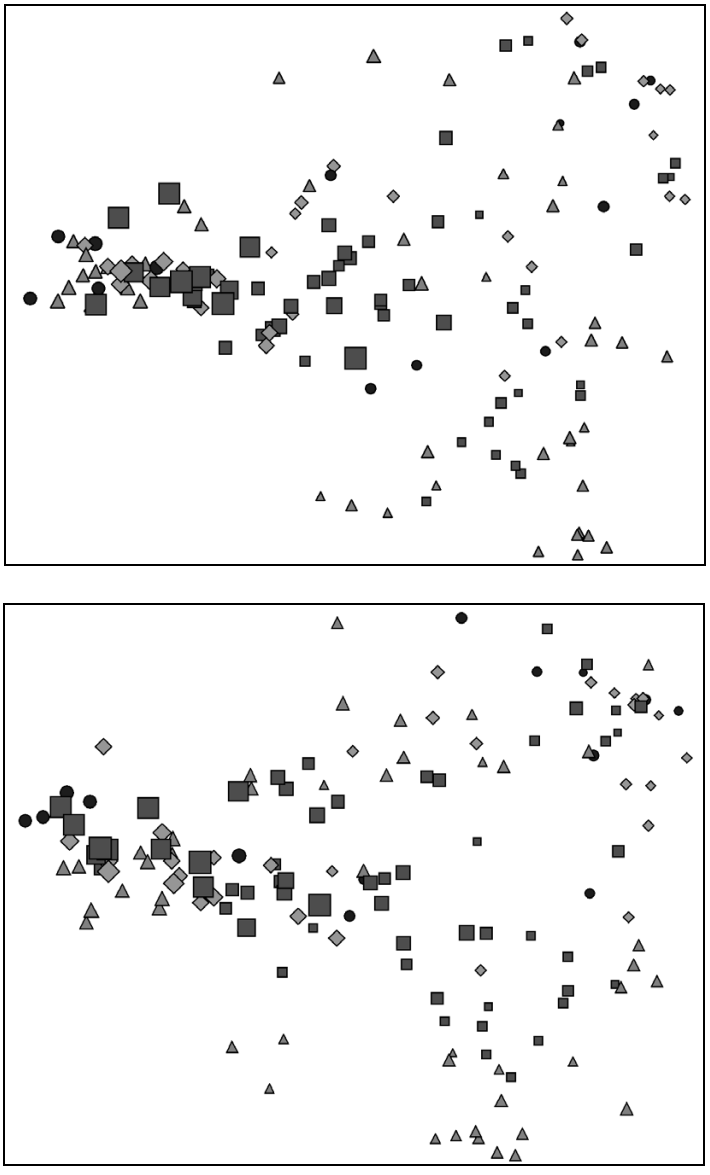
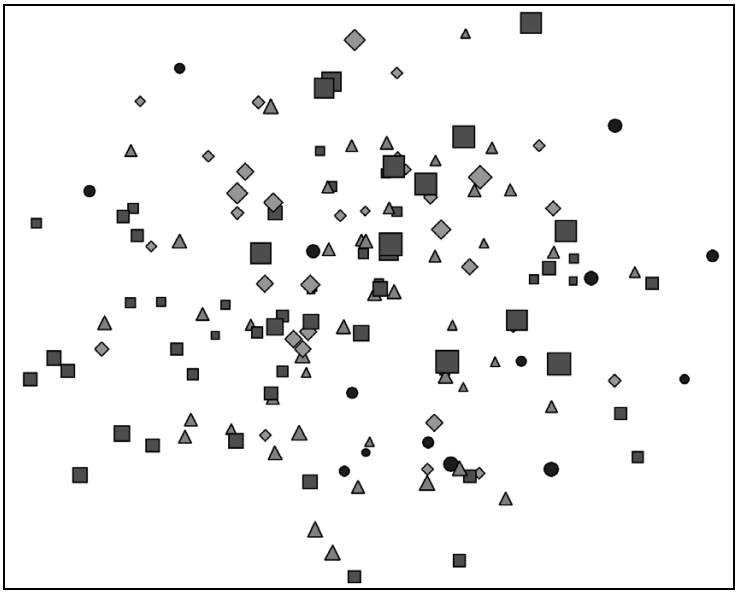
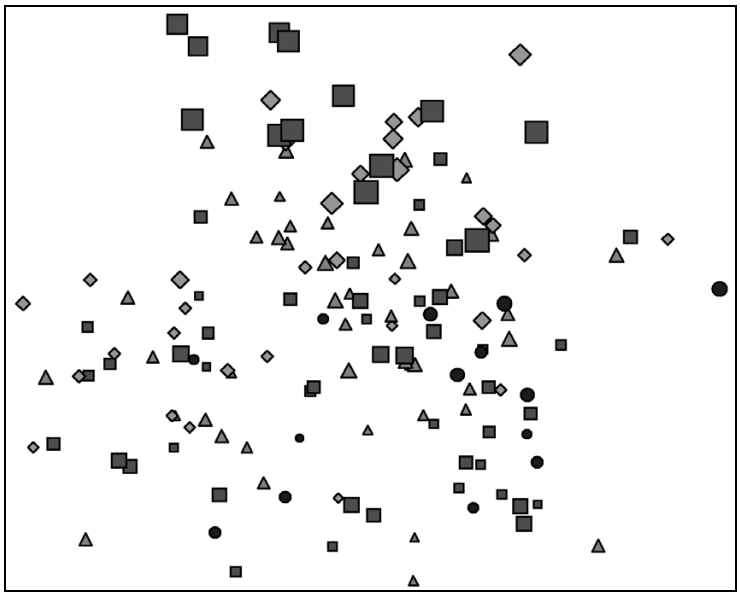


