Recommender systems based on detection community in academic social network

Abstract— The speed with which new scientific articles are published and shared on academic social networks generated a situation of cognitive overload and therefore a major challenge for researchers. In this context, we propose a scientific article recommendation approach based on the discovery of thematic community structures, it focuses on the topological structure of the network combined with the analysis of the content of the social object (scientific article), a strategy that aims to mitigate the cold start problems and sparsity data in scoring matrix. A key element of our approach is the modeling of the researcher’s thematic centers of interest derived from his corpus (a set of articles that interested him). In this perspective we use the technique of semantic exploration and extraction of latent topics in document corpora, LDA (Latent Dirichlet Allocation), an unsupervised learning method which offers the best solution of scalability problem compared to other techniques of topic modeling, this technique allows us to build a profile model in the form of vectors in which the components are the probabilistic distributions on topics that reflect the interests of the researcher. The profile models thus constructed will be grouped into thematic clusters based on dominant topics using the fuzzy clustering algorithm, since the same topic can be treated in different scientific fields. Will follow a step of detection of community structures in thematic clusters to identify significant communities, the aim of this step is to project the recommendation process in a small space allowing better performance by reducing the computation time and the storage space for researcher/article data. The preliminary results of the experience of our approach on a population of 13 researchers and 60 articles shows that the articles generated by the recommendation process are very relevant to the target researcher or his community.

Keywords— LDA, Recommender system, academic social network, community detection, the fuzzy clustering algorithm.

I. INTRODUCTION

Academic social networks like generalist social networks provide millions of researchers with functionalities that allow them to promote their publications, to find relevant articles and to discover trends in their areas of interest. However, the rapidity with which new articles are published and shared, especially on these academic social networks, generates a situation of cognitive overload and is therefore a major challenge for the researcher in search of relevant and recently published information. It is in this context that scientific article recommendation systems are used to filter the huge amount of articles shared on these platforms. In recent years many researchers have become interested in recommending scientific articles [1][2][3][4][5].

The recommendation of scientific articles, aims at recommending relevant articles in correlation with the interests of a researcