4th International Conference on Advanced Research in SOCIAL SCIENCES
26-28 November , 2021 Oxford – United Kingdom

Referencing Scientific Articles by LDA Technology

Boussaadi Smail[1], Hassina Aliane[2], Pr Abdeldjalil Ouahabi[3]

[1]Researcher scholar, LIMPAF Laboratory, University of Bouira, Algeria.

[2]Director of Information Sciences R&D Laboratory, Cerist,Algiers, Algeria.

[3]List LIMPAF laboratory, University of Bouira, Algeria.

Abstract

Finding relevant scientific articles is a laborious process, which requires a lot of time and effort, especially on online journal databases, where article research is based on metadata such as keywords or authors' names. In this context we are interested in the referencing of articles by topics they address, thanks to topic modeling technology as LDA applied, on article metadata (Title, abstract, keywords). Based On the preliminary results, it is shown that the combination of some metadata generates a distribution of words on topics of high coherence and at the same time allows to identify the subjects of interest of each scientific article with a better interpretability

Keywords : Latent Dirichlet Allocation, Scientific Article, Topic Modeling.

1. Introduction

The extraction of topics from a scientific corpus is essential for many applications, such as the personalization of content (Boussaadi et al,2020) or the classification of documents (Yi,& Allan,2009). In the rest of this publication, we refer to a scientific corpus any database of scientific publication of journal-style whose structure is subdivided into sections as follows: Title, Authors and Affiliation, Abstract, Introduction, Methods, Results, Discussion, Acknowledgments, and References (Literature Cited).In this context to understand the content of a scientific publication, it is necessary to extract and analyze its topics (Roy,2021). a topic is a list of words that occur in statistically significant methods.

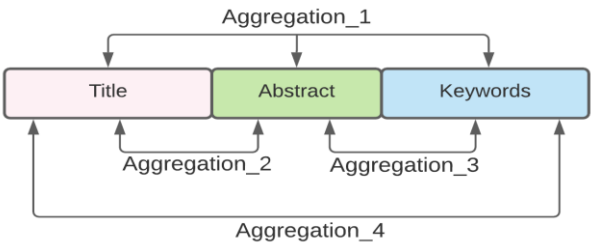
The process of learning and identifying these topics is called topic modeling. The search for a relevant scientific publication, through topics, is less laborious than the classic keyword search, for a researcher, in this context to understand the content of a scientific publication, it is necessary to extract and analyze its topics (Roy,2021). The process of learning and identifying these topics is called topic modeling. The search for a relevant scientific publication, through topics, is less laborious than the classic keyword search, for a researcher, the first method greatly reduces the amount of information returned, which is again in time and productivity. Topic modeling is a semantic process based on hidden (latent) variables, the objective of thematic modeling is to discover these variables.

We are interested in this article in the application of the LDA algorithm (Blei et al,2003), for the extraction of topics from different combinations of metadata (Title, abstract, keywords) in a scientific corpus to reference each article of the corpus by these own topics.

This work presents an experimental study of theme extraction from scientific publications, specifically, we apply the LDA (Latent Dirichlet Allocation) algorithm on different combinations of scientific publication metadata (Fig.1) and compare the results of each combination in terms of coherence.

After presenting in section 2 the previous work on publication representation by topics, we describe in section 3 the methodology we followed to create the different learning models to generate the topics. In section 4 we discuss the experimental results before ending with a conclusion and perspectives in section 5.

Figure 1: Metadata aggregation



2. Related Work

The topic modeling methods are generally used for automatically organizing, understanding, searching, and summarizing large electronic archives. Most of the work in topic extraction started with latent semantic analysis, a mathematical and statistical method proposed in the late 1980s by (Deewester et al,1990). The method was improved by a probabilistic p-LSA approach proposed by (Hofmann,2004) and has been widely used in many related fields such as information retrieval, information filtering, NLP, and machine learning, but the approach has been prone to the problem of overfitting. the LDA approach is proposed to correct the limitations of the two approaches, LDA algorithm is a very popular technique for semantic exploration and inference of subjects (Boussaadi,&Aliane,2020), notably in the system's recommendations of scientific publications (Boussaadi,&Aliane,2020).

