Using an Explicit Query and a Topic Model for Scientific Article Recommendation

Boussaadi smail (sboussaadi@gmail.com)
university of Bouira- Algeria https://orcid.org/0000-0002-7033-4183

Research Article

Keywords: Topic modeling, Latent Dirichlet Allocation, Non-negative Matrix Factorization, Recommendation, scientific article

Posted Date: April 1st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1506014/v1

Abstract

The search for relevant scientific articles is a crucial step in a scientific research project. However, the considerable number of articles published and uploaded online in digital databases (such as google scholar, semantic scholar, etc.) makes this task tedious and negatively impacts the researcher's productivity. This article proposes a new method of recommending scientific articles, taking advantage of content-based filtering. The challenge is to target the relevant information to meet the researcher's needs regardless of his research's domain. Our recommendation method has based on a semantic exploration using the latent factors. Our goal is to achieve an optimal topic model, which will serve as the basis for the recommendation process. Experiences confirm our performance expectations by showing relevance and objectivity in the results.

1 Introduction

Recommender systems (RS) of scientific articles are applications designed with modern engineering methods and artificial intelligence algorithms. Their roles are to filter enormous scientific publications in online scientific databases (such as google scholar, semantic scholar). Indeed, one of the obstacles that researchers face is the problem of cognitive overload, and scientific recommendation systems are an effective solution to this problem [1]. In the literature, we distinguish four approaches for recommending scientific articles. The classic approaches are based on collaborative filtering (CF) [1] and based on content filtering (CBF) [1] and recently appeared approaches based on graphs (GB) [2], and approaches based on the topic modeling [3]. Hybrid approaches combine two or more techniques from previous approaches to improve the quality of recommendations [4]. Content-based SR is similar to machine learning classifiers. The underlying logic relies on the knowledge of the researcher's interactions, which causes the cold-start problem and the overspecialization problem that prevents the discovery of new publications [2].SRs based on collaborative or social filtering relies on social opinions. Although they can theoretically adapt to any domain, they suffer from the double problem of cold-start (new researcher / new article). Indeed, the recommendation process cannot identify similar researchers for a new researcher without a history of evaluations or very few ratings; the same problem occurs with

articles [4]. Graph-based approaches mainly focus on building a graph to model a citation network (Citation Graph) or a social network; nodes and links, respectively, model researchers and articles. A major limitation of this approach is that the recommendation process does not consider the articles' content and the researchers' topic interests [2]. Topic model-based SRs use latent factors to model the topic interests of researchers or identify communities of researchers in the recommendation process [5]. In this article, we are interested in a new method based on an explicit query, combined with latent factors, to minimize the limitations of the first-mentioned approaches by translating the explicit requirements of the researcher into topics.

2 Related Work

This section briefly reviews the scientific literature that addresses the problem of topic model-based scientific article recommendation. The topic model algorithms most used in scientific article recommendation systems design are LDA (latent Dirichlet allocation) [6]. In [5], the authors introduced the Collaborative Topic Regression (CTR) recommendation model, which combines matrix factorization and LDA into a single generative process, where articles latent factors are obtained by adding an offset latent variable to the topic proportion vector. In [7], the authors proposed a hybrid approach combining latent factors produced by LDA and relevance-based language modeling and using the clustering of searchers based on latent topics of interest as reliable sources to produce recommendations; the approach suffered from consistency problems and interpretability of inferred topics. The authors [8] proposed a topic model based on researcher interactions rather than article content to extract consistent and interpretable topics.

In [9], the authors combined collaborative filtering and latent topics to detect community researchers based on the dominant topic in an academic social network; the approach significantly improves the relevance of the recommendations compared to existing approaches. However, the recommendation process relied on a single source of information to identify the thematic interests of the researchers, which reduced the accuracy and interpretability of the researchers' profiles. For more details on the recommendation of scientific articles, we recommend the extensive survey work done by [3]. The major challenge with previous work is that the inferred topics suffered from consistency and interpretability, which limited the precision in identifying the researcher's thematic interests and therefore reduced the quality of the recommendations. This paper proposes a new method based on an explicit query formalism combined with latent topics. The proposed method does not suffer from the dual researcher/article cold-start problem or the overspecialization problem.

2.1 Problematic of Topic Modeling Technique

The major challenge and ambiguity involved in topic modeling is model validation. Both modeling techniques, LDA and NMF, have been the subject of several research studies to demonstrate the performance of each approach. In [9], the authors demonstrated that NMF obtains better results in the classification task, but LDA produces topics with better coherence. According to the study's authors, it is more likely to work well for humans because of the stable coherence of LDA.

In [10], the authors used a corpus of 13,000 citations on the subject of COVID-19 to compare the consistency of the two techniques and found that, for applications in which a human end-user will interact with learned topics, the flexibility of LDA and the coherence advantages of LDA warrant strong consideration. They conclude that the LDA model is more relevant than the NMF model in the case of the large corpus. In [11], the authors explored human concerns towards the COVID-19 vaccine using

Twitter data and concluded that NMF performed better than LDA in the coherence score.

