

Concept symbols revisited: Naming clusters by parsing and filtering of noun phrases from citation contexts of concept symbols¹

JESPER W. SCHNEIDER

Royal School of Library and Information Science, Department of Information Studies, Aalborg (Denmark)

The present study presents a semi-automatic method for parsing and filtering of noun phrases from citation contexts of concept symbols. The purpose of the method is to extract contextual, agreed upon, and pertinent noun phrases, to be used in visualization studies for naming clusters (concept groups) or concept symbols. The method is applied in a case study, which forms part of a larger dissertation work concerning the applicability of bibliometric methods for thesaurus construction. The case study is carried out within *periodontology*, a specialty area of dentistry. The result of the case study indicates that the method is able to identify highly important noun phrases, and that these phrases accurately describe their parent clusters. Hence, the method is able to reduce the labour intensive work of manual citation context analysis, though further refinements are still needed.

Introduction

A common challenge in literature visualization studies is how to interpret the actual mappings of document entities. Dimensionality and link reduction algorithms are typically applied to investigate co-citation networks for their salient structures (e.g., BÖRNER, CHEN, & BOYACK, 2003). However, it greatly enhances the interpretability of the resulting mappings, if co-citation networks are somehow transformed into some sort of conceptual networks (e.g., SMALL, 1986). In this paper, we introduce a semi-automatic parsing method designed to transform a document co-citation network into a conceptual network of noun phrases.

Most often in document co-citation analyses, the aggregate clusters of cited references are named by single words (WHITE & MCCAIN, 1989; 1997; WILSON, 1999). The process of naming clusters is usually automatic. Specific entities, from documents *citing* the individual members of a cluster, are extracted and subsequently subjected to a frequency analysis. Consequently, the most frequently occurring citing document entities in the research front are used to name the topic(s) or concept(s) of the cluster (e.g., WHITE & MCCAIN, 1989; WILSON, 1999). It is important to emphasize that the automatic extraction of entities typically means extraction of single word entities, and not multiple word entities, such as noun phrases. A notable exception is the studies by NOYONS and colleagues (e.g., 1999), where noun phrases are extracted from titles and abstracts of citing papers.

For practical reasons, the composition of document representations in the citation databases of ISI, usually determine the entities available for naming clusters (WHITE & MCCAIN, 1989). The most commonly used of these entities are title words or ISI's special subject categories. Conversely, domain dependent databases are often used in co-word studies (HE, 1999). Further, domain dependent databases can also be used in conjunction with citation databases in document co-citation studies. In the latter case, the same bibliographic reference is identified in the citation database, as well as in the domain dependent database (INGWERSEN & CHRISTENSEN, 1997). The document representation of the citation database provides the references needed for the co-citation analysis, whereas, the domain specific indexing descriptors and classification codes, for the same document

¹ The present article is an extended and revised edition of a paper presented at the ISSI 2005 conference in Stockholm, Sweden, July 24-28, 2005.

representation, can be obtained from the domain dependent database (INGWERSEN & CHRISTENSEN, 1997). As a result, the latter can be used for naming or evaluation of the generated co-citation clusters.

The seminal work by SMALL (1978) is a more sophisticated approach to the transformation of document co-citation networks into conceptual networks. SMALL (1978) established that highly cited documents symbolize concepts to those who cite them. While it has long been known, that when references are turned into citations they can be construed as subject headings (e.g., GARFIELD, 1974), different people may construe the same cited document differently. SMALL (1978) showed, however, that citing authors in chemistry tend to be both specific and highly uniform in the meanings they assign to cited documents, as revealed by the contexts of the references. Scientists tend to give earlier works consensual meaning by 'piling up' identical or similar words and phrases in the sentences in which their citation markers are embedded (SMALL, 1978). Consequently, when citation contexts show that citing authors have used a cited document to stand for a given idea more or less uniformly over many papers, the document has, according to SMALL (1978), attained the status of a *concept symbol*. Accordingly, the highly cited document communicates a specific topic and resembles a subject heading or descriptor. The focus of *citation context analysis* is naming of individual cited references. As a result, the basis for naming their aggregate clusters is the common concept(s) identified among the member concept symbols. It is believed that naming references and clusters by use of citation context analysis ensures contextual and pertinent phrasal concepts (e.g., SMALL, 1986).

Unfortunately, citation context analysis is usually labour intensive work, where the full text of citing documents is manually scanned to identify citation contexts in order to select key phrasal concepts that capture major aspects of the cited documents (e.g., SMALL, 1978; 1986; REES-POTTER, 1989). Research by O'CONNOR (1982; 1983) shows that citation contexts can be identified automatically within the structure of full text documents. Nevertheless, O'CONNOR (1982; 1983) only extracted single words from the citation contexts. Furthermore, SMALL (1979) has pointed out that it is unlikely that the identification process of concept symbols can be done entirely automatically. According to SMALL (1979), a computer cannot recognize unforeseen synonymy, thus, the words and phrases that show consensus on what documents symbolize must therefore be recognized by a human reader.

The purpose of the present paper is to introduce a semi-automatic method for extraction of noun phrases by natural language parsing of citation contexts in citing documents. This is different from the studies by NOYONS and colleagues (e.g., 1999) mentioned above, where phrases are extracted from titles and abstracts of citing papers. A subsequent frequency analysis and filtering procedure create a portfolio of important noun phrases, which is attached to each of the highly cited references. The portfolio of noun phrases constitutes the basis used to characterize the cited references and eventually naming their parent clusters. Accordingly, the method transforms a document co-citation network into a conceptual network of noun phrases.

The aim of the present work is to improve the single word naming of clusters, by use of noun phrases instead. We believe that noun phrases more accurately describe the topic(s) and concept(s) of a cluster and its individual members. As demonstrated by SMALL (1978), such noun phrases should be extracted from citation contexts of citing documents. Further, since these noun phrases are contextual, they probably reflect consensus usage of terminology, accordingly, they resemble agreed upon indexing descriptors. The aim is therefore to develop and explore a method that can reduce the labour intensive work in connection with citation context analysis. We believe that parsing of citation contexts reduces the workload and eventually improves the process of naming clusters and their individual members in mapping studies.

The paper is composed of four main sections. Section two presents the integrated method of document co-citation analysis, visualization, and citation context analysis used for construction of the conceptual network. Section three outlines the main results. Finally, in section four we discuss the main findings.

Method

The present study derives from a comprehensive dissertation work that investigates the applicability of bibliometric methods for semi-automatic thesaurus construction (SCHNEIDER, 2004; SCHNEIDER & BORLUND, 2004). Only the basic methodical steps and results are outlined here. Readers are referred to SCHNEIDER (2004) for a more detailed description of the method and its results.

In this study, we apply an integrated method of *document co-citation analysis*, *visualization*, including complete-link cluster analysis and Pathfinder network scaling, and *citation context analysis*, including semi-automatic parsing of citation contexts. The study is based on bibliographic data retrieved and downloaded from *Science Citation Index*® (SCI®) hosted by *Dialog*®. The bibliographic data contain 801 research and review papers published within *periodontology* in 2001. *Periodontology* is a specialty area within dentistry.

For practical reasons pertaining to the dissertation work, a citation threshold value of 13 is decided upon. Due to the chosen threshold value, the 64 most highly cited references within *periodontology* in 2001 are selected for the co-citation analysis.

Co-citation analysis and visualization

A central point of the present study is to create coherent *concept groups* that contain semantically related concept symbols. That is, clusters of significantly highly co-cited reference pairs. As demonstrated by SMALL & GREENLEE (1980), highly cited references that are not co-cited a significant number of times with other selected cited references, can cause problems to the subsequent co-citation clustering. To minimize the possibility of singletons occurring during the clustering procedure, relative co-citation strengths are calculated by use of the Jaccard proximity measure (SMALL & GREENLEE, 1980). The threshold value is set to 0.16. The threshold reduces the co-citation matrix to 45 cited references. All references thus participate in at least one significant co-citation relation on or above the 0.16 threshold. The reduced co-citation matrix forms the basis for the subsequent complete-link cluster analysis, as well as the *Pathfinder network* scaling. Complete-link clustering is preferable if conceptually solid clusters are wanted (e.g., SPARCK JONES, 1971).

The purpose of the cluster analysis is to highlight the salient topics within *periodontology*, as reflected in the citing literature for the year 2001. In the present case, such clusters are termed *concept groups*, as we assume that they are an aggregate of their member concept symbols. To obtain a more intuitive understanding of concept group compositions, and their structural relations, the *Pathfinder network* algorithm (SCHVANEVELDT, 1990) is applied. *Pathfinder network* scaling simplifies the co-citation matrix by retaining the strongest co-citation links with reference to the triangle inequality condition (SCHVANEVELDT, 1990).

Citation context analysis and parsing

The assumption behind the present study is that terminology used in the citation contexts of citing papers reflect upon concepts in a specialty area due to the notion of concept symbols (SMALL, 1978). In the dissertation work, we focus on the ability of citation context analysis

and noun phrase parsing for the selection of candidate thesaurus descriptors (SCHNEIDER, 2004). Descriptors reflect standardized, agreed upon, terminology in a specialty area; hence, they might as well be used to name concept symbols (cited references) and concept groups (cluster) in traditional literature mapping studies. Consequently, a prerequisite for the present study is that highly cited references within *periodontology* largely act as concept symbols.

The basic procedures of the present citation context analysis include: 1) determination and identification of citation contexts, 2) parsing of citation contexts in order to extract portfolios of noun phrases, 3) Filtering of noun phrases and naming of concept groups.

A set of rules is developed to define a citation context within a citing paper for the purpose of incorporation into the present study. Two conditions should be accentuated; for an extensive description see SCHNEIDER (2004). First, conventionally a citation context is designated according to a certain sentence limit from or surrounding the reference marker (SMALL, 1978; 1986; O'CONNOR 1982; 1983). We apply a less rigid citation context limitation, as all citing documents from the 2001 sample used in the present analysis are available in electronic form. The electronic form makes the manual process of citation context selection less cumbersome and swifter. As a rule of thumb, we aim at citation contexts with the least possible number of sentences, but still sufficient for construction of a meaningful citation context. Most of the time, one to three sentences suffice, but in some cases we surpass three sentences in order to construct a meaningful citation context. This procedure can be automated though fixed rules are needed, which makes the process less flexible (O'CONNOR, 1982; 1983). Since our study is exploratory, we want to be sure that the contexts indeed reflect upon the content of the cited document.

Second, contrary to REES-POTTER (1989) and in accordance with SMALL (1986), we do not discard so-called 'perfunctory' or redundant references. 'Perfunctory' references are cited references mentioned or acknowledged but not explicitly described in a citation context (MORAVCSIK & MURUGESAN, 1975). We believe that 'perfunctory' or redundant references can serve a function in relation to the present study. Such references may, for example, signal simultaneous and independent discovery, or indicate the availability of more than one good source for the same concept (SMALL, 1982). For example, the 2001 sample of citing papers from the specialty area of *periodontology* have a mean of 39 references per paper, with a mean paper size of nine pages (median of eight). This indicates a high consumption of cited references within *periodontology*. There seems to be a logical pattern in the consumption, as especially the reporting of experimental studies is divided into several small-sized journal papers. For that reason, an active publication pattern can be observed. It is therefore apparent that some of these papers are so closely related in origin, that when later citing authors conceptualize their findings and express them in more general terms, he or she may very well need to refer to several of these cited references. This leads to citation contexts containing what would immediately be regarded as 'perfunctory' or redundant references. Consequently, it may very well be that two or more related cited references jointly denote the same concept symbol. The purpose of the present citation context analysis is to identify terminology preferably attached to individual cited references. By allowing 'perfunctory' references, it becomes difficult to distinguish which words and phrases belong to which references in a citation context. But, it is assumed that the sample of citation contexts eventually yield sufficient information concerning the individual cited references, in order to judge whether they are concept symbols or not. If a reference is 'perfunctory' in several citation contexts, but the terminology used in these contexts are agreed upon, then it is fair to assume that the cited reference is semantically related to this terminology usage, 'perfunctory' or not. Most importantly, by allowing 'perfunctory' references, we save time in an otherwise time consuming analysis.

For the citation context analysis, a minimum number of five citing papers are chosen for each cited reference. This entails that at least five citation contexts are investigated and parsed for each cited reference, in order to identify concept symbols and extract contextual and pertinent noun phrases. Notice, the procedure deliberately does not take into account multiple citing of a reference within a document. For example, initially five different citing papers for cited reference *X* are located. The citation contexts to reference *X* is subsequently identified in these five citing papers. However, some of the papers may have cited reference *X* several times, thus, the number of citation contexts to cited reference *X* extends beyond the initial five. The threshold value of *at least* five implies a larger probability of highly co-cited references to emerge in several ‘extra’ citation contexts, contrary to ‘lower’ co-cited references. The rationale behind the threshold of at least five citation context is twofold. First, we expect that the majority of cited references eventually will attain more than five citation contexts due to the sampling procedure. Secondly, we suppose that a minimum of five citation contexts may be a sufficient number for identification of concept symbols. Consequently, the present sampling procedure is less time consuming compared to previous attempts (e.g., SMALL, 1978; REES-POTTER, 1989).

Eventually, 88 citing documents were needed to obtain at least five citation contexts for each cited reference. Consequently, all citation contexts in the 88 citing documents that referred to at least one of the study’s 45 highly cited references were manually extracted from the electronically stored documents. The 88 citing documents produce a sample of 580 citation contexts. That is an average of 12.9 citation contexts per cited reference with a median of 11. The highest number of contexts to one cited reference is 34 and the lowest is five. Further, only two cited reference ended up having the minimum number of five citation contexts attached to it.

Two important assumptions pertaining to the present method are the consensus usage of terminology in the citation contexts of concept symbols, and the conceptual importance of this terminology. In order to extract pertinent noun phrases from the citation contexts, we therefore need to verify whether the cited references in fact act as concept symbols within the investigated specialty area. For this purpose, we apply a modified version of the ‘consensus passage’ procedure introduced by SMALL (1986). The ‘consensus passage’ procedure investigates single word usage in citation contexts. A ‘consensus passage’ score expresses the degree of terminological consensus usage of single content words among a cited reference’s citation contexts. No attention is paid to the actual meaning of these content words. A good score indicates agreement on terminology, which indicates that the cited reference acts as a concept symbol. Consequently, the cited reference’s citation context with the highest score, i.e. its ‘consensus passage’, very likely contains phrases or sentences that express the conceptual meaning (SMALL, 1986). For a detailed description of this comprehensive examination, see SCHNEIDER (2004).

Of interest to the present study, is whether highly cited references within the specialty area of *periodontology* show an inclination to act as concept symbols, similar to a number of other scientific domains (e.g., SMALL, 1978). The ‘consensus passage’ score is therefore used to determine if a cited reference acts as a concept symbol and the ‘consensus passage’ is used to help clarifying its conceptual meaning.

The purpose of the present study is to investigate whether noun phrase parsing of citation contexts, where a portfolio of the most frequently occurring noun phrases are extracted and attached to a cited reference, can improve the otherwise cumbersome process of determining the meaning of concept symbols. The most profound extension of the present study from former visualization studies is therefore the novel step of natural language parsing of noun phrases from the citation contexts. The rationale for the use of noun phrases as key

conceptual phrases is that they represent more meaningful concepts than individual words (e.g., ANICK & VAITHYANATHAN, 1997). Noun phrases are believed to be content bearing units and thus good indicators of the content of a text. As an example, noun phrases are widely used across sublanguage domains to describe concepts succinctly. It is therefore appropriate to identify and extract such phrases from the citation contexts of citing papers, in order to identify agreed upon terminology to be used for naming concept symbols or concept groups. We use the advanced noun phrase parser *Connexor* (www.connexor.com)² to select phrases from the citation contexts of citing documents. *Connexor* is a shallow syntactic parser based on a functional dependency grammar that produces part-of-speech tags, noun phrase markers, and relational dependencies between constituents. *Connexor* is a continuation of the older *NPtool* noun phrase parser, which was specifically developed for automatic indexing purposes (VOUTILAINEN, 1993). The application of parsing techniques for phrase extraction in connection with citation context analysis is very different from former approaches (SMALL, 1978; 1986; O'CONNOR, 1982; 1983). The application of a noun phrase parser ensures identification of phrases that represent concepts in a more meaningful way than individual words alone, and it reduces the workload of manual citation context analysis.

The citation context analysis and the parsing procedure are illustrated in Figure 1.

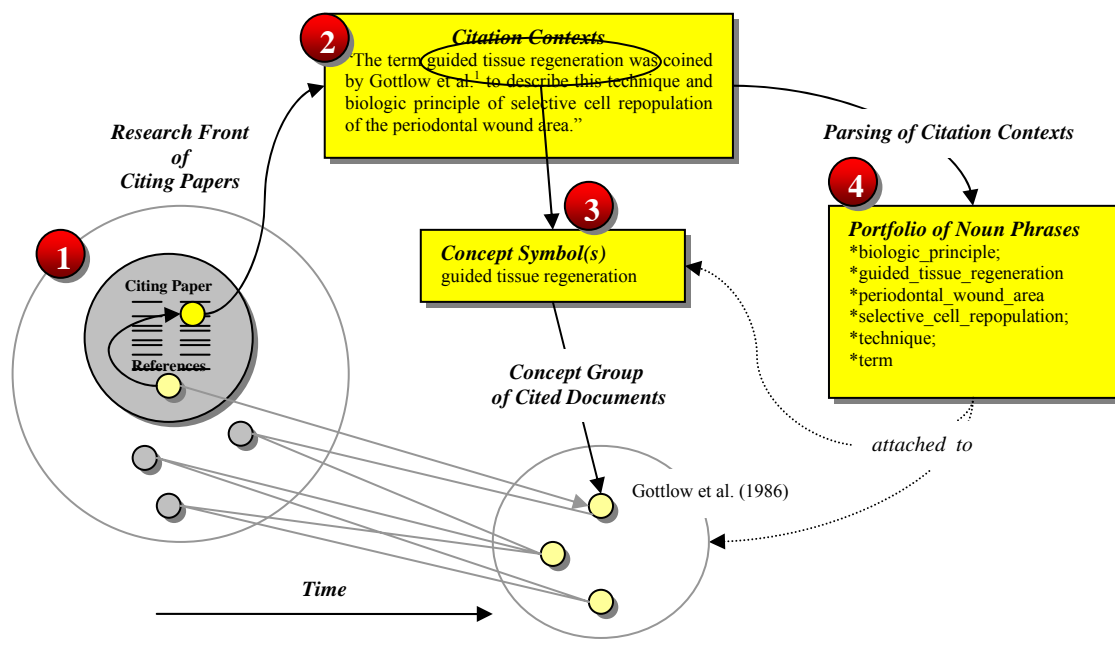


Figure 1. Citation context analysis and parsing process.

The numbered circles indicate how to read the successive steps. Figure 1 depicts the relationship between citing papers in a research front and their intellectual base of earlier cited references in a cluster. The highlighted citing paper (1) in the research front refers back to a highly cited reference (i.e., GOTTLOW ET AL. (1986)). The citation context for this reference is identified in the citing paper (2). By comparing at least five citation contexts for the same cited reference, it is possible to establish whether this reference is a concept symbol to later citing authors. The degree of consensus usage of terminology is what determines the status of the cited reference. The concept symbol is marked in the citation context; in this case, the cited reference (GOTTLOW ET AL., 1986) symbolizes the concept of *guided tissue regeneration* (3). Moreover, the concept symbol 'is brought back' to its cited reference in the cluster in

² Parsing was carried out at the University of Tampere, Department of Information Studies in Finland.

order to transform the cluster into a group of concepts, where the meanings common to all concept symbols give name to the concept group. Finally, the citation contexts of the concept symbols are parsed in order to extract their attached noun phrases (4). Eventually, each concept symbol will have a number of noun phrases attached to them based on frequency analyses.

As a final step, a filtering procedure designates the most prevalent noun phrases among the portfolios within a concept group. The filtering procedure is essentially a normalization procedure that identifies primary and secondary noun phrases. Accordingly, the most prevalent noun phrases are used to name a concept group.

Eventually, the aim is to investigate to which extent, the cumbersome process of manual identification of concept symbols can be substituted by noun phrase parsing and creation of portfolios. In order to verify the appropriateness of the extracted noun phrases in naming the concept groups, we use the quantitative indicator of semantic coherence introduced by BRAAM, MOED & VAN RAAN (1991).

Results

This section presents the result of the cluster analysis, which is visualized as a *Pathfinder network*. The section also exemplifies the process of naming concept groups, as we use concept group 1 in an illustrative case. Finally, because of the naming procedure, the citation network illustrated in Figure 2 is transformed into a conceptual network in Figure 5.

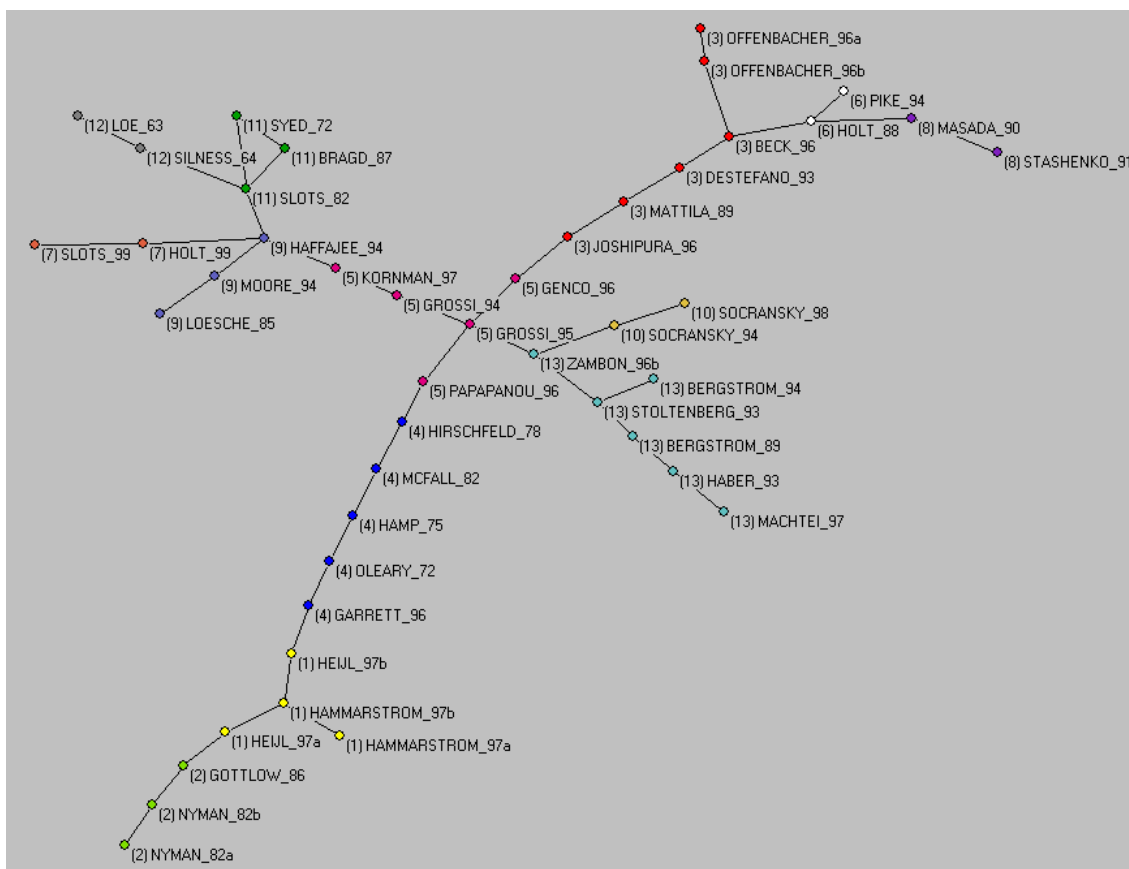


Figure 2. Pathfinder network visualization of the clustered co-citation network.

For the sake of simplicity, only concept group names are visualized. As indicated, one can zoom in on the individual concept symbols within a group to see their attached portfolio

of noun phrases. The hierarchical cluster analysis is truncated at 13 clusters. The smallest clusters contain two references and the largest clusters contain six references. The cluster result is visualized below in Figure 2 by use of *Pathfinder network* scaling and Pajek (BATAGELJ & MRVAR, 1998).

As mentioned above, we apply the ‘consensus passage’ procedure in order to verify whether the individual cited references (illustrated in Figure 2) act as concept symbols (SCHNEIDER, 2004). In the present case study, of 45 cited references, 42 act as concept symbols. Three cited references were discarded for further analysis: OFFENBACHER_96a in concept group 3, as well as HIRSCHFELD_78 and MCFALL_78 from concept group 4. We have nonetheless, strong indications that consensus terminology can be identified for these references *if* the sample of citation contexts is enlarged. Eventually, the common concepts referred to by all members of the two concept groups in question are indicated to some degree in several citation contexts to these references; however, not sufficiently to warrant a designation as concept symbols from the outlined rules in SCHNEIDER (2004).

For validation purposes, the conceptual meaning of these concept symbols is deduced from their ‘consensus passage’ in combination with their portfolio of agreed upon noun phrases. Appendix 1 outline the conceptual meaning of the 42 concept symbols.

A *mean* ‘consensus score’ is often used to characterize how well the cited references as a set act as concept symbols. The mean ‘consensus score’ measures the degree of consensus among the citing terminology. The mean ‘consensus score’ for the present study is 0.52 (median is 0.56) (SCHNEIDER, 2004, p. 255). This we consider a good score that indicates consensual usage of terminology, as it corresponds to the exemplary result of 0.48 from the study of *leukaemia viruses* by SMALL (1986, p. 102). The result demonstrates that highly cited references within the specialty area of *periodontology* show an inclination to act as concept symbols, thus, the assumptions for the present study is fulfilled. We utilize this knowledge to compare the naming of concept groups by use of noun phrase parsing in relation to the manually deduced interpretations from the list of concept symbols illustrated in Appendix 1..

The distribution of citation contexts and ‘consensus passages’ within the document structure

A cited reference’s ‘consensus passage’ is the most characteristic citation context in its sample. Here we find the most agreed upon single content words among the sample citation contexts. No attention is paid to the actual meaning of these content words. However, if a cited reference acts as a concept symbol, then the chosen ‘consensus passage’ very likely contains phrases or sentences that express the concept. This makes the ‘consensus passage’ attractive for noun phrase parsing and eventually identification of concept symbols.

It is therefore interesting to study where ‘consensus passages’ are identified in the document structure, and to compare the proportion of ‘consensus passages’ to the sample of citation contexts distributed among the individual document sections. What is interesting is whether the distribution of the sample of citation contexts among the different document sections, resembles the distribution of the 45 statistically derived ‘consensus passages’. Such a comparison can give an indication of where we can expect to recognize the most agreed upon concept symbols in the document structure. Figure 3 below shows the proportion of citation contexts to the 45 ‘consensus passages’, distributed among the individual document sections.

