

Measuring and analysing the internal, topical coherence of Web of Science Subject Categories

Stephan Stahlschmidt¹ Marion Schmidt²

¹ *stahlschmidt@dzhw.eu*

German Centre for Higher Education Research and Science Studies (DZHW), Berlin (Germany)

² *schmidt@dzhw.eu*

German Centre for Higher Education Research and Science Studies (DZHW), Berlin (Germany)

Abstract

The Web of Science subject categories constitute a ubiquitous standard partition of disciplines in the Web of Science database and many bibliometric analyses. Due to the classification of complete journal issues, they have been criticized as being too broad, while entailing substantial implications on normalization exercises. Several alternatives have been proposed, but due to the unknown ground truth no proposal has found widespread application. Furthermore an empirical analysis of the Subject Categories' shortcomings on a comprehensive scale has not been conducted.

In this research-in-progress paper we thus measure the topical coherence of all Web of Science Subject Categories via an extensive clustering exercise including all articles indexed in the Web of Science between 2003 and 2012. This descriptive analysis will be supplemented by a regression analysis explaining the influence of several explanatory factors on the varying Subject Categories' coherence levels.

Given the current state of the ongoing analysis, we can verify the coarseness of the Web of Science Subject Categories and additionally observe an increase in topical heterogeneity over time, which we will analyse in the upcoming regression.

Conference Topic

indicators, metrology, databases, methods and techniques

Introduction

Publication and citation indicators are usually normalized in order to compensate for different publication and citation behaviour. For bibliometric studies that are based on the Web of Science (WoS) the overlapping classification system of subject categories (SCs) as provided by the database producer is mostly used. Citation indicators such as the mean normalized citation score (Waltman, van Eck, van Leeuwen, Visser & van Raan, 2011) normalize the citation rate of a given publication body with expected values calculated on the basis of the subject categories. The observed values are therefore related to their subject contexts in order to ensure comparability.

The WoS SCs are often criticized within the bibliometric community, especially because the classification is not carried out at the level of the individual articles, but rather at the level of journals issues. This is generally considered to be relatively inaccurate, particularly regarding increasing interdisciplinarity (Gómez, Bordons, Fernández & Méndez, 1996). New journals are assigned to existing subject categories by partially manual, partially algorithmic methods on the basis of citation data (Leydesdorff & Rafols, 2009, Pudovkin & Garfield, 2002). These methods are not publicly documented.

In general, a wide range of alternative clustering and community detection algorithms is available (Kou, Peng & Wang, 2014, Subelj, van Eck & Waltman, 2016); these algorithms can be based on a similarity measure between publications or entire journals. Different principles to infer the similarity have been proposed in the literature: they can be determined

either on the basis of citation data (usually bibliographic coupling or cross-citation) or the actual text (e.g. shared use of terms or phrases). In recent years, hybrid approaches combining citation and textual approaches were propagated (Thijs, Schiebel & Glänzel, 2013, Thijs, Zhang & Glänzel, 2015), as they compensate for both the lack of recall of the citation perspective and the lack of precision of the textual approaches.

While some comparative analyses of the WoS SC classification with alternative science maps based on inter-journal citation relationships have been proposed (Boyack, Klavans & Börner, 2005, Leydesdorff, 2006), these studies ultimately lack a gold standard to evaluate the diverse mappings. To partly address this issue and control for the randomness arising from the application of varying data sets, a comparative analysis of different methods on the basis of an identical data set is proposed in a current special edition of *Scientometrics* (Gläser, Glänzel & Scharnhorst, 2017). But without assured knowledge on the ground truth the benefit of entire alternatives to the proprietary SC classification (Glänzel & Schubert, 2003, Ruiz-Castillo & Waltman, 2014) cannot be assessed. Consequently none of these proposals have been adopted widely by the community until now, and the subject categories, probably due to their ubiquitous availability, are still the standard in the bibliometric community.

Nevertheless the subject categories of the WoS obviously vary in size and specificity. Van Eck, Waltman, Van Raan, Klautz, & Peul (2013) show evidence of heterogeneity within several medical subject categories along the dimensions of clinical and experimental research. Publications that are identifiable as experimental within the graphical representation exhibit higher citation counts. This finding poses, especially on the micro level, severe problems for the normalization with subject categories. However, due to its focus on medicine and a certain period of time, a comprehensive analysis on the subject is still missing, while being essential for an informed normalization in the computation of the MNCS and any percentile methods due to potential heterogeneous citation levels.