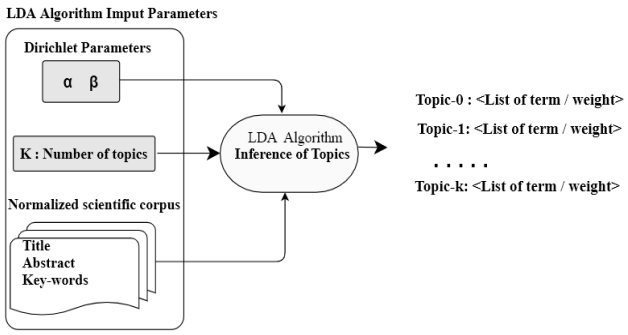
2.1 LDA Model

Latent Dirichlet Allocation is an unsupervised method of topic modeling. The basic idea is that the documents are represented as random mixtures over latent topics, where a topic is characterized by a distribution over words. The LDA inference is a list of topics: This contains relevant words and their occurrences in these obtained topics Fig.2.

From a methodological point of view, LDA extends the p-LSA model by introducing two variables alpha and beta (α , β) of Dirichlet types called hyperparameters and graces to which the method becomes a completely generative model, capable of inferring the topics on new documents.

- The “α” hyperparameter controls the number of topics expected in the document.
- The ‘β’ hyperparameter controls the distribution of words per topic.

Figure .2 Plate Notation of Latent Dirichlet Allocation (LDA)



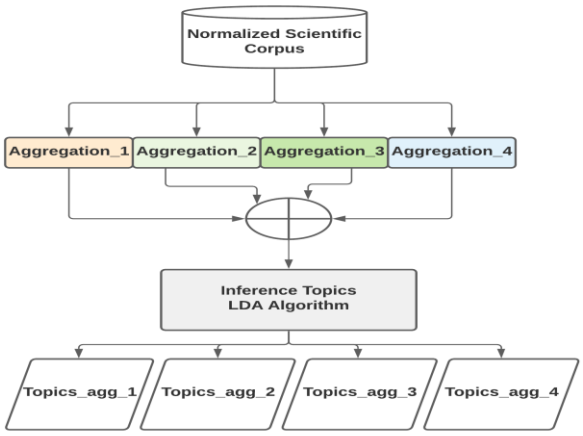
Various methods have been proposed to estimate LDA parameters α and β, such as the variational method (Blei et al,2003), expectation propagation (Minka, et al,2002), and Gibbs sampling (Griffiths et al,2004).

3. Method

This section describes the different steps of our method for topic inference, to answer the problem of topic extraction in a scientific corpus, our method is organized in three different steps Fig.3, namely: the collection of unstructured and semi-structured data from sites specialized in the publication of scientific articles, which will undergo a series of pre-processing to clean them and reduce the vocabulary used in the corpus. We then proceed to the selection of the features that interest us (title, abstract, keywords), to combine them in three aggregations (aggregation_1, aggregation-2, aggregation_3), which will constitute the inputs of the LDA algorithm.

The next step is the training of the data for the training of the model, so for each selected aggregation, the LDA algorithm infers topics, the objective of this step is the inference of three sets of topics that will be compared based on several metrics. The last step is the selection of the model that presents the best characteristics in terms of accuracy and human interpretability. The model in question will be used to reference each publication in our scientific corpus.

Figure.3 Workflow and data transfer scheme



3.1 Experiments setup

3.1.1 Experimental data

In our experience, we use the publicly available dataset presented by (Kershaw et al,2020). This is the first open corpus of interdisciplinary scientific research papers. This corpus includes not only the full text of the article but also the metadata of the documents, as well as title, abstract, keywords, and bibliographical references.

For our experiment, we have selected only the data related to the features we are interested in such as title, abstract, and keywords, as well as the data related to the authors and co-authors for analysis and evaluation purposes. The final dataset consists of 1040 scientific publications. We split our dataset into 80 percent data taken in train and 20 percent data were taken in a test. The detail of dataset statistics is shown in Tab.1.