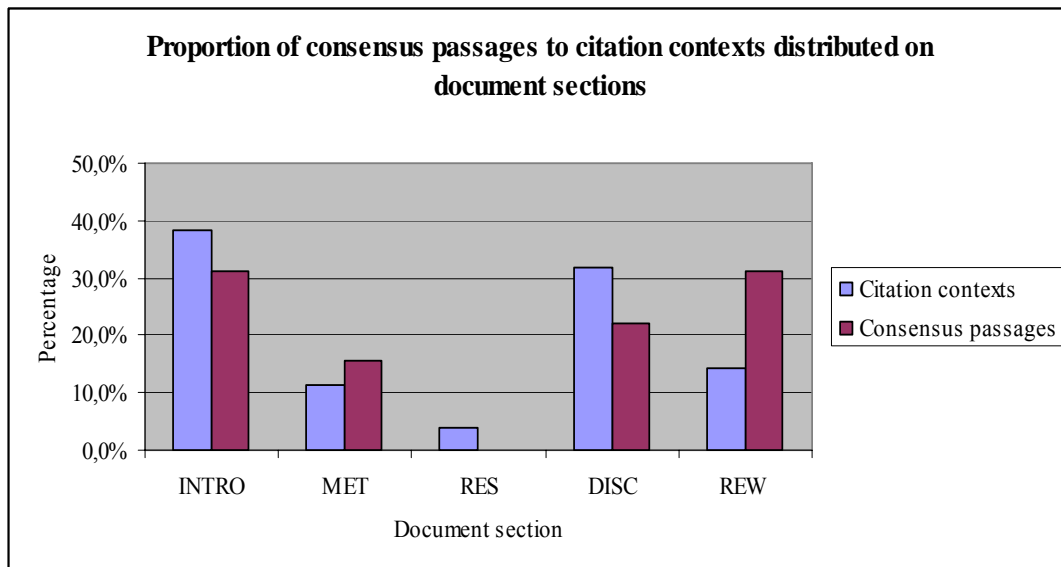


Figure 3. Proportion of consensus passages to citation contexts distributed among document sections (SCHNEIDER, 2004).

The distribution pattern of the sample of citation contexts among the individual document sections is expected, as most citation contexts are found in the *introduction* section (cf. MCCAIN & TURNER, 1989). From Figure 3, we can observe that the *method* and *review*³ sections produce relatively more ‘consensus passages’ than we would expect from the distribution of citation contexts. Conversely, the *introduction* and *discussion* sections produce fewer ‘consensus passages’ than we would expect. The reason for these findings is that higher ‘consensus’ scores are obtained from citation contexts extracted from *method* and *review* sections. Consequently, these sections produce the most recognizable concept symbols.

Figure 4, illustrates the *mean* ‘consensus passage’ scores distributed among the individual document sections.

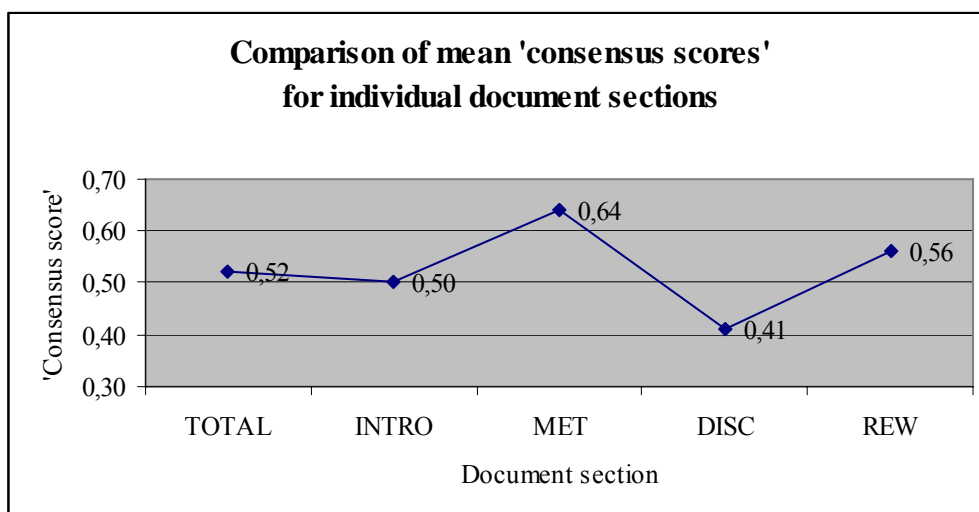


Figure 4. Comparison of mean ‘consensus scores’ for individual document sections. TOTAL corresponds to mean for all document sections combined.

³ In this context, ‘review section’ denotes the review section in review papers. Review papers in periodontology do not follow the traditional document form used in experimental and theoretical papers; instead, they have a large review section.

The total mean score is 0.52, whereas the mean ‘consensus scores’ for *method* is 0.64 and *review* 0.54. Figure 4 substantiates the findings presented in Figure 3, however, the mean score for ‘consensus passages’ in the *review* section is not markedly higher than the total score. Consequently, Figure 3 and 4 indicate that the most recognizable concept symbols are located in the *method* section of citing papers in the present sample. This corresponds to the seminal findings by SMALL (1978). The *method* sections do produce very clear concept symbols that denote methods, techniques, or instruments applied in experimental studies. These concept symbols are not only easy to recognize from the ‘consensus passages’, they also appear very clearly in the portfolio of frequently occurring noun phrases (SCHNEIDER, 2004).

Notice that the *discussion* sections produce fewer ‘consensus passages’ than expected. Likewise, the mean ‘consensus score’ obtained from ‘passages’ extracted from the *discussion* section is markedly lower than the total mean. In the present study, the *discussion* sections typically consider very specific aspects of a cited reference, usually in relation to findings reported on in the citing paper. This implies that cited references often appear exclusively in the citation contexts appearing in the discussion sections. Such citation contexts do not always refer directly to the agreed upon concept symbol due to their specificity (SCHNEIDER, 2004). Accordingly, the ‘consensus passage’ from the *discussion* section is less useful as a target for phrase extraction in the present case.

Naming of concept groups

We use concept group 1 as an illustrative example of the naming procedure. Consider Table 1 below. The table contains data for concept group 1, which corresponds to cluster 1 in Figure 2. The concept group contains four significantly co-cited references. As indicated above from the ‘consensus passage’ investigation, these cited references act as concept symbols. The second column shows the portfolio of noun phrases extracted and attached to each of the four cited references. A citation context frequency is shown in connection with each noun phrase. The citation context frequency is a binary count of the number of citation contexts in which the noun phrase occurs. A threshold value of around $\frac{1}{3}$ of the sample size is applied. This implies that in order to be selected for the portfolio of frequently occurring phrases, a phrase must occur at least in $\frac{1}{3}$ of the citation contexts to a particular concept symbol (SCHNEIDER, 2004).

Table 1. Concept symbols of concept group 1 and their portfolios of noun phrases.

Concept group 1: Enamel matrix proteins	
Concept symbols	Portfolio of extracted noun phrases
HAMMARSTROM_97a	5 enamel_matrix_protein 4 enamel_matrix_derivative 3 periodontal_regeneration
HAMMARSTROM_97b	7 enamel_matrix_protein 2 regenerative_therapy
HEIJL_97a	7 enamel_matrix_protein 3 periodontal_regeneration
HEIJL_97b	16 enamel_matrix_derivative 14 clinical_attachment_level 12 gain 11 treatment

Finally, a filtering procedure is applied, which separates the noun phrases from the individual portfolios in two categories, primary and secondary phrases. The filtering procedure is based on a simple ‘document frequency analysis’. The number of citation contexts attached to the individual concept symbol influences the frequency count of noun phrases in the portfolios. Hence, the application of ‘document frequency’ at the portfolio level normalizes for the variances in frequencies of noun phrases between the portfolios. Consequently, primary phrases have higher frequencies across the portfolios than secondary phrases. This implies that primary phrases most likely appear in several portfolios within a group. Primary phrases, therefore, most likely reflect upon the common concept of the group. Conversely, secondary phrases are likely to reflect upon specific aspects of the common concept within a group. The result of the filtering procedure for concept group 1 is illustrated below in Table 2.