In this research-in-progress paper we analyse the topical coherence of the WoS SC classification via an internal clustering of articles within the SCs. By predominantly focusing on the subject-matter of an article we follow the general principle of scientific writing, namely progressing knowledge on a particular subject, and attach less value to diverging perspectives, which would result in a different partition of the articles (Wang & Rohe, 2016). The main research question is to what extent potential heterogeneities affect the normalization with SCs, whether more or less problematic areas of the classification with respect to coherence can be identified and how the phenomenon develops over time. Furthermore, we test approaches (such as citation autarky) to explain heterogeneous content structures with differing citation patterns within SCs in order to obtain informative background knowledge for any attempt to construct a more diverse alternative to the WoS SC classification.

Method

As stated, we assess the internal coherence of the WoS SCs by employing a hierarchical clustering approach and then by interpreting the number of resulting cluster as well as varying cluster specific citation levels as measures of coherence. Subsequently we analyse the observed level of coherence via explanatory factors. Consequently the clustering technique is not understood as resulting in a natural partition of the publications representing the unknown ground truth. It is rather applied as an operational measuring device (Dingle, 1950) to facilitate a valid comparison of the SCs' internal coherence, because validating a representational measure with an unknown ground truth seems rather challenging (Hand, 1996). Indeed, the main interest of our work is a comparison of the operationally measured

coherence values across the SCs and a supplemental regression analysis explaining the influence of several explanatory factors on the varying SCs' coherence levels.

To this end, we firstly define the relations among all articles in a particular SC and year according to their topical accordance and supply the resulting distance to the clustering algorithm. The subject-matter of an article is induced, albeit imperfectly, via the corresponding author keywords and reference list and we match any two articles based upon jointly applied author keywords and commonly referred literature (bibliometric coupling). Citations are predominantly understood as carrying cognitive information (Merton, 1957), marking concepts (Small, 1978) and consequently establishing a priority claim on the particular content (Kaplan, 1965). Given this perspective, citing a document might be understood as an act of acknowledgment and reward towards the cited document (Ravetz, 1971). Persuasive and perfunctory citations (Kaplan, 1965) diverge from this Mertonian ideal and illustrate the social facet of citations. In contrast, author keywords predominantly refer to concepts without any link to one or several persons. They thus reduce to some extent the social nuisance in this reference system.

A local cosine similarity $sim_loc_X^{(A,B)}$ between the articles A and B on the elements of the set $X \in \{author\ keywords, referred\ literature\}$ is computed separately for keywords and references via

$$sim_loc_X^{(A,B)} = \frac{|X^{(A)} \cup X^{(B)}|}{\sqrt{|X^{(A)}| * |X^{(B)}|}},$$

where $|\cdot|$ denotes the cardinality of a set. In a pre-processing step, the set of keywords is generalized by a stemming procedure to neutralize flexional variants used by the respective authors. At the same time the bibliometric coupling relation between any two articles is sharpened by excluding review articles, publications with a time lag of more than 10 years and non-source items, that is publications not published in the set of journals indexed in the WoS, but solely referred to by an indexed article.

Furthermore we extend this local perspective to a global level by transferring the local cosine similarity into a global one, $sim_glo_X^{(A,B)}$ (Ahlgren, Jarneving, & Rousseau, 2003). Thereby we consistently and throughout our study apply a global perspective including all publications in a particular SC and year at every step. We construct a similarity on such a global level by comparing the local similarities of article A to all other articles \bar{A} and itself with the vector of local similarities of article B with \bar{B} and itself. Apart from this theoretical argument of a consistent approach taking into account the whole SC, the global approach also incorporates important empirical advantages, like a decrease in the sparsity and a higher level of variance in the distance matrix decreasing the likelihood of ties and their severe consequences for clustering.

However, the cosine distance $dist_X^{(A,B)} = 1 - sim_glo_X^{(A,B)}$ does not denote a proper distance metric, as neither the Cauchy-Schwarz inequality nor the coincidence axiom hold up. Hence, we transfer the cosine similarity to the angular distance via the inverse cosine and map the values in a normalization step into the interval $[0,1]$. Lastly we take an average on the distances resulting from author keywords and bibliometric coupling with equal weights in order to combine them to a hybrid distance measure (Braam, Moed & van Raan, 1991a, Braam, Moed & van Raan, 1991b)

These hybrid distances between all articles of a SC in a specific year constitute the basis for the subsequent clustering approach. We opt for an implementation based on agglomerative clustering foregoing the popular modularity based community detection algorithms due to their restriction to local and not global steps to improve the objective function. A local focus in the greedy optimization increases the likelihood to reach varying local optima across the SCs which would consequently diminish the comparability across SCs. In contrast, this comparability is ensured in our hierarchical clustering setting via a uniform application of the same linkage criteria and cut-off across SCs and years. We apply the so-called “ward linkage criterion”, following Bagatelj (1988), who justified the use of any dissimilarities instead of squared Euclidean distances for the Ward method.

