ANALYSING THE EFFECT OF DIFFERENT DESIGN CHOICES IN NETWORK-BASED TOPIC DETECTION

Carla Galluccio[1], Matteo Magnani[2], Davide Vega[2],
Giancarlo Ragozini[3] and Alessandra Petrucci[1]

[1] Department of Statistics, Computer Science, Applications "G. Parenti", (e-mail: carla.galluccio@unifi.it, alessandra.petrucci@unifi.it)

[2] Department of Information Technology, Division of Computing Science, (e-mail: matteo.magnani@it.uu.se, davide.vega@it.uu.se)

[3] Department of Political Sciences, (e-mail: giancarlo.ragozini@unina.it)

ABSTRACT: In the literature on topic modelling, network-based procedures for topic detection have become popular as an alternative to classical topic models, showing promising results. However, the lack of a systematic analysis of how the design choices made in text processing and network definition affect the results in terms of topics detected makes using these procedures demanding. Therefore, this work aims to fill this gap by showing how and to what extent the choices made during the analysis influence the features of the topics discovered.

KEYWORDS: text network analysis, community detection, topic detection

1 Introduction

Network-based procedures for topic detection are based on the idea that any text can be represented as a word co-occurrence network, where topics are defined as groups of strongly connected words (Hamm & Odrowski, 2021).

More specifically, a network-based topic discovery process is made up of different steps that could be summarised as follows: i) text preprocessing; ii) definition of the word co-occurrence matrix; iii) network definition and selection of the community detection algorithm.

Even if many works have applied network-based procedures for analysing textual data and discovering topics, none of them focused on how the choices made in the design phase affect the final result in a systematic way.

Thus, this work aims to start filling this gap by studying how and to what extent some of the choices made during the analysis influence the features of topics discovered. In particular, in this work, we focused primarily on the definition of the word co-occurrence matrix and the selection of the community detection algorithm, as these steps are unique to network-based approaches.

2 Method and Materials

We conducted the analysis employing the BBC news article collection, a widely used corpus in the context of textual analysis and topic detection. The collection comprises 2,225 complete news articles collected from 2004 to 2005 regarding five topics: business, entertainment, politics, sport and technology (Greene & Cunningham, 2006).

As text preprocessing, we removed non-alphanumeric characters, numbers and words composed of 1 or 2 characters, divided the text into tokens (unigrams), removed the stopwords using a stoplist provided with the dataset, and finally stemmed the text. Then, we removed words with a value of tf-idf less than 0.01 (Allahyari *et al.*, 2017). At the end of the preprocessing step, the number of unique word tokens was equal to 18,422.

The word co-occurrence matrices were generated by counting the number of times two words co-occur in the same document within a specific window size, that is a set of neighbouring words within a specified distance, respectively equal to 2, 5, 10, 15 and 20 words on the right of the baseline word.

Afterwards, we defined different filters and weighting schemes on the word co-occurrence matrices. The first aspect was examined by removing from the word co-occurrence matrix the 100, 500, and 1000 words with the lowest co-occurrence values and the 50, 100, and 500 words with the highest co-occurrence values.

On the other hand, the second aspect was tested by considering an additional weighting scheme based on word proximity. In this case, we assigned more weight to the words nearest the target one inside the window. For example, for a window size equal to 5, we set a weight equal to 1 to the word adjacent to the target word, a weight equal to 4/5 to the next word and so on, until the last word, which takes a weight equal to 1/5.

Finally, we employed the Louvain community detection algorithm, Newman's leading eigenvector algorithm and the SLPA algorithm to discover topics in text networks obtained from the word co-occurrence matrices (interpreted as weighted adjacency matrices). The first two algorithms are non-overlapping community detection algorithms based on modularity maximisation, while the third is an overlapping community detection algorithm (Blondel *et al.*, 2008, Newman, 2006, Xie *et al.*, 2013).

The choice of using an overlapping community detection algorithm lies in the hypothesis that while non-overlapping community detection algorithms could correctly assign topics' characteristic words, multi-topic words could be arbitrarily assigned to one of the communities they should have been included.

3 Results

Our findings showed that with the increase in the window sizes, the number of communities found by the three algorithms decreases, remaining stable for window sizes greater than 5. In particular, for window sizes greater than 2, the number of communities found by the Louvain algorithm and Newman's algorithm is always greater than the number of communities identified by SLPA, which finds only one community with these settings.

Further to this point, Newman's algorithm generally finds three communities in the different experimental settings, while the Louvain algorithm finds almost always five communities for window sizes greater than 5. Notice that the communities found by the Louvain algorithm are coherent in number and content with the BBC news articles collection's topics. Interestingly, when the Louvain algorithm finds a number of communities greater than five for window sizes greater than 5, they are pretty unbalanced, with five bigger communities coherent with the original topics.

Computing the ARI (Hubert & Arabie, 1985) on the communities found by the Louvain algorithm under different settings, we observed that the ARI is generally high for different window sizes, particularly between the partitions obtained for window sizes greater than 5 (ranging from 0.604 to 0.878). This result shows that even if the algorithm finds the same number of communities, they are not identical.

Filters on words with the lowest degree from the word co-occurrence matrix do not affect the results. Conversely, removing words with the highest word co-occurrence remarkably increases the number of communities found for a window size equal to 2 (ranging from 27 to 112).

Similarly, using a different weighting scheme does seem to affect the number of communities found, which is noticeably higher when we use the proximity weighting scheme. However, also in this case, increasing the window sizes decreases the number of communities found.

4 Conclusions

The results obtained show that different design choices during text preprocessing and network definition affect the features of topics detected, mainly in terms of the number of topics discovered. For future work, we aim to focus on extending the assessment of the effects of these design choices on different kinds of texts, such as textual social media (like Twitter or Facebook).

References

ALLAHYARI, M., POURIYEH, S., ASSEFI, M., SAFAEI, S., TRIPPE, E.D., GUTIERREZ, J.B., & KOCHUT, K. 2017. A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv:1707.02919*, 1–13.

BLONDEL, V.D., GUILLAUME, J., LAMBIOTTE, R., & LEFEBVRE, E. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 1–12.

GREENE, D., & CUNNINGHAM, P. 2006. Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering. *Pages 377–384 of:* COHEN, W., & MOORE, A. (eds), *Proc. 23rd International Conference on Machine learning (ICML'06)*. ACM Press, New York.

HAMM, A., & ODROWSKI, S. 2021. Term-Community-Based Topic Detection with Variable Resolution. *Information*, **12**, 221–252.

HUBERT, L., & ARABIE, P. 1985. Comparing partitions. *Journal of classification*, **2**, 193–218.

NEWMAN, M.E.J. 2006. Finding community structure in networks using the eigenvectors of matrices. *Physical review E*, **74**, 1–12.

XIE, J., KELLEY, S., & SZYMANSKI, B.K. 2013. Overlapping community detection in networks: The state-of-the-art and comparative study. *Acm computing surveys (csur)*, **45**, 1–35.