# Large-scale topic networks: can we improve efficiency and obtain similar results using LSH?

Bart Thijs bart.thijs@kuleuven.be Belgium ECOOM - FEB – KULeuven

Diane Gal Diane.gal@kuleuven.be Belgium Department of Cardiovascular Sciences, KULeuven

Mehmet Abdulhayoglu mehmetali.abdulhayoglu@kuleuven.be Belgium ECOOM, FEB, KU Leuven

Wolfgang Glänzel wolfgang.glanzel@kuleuven.be Belgium ECOOM and Dept. MSI, FEB, KU Leuven & Dept. Science Policy & Scientometrics, Library of the Hungarian Academy of Sciences

## Introduction

Science mapping and topic detection  have a long tradition in the field on scientometrics resulting in a broad toolbox of methodologies and approaches to tackle the common challenges associated with these endeavours. Recent developments in network based community detection, using either modularity or the map equation as the fundament for clustering, are allowing much larger networks than ever before. Bibliometricians have already integrated these techniques and sought to deploy them to their fullest potential. Large document networks have been built covering several million of papers using direct citations as edges between nodes,, as input for clustering and topic detection. Other types of networks using bibliographic coupling or lexical similarties or even the combination of both in hybrid approaches have been used to model and identify topics in document sets (Velden et al., 2017). Topic similarity between documents is then based on identifying common lexical items, such as extracted noun phrases or thesaurus terms in each document; in addition to being based on common references or shared references among documents. Once a similarity score is calculated between all documents, then the above mentioned network-clustering algorithm is applied to identify clusters of documents that are more closely related to one-another than to the documents in other clusters. An important property of these noun-phrase based networks, but to some extent also in bibliographic coupling networks, is their high density compared to direct citation networks. Consequently, the creation of these networks demands much more storage, memory and computational power from the processing systems.

As part of a collaborative research project in the cardiovascular field, we are working to obtain a global view of what topics have been published in cardiovascular research over a 21-year period. In the health field, large scale network analysis has been undertaken using PubMed, testing a variety of data inputs (Boyack et al., 2011; Boyack & Klavans, 2010). Using a large corpora of documents we wish to map and identify topic networks to better delineate and build understanding of the various elements and sub-topics within the field. However, noun-phrase based data sets result in hyperdimensional spaces for representing document sets and when turned into document-by-document similarities they are often weak or even results in inexistent ties. The most natural solution is to apply thresholds to the weight of the similarities and to make the analysis more feasible or efficient. But this approach does not reduce the memory and computational requirements for the calculation of the similarities, as the thresholds can only be applied after the complete processing of all possible links.

Another promising method to reduce high-dimensionality is called Locality Sensitive Hashing (LSH), which uses calculated signatures for each item and then places similar items together in 'buckets' or hyper-planes allowing to approximate the nearest or most similar neighbours and reduce overall computational needs (Leskovec, Rajaraman, & Ullman, 2014). The application of this technique requires the selection of an appropriate set of parameters in order to obtain the most suitable result with a minimum of hardware resources. Earlier Adbulhayoglu and Thijs (2017) used LSH to match references to items in bibliographic databases where only exact matches should be retained. In the current application, many links between documents are desirable. The iterative testing of LSH parameters is, however, not often reported in detail and we were not able to locate any bibliometric publication that compared the application of different LSH parameters on the subsequent outcome in terms of document set topic networks.

As previously presented, a dataset of 804,525 publications in the cardiovascular field from 1993-2013 have been identified and meta-data collected from Clarivate Analytics Web of Science Core Collection (WoS) (Gal, Sipido, & Glänzel, 2015; Gal, Glänzel, & Sipido, 2016). In this study, we will investigate how altering the parameters in Locality Sensitive Hashing (LSH) would affect efficiency in terms of resources used (time, cost, storage), using high-powered cloud-based parallelized computing. Secondly, and more importantly, we investigate how the network analysis results change when altering the LSH parameters, to compare whether we can identify the most stable communities, and therefore topics, across the different LSH parameters implemented.