LDA tends to produce more coherent and human-interpretable topics than NMF. However, NMF tends to give better classification potential regardless of data size. Both techniques have compelling arguments

depending on the context of their use. These arguments serve as a basis for comparing the results of the two techniques in our study and exploiting the better-performing algorithm for the recommender process.

3 Concept Related To Our Research

2.2 Topic Model (TM)

Topic models (TM) are an essential machine learning technology that is used

in different fields such as text analysis [6], recommendation systems [12], image

analysis [13], medical sciences [14]. There are many articles in this domain, and we definitely cannot mention all of them. TM can be used to discover hidden topics in a text collection such as documents, short texts, chats, Twitter and Facebook posts, user comments on news pages, blogs, and emails.

2.3 Topic coherence Measures

The Coherence Score is a widely used performance metric to evaluate topic modeling methods. It gives a realistic way to figure out how many different topics there are in a document. Each generated topic has a list of words, such as a cluster of words. This measure looks at the average pairwise word similarity scores of the words that are linked to the subject. The topic model with high Coherence Measure value is considered a suitable topic model.

Let's take a quick look at the most popular coherence measures and how they are calculated:

C_v is a measure based on a sliding window, one-set segmentation of the top words, and an indirect confirmation measure that uses normalized pointwise mutual information (NPMI) and the cosine similarity.

C_umass is based on document co-occurrence counts, a one-preceding segmentation, and a logarithmic conditional probability as confirmation measure.

2.4 Finding Semantic Similarity Articles

The vectors representing the articles are probability vector representations of topic distributions. We select for a modified Jensen-Shannon divergence (JSD) measure [15]. Indeed, when employed for similarity computation, JS Distance cannot determine the semantic link between subjects. To address this issue, [16] proposes a novel technique for assessing similarity that considers the semantic correlation of subjects from the perspective of word co-occurrence and augments the original JS measurement method by computing the semantic correlation of topic feature words.

Assume z_i is the topic of the article d_i, word set $W = \{w_{i1}, w_{i2}, \ldots, w_{in}\}$ is the feature word of the topic. According to the co-occurrence Eq. (1).

$$p\left(w_{im}, w_{in}\right) = p\left(w_{im} \mid z_i\right) \times p\left(w_{in} \mid z_i\right)$$

1

Where $p_{11}, p_{12}, \ldots, p_{nn}$ is the co-occurrence probability of feature word. The similarity computing equation of w_{im} and w_{jn} is as shown in Eq. (2).

$$Correlation\left(w_{im}, w_{jn}\right) = \frac{p_{mn}}{p_{im} + p_{jn} - p_{mn}}$$

2

Hence, the semantic similarity between two articles d_i and d_j is shown in Eq. (3).

$$Similarity\left(d_i, d_j\right) = \lambda D_{js}\left(d_i, d_j\right) + (1 - \lambda)\left[\frac{\sum_{m,n=1}^{v_d}\left(1 - correlation\left(w_{im}, w_{jn}\right)\right)}{\left(v_d\left(v_d - 1\right)\right)}\right]$$

3

v_d denotes the number of feature words of this selected article. $\lambda \in [0, 1]$ denotes a correlation coefficient assigned to this article. D_{js} is the Jensen-Shannon distance.

4 The Proposed Approach

This section presents our method for scientific article recommendations to solve the limitations mentioned in section 2. Our proposal includes first a selection step of the algorithm that generates the most coherent topic model, which will serve as a basis for the recommendation process, the purpose of this step is a reference matrix of articles/topics; to achieve this goal, we use, the most popular topic modeling algorithms, namely LDA and non-negative matrix factorization (NMF) [17].

In the second step, the recommendation process performs a semantic similarity calculation to generate a list of relevant articles, which will present to the target researcher Figure.1

Notation and approach

We denote by $A = \left\{A_1, A_2, \ldots, A_n\right\}$ the set of target researchers, and by A_i a generic researcher in A, and by $D = \left\{d_1, d_2, \ldots, d_m\right\}$, our corpus, that contains the articles that could be potentially interesting to our generic researcher.

The target article d_i is represented by a topic distribution associated with the predominant topic (obtained by applying the best-performing algorithm between LDA and NMF).

Our recommendation algorithm aims to present to a target researcher (A_i), a list of the most relevant and similar articles to the target article. The proposed approach has the following two steps.

Step 1:

- Application and evaluation of the LDA and NMF algorithms on the experimentation corpus (with different combinations of hyperparameters), the goal is to select the best performing algorithm, which we call algorithm_1.
- Referencing the recommendation corpus, by applying algorithm_1. each article will be represented by its predominant topic.

Step 2:

We accept the researcher's query to identify the target article (designated d_j).

- Retrieval all the meta data set of the target article d_j (from google scholar to be precise) and we apply algorithm_1. it is assumed that the target article d_j is referenced by its predominant topic designated by topic_d.
- For each article d_i $\left(d_i \in D, d_i \neq d_j\right)$ and all articles referenced by topic_d do:

Computing $Similarity\left(d_i, d_j\right)$ using Eq. (3).

- Ranking of the similar articles in descending order of relevance, and recommendation to the generic researcher A_i the top-ranked articles.