Table 2. Significant noun phrases in Concept group 1 after filtering.

Primary phrases:	3	enamel_matrix_protein
	2	periodontal_regeneration
	2	enamel_matrix_derivative
Secondary phrases:	1	treatment
	1	regenerative_therapy
	1	clinical_attachment_level
	1	gain [discarded]

Notice, an ‘inverse document frequency’ threshold is also invoked. This threshold creates a ‘negative dictionary’ of non-significant descriptors that occurs frequently across several of the concept groups. Such descriptors are discarded from the process of naming the concept groups. Accordingly, the noun phrases used to name concept group 1 are: enamel_matrix_protein, periodontal_regeneration, and enamel_matrix_derivative. The most significant noun phrase is **enamel_matrix_protein**.

In order to substantiate the naming of the concept groups, a quantitative indication of semantic coherence is also investigated (BRAAM, MOED, & VAN RAAN, 1991). The degree of semantic coherence is measured by comparing the individual portfolios of noun phrases attached to the concept symbols in a group, with an ‘aggregate portfolio’ of noun phrases that represents the entire concept group. The ‘aggregate portfolio’ is represented as a vector that comprises all different noun phrases appearing in the individual portfolios of the group’s concept symbols. The individual portfolios are likewise represented as vectors with a length that corresponds to their number of different noun phrases. A binary count is used to indicate whether a noun phrase is present or absent in the vector representations. Notice the ‘aggregate portfolio’ representing the concept group only contains presence counts, as it is a representation of all different noun phrases in the group. The binary form of the cosine measure is used to determine the similarity between the individual portfolios and the ‘aggregate portfolio’ of the concept group (OCHIAI, 1957). This implies that a similarity result is obtained for each portfolio attached to a concept symbol within a concept group. Eventually, the *average value* of the individual similarity scores indicates the semantic coherence within the concept group.

In their study, BRAAM, MOED & VAN RAAN (1991) obtained average cosine similarities in the range of 0.36 to 0.44, which they considered sufficient in order to conclude that their groups were coherent. Thus, we use their results as a baseline for evaluating the average cosine similarities computed for the present concept groups. Table 3 below, outlines the

semantic coherence scores for the individual concept groups, as well as manually interpreted names for the concept groups.

Table 3. The semantic coherence scores

Concept group no.	Name of concept groups	Semantic coherence
1	Enamel matrix proteins	0.620
2	Guided tissue regeneration	0.902
3	Complications of periodontal disease	0.650
4	Furcation involvement	0.548
5	Risk factors for periodontal disease	0.585
6	Periodontitis progression, p. gingivalis	0.762
7	Periodontal pathogen, p. gingivalis	0.787
8	Cytokines	0.707
9	Periodontal pathogens	0.623
10	Classification of bacteria	0.774
11	Periodontal pathogen, A. actinomycetemcomitans	0.638
12	Periodontal index	0.707
13	Risk factor of smoking	0.592

These names are deduced from the manually inferred concept symbols identified from the ‘consensus passage’ investigation in SCHNEIDER (2004). The manually inferred names are included in the present study only to indicate the resemblance or difference between the parsed and deduced phrases.

The result of the evaluation and naming of the concept groups is very clear. All groups are semantically coherent. This implies that the concept symbols within the concept groups unequivocally refer to some common concept. The ‘coherence scores’ in the present analysis, ranging from 0.585 to 0.902, are far above the results obtained by BRAAM, MOED & VAN RAAN (1991, pp. 240-241). Thus, compared to their results and subsequent conclusions, the present quantitative coherence results are very convincing. We believe that the aforementioned successive steps ensure that the extracted noun phrases are contextual, agreed upon, and therefore assumed to be important and semantically coherent.

The conceptual network visualized in Figure 5 below, shows the most important noun phrases for each concept group resulting from the parsing and filtering procedures. Noun phrases marked with bold in Figure 5, are the most significant among the primary phrases.

If we compare these noun phrases to the manually deduced names depicted in italic, we can see that there is a considerable overlap. However, the manually deduced names tend to be a bit broader in their conceptions. For example, concept group 12 is named *periodontal index* from the manual analysis. This is a broader term, which incorporates the parsed phrases *gingival index* and *plaque index*. Likewise, concept group 8 is named *cytokines*. This is also a broader term that incorporates the parsed phrases *tnf-alpha* and *interleukin-1beta*.

