
# Clustering of scientific publications at the KTH Library

Tobias Jeppsson & Peter Sjögårde

TRITA-LIB 2018:1
DOI: 10.30746/Trita-LIB-2018:1
ISBN: 978-91-7729-669-0

# Clustering of scientific publications at the KTH Library

Tobias Jeppsson & Peter Sjögårde

KTH Library, Unit for Publication Infrastructure

26 October 2017

## Background

Classification of scientific publications into categories and groups can fill many purposes. Within bibliometrics, such classifications can, for instance, be used to field-normalize citation rates or to extract sets of publications that deal with particular topics. However, the current article categorizations in the major citation databases (*Web of Science* and *Scopus*) are relatively crude and based on journal classifications. Therefore, they are often insufficient to properly categorize research activities. At the same time, many methods have been developed to partition networks into clusters, and these methods can be applied to networks of scientific publications to create classifications that are independent of the static categories that are provided by citation databases.

The purpose of this document is to describe how an algorithmic classification of publications have been implemented in *Bibmet* (the publication database at KTH containing Web of Science records). The goal of this methodology is to create a hierarchical cluster-based classification of research publications, based on citation links between documents. The classification can be used for subject differentiated analyses, to answer questions such as:

- What is the subject profile of university *X*?
- Which are the most rapidly growing research fields?
- What is the citation relation between subject field *c* and *d*?

The classification can also be used for e.g. citation normalization, to discover and select related sets of publications, and many other purposes.

## Method and implementation

The clustering of publications into classes is based on the CWTS method for partitioning of publications (Li & Ruiz-Castillo, 2013; Ruiz-Castillo & Waltman, 2015; Waltman & Eck, 2013). The general methodology is based on optimization of a modularity function

(Newman, 2004). The software Modularity Optimizer have been used, which can be downloaded at http://www.ludowaltman.nl/slm.

The software creates a partitioning of nodes in a network. In our implementation, nodes (vertices) are publications and the relations (edges) are citations. Links between publications can be calculated in different ways, as direct citations, bibliographic coupling (the number of common references between two publications) or co-citations (when two publications have been cited by the same publication). Direct citation have been used here for two reasons; 1) efficiency and, 2) they perform well compared to co-citations and bibliographic coupling for clustering of research publications (Klavans & Boyack, 2017).

In most aspects, the implementation follows the method described by Ruiz-Castillo & Waltman (2015) and Waltman & Eck (2013). This includes 1) the software used, 2) the use of direct citations, 3) normalization of links between publications, and 4) treatment of small publication clusters based on the clustering algorithm. The resolution parameter has been adjusted to aim for approximately the same levels of aggregation as the CWTS implementation, i.e. about 20, 700 and 35000 classes per level. In addition, an intermediate level with about 4000 classes has been created. This level of aggregation has been proposed for field normalization by Ruiz-Castillo and Waltman (2015). So in total, the KTH implementation of modularity-based clustering is using 4 hierarchical levels, with level 1 as the lowest and level 4 as the highest level of aggregation.

The hierarchical implementation at KTH differs in one aspect from the implementation at CWTS, and that is in how clusters are clustered at level 2 and above (i.e. the method to cluster level 1 to level 2, level 2 to level 3 and so forth). The software does not include the possibility to cluster publications into a hierarchal classification in one step. Instead, the program has been executed independently for each new level of aggregation. At all levels above the lowest publication-level classification, the relations between classes at a given level have been calculated as the weighted sum of the citation relations between publications. For example: The publications *P1* and *P2* have a relation of the weight *r*. *P1* has been clustered into class *C1* and *P2* has been clustered into class *C2*. *C1* has the size of $N_{C1}$ and *C2* have the size of $N_{C2}$ then the pair of *P1*-*P2* contributes with a relation strength of $r/(N_{C1} * N_{C2})$ between *C1* and *C2*. These contributions are then summed for all combinations of classes at a particular level, to produce relations between classes. These relations are subsequently used to create a classification at the next level of aggregation.