Figure 2. Results from bibliographic coupling. Cosine Index (a) and Inclusion Index (b). (MDS: (a) stress: 0.112, (b) stress: 0.121; 9-D). Older papers appear larger. Squares represent characterizations; diamonds represent LEDs, triangles lasers, and circles both LEDs and lasers

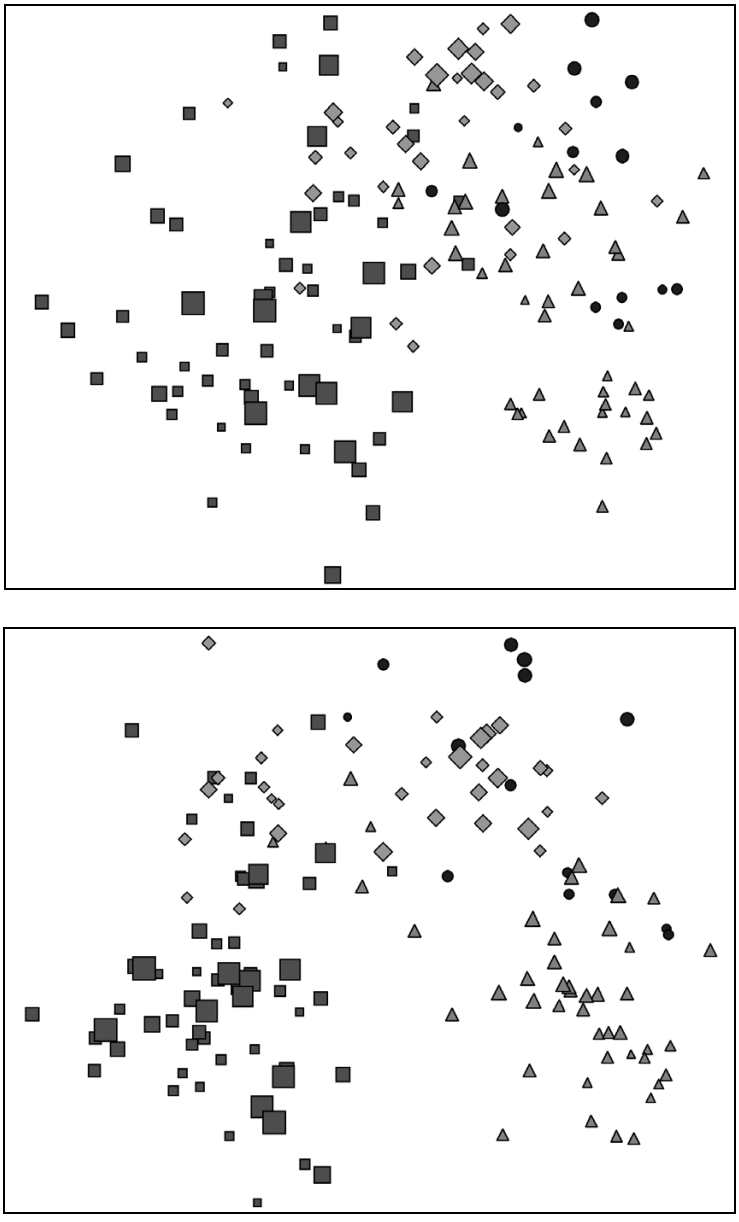


(a)