Table 1: The statistics of the dataset

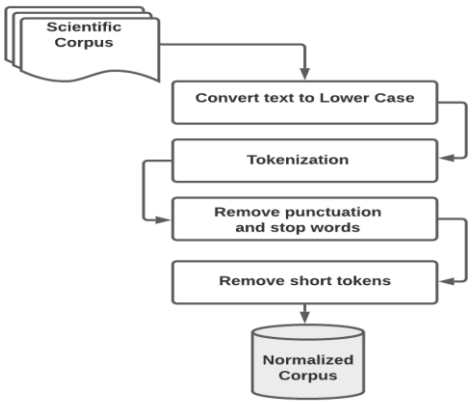
Number of documents	1040
Vocabulary size	10057
Number of a training set	832
Number of a test set	208

3.1.2 Preprocessing

The Data preprocessing phase is crucial for generating a util topic model Fig.4, and several technical issues from NLP are required according to our objectives, such as: Convert text data into lower case, Tokenization, Removal of stop words, and lemmatization by applying the NLTK python package. this step provides a cleaned and normalized corpus for training the model.

- Tokenization: the process of segmenting text into words, clauses, or sentences (here we will separate words and remove punctuation).
- Removal stop words: removal of commonly used words unlikely to be useful for learning, such as "the", "if", "and"...These words don't add any extra information in a text.
- Lemmatization: The goal of lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form.

Figure.4 Data Preprocessing



3.1.3 Parameter Settings

3.1.3.1 Selection of hyperparameters α and β:

For the choice of the Dirichlet hyperparameters α and β, we are inspired by the work of [12], adapted to a scientific corpus. The authors have fixed the values of α = 50/(number of topics) and β = 0.01 for a vocabulary of 20000 words, and the number of iterations fixed at 2000 to infer models with stable convergence.

3.1.3.2 Selection optimal number of the topic (K):

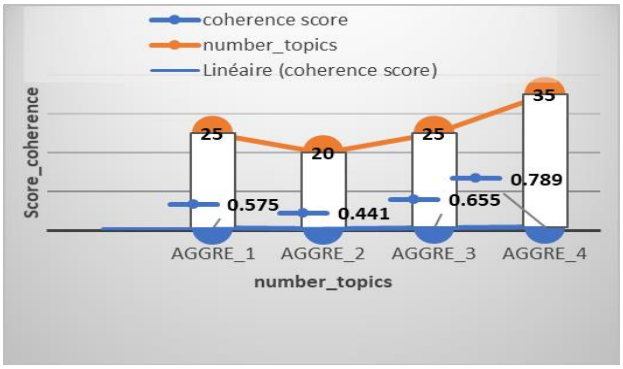
We set K in [10,20,30,40,50,60,70,80] empirically and choose a best one via several experiments Tab.2. The coherence metric is used to evaluate the quality of modeling ability of the model for each data aggregation.

A higher coherence score offers meaningful and interpretable topics. Optimal number of topics are selected based on the highest coherence score Fig.5.

Table 2: Optimal number of topics per aggregation

Training data	***Score coherence***	***Number topics***
Aggregation_1	*0.575*	*25*
Aggregation_2	*0.440*	*20*
Aggregation_3	*0.650*	*25*
Aggregation_4	***0.879***	***35***

Figure.5 The coherence scores and number of topics for each aggregation.



3.1.3.3 Model training and topic inference

The LDA calculation module is implemented by an open-source library called MALLET (Machine Learning for Language Toolkit) (McCallum A,2002).

4 Results and analysis

Tab. 3 and Tab.4, show the results achieved by each aggregation of data, using LDA with the optimal parameters for several topics (K) and the hyperparameters α and β, and for space reasons, we have selected only the first 5 topics for each aggregation.

By examining the distributions of the most important words for each topic, we find that the topics inferred by the data of aggregation_4 are of high precision and great interpretability by humans. The high consistency score of the model relating to the aggregation_4, once again demonstrates that the choices of the alpha and beta parameters as well as an adequate number

of topics, can identify the topics of a scientific corpus with high precision and make it possible to reference each scientific publication by the topics that correspond to it.