5 Experimental Design

5.1. Experimental Dataset and Tools

We use the dataset of Elsevier; this is a corpus of 40 001 open access articles from across Elsevier's journals representing a large scale, cross-discipline set of research data to support Natural Language Process (NLP) and Machine Learning (ML) research Fig. 2. The corpus includes the articles' full text and metadata (title, abstract, keywords). We split the dataset into two smaller datasets; the first is designated DM, we use it for the selection of the best performing algorithm. The second dataset is designated by DR; we use it for the recommendation process; the statistical details of our datasets are as follows Table 1.

Table 1
The statistics of the dataset

Dataset	Number of Articles	Vocabulary size	Training set size	Training set test
DM	1600	400k	70%	30%
DR	2400	675k	80%	20%

5.2 Data preprocessing

Text document preprocessing is an essential task for text exploring semantics. Data preprocessing is used to extract information into a structured format to analyze the patterns (hidden) within the data. The preprocessing steps are supported in Stanford's NLTK Library [18] and contain the following patterns:

- Convert Text Data into Lower Case: Text datasets are converted into lower case for preventing the various words differences.
- Punctuation: punctuation such as (“.”, “,”, “–”, “!” etc.) are eliminated from the text datasets.
- Stop-word elimination: removal of the most common words in a language such as “and”, “are”, “this” etc, that is not helpful and in general unusable in text mining and that do not contain applicable information for the study.
- Stemming: the conversion of words into their root, using stemming algorithms such as Snowball Stemmer.
- Tokenization: Text datasets are converted into tokens (words). The tokenization identifies the meaningful keywords from the text data. The outcome of tokenization is a sequence of tokens.

5.3 Topic Model Algorithm

The NMF implementation was done by scikit-learn's NMF (version Optimized for the Frobenius Norm) functionality. For LDA, we use two different implementations as follows:

- The Gensim algorithm was implemented with the Gensim library, a Python library for topic modeling; the algorithm utilizes the inference technique online variational Bayes [19].
- The Mallet topic model package incorporates a rapid and highly scalable implementation of Gibbs sampling.

The reason to choose these different implementations was how they differ in inference technique. There is no standard best solution for making inferences since it heavily depends on the dataset. Since the inference method most suitable for the scientific text was unknown, both implementations were reasonable to evaluate. For both implementations, the output of the models is in the same format as in NMF.

5.4 NMF and LDA Algorithm selection

Figure.3, shows the process applied to the DM dataset and evaluated by the quantitative metrics (section This procedure aimsure is to select the algorithm that meets our objective. Figures 4 and 5 clearly show that the LDA Mallet algorithm provides a better graphical quantitative evaluation for several subjects $k = 50$, compared to the other algorithms. for the rest of our work, we opt for the subject model produced by the LDA Mallet version.

5.5 Baseline Methods

In assessing the effectiveness of our proposed method, we compare the recommendation results with two methods representing baseline CBF methods. In the first method (denoted by LDA_ASPR), the authors [20] have exploited the topics related to the researcher's scientific production (authored articles) to define her/his profile formally; in particular, they employed the topic modeling to represent the user profile formally, and language modeling to formally represent each unseen paper. The recommendation technique they proposed relies on assessing the closeness of the language used in the researcher's papers and the one employed in the unseen papers. The authors proposed the PRPRS (Personalized Research Papers Recommender System) in the second method, which extensively designed and implemented a user profile-based algorithm to extract keyword by keyword and keyword inference [21].

5.6 Evaluation Metrics

We evaluate the general performance of our method using the two most commonly used evaluation metrics in recommender systems: Precision and recall. Precision or true

positive accuracy is calculated as the ratio of recommended articles relevant to the total number of recommended articles; Precision has given by Eq. (4).

$$Precision = \frac{\sum (relevant\_articles) \cap \sum (retrieved\_articles)}{\sum (retrieved\_articles)}$$

4

Recall or true positive rate is calculated as the ratio of recommended articles that are relevant to the total number of relevant articles, recall given by Eq. (5),

$$Recall = \frac{\sum (relevant\_articles) \cap \sum (retrieved\_articles)}{\sum (retrieved\_articles)}$$

5

6 Results And Discussions

Our method is designed to favor precision which has more influence on the satisfaction of the target researcher than recall. Indeed, precision reflects the performance of the recommendation system in satisfying the target researcher's need for valuable articles. The different experiments we have conducted

show that the scientific article recommendation method we have proposed provides the target researcher with scientific articles of high relevance compared to the state-of-the-art methods that rely on a single topic modeling technique. Our method first selects the best performing algorithm considering the textual nature (scientific text), which provides an optimal topic model, which serves as a basis for the recommendation process hence the accuracy obtained as illustrated in Fig. 6. Indeed, the accuracy of our method surpasses that of other methods, even if it is significantly close to the LDA_ASPR method for the values of N = 20. However, when the value of N increases beyond the value N = 25, it seems clear that the accuracy increases and surpasses that of the LDA_ASPR and PRPRS methods.