(b)

Figure 3. Results from co-citation analysis. Cosine Index (a) and Inclusion Index (b). (MDS: (a) stress: 0.147, (b) stress: 0.150; 9-D). Older papers appear larger. Squares represent characterizations; diamonds represent LEDs, triangles lasers, and circles both LEDs and lasers



(a)

(b)

Figure 4. Results from Co-word analysis. Cosine Index (a) and Inclusion Index (b). (MDS: (a) stress: 0.108, (b) stress: 0.112; 9-D). Older papers appear larger. Squares represent characterizations; diamonds represent LEDs, triangles lasers, and circles both LEDs and lasers

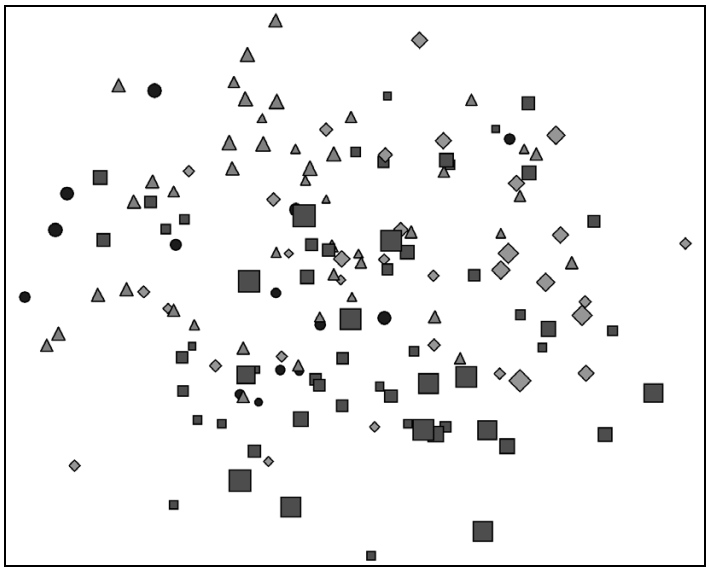


Figure 5. Results from SAO analysis. (MDS: stress: 0.150, 9-D). Older papers appear larger. Squares represent characterizations; diamonds represent LEDs, triangles lasers, and circles both LEDs and lasers

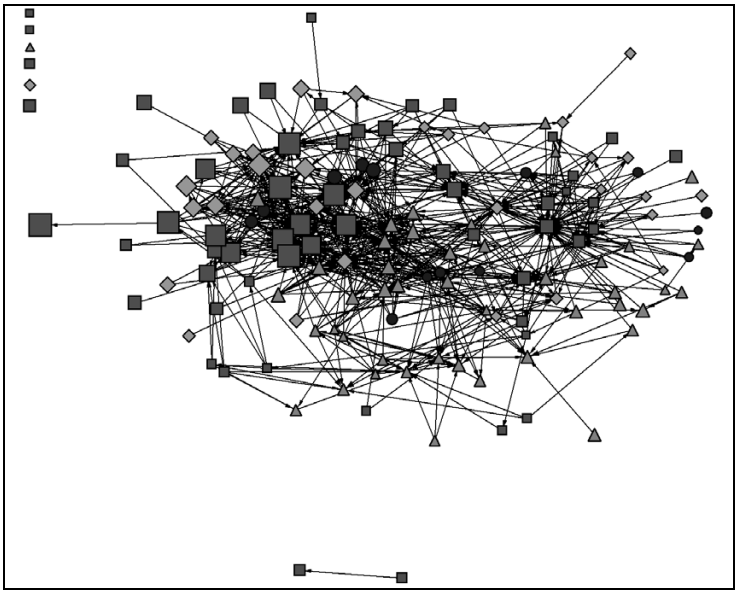


Figure 6. Results from self-citation network. Drawn with spring embedding algorithm. Older papers appear larger. Squares represent characterizations; diamonds represent LEDs, triangles lasers, and circles both LEDs and lasers. Papers that are not connected via citation links are situated in the upper left corner

Factor analysis

In this section, the data from the similarity matrices is investigated by means of factor analysis in order to enhance the results provided by the visualizations in the previous section. The goal is to investigate which methodologies (co-word analysis, co-citation analysis, etc.) describe, more or less, the same type of similarity.