To make the resulting classification more useable and searchable, several descriptors of the classes are also produced. Most importantly, a label is attached to each cluster, constructed from the three terms that best describes the cluster, based on an algorithmic weighting of several terms. The candidate terms come from four bibliographic fields; Author keywords, journal names, Web of Science subject categories and author addresses.

# Data

As data source we have used the bibliometric database *Bibmet*, which is a relational database based on Web of Science-records from 1980. Only publication types "article" and "review", based on Web of Science categories, have been used in the classification. Proceedings were not used since we have experienced data quality problems with such records regarding

citation matching algorithms. Publications without citation links have been excluded. A minimum publication class size has been set at each level of aggregation. Publication classes smaller than the set limit have been referred to other classes based on their relational strengths (the size limits currently used are shown in Table 1).

If a class with a publication count below the class size limit has no relations to any class above the class size limit, the publications in the class have been disregarded. This may exclude publications with few references to other publications in the database that also have few citations. Such publications are not likely to be of high interest for bibliometric analyses.

## Results

The methodology for creating a cluster-based publication-level classification of scientific publications has been used several times on *Bibmet*, as the publication database has been updated. To date, a full classification across all hierarchical levels has been created four times, based on data following *Bibmet* updates 2015Q3, 2016Q3, 2016Q4 and 2017Q1. The most recent classification was based on publication data until the first quarter of 2017 (*Bibmet* is updated quarterly). 30.7 million articles and reviews were given as input. Out of these, 133 741 publications (approximately 0.43%) could not be classified. Number of classes at each level is given in table 1.

*Table 1. The number of classes at each classification level, along with class size thresholds.*

| | Num. classes above threshold | Threshold | Num. classes below threshold | Num. publ. below threshold |
|---|---|---|---|---|
| **Level 1** | 37727 | 50 | 59413 | 192286 |
| **Level 2** | 4190 | 500 | 392 | 127609 |
| **Level 3** | 722 | 5000 | 87 | 101654 |
| **Level 4** | 25 | 120000 | 3 | 35387 |

The number of publications in each class sizes differ substantially within levels (reflecting the sizes of the different research fields), ranging between approximately 4.1 million and 0.12 million at level 4, and between 6809 and 50 at level 1.

The hierarchical classification of publications, based on data until the first quarter 2017 and showing levels 2 to 4 is shown in Fig. 1, and can be navigated online at: http://www.kth.se/bibliometrics/classification/2017Q1/

## Future developments

The current implementation of the hierarchical clustering has been used in several projects so far, but much evaluation remains to be done. The overall categorization of publications appears to capture topics and specializations well. However, further systematic evaluation of

how well areas of research are delineated by the classification is needed, and whether the classification seems to work better on some fields of science than others.

We have also observed that the turnover of publications between subsequent clusterings, based on new quarterly data, appears to be relatively high. This issue should be investigated further – and more generally, how the cluster network evolves over time – and can have important implications on how the cluster-based classification can be used for field-normalization.

In the current implementation, creating a new hierarchical clustering in *Bibmet*, including all related programs to extract labels and other descriptive information, is relatively time consuming. The workflow is using SAS and Modularity Optimizer in particular steps back and forth, and takes at about 1.5-2 days to execute in total. However, much of this time is used by SAS-processes to reclassify small clusters and extract information about the created classes, and there is room to re-write these programs to make them more efficient.

