PRESENTATION

Same Data, Different Results? On a Comparative Topic Extraction Exercise

Theresa Velden*, Kevin Boyack, Nees van Eck, Wolfgang Glänzel, Jochen Gläser, Frank Havemann, Michael Heinz, Rob Koopman, Andrea Scharnhorst, Bart Thijs, Shenghui Wang

Topic extraction from scientific literature seems to be as much art as science. Different teams within the field of scientometrics use different approaches based on their familiarity with specific methods, investment in the development of specific tools, their long-term experience with the mapping of scientific fields, and in-house experimentation to optimize an approach. Rarely are results published that apply alternative approaches to the same data set and compare the results, and there is a lack of understanding how approaches differ and how these differences affect the results obtained. Is one approach better than another? In what ways do the solutions that they produce differ from one another? What are the 'screws and levers' of each approach, and how do they affect the results?

To shed light on these questions, our collaboration of researchers from different teams in scientometrics has applied topic extraction approaches to the same data set. In this paper we provide a comparative overview of the properties of these approaches and use a variety of methods to compare the results and understand differences in the solutions they produce. The selection of approaches compared in this paper is opportunistic in that these are the approaches developed and used by the teams that have come together for this collaboration, which has evolved out of discussions at the annual meeting of the members of the advisory board of a recent BMBF funded project on the diversity of science. This means that for each approach used in the comparison there is one member or team in our collaboration who is intimately familiar with that approach. Each of us has made informed and sometimes pragmatic decisions on what approach to pursue and how to tweak it to meet the specific objectives of our respective research and tool development projects. Importantly for our purpose here is that this set of approaches covers a wide range of ways to model data and a number of different clustering algorithms to exploit the structure of the modeled data to extract topics. This variety makes this set of approaches suitable as a first set to explore the question of how approaches and their results differ.

When comparing approaches it can be useful to distinguish conceptually two aspects of each approach: the way the data is modeled (what features of the articles in the data set are extracted and used to represent the data) and the clustering algorithm that is used to detect regularities in the data and extract groups of articles that represent candidates for 'topics'. The data models range from citation based models over hybrid (citation and text based models) to purely semantic models. The algorithms used include four of the most popular clustering algorithms (k-means [1] , Infomap [2] , Louvain [3],and Smart Local Moving Algorithm [4], which is an improved variant of the Louvain algorithm) along with a new `memetic type' algorithm [5] designed specifically for the extraction of overlapping, poly-hierarchical topics in the scientific literature.

The data set used in this collaboration was downloaded from the Web of Science using the astronomy and astrophysics subject category to identify 59 major journals in this field. The

---

* presenter

final data set includes 111,616 articles, letters and proceedings papers covering the years 2003-2010.

We use quantitative, set based measures of overlap between clustering solutions as well as visualizations to study differences and commonalities between the topics extracted. The visualizations include topic affinity networks, VOSViewer term maps, and a new, interactive tool developed by OCLC, 'Ariadne'. With their help we identify interesting commonalities and differences that we further investigate in case studies. For example, one pair of solutions that is based on the same data model (a direct citation network) but uses two different algorithms for clustering, exposes great commonality with many clusters strongly overlapping. However, one cluster is split in one of the solutions, and investigation of the VOSViewer term maps reveals that the topic area (solar system) is split according to the type of objects studied, asteroids and comets on the one hand, and planets on the other hand. In another case study we are investigating the difference between approaches that build on citations in their data model (and ignore the roughly 8% of the publications that are not part of the giant component of the direct citation network) and an approach that uses a semantic matrix derived from the metadata of each publication and includes all publications in the data set. Interestingly, about 50% of the publications in the second largest cluster produced by the latter approach, are outside the giant component of the direct citation network. This may imply that it is able to identify a larger topic that the direct citation based approach by construction is bound to miss.

We are drawing the following, tentative conclusions from our investigation: Difference of perspectives matters, namely that the clustering that is imposed from 'outside' by an entire science map clustering differs significantly from a clustering derived from the data set alone. It appears that data models matter more than differences between clustering algorithms of one class (hard clustering). Further, algorithm classes matter, namely local clustering vs. global, hard clustering, provides very different results. And finally, the comparison of cluster approaches in absence of an unambiguous ground truth is itself a research agenda that needs to be developed if comparison and evaluation are not directed primarily at technical competitiveness (e.g. efficiency) but at how well topic extraction approaches serve specific substantial purposes.

A special issue of the journal *Scientometrics* about this work is in preparation.

References

[1] Witten, I. H., & Frank, E. (2005). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

[3] Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008(10), P10008.

[2 Rosvall, M., & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, *105*(4), 1118-1123.

.

[4] Waltman, L., & van Eck, N. J. (2013). A smart local moving algorithm for large-scale modularity-based community detection. The European Physical Journal B, 86(11), 1-14.

[5] Havemann, F., Gläser, J., & Heinz, M. (2015). A link-based memetic algorithm for reconstructing overlapping topics from networks of papers and their cited sources. In *15th International Conference on Scientometrics and Informetrics.*