Table 3 shows the correlations between the different indicators. As expected, they are relatively high for the same indicator when computed as Cosine or Inclusion Index. As one would expect, the correlation is highest for the co-word analyses due to the Gaussian shape-like distribution of the words, and lower for citation based measures with the rather skewed distributions. In addition, it can be seen that co-citations and the citation link have a very low correlation with bibliographic coupling, co-citation analysis and SAOs.

Table 3. Correlation matrix

(1) Bibliographic Coupling (Cosine Index)	(1)	(2)	(3)	(4)	(5)	(6)	(7)
(2) Bibliographic Coupling (Inclusion Index)	0.889						
(3) Co-Citations (Cosine Index)	0.360	0.291					
(4) Co-Citations (Inclusion Index)	0.253	0.200	0.789				
(5) Co-Words (Cosine Index)	0.515	0.455	0.277	0.213			
(6) Co-Words (Inclusion Index)	0.492	0.448	0.250	0.191	0.966		
(7) SAOs	0.564	0.457	0.254	0.172	0.579	0.551	
(8) (self) citation link (dummy)	0.264	0.222	0.306	0.298	0.161	0.149	0.137

Using Cronbach's alpha, we test whether all five measures, namely bibliographic coupling, co-citation analysis, co-word analysis, SAO structures, and citation links can be described as a composite indicator of similarity. Two datasets are created: one for the computations involving the Cosine Index, another for those involving the Inclusion Index. As it can be seen in Table 4, Cronbach's alpha is in the order of 0.5 for the five similarity measures, indicating that the reliability of a composite indicator of similarity, comprising all five different measures, would be relatively low. This holds true regardless of the dataset (i.e. similarity index) chosen.

Exploratory factor analysis is carried out next for the similarity measures. We employ Maximum Likelihood estimation and use, due to the skewed nature of our data, robust standard errors and a mean- and variance-adjusted chi-square test statistics implemented in the software package Mplus. The results differ depending on the similarity measure employed. When using the Cosine Index, SAO structures represent one factor, and co-word analysis, co-citation analysis, and citation links represent another, while bibliographic coupling seems to represent something different. Looking at the Inclusion Index data, SAO structures and co-word analysis are grouped as one factor, and citation links represent the other factor, while bibliographic coupling and co-citation analysis cannot clearly be assigned to any factor. These results are puzzling

because neither the inspection of Figures 2–6 nor of the correlation data in Table 3 suggests such a large difference between the two similarity measures. The implication is that the differences between using the two similarity measures are larger than one would expect.

Table 4. Rotated components matrix with factor loadings of the exploratory factor analysis for similarity measures

	Cosine Index		Inclusion Index	
	factors		factors	
	1	2	1	2
SAO structures (separate index)	0.751	–0.102	0.101	0.709
Co-word analysis	–0.004	0.644	0.096	0.535
Bibliographic coupling	0.131	0.002	0.238	0.485
Citation link (dummy)	–0.095	0.625	0.816	0.061
Co-citations	–0.010	0.485	0.351	0.203
Chi-square (mean and variance adjusted)	0.000		0.000	
df	1		1	
p-value	0.9968		1	
Cronbach's alpha (all five variables)	0.499		0.523	

Estimation: Maximum Likelihood with robust standard errors and mean- and variance-adjusted chi-square test statistic for not normally distributed data.
Rotation method: Varimax.

In general, both backward and forward citations suffer from an informant bias: while in this case backward citations (as well as the content written by the common author and expressed by words and SAO structures) reflect the knowledge base of the common author under consideration, forward citations reflect a heterogeneous knowledge base of the scientific community within the field: different scientists have divergent knowledge about the scientific progress in their (and adjacent) fields, hence they should show heterogeneity in the propensity to cite existing literature. Under this assumption, the citation links presenting forward self-citations should not suffer from an impact of knowledge heterogeneity, but in fact, they show a somewhat similar pattern as the total forward citations in the co-citation analysis do.