Table 3: The most frequent words for each topic

Aggregation_1					Aggregation_2				
Topic1	Topic2	Topic3	Topic4	Topic5	Topic1	Topic2	Topic3	Topic4	Topic5
Cortical Human Sky Order axon	Health Car Program Age process	data science Process Mind old	Title Blood Hospital Mean health	Security Machine Risk Program list	Cold Brain Human Car bad	Risk Second Machine Analysis server	Human Age Science Mind program	Axon Adult Age Brain clinical	Security Data Old Server axis

Table 4: The most frequent words for each topic

Aggregation_4					Aggregation_3				
Topic1	Topic2	Topic3	Topic4	Topic5	Topic1	Topic2	Topic3	Topic4	Topic5
Brain **Axon** **Neural** **Cortical** **Survival**	**Human** **Age** **Adult** **Older** **health**	**Science** **Computer** **Program** **Language** **java**	**Health** **Cancer** **Clinical** **Human** **blood**	**Information** **security** **management** **analysis** **risk**	Human Blood Brain Sky old	Science Data Risk Information security	Cancer Blood Clinical Age hospital	Axon Neuron Pain Human transit	Program Java Code Science student

5 Conclusion and perspective

In this article, we have presented a method of topic inference from a scientific corpus. We have selected the metadata such as the title, the abstract, and the keywords to compose several aggregations, to create LDA models for each aggregation, the objective being to select the most precise and interpretable topics, to use the corresponding model to reference the publications of the corpus. The results of the experiment have shown that the combination (title and keywords) constitutes a better source of topic modeling for referencing the articles of the corpus. a perspective opens in this work which the use of the full text of each publication to increase the semantics between the words of the topics.

References

Boussaadi, Smail, Aliane, H., & Abdeldjalil, O. "Recommender systems based on detection community in academic social network." *2020 International Multi-Conference on: "Organization of Knowledge and Advanced Technologies*" (OCTA) (2020): 1-7.

Yi, X., & Allan, J. (2009). A Comparative Study of Utilizing Topic Models for Information Retrieval. *ECIR*.

Roy, A. (2021). Recent Trends in Named Entity Recognition (NER). ArXiv, abs/2101.11420.

Blei D.M., Ng A.Y., Jordan M.I.: Latent Dirichlet Allocation, *The Journal of Machine Learning Research*, vol. 3(1), pp. 993–1022, 2003.

Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnas, G.W., & Harshman, R.A. (1990). Indexing by Latent Semantic Analysis. *J. Am. Soc. Inf. Sci*., 41, 391-407.

Hofmann, T. (2004). Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning, 42*, 177-196.

Boussaadi, S., Aliane, H., & Abdeldjalil, O. (2020). The Researchers Profile with Topic Modeling. *2020 IEEE 2nd International Conference on Electronics, Control, Optimization and Computer Science* (ICECOCS), 1-6.

Boussaadi, S., Aliane, H., Abdeldjalil, O., Houari, D., & Djoumagh, M. (2020). Recommender systems based on detection community in academic social network. 2020 International Multi-Conference on: "*Organization of Knowledge and Advanced Technologies" (OCTA)*, 1-7.

Minka, T.P., & Lafferty, J.D. (2002). Expectation-Propogation for the Generative Aspect Model. *UAI*.

Griffiths, T.L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101, 5228 - 5235.

Kershaw, D. (Creator), Koeling, R. (Creator) (Aug 4 2020). Elsevier OA CC-BY Corpus. Mendeley Data. 10.17632/zm33cdndxs.2

Griffiths TL, Steyvers M. Finding scientific topics. Proc Natl Acad Sci U S A. 2004;101 Suppl 1(Suppl 1):5228-5235. doi:10.1073/pnas.0307752101.

McCallum A. Mallet: A machine learning forlanguage toolkit; 2002. Available from: *ttp://mallet.csumass.edu*.

M. Young, The Technical Writer's Handbook. *Mill Valley*, CA: University Science, 1989.