## Data and methods

This study is based on the publication meta-data in the cardiovascular field (1993-2013) obtained from the Web of Science. Noun phrases are extracted using the natural language processing framework developed at Stanford (Chen & Manning, 2014). For the localilty sensitive hashing we opted for the implementation of an LSH version described by Ravichandran et al. (2005) made available by Souncloud through github.com[1]. This implementation requires three main parameters:

- The size of the set of hashing vectors or functions which are used to calculate the bit-based signatures
- The number of permutations for changing the order of the calculated bits in the signatures and subsequent sorting of the complete document set
- The number of selected neighbours within each bucket.

From an earlier paper (Abdulhayoglu & Thijs, 2017) we learned that increasing the first two parameters has a beyond linear effect on the time needed to finish the linking. Increasing the number of selected neighbours has only marginal effect on the time but increases the density and hence the storage and complexity of the following clustering exercise.

We will run twenty-seven different combinations of these parameters, each at three levels: 20-50-100 for the number of hashing vectors, 5-20-50 for the number of permutations and 50-100-300 for the number of neighbours.

## Expected results and conclusions

We expect that increasing the first two parameters will result in more accurate approximations of the actual strength of the link between documents. The effect of the third parameter is expected to interact with the levels applied for the two others. Locality Sensitive Hashing is prone to false positives at too low levels of parameters set and this possible error is leveraged by extending the number of neighbours.

[1] See https://github.com/soundcloud/cosine-lsh-join-spark

We hope that the study of these parameters will enable bibliometricians to build large scale lexical based document networks for topic detection and science mapping.

## References

Abdulhayoglu, MA. & Thijs, B. (2017). Use of Locality Sensitive Hashing (LSH) Algorithm to Match Web of Science and SCOPUS. *Proceedings of the Fifth Workshop on Bibliometric-enhanced Information Retrieval {(BIR)} co-located with the 39th European Conference on Information Retrieval {(ECIR} 2017), Aberdeen, UK*. 30-40.

Boyack, K. W., & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? Journal of the American Society for Information Science and Technology, 61(12), 2389–2404. https://doi.org/10.1002/asi.21419

Boyack, K. W., Newman, D., Duhon, R. J., Klavans, R., Patek, M., Biberstine, J. R., Börner, K. (2011). Clustering More than Two Million Biomedical Publications: Comparing the Accuracies of Nine Text-Based Similarity Approaches. PLoS ONE, 6(3), e18029. https://doi.org/10.1371/journal.pone.0018029

Chen, D., Manning. C.D. (2014) A Fast and Accurate Dependency Parser using Neural Networks. Proceedings of EMNLP.

Gal, D., Sipido, K. R., & Glänzel, W. (2015). Using Bibliometrics-aided Retrieval to Delineate the Field of Cardiovascular Research. In Proceedings of ISSI 2015 Istanbul (pp. 1018–1023). Istanbul: Bogaziçi University Printhouse. Retrieved from http://issi2015.org/en/Proceedings-of-ISSI-2015.html

Gal, D., Glänzel, W., & Sipido, K. R. (2016). Mapping cross-border collaboration and communication in cardiovascular research from 1992 to 2012. European Heart Journal, ehw459. https://doi.org/10.1093/eurheartj/ehw459

Leskovec, J., Rajaraman, A., & Ullman, J. D. (2014). Mining of massive datasets / Jure Leskovec, Standford University, Anand Rajaraman, Milliways Labs, Jeffrey David Ullman, Standford University (Second edition). Cambridge: Cambridge University Press.

Ravichandran, D., Pantel., P., & Hovy., E. (2005). Randomized algorithms and nlp: using locality sensitive hash function for high speed noun clustering. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. 622-629.

Velden, T., Boyack, K. W., Gläser, J., Koopman, R., Scharnhorst, A., & Wang, S. (2017). Comparison of topic extraction approaches and their results. Scientometrics. https://doi.org/10.1007/s11192-017-2306-1