Precision and Recall in Classifying Scientific Literature: Comparing Topic Modelling to Kernel-Based Spectral Clustering

Arho Suominen[1], Stephen Carley[2], Hannes Toivanen[1], Alan Porter[2,3]

[1]VTT Technical Research Centre of Finland, Finland
[2]Technology Policy and Assessment Center, Georgia Tech, Atlanta, GA
[3]Search Technology, Inc., Norcross GA 30992, USA
Email: arho.suominen@vtt.fi, Tel: +358 50 5050 354

Abstract

The availability of methods that can be applied directly to text, such as topic modelling (Blei & Lafferty 2009) and string kernels (Karatzoglou & Feinerer 2007), have shown promise as a tool for text mining. Studies show that the text-based clustering methods can differentiate between document groups with high accuracy (Karatzoglou & Feinerer 2010; Wei X. & Croft 2006). Recently Yau et al. (2013) showed that topic modelling algorithms, although dependant on the method, showed excellent precision and recall values for scientific documents.

The introduction of kernels that can be used directly, without the need for feature extraction prior to analysis, makes kernels a viable method for analysing textual data. The most significant drawback of the methods has been the complexity of the algorithms, which has been reduced by the use of suffix trees. In this paper, we use spectral clustering to classify the set scientific abstracts. We apply the R implementation, through the kernlab package in R, based on the algorithm by Ng et al. (2002).

The study extends the earlier work by Yau et al. (2013), who used topic modeling, to consider kernel-based spectral clustering in classifying a set of selected scientific papers. The sample used consists of seven technologies that were merged to a single corpus (N = 1254), seen in Table 1, for which the algorithm was used to distinguish between the documents on different technologies. The analysis was done with three levels of pre-processing, with increasing intensity. The algorithms were run with each of the three pre-processed corpuses, to which the classification accuracy was calculated as precision, recall and F-score.

Table 1 Diverse Test Set from the Web of Science.

Category(Keyword)	Document #
MEMS	345
Solar cell or photovoltaic	178
Tissue Engineering	217
RNAi or RNA inference	127
Graphene	180
Genetic Algorithm	114
(stochastic or non-linear) programming	99

The results show that kernel-based spectral clustering is able to classify documents with a high accuracy – highest F-score average for the seven technologies being 0,721. The variance between the F-scores of technologies is however significant, from a high of 0,874 to a low of 0,217. The results also suggest that increasing pre-processing intensity lowers the algorithm's capability to distinguish between the technologies. The F-score average diminishes from 0,721, with the minimal pre-processing, to 0,606 as pre-processing is increased.

Comparing to the results by Yau et al. (2013), the approach used in this study performs worse by the values of the F-scores and their variance across technologies. Although the direct use of text, without feature extraction, is a promising option, some data considerations that could better the results of the kernel-based spectral clustering should be taken into account before applying the method.

References

Blei, D.M. & Lafferty, J.D., 2009. Text Mining: Classification, Clustering, and Applications. In A. N. Srivastava & M. Sahami, eds. Taylor and Francis, pp. 71–94.

Karatzoglou, A. & Feinerer, I., 2010. Kernel-based machine learning for fast text mining in R. *Computational Statistics & Data Analysis*, 54(2), pp.290–297.

Karatzoglou, A. & Feinerer, I., 2007. *Text clustering with string kernels in R*, Springer.

Ng, A.Y., Jordan, M.I. & Weiss, Y., 2002. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2, pp.849–856.

Wei X. & Croft, W.B., 2006. LDA-based Document Models for Ad-hoc Retrieval. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. p. 178-185.

Yau, C.-K. et al., 2013. Clustering scientific documents with topic modeling.Unpublished