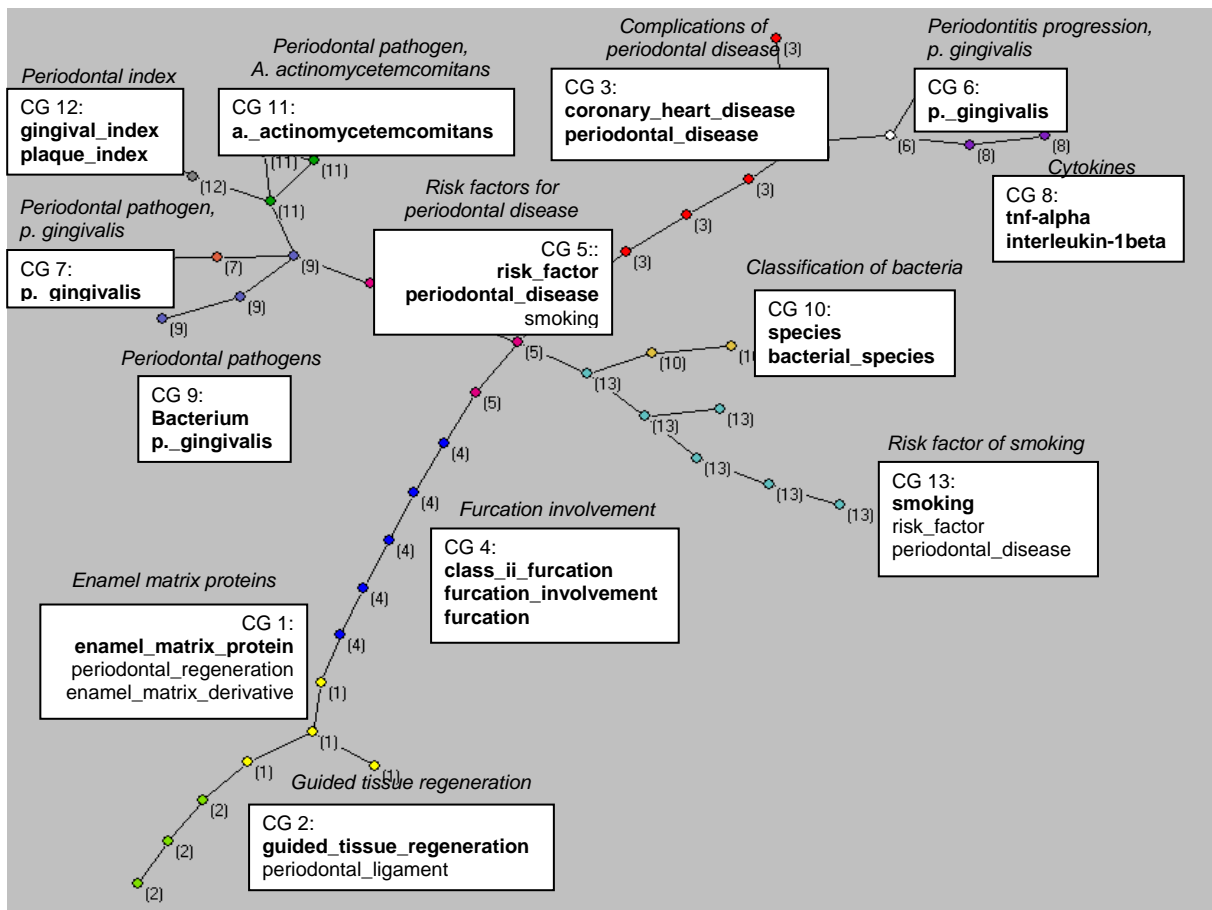


Figure 5. Visualization of conceptual network, concept groups are named by their most important noun phrases. Primary descriptors are indicated in bold and manually named concept groups depicted in Table 3 are indicated in italic.

Consequently, the parsed and filtered noun phrases are very specific, contextual, and thus pertinent descriptors that characterize their parent concept groups adequately. This can also be surmised from an investigation presented in SCHNEIDER (2004), where it is found that three out of every four parsed primary phrase corresponded to a MeSH® descriptor. This overlap is statistically significant. In addition, for validation purposes in SCHNEIDER (2004), concept groups are also named by single words extracted from titles of citing documents in the research fronts. Likewise, concept groups are also named by the most frequent *Medical Subject Headings® (MeSH®)* of the citing documents in the research fronts. The results of these two naming procedures are compared to naming concept groups by important noun phrases parsed from citation contexts. Not surprisingly, single words produce indistinct concept group names, and the most frequently occurring *MeSH®* descriptors most often produce too broad conceptions. The latter means that several concept group names are hard to distinguish (SCHNEIDER, 2004).

As a result, citation contexts of a concept symbol within *periodontology* contain agreed upon and pertinent descriptors. Most likely, such descriptors can be identified and interpreted in the statistically identified ‘consensus passages’. Such descriptors are very suitable for naming concept groups generated by document co-citation analyses, as they are very specific and agreed upon among the users of the domain.

Discussion

The present study has briefly introduced a semi-automatic method for parsing and filtering of noun phrases from citation contexts. The purpose of the method is to improve the naming of clusters or concept groups in literature visualization studies. The method aims at reducing the labour intensive manual citation context analysis since it automatically extracts phrasal concepts. At the same time, these noun phrases are considered more useful than single words for naming clusters or concept symbols in a co-citation network. Noun phrases extracted by the semi-automatic method are contextual, agreed upon, and pertinent, at least in the case study of *periodontology* presented in this paper. In fact, the extracted noun phrases can be considered as domain descriptors.

The result of the case study indicates that the method is able to identify highly important noun phrases, and that these phrases accurately describe their parent concept groups (clusters). However, such phrases have tendency to be more specific than manually inferred names.

Refinements of the method are still needed. It is very important to reduce the manual workload further. The most obvious next step is therefore to devise an algorithm, based on the experiences from present study that can identify suitable citation contexts automatically. However, a basic weakness of the present approach is the dependency on full text documents.

As indicated in the introduction, SMALL (1979) pointed to the problem of synonymous terminology, and how this influences an automatic frequency analysis needed for semi-automatic citation context analysis. Consider concept group 1. The two phrases *enamel matrix protein* and *enamel matrix derivative* are in a so-called definitional relationship to each other; either synonymous, near synonymous, or a class inclusion relationship (SOERGEL, 1974). A definitional relationship between two phrases typically manifests itself in such a way that the phrases rarely co-occur together in the same text window (SOERGEL, 1974). However, as demonstrated above, the extracted noun phrases in the concept groups are semantically related to each other, they all refer to the common concept of the group. These noun phrases appear together because they frequently share the same textual surroundings, that is, the context that surrounds a specific concept symbol in citing papers. The citation context is likely to be contextual in relation to subject matter. It is therefore assumed that the selected set of noun phrases is important and semantically related. The conjecture is that some of these noun phrases will co-occur directly with each other in the citation contexts. While other noun phrases will not co-occur at all with each other in these contexts, except that they do co-occur with the concept symbol, or rather, with its reference marker. Thereby, all noun phrases become related to each other, either directly by occurring in the same context, or indirectly through their common co-occurrence with the concept symbol. For example, noun phrase *A* and noun phrase *B* are indirectly related if they both co-occur with concept symbol *X*, but not directly with each other. Theoretically, this opens up the possibility of bringing synonymous or near-synonymous phrases into the analysis. These phrases rarely co-occur together, but they often share common textual contexts.

In SCHNEIDER (2004), we demonstrate how second-order co-occurrence analysis can be used in citation context analyses to identify definitional relationships among the important noun phrases in the concept groups.

*

The author wish to thank Peter Ingwersen and Pia Borlund for their valuable supervision during his doctoral work; Olle Persson for help with *Bibexcel*, and Eija Airio for assist and support with *Connexor*.

References

- Anick, P. G. and Vaithyanathan, S. (1997). Exploiting clustering and phrases for context-based information retrieval. *Proceedings of the ACM/SIGIR Conference on Research and Development in Information Retrieval, Philadelphia, PA, 1997*, 314-323.
- Batagelj, V. and Mrvar, A. (1998). Pajek: A program for large network analysis. *Connections*, 21(2), 47-57.
- Börner, K., Chen, C. and Boyack, K. W. (2003). Visualizing knowledge domains. In B. Cronin (Ed.), *Annual Review of Information Science and Technology* (Vol. 37, pp. 179-255). Medford, NJ: Information Today, Inc.
- Braam, R. R., Moed, H. and van Raan, A. F. J. (1991). Mapping of Science by combined Co-Citation and Word Analysis. I. Structural aspects. *Journal of the American Society for Information Science*, 42(4), 233-251.
- Garfield, E. (1974). The citation index as a subject index. *Current Contents*, May(18), 5-7.
- He, Q. (1999). Knowledge discovery through co-word analysis. *Library Trends*, 48(1), 133-159.
- Ingwersen, P. and Christensen, F. H. (1997). Data set isolation for bibliometric online analyses of research publications: Fundamental methodological issues. *Journal of the American Society for Information Science*, 48(3), 205-217.
- McCain, K. W. and Turner, K. (1989). Citation context analysis and ageing patterns of journal articles in molecular genetics. *Scientometrics*, 17(1-2), 127-163.
- Moravcsik, M. J. and Murugesan, P. (1975). Some results on the function and quality of citations. *Social Studies of Science*, 5, 86-92.
- Noyons, E. C. M. (1999). *Bibliometric mapping as a science policy and research management tool*. Leiden University, The Netherlands: DSWO Press.
- Ochiai, A. (1957). Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions. *Bulletins of the Japanese Society for Scientific Fisheries*, 22, 526-530.
- O'Connor, J. (1982). Citing statements: Computer recognition and use to improve retrieval. *Information Processing and Management*, 18, 125-131.
- O'Connor, J. (1983). Biomedical citing statements: computer recognition and use to aid full-text retrieval. *Information Processing and Management*, 19(6), 361-368.
- Rees-Potter, L. K. (1989). Dynamic thesaural systems: a Bibliometric study of terminological and conceptual change in sociology and economics with the application to the design of dynamic thesaural systems. *Information Processing and Management*, 25(6), 677-691.
- Schneider, J.W. (2004) *Verification of bibliometric methods' applicability for thesaurus construction*. PhD dissertation. Aalborg: Royal School of Library and Information Science. Available: <http://biblis.db.dk/uhtbin/hyperion.exe/db.jessch04>
- Schneider, J.W. and Borlun, P. (2004). Introduction to bibliometrics for construction and maintenance of thesauri: Methodical considerations. *Journal of Documentation*, 60(5), 524-549.
- Schvaneveldt, R. W. (Ed.). (1990). *Pathfinder Associative Networks: Studies in Knowledge Organization*. Norwood, New Jersey: Ablex Publishing Corporation.
- Small, H. (1978). Cited documents as concept symbols. *Social Studies of Science*, 8(3), 327-340.
- Small, H. (1979). Co-citation context analysis: The relationship between bibliographic structure and knowledge. In R. D. Tally & R. R. Deultgen (Eds.), *Proceedings of the American Society for Information Science Annual Meeting, Minneapolis, USA, 1979* (Vol. 16, pp. 270-275). White Plains, NY: Knowledge Industry Publications.