The main reason for the discrepancy found in our analysis should be that citations represent knowledge flows describing topics based on the papers cited, but further developed to new concepts through combination of knowledge from various sources. Such further development represents another type of similarity that cannot be expressed by semantics. HARTER & AL. [1993] came to the same conclusions after comparing similarities calculated on the basis of descriptors and measured by the Jaccard Index with citation links.

To date, semantic analyses encounter substantial difficulties in comparing the content of documents. A simple co-word analysis is only able to recognize a superficial level of similarity since it is limited to the exact type of words used by the author.

Including linguistic rules and thesauri will enhance the capabilities of the tools, but certainly they will be less able to describe similarities as can be recognized by human beings. If an author cites an article, he or she can transfer the content to a meta-level and compare the documents, making a reference that describes a certain degree of similarity when appropriate.

In contrast to co-citation analysis and citation links, citation networks drawn with spring-embedding algorithms should only then represent similarities comparable to semantic analysis if they comprise a relatively high network density, meaning that the papers are grouped relative to other ones within the citation space, while this relative relationship turns into a similarity measure. This view is supported by SMALL & GRIFFITH [1974], who could link dense areas within a citation network to scientific specialties.

Why do the findings discussed so far not hold true for bibliographic coupling to the same extent? This backward citation-based methodology lies somewhere between the two forward citation-based analyses and the semantic approaches. The reason lies in the skewed nature of citation data. Many documents are never cited, and many receive only very few citations. So it is not possible to properly calculate similarities between them based on co-citation data. However, often they contain backward citations referring to more highly cited papers. Therefore, there is simply much more data available for computing similarities based on such backward references, in a similar order of magnitude than for co-word analysis. This should be the major reason why bibliographic coupling tends to yield results sharing characteristics of both semantic and citation-based approaches.

Conclusions

It could be shown that for identifying similar contents in a variety of documents the Inclusion Index should be preferred over the Cosine or Jaccard Index. This holds true not only when computing the similarity based on words, etc., but for citation data in particular. Additionally, different similarity measures were compared graphically, including citation networks. It could be seen that the different methodologies clearly reveal different pictures of the research landscape. Factor analysis uncovered that the similarity measures used in this paper relate to two different constructs: assuming that the Inclusion Index is preferred, on the one hand co-word analysis and SAO structures seem rather to represent semantic similarity, while on the other hand there is substantial heterogeneity with the citation-based measures that are based on knowledge flows.

Future research could not only expand the scope of this paper towards a larger dataset comprising papers of different authors and scientific fields, it could also test the difference in the results when applying similarity measures on parts of a document such as abstracts (available, for instance, in the SCI), conclusions, or full-texts. A co-word

analysis based on a papers' full-text, not solely the abstract, would be an example. In addition, various other similarity indices could be used that take into account the occurrence of single words, e.g. comparing ceteris paribus the results of the Inclusion Index with the index introduced for the SAO structures. Future work could also address the "optimal" item threshold level for co-word analysis, co-citation analysis, and bibliographic coupling under given similarity indices.

*

The authors would like to thank Martin G. Moehrle for discussions on similarity measures and Adam Bartkowski and two anonymous referees for comments on an earlier draft of this paper.

References

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,* 54 : 550–560.

BARTKOWSKI, A., HILL, J., LÜHR, C., SCHRAMM, R. (2004), Rationelle Patentrecherche und Patentanalyse. In: R. SCHRAMM, S. MILDE (Eds), *PATINFO 2004 Patentrecht und Patentinformation – Mittel zur Innovation.* pp. 177–204.

BERGMANN, I., BUTZKE, D., WALTER, L., FUERSTE, J. P., MOEHRLE, M. G., ERDMANN, V. A. (2007), Evaluating the Risk of Patent Infringement by Means of Semantic Patent Analysis: The Case of DNA Chips, *Proceedings of the R&D Management Conference,* Bremen, July 4-6, 2007.

BLANCHARD, A. (2007), Understanding and customizing stopword lists for enhanced patent mapping, *World Patent Information,* 29 : 308–316.

BOERNER, K., CHEN, C., BOYACK, K. W. (2003), Visualizing knowledge domains, *Annual Review of Information Science and Technology,* 37 : 179–255.