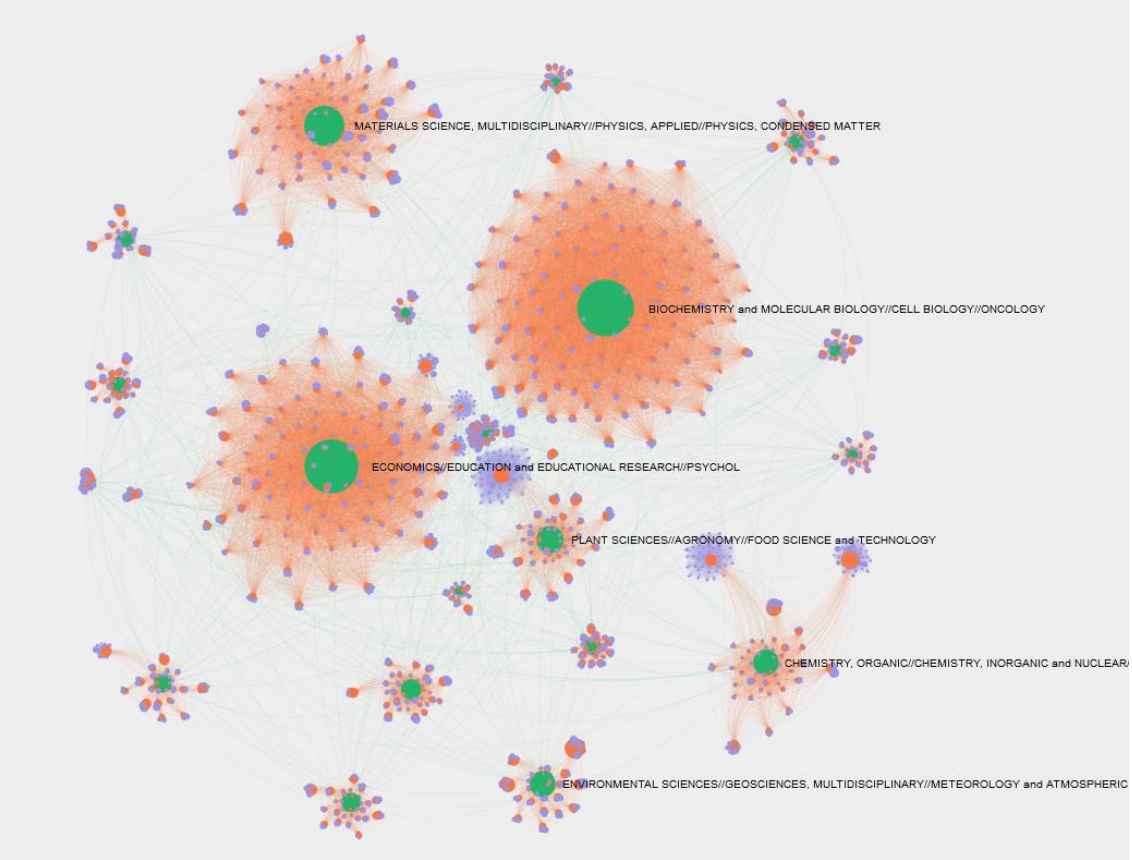


*Fig 1. A network layout of the hierarchical classification, showing cluster levels 2 to 4. In the online version, classes can be accessed to show summary information that describes the contents of the class, as well as publication volume over time.*
*Also see: www.kth.se/bibliometrics/classification/2017Q1/network/index.html.*

## References


Klavans, R., & Boyack, K. W. (2017). Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge? *Journal of the Association for Information Science and Technology*, 68: 984–998. https://doi.org/10.1002/asi.23734

Li, Y., & Ruiz-Castillo, J. (2013). The comparison of normalization procedures based on different classification systems. *Journal of Informetrics*, 7(4), 945–958. https://doi.org/10.1016/j.joi.2013.09.005

Newman, M.E.J. (2004). Fast algorithm for detecting community structure in networks. *Physical Revew E*, 69, 066133. https://doi.org/10.1103/PhysRevE.69.066133

Ruiz-Castillo, J., & Waltman, L. (2015). Field-normalized citation impact indicators using algorithmically constructed classification systems of science. *Journal of Informetrics*, 9(1), 102–117. https://doi.org/10.1016/j.joi.2014.11.010

Waltman, L., & Eck, N. J. van. (2013). A smart local moving algorithm for large-scale modularity-based community detection. *The European Physical Journal B*, 86(11), 471. https://doi.org/10.1140/epjb/e2013-40829-0