- Small, H. (1982). Citation context analysis. In B. Dervin & M. J. Voigt (Eds.), *Progress in Communication Sciences* (Vol. 3, pp. 287-310). Norwood, N.J.: Ablex.
- Small, H. (1986). The synthesis of specialty narratives from co-citation clusters. *Journal of the American Society for Information Science*, 37(3), 97-110.
- Small, H. and Greenlee, E. (1980). Citation context analysis of a co-citation cluster: Recombinant DNA. *Scientometrics*, 2(4), 277-301.
- Soergel, D. (1974). *Indexing languages and thesauri: construction and maintenance*. Los Angeles, CA: Melville.
- Sparck Jones, K. (1971). *Automatic keyword classification for information retrieval*. London: Butterworths.
- Voutilainen, A. (1993). NPTool, a detector of English noun phrases. In *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives, Ohio State University, Columbus, Ohio, 1993* (pp. 48-57).
- White, H. D. and McCain, K. W. (1989). Bibliometrics. In M. E. Williams (Ed.), *Annual Review of Information Science and Technology* (Vol. 24, pp. 119-186). Amsterdam: Elsevir.
- White, H. D. and McCain, K. W. (1997). Visualization of literatures. In M. E. Williams (Ed.), *Annual Review of Information Science and Technology* (Vol. 33, pp. 99-168).
- Wilson, C.S. (1999). Informetrics. In M. E. Williams (Ed.) *Annual Review of Information Science and Technology* (Vol. 34, pp. 107-247). Amsterdam: Elsevir.

Appendix 1: Manually deduced conceptual meaning of concept symbols

Concept group no.	Cited reference	Concept symbol of
1	HAMMARSTROM_97a	The role of enamel matrix proteins in periodontal regeneration
1	HAMMARSTROM_97b	The use of enamel matrix proteins for regenerative therapy
1	HEIJL_97a	The use of enamel matrix proteins for periodontal regeneration
1	HEIJL_97b	Enamel matrix derivative treatment leads to gains in clinical attachment level
2	GOTTLOW_86	Guided tissue regeneration
2	NYMAN_82a	Guided tissue regeneration
2	NYMAN_82b	Guided tissue regeneration
3	BECK_96	Association between periodontal disease and coronary heart disease
3	DESTEFANO_93	Association between periodontitis and coronary heart disease
3	JOSHIPURA_96	Association between periodontal disease and coronary heart disease
3	MATTILA_89	Association between oral health and coronary heart disease
3	OFFENBACHER_96a	Not identified
3	OFFENBACHER_96b	Periodontal disease as a risk factor for pre-term low birth weight
4	GARRETT_96	Periodontal regeneration techniques for class II furcations
4	HAMP_75	Degrees of furcation involvement
4	HIRSCHFELD_78	Not identified
4	MCFALL_82	Not identified
4	OLEARY_72	Modified O'Leary plaque record score for oral hygiene
5	GENCO_96	Risk factors for periodontal disease
5	GROSSI_94	The smoking risk factor for periodontal disease and attachment loss
5	GROSSI_95	Smoking as a risk factor for periodontal disease and alveolar bone loss
5	KORNMAN_97	Polymorphisms of the interleukin-1 genotype as markers for an increased risk of chronic periodontitis
5	PAPAPANOU_96	Smoking, age, and diabetes as risk factors for periodontal disease
6	HOLT_88	<i>P. gingivalis</i> initiates progression of periodontitis
6	PIKE_94	Gingipains of <i>P. gingivalis</i>
7	HOLT_99	<i>P. gingivalis</i> bacterium
7	SLOTS_99	Association between periodontal disease and <i>A. actinomycetemcomitans</i> and <i>P. gingivalis</i>
8	MASADA_90	Interleukin-1beta
8	STASHENKO_91	Tnf-alpha
9	HAFFAJEE_94	The bacteria <i>P. gingivalis</i> , <i>A. actinomycetemcomitans</i> , and <i>B. forsythus</i> are pathogens for periodontitis
9	LOESCHE_85	High proportions of <i>P. gingivalis</i> and spirochetes in patients with 'early-onset periodontitis
9	MOORE_94	Bacterial species of periodontal disease
10	SOCRANSKY_94	Checkerboard DNA-DNA hybridization technique
10	SOCRANSKY_98	The microbial complex of <i>B. forsythus</i> , <i>P. gingivalis</i> , and <i>T. denticola</i>
11	BRAGD_87	Increased level of <i>A. actinomycetemcomitans</i> in 'localized early-onset periodontitis
11	SLOTS_82	Identification of <i>A. actinomycetemcomitans</i> by use of TSBV
11	SYED_72	Reduced transport fluid

12	LOE_63	Gingival index
12	SILNESS_64	Plaque index
<hr/>		
13	BERGSTROM_89	Cigarette smoking as a risk factor for periodontitis
13	BERGSTROM_94	Smoking as a risk factor for periodontal disease
13	HABER_93	Smoking as a risk factor for periodontal disease
13	MACHTEI_97	Smoking as a risk factor for periodontal disease
13	STOLTENBERG_93	No influence of smoking on the prevalence of bacterial species involved in periodontal disease
13	ZAMBON_96b	Difference in the prevalence of periodontal pathogens, such as <i>B. forsythus</i> , between smokers and non-smokers
<hr/>		

*Cited references indicated with grey do *not* act as concept symbols as prescribed from the rules in SCHNEIDER (2004).