Figure.7 illustrates the comparison based on the recall; as we can see, the difference in performance between our method and LDA_ASPR is very insignificant. Indeed, the LDA_ASPR method performs slightly for N = 15 and N = 20; however, our method shows a significant difference as the number N increases, especially when N's value is greater than 25. The low recall-based performance of our method is the result of the qualification constraints of the candidate items. Thus, our method selects only those articles whose content is semantically very close to the content of the target article and thus leaves many other less relevant articles unrecalled. These improvements are mainly due to the strictness in qualifying a candidate paper which removed less relevant papers to the target paper. This, therefore, increases the system's ability to return relevant and practical recommendations at the top of the recommendation list.

7 Conclusion And Future Work

This article uses latent factors to optimize content-based filtering (CBF). Our approach analyzed the two most widely used topic modeling techniques to select the most optimum topic model to serve as the foundation for the recommendation process. There is no consensus in the scientific literature on which modeling approach performs best, and our strategy is based on this knowledge.

Our solution outperforms the reference methods on the most generally used performance criteria, as proven using a publicly accessible dataset. Our following study will examine other criteria that best characterize the content of scientific articles, such as citations, to improve performance.

References

1. Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE transactions on knowledge and data engineering, 17(6), 734–749.
2. Bai, X., Wang, M., Lee, I., Yang, Z., Kong, X., & Xia, F. (2019). Scientific paper recommendation: A survey. Ieee Access, 7, 9324–9339.
3. Beel, J., Gipp, B., Langer, S., & Breitinger, C. (2016). paper recommender systems: A literature survey. International Journal on Digital Libraries, 17(4), 305–338.
4. Sakib, N., Ahmad, R. B., & Haruna, K. (2020). A collaborative approach toward scientific paper recommendation using citation context. IEEE Access, 8, 51246–51255.

5. Wang, C., & Blei, D. M. (2011, August). Collaborative topic modeling for recommending scientific articles. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 448–456).
6. Amami, M., Faiz, R., Stella, F., & Pasi, G. (2017, August). A graph-based approach to the scientific paper recommendation. In Proceedings of the international conference on web intelligence (pp. 777–782).
7. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. the Journal of Machine Learning Research, 3, 993–1022.
8. Rossetti, M., Vargas, S., Magatti, D., Pettit, B., Kershaw, D., Hristakeva, M., & Jack, K. (2017, February). Effectively identifying users' research interests for scholarly reference management and discovery. In Proceedings of the 1st Workshop on Scholarly Web Mining (pp. 17–24).
9. Stevens, K., Kegelmeyer, W.P., Andrzejewski, D., & Buttler, D.J. (2012). Exploring Topic Coherence over Many Models and Many Topics. EMNLP.
10. Ben Msik Casablanca (2020). \Topic Modeling Coherence: A Comparative Study between LDA and NMF Models using COVID'19 Corpus". In: International Journal 9.4.
11. Khanjari Nezhad Jooneghani, Zeinab. COMPARISON OF TOPIC MODELING METHODS FOR ANALYZING TWEETS ON COVID-19 VACCINE. Master's Thesis. East Carolina University, July 2021.
12. Smail Boussaadi, Hasina Aliane, & Abdeldjalil Ouahabi. (2021). Recommender Systems Based on Detection Community in Academic Social Network. International Journal of Engineering Science, 171. https://doi.org/10.5281/zenodo.5801853
13. Geman, S., & Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-6, 721–741.
14. Zhang, X. P., Zhou, X. Z., Huang, H. K., Feng, Q., Chen, S. B., & Liu, B. Y. (2011). Topic model for Chinese medicine diagnosis and prescription regularities analysis: case on diabetes. Chinese journal of integrative medicine, 17(4), 307–313. https://doi.org/10.1007/s11655-011-0699-x
15. Lin, J. (1991). Divergence measures based on the Shannon entropy. IEEE Transactions on Information theory, 37(1), 145–151.
16. Shao, M., & Qin, L. (2014, March). Text similarity computing based on LDA topic model and word co-occurrence. In Proceedings of the 2nd International Conference on Software Engineering, Knowledge Engineering and Information Engineering (SEKEIE 2014) (pp. 199–203).
17. Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. Nature, 401(6755), 788–791
18. Albalawi, R., Yeap, T.H., & Benyoucef, M. (2020). Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis. Frontiers in Artificial Intelligence, 3.
19. Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. Nature, 401(6755), 788–791.

20. Amami, M., Pasi, G., Stella, F., & Faiz, R. (2016). An LDA-Based Approach to Scientific Paper Recommendation. NLDB.
21. Hong, K., Jeon, H., & Jeon, C. (2013). Personalized Research Paper Recommendation System using Keyword Extraction Based on UserProfile.

Declarations

The authors declare no competing interests.

Figures

Figure 1

Flowchart of proposed recommendation approach.