Preliminary Results

Due to the extensive scope of the analysis incorporating all WoS SCs for the years 2003 - 2012, we currently can present preliminary results for a non-random sample of 19 SCs.

All SCs possess several clusters at some point of the observation period and can be classified as heterogeneous according to our measure. Taking into account the applied keywords and reference patterns, these SCs contain several subgroups differing by the respective subject-matter. The left panel of Figure 1 depicts the corresponding number of clusters per SC and year. The mean (90% quantile) number of clusters increases from 48 (122) to 95 (238) clusters in the observed time period, thus exhibiting a notable growth. This increase over time might mirror a potential rise in the level of specialisation in the WoS indexed articles, even when the actual amount of differentiation – measured by the number of clusters – depends on the applied configuration of both dissimilarity measures and clustering which might still be improved.

Whereas the observed increase in topical heterogeneity is interesting in itself, this scientific development entails a problem for bibliometrics only if the cluster specific citation levels vary within the observed clusters of a SC. The right panel in Figure 1 reports on the coefficient of variation on the cluster specific means:

$$CoV = \frac{SD(\overline{citations}_{cluster_c|SC})}{\overline{citations}_{SC}},$$

where $\overline{citations}_{cluster_c|SC}$ denotes a vector of average numbers of citations received by articles in the cluster $c \in \{1, \dots, C\}$ in a certain SC. The magnitude of the variation in mean citations across the clusters amounts on average to 75% to 150% of the mean citations received by the whole SC. This mean appears stationary without any trend, but is accompanied by a high volatility and demonstrates a substantial difference between the average citations of articles in the clusters and the SC as a total.

Given the current state of 19 SCs, the WoS SC classification seems too coarse to reflect the topical heterogeneity observed in the articles indexed in the WoS. Furthermore the heterogeneity seems to increase over time. However, a more detailed and comprehensive verdict will be reported once the processing of the remaining SCs has been completed.

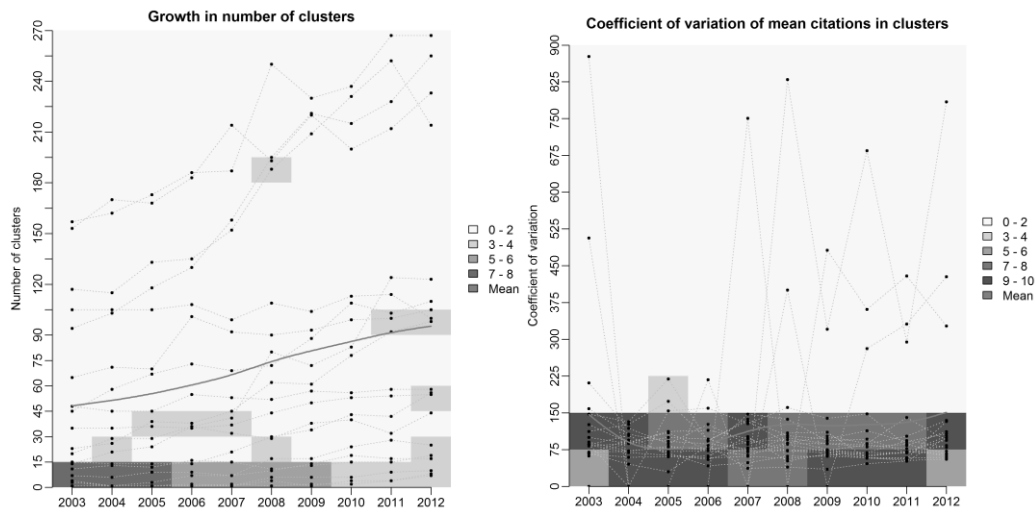


Figure 1: Number of cluster per SC and year (left) and variation coefficient of citation means (right). Every line denotes a separate SC.

Outlook

This research-in-progress paper will result in a descriptive analysis of the topical coherence of all SCs in the WoS for the years 2003-2012, an endeavour encompassing more than 10 million articles, and secondly will bring up explanatory models which explore the reasons for the observed variation in topical coherence across the SCs. Consequently we will present a descriptive analysis on the whole set of SCs and extract explanatory variables from the WoS database to conduct an inferential analysis which ideally results in informative background knowledge to construct a more diverse alternative to the WoS SCs classification.