BORGATTI, S. P., EVERETT, M. G., FREEMAN, L. (1999), Ucinet 6 for Windows - Software for Social Network Analysis, Harvard, MA: Analytic Technologies.

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 : 155–205.

CLARKSON, G. (2004), Objective Identification of Patent Thickets: A Network Analytic Approach, Harvard Business School Doctoral Thesis http://www.si.umich.edu/stiet/researchseminar/Fall%202004/Patent%20Thickets%20v3.9.pdf.

DREßLER, A. (2006), *Patente in technologieorientierten Mergers und Acquisitions,* Dt. Univ.-Verl, Wiesbaden.

GOLBECK, J., MUTTON, P. (2006), Spring-embedded graphs for semantic visualization. In: V. GEROIMENKO, C. CHEN (Eds), *Visualizing the Semantic Web - XML-based Internet and Information Visualization.* Springer, pp. 172–182.

HAMERS, L., HEMERYCK, Y., HERWEYERS, G., JANSSEN, M., KETERS, H., ROUSSEAU, R., VANHOUTTE, A. (1989), Similarity measures in scientometric research: the Jaccard index versus Salton's cosine formula, *Information Processing and Management,* 25 : 315–318.

HARTER, S. P., NISONGER, T. E., WENG, A. (1993), Semantic relationships between cited and citing articles in library and information science journals, *Journal of the American Society for Information Science,* 44 : 543–552.

INVENTION MACHINE CORPORATION (NO DATE), Accelerating the speed of knowledge, White Paper, http://lsdis.cs.uga.edu/SemWebCourse_files/WP/Invention_Machine.pdf (March 09, 2007).

JACCARD, P. (1901), *Bulletin del la Société Vaudoisedes Sciences Naturelles,* 37 : 241–272.

JARNEVING, B. (2005), A comparison of two bibliometric methods for mapping of the research front, *Scientometrics,* 65 : 245–263.

KAMADA, T., KAWAI, S. (1989), An algorithm for drawing general undirected graphs, *Information Processing Letters,* 31 : 7–15.

KESSLER, M. M. (1963), Bibliographic coupling between scientific papers, *American Documentation,* 14 : 10–25.

LEYDESDORFF, L. (1987), Various methods for the mapping of science, *Scientometrics,* 11 : 295–324.

MARSHAKOVA, I. V. (1973), System of document connections based on references, *Scientific and Technical Information Serial of VINITI,* 6 : 3–8.

MOEHRLE, M. G., WALTER, L., GERITZ, A., MÜLLER, S. (2005), Patent-based inventor profiles as a basis for human resource decisions in research and development, *R & D Management,* 35 : 513–524.

PETERS, H., BRAAM, R., RAAN, A. (1995), Cognitive resemblance and citation relations in chemical engineering publications, *Journal of the American Society for Information Science,* 46 : 9–21.

PORTER, M. (1980), An algorithm for suffix stripping program, *Program,* 14 : 130–137.

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 : 166–180.

RAMLOGAN, R., MINA, A., TAMPUBOLON, G., METCALFE, J. (2007), Networks of knowledge: The distributed nature of medical innovation, *Scientometrics,* 70 : 459–489.

RIJSBERGEN, C. V. (1979), *Information Retrieval,* Butterworth, London.

RIP, A., COURTIAL, J. (1984), Co-word maps of biotechnology: An example of cognitive scientometrics, *Scientometrics,* 6 : 381–400.

SALTON, G., MACGILL, M. J. (1983), *Introduction to Modern Information Retrieval,* McGraw-Hill, New York.

SHARABCHIEV, J. T. (1989), Cluster analysis of bibliographic references as a scientometric method, *Scientometrics,* 15 : 127–137.

SMALL, H., GRIFFITH, B. C. (1974), The structure of scientific literatures I: Identifying and graphing specialties, *Science Studies,* 4 : 17–40.

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 : 265–269.

STERNITZKE, C., BARTKOWSKI, A., SCHRAMM, R. (2007), Regional PATLIB centres as integrated one-stop service providers for intellectual property services, *World Patent Information,* 29 : 241–245.

TSOURIKOV, V. M., BATCHILO, L. S., SOVPEL, I. V. (2000), Document semantic analysis/selection with knowledge creativity capability utilizing subject-action-object (SAO) structures, *United States Patent No. 6167370*.