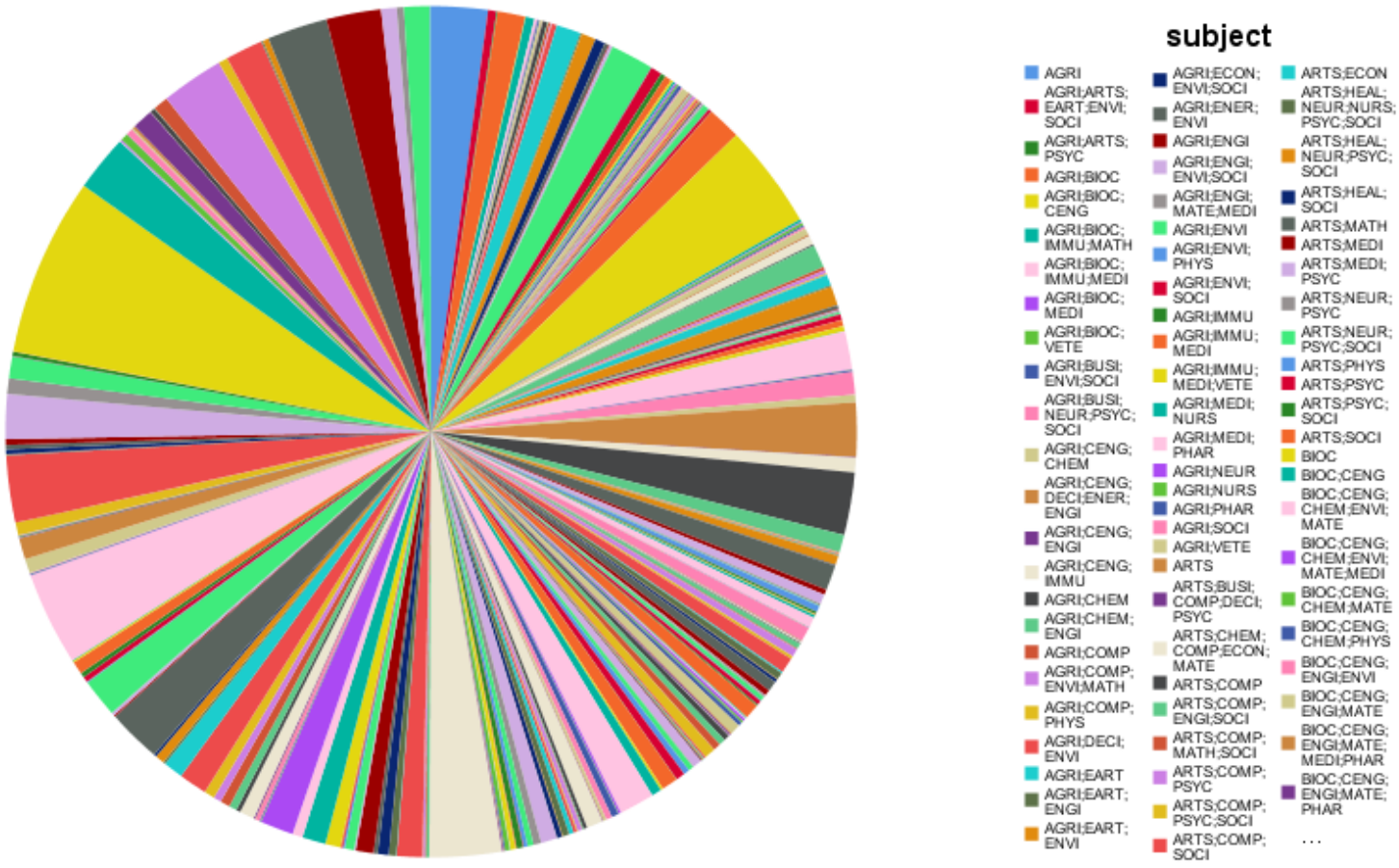

Figure 2

Subject distribution diagram

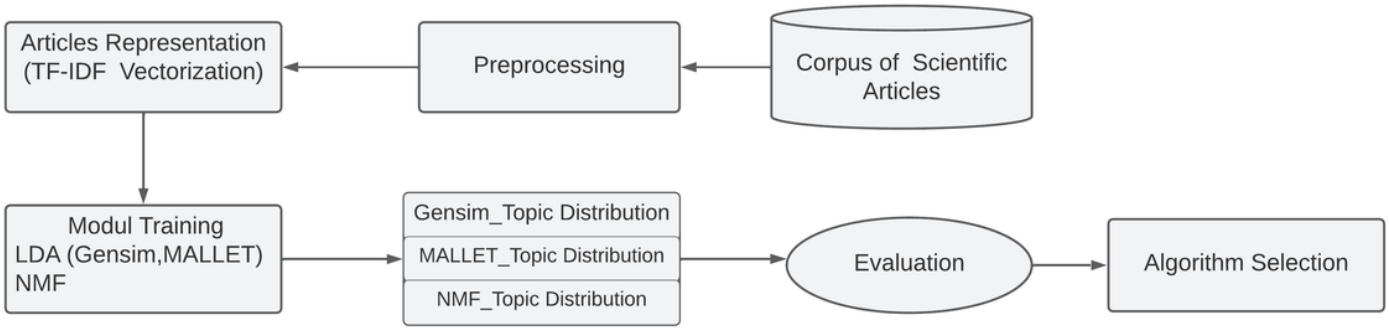

