Non-negative Matrix Factorization for Topic Modeling

Alberto Purpura
University of Padua
Padua, Italy
purpuraa@dei.unipd.it

ABSTRACT

In this abstract, a new formulation of the Non-negative Matrix Factorization problem for topic modeling will be presented. It allows the user to iteratively improve the topic model with a higher level of detail than current NMF-based approaches such as [2], achieving a higher performance than other weakly-supervised LDA-based methods.

CCS CONCEPTS

• **Computing methodologies** → **Topic modeling**; *Natural language processing*;

KEYWORDS

Topic modeling, Non-negative Matrix Factorization, Natural Language Processing

1 INTRODUCTION

Topic modeling is a convenient way to analyze and classify large quantities of documents. The most successful techniques for document classification (DC) rely traditionally on Support Vector Machines [4]. However, the availability of large datasets to train these classifiers is limited, especially for domain-specific tasks. Therefore, unsupervised approaches such as Latent Dirichlet Allocation (LDA) [1] or Non-negative Matrix Factorization (NMF) [6] have received an increasing attention over the years. NMF, in particular, is a fast [8] and easily implementable [5] method for unsupervised or weakly-supervised [2] classification of documents with relatively few hyperparameters to tune and easily-interpretable results.

2 NON-NEGATIVE MATRIX FACTORIZATION

NMF has two main advantages when compared to LDA. The first is that there are completely deterministic algorithms for its resolution [5]. Second, NMF allows for an easier tuning and manipulation of its parameters [9]. The UTOPIAN system [2] is an example of a topic modeling framework based on this technique that allows users to interact with the topic model and steer the result in an user-driven manner without any knowledge of how topic models work. This way of interacting with the data allows for a combined exploration and improvement of the results. The factorization problem tackled in UTOPIAN, however, can be improved in order to allow the user to express his/her opinion on the model and manipulate it in a more detailed fashion, using a prior matrix P – with the same shape of the term-topic matrix W – as shown in (1). Matrix P is used to add a few constraints on the NMF optimization problem, concerning the relations between the terms in the documents and a set of topics:

$$\min_{W,H\geq 0} ||A - WH||_F^2 + \phi(\alpha_p, W, P) + \psi(H), \quad (1)$$

where $\psi(H) = \beta \sum_{i=1}^{n} ||h_i.||_1^2$, and $\phi(\alpha_p, W, P) = \sum_{i=1}^{m} \sum_{j=1}^{k} \alpha_{p\,ij} (w_{ij} - p_{ij})^2 + \alpha||W||_F^2$. Here, α (scalar) and α_p (matrix shaped like W) are two regularization parameters, while – using the same notation of [9] – H and A are respectively the topic-document and the term-document matrix. This formulation allows the user to express more detailed preferences on how the classification should be performed by specifying which terms in the documents imply certain topics. This method achieves, with a weaker supervision, comparable results in DC tasks [7] to other LDA-based weakly-supervised models such as [3].

3 CONCLUSIONS

The presented formulation of the NMF optimization problem allows adding some constraints, in a weakly-supervised fashion, in order to improve the quality of the results of the DC system. For this reason, it can be a great choice as a weakly-supervised method for DC in many real-world scenarios. In order to fully exploit the inclusion of user-feedback in the classification process, the development of an online update algorithm of the classification – still missing, to our knowledge, in the literature – would be extremely helpful to immediately assess its quality after each interaction of the user.

REFERENCES

[1] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.
[2] Jaegul Choo, Changhyun Lee, Chandan K Reddy, and Haesun Park. 2013. Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization. *IEEE transactions on visualization and computer graphics* 19, 12 (2013), 1992–2001.
[3] Aitor García-Pablos, Montse Cuadros, and German Rigau. 2018. W2VLDA: almost unsupervised system for aspect based sentiment analysis. *Expert Systems with Applications* 91 (2018), 127–137.
[4] Thorsten Joachims. 2002. *Learning to classify text using support vector machines: Methods, theory and algorithms*. Vol. 186. Kluwer Academic Publishers Norwell.
[5] Jingu Kim, Yunlong He, and Haesun Park. 2014. Algorithms for nonnegative matrix and tensor factorizations: A unified view based on block coordinate descent framework. *Journal of Global Optimization* 58, 2 (2014), 285–319.
[6] Pentti Paatero and Unto Tapper. 1994. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics* 5, 2 (1994), 111–126.
[7] Alberto Purpura. 2018. *Weakly and Semi-Supervised Approaches for Aspect-Based Sentiment Analysis.* Master's thesis. University of Padua.
[8] Stephen A Vavasis. 2009. On the complexity of nonnegative matrix factorization. *SIAM Journal on Optimization* 20, 3 (2009), 1364–1377.
[9] Fei Wang, Tao Li, and Changshui Zhang. 2008. Semi-supervised clustering via matrix factorization. In *Proceedings of the 2008 SIAM International Conference on Data Mining*. SIAM, 1–12.

This work was supported by the CDC-STARS project and co-funded by UNIPD.

DESIRES 2018, August 2018, Bertinoro, Italy