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 and Technology*, 54(6), 550–560. <https://doi.org/10.1002/asi.10242>
- Batagelj, V. (1988). Generalized Ward and related clustering problems. In: *Classification and related methods of data analysis* (ed. Bock, H.), North-Holland: Amsterdam.
- Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics*, 64(3), 351–374. <https://doi.org/10.1007/s11192-005-0255-6>
- Braam, R. R., Moed, H. F., & van Raan, A. F. J. (1991a). 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.
- Braam, R. R., Moed, H. F., & van Raan, A. F. J. (1991b). Mapping of science by combined co-citation and word analysis. II: Dynamical aspects. *Journal of the American Society for Information Science*, 42(4), 252–266.
- Dingle, H. (1950). A theory of measurement. *British Journal for the Philosophy of Science*, 1(1), 5–26.
- Glänzel, W., & Schubert, A. (2003). A New Classification Scheme of Science Fields and Subfields Designed for Scientometric Evaluation Purposes. *Scientometrics*, 56(3), 357–367. <https://doi.org/10.1023/A:1022378804087>
- Gläser, J., Glänzel, W., & Scharnhorst, A. (2017). Same data—different results? Towards a comparative approach to the identification of thematic structures in science. *Scientometrics*. <https://doi.org/10.1007/s11192-017-2296-z>

- Gómez, I., Bordons, M., Fernández, M. T., & Méndez, A. (1996). Coping with the problem of subject classification diversity. *Scientometrics*, 35(2), 223–235. <https://doi.org/10.1007/BF02018480>
- Hand, D. J. (1996). Statistics and the Theory of Measurement. *Journal of the Royal Statistical Society: Series A*, 159(Part 3), 445–492.
- Kaplan, N. (1996). The norms of citation behavior: Prolegomena to the footnote. *American documentation*, 16(3), 179–184.
- Kou, G., Peng, Y., & Wang, G. (2014). Evaluation of clustering algorithms for financial risk analysis using MCDM methods. *Information Sciences*, 275, 1–12. <https://doi.org/10.1016/j.ins.2014.02.137>
- Leydesdorff, L. (2006). Can scientific journals be classified in terms of aggregated journal-journal citation relations using the Journal Citation Reports? *Journal of the American Society for Information Science and Technology*, 57(5), 601–613. <https://doi.org/10.1002/asi.20322>
- Leydesdorff, L., & Rafols, I. (2009). A global map of science based on the ISI subject categories. *Journal of the American Society for Information Science and Technology*, 60(2), 348–362. <https://doi.org/10.1002/asi.20967>
- Merton, R. K. (1957). Priorities in Scientific Discovery: A Chapter in the Sociology of Science. *American Sociological Review*, 22(6), 635–659.
- Pudovkin, A. I., & Garfield, E. (2002). Algorithmic procedure for finding semantically related journals. *Journal of the American Society for Information Science and Technology*, 53(13), 1113–1119. <https://doi.org/10.1002/asi.10153>
- Ravetz, J. R. (1971). *Scientific knowledge and its social problems*. Oxford: Clarendon Press.
- Ruiz-Castillo, J., & Waltman, L. (2014). Field-Normalized Citation Impact Indicators Using Algorithmically Constructed Classification Systems of Science. *Working Paper*. Abgerufen von <http://e-archivo.uc3m.es/handle/10016/18385>
- Small, H. (1978). Cited Documents as Concept Symbols. *Social Studies of Science*, 8(3).
- Subelj, L., van Eck, N. J., & Waltman, L. (2016). *PLoS ONE*, 11(4), e0154404.
- Thijs, B., Schiebel, E., & Glänzel, W. (2013). Do second-order similarities provide added-value in a hybrid approach? *Scientometrics*, 96(3), 667–677. <https://doi.org/10.1007/s11192-012-0896-1>
- Thijs, B., Zhang, L., & Glänzel, W. (2015). Bibliographic coupling and hierarchical clustering for the validation and improvement of subject-classification schemes. *Scientometrics*, 105(3), 1453–1467. <https://doi.org/10.1007/s11192-015-1641-3>
- Van Eck, N. J., Waltman, L., Van Raan, A. F. J., Klautz, R. J. M., & Peul, W. C. (2013). Citation Analysis May Severely Underestimate the Impact of Clinical Research as Compared to Basic Research. *PLoS ONE*, 8(4), e62395. <https://doi.org/10.1371/journal.pone.0062395>
- Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., & van Raan, A. F. J. (2011). Towards a new crown indicator: an empirical analysis. *Scientometrics*, 87(3), 467–481. <https://doi.org/10.1007/s11192-011-0354-5>
- Wang, S. & Rohe, K. (2016). Discussion of “Coauthorship and citation networks for statisticians”, *The Annals of Applied Statistics*, 10, 1820-1826.