Figure 3

Block Diagram for Algorithm Selection

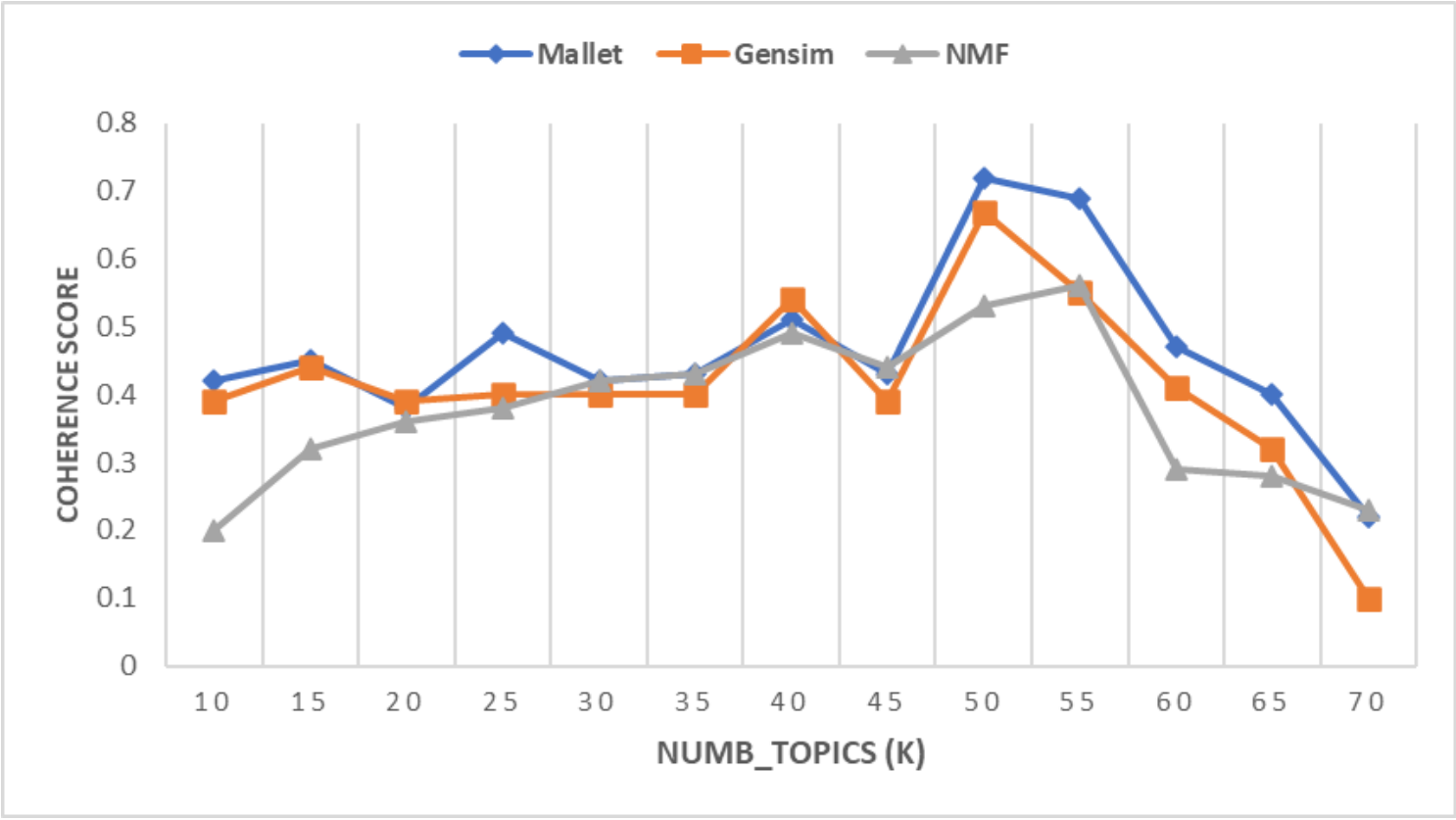

Figure 4

Quantitative (C_v score) results for LDA and NMF

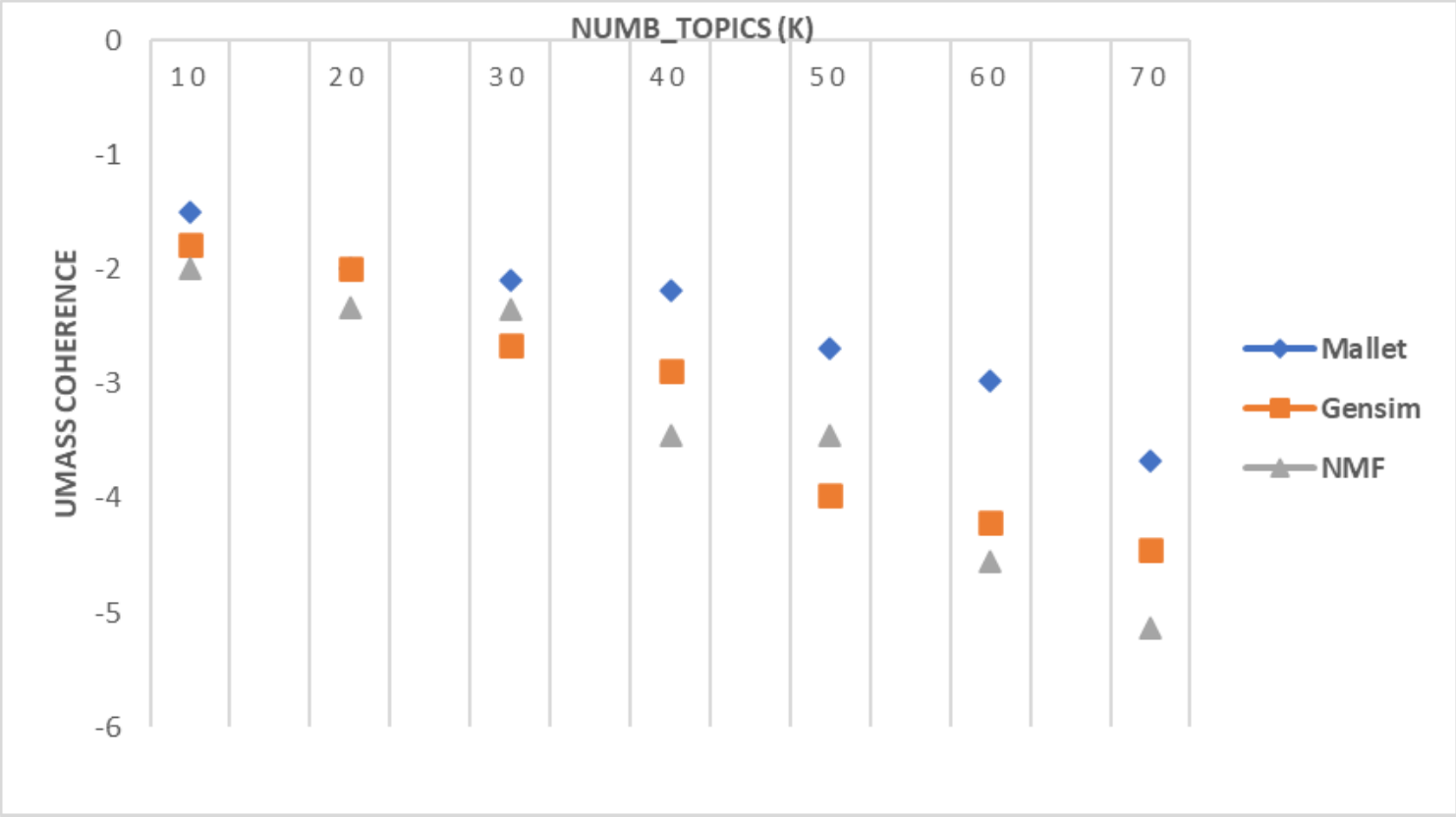

Figure 5

Quantitative (UMass score) results for LDA and NMF.

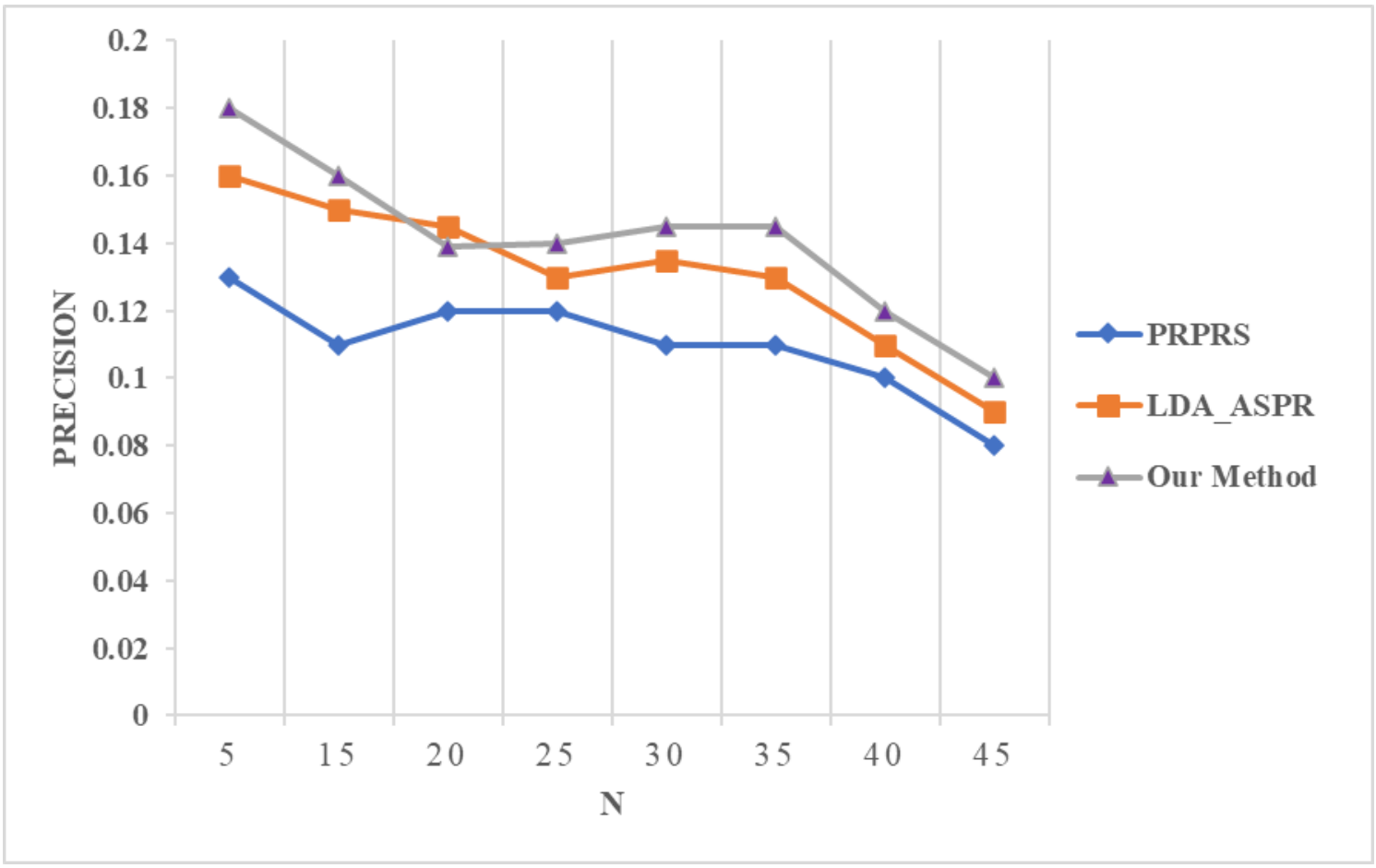

Figure 6

Precision performance on the dataset.

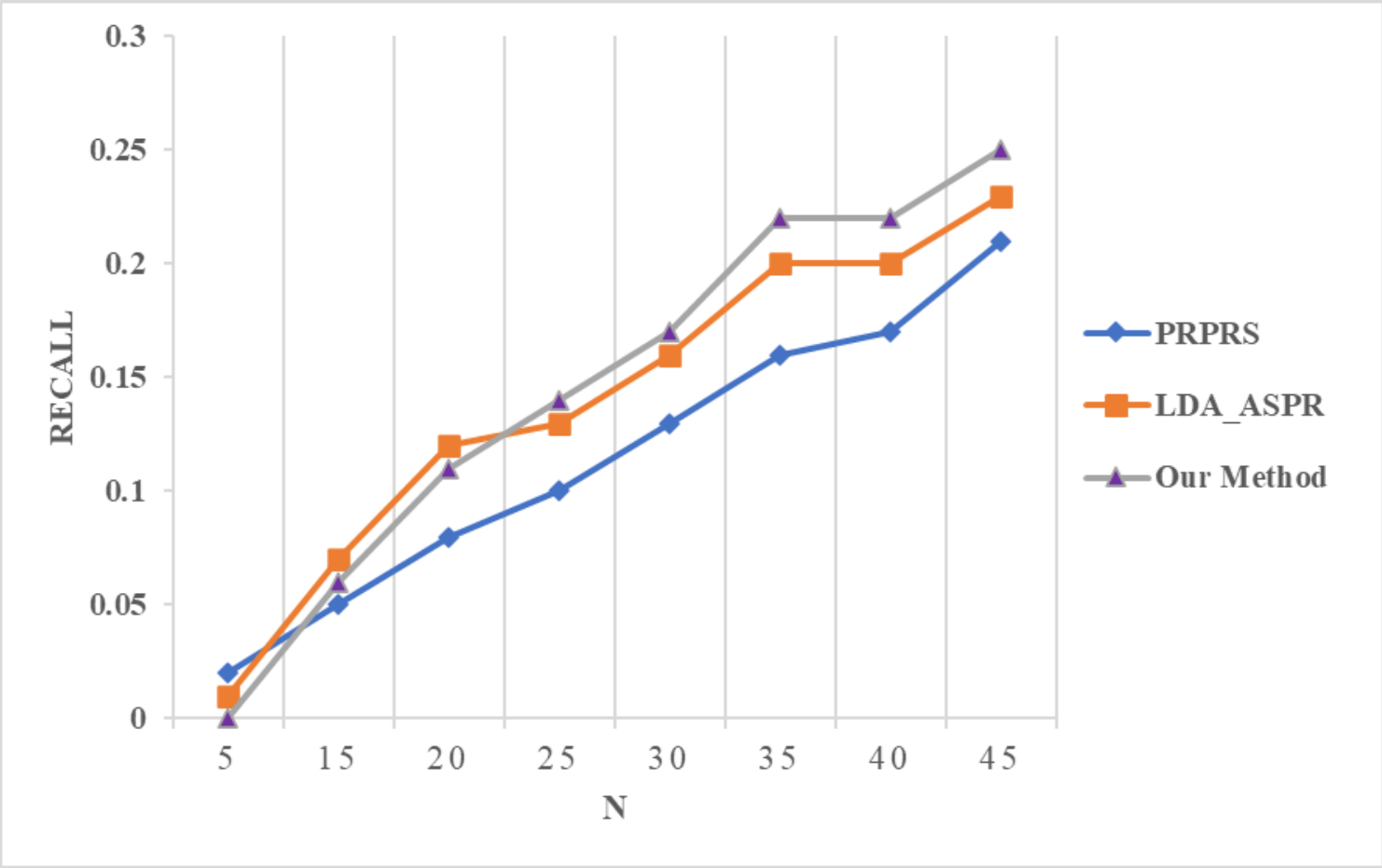

Figure 7

Recall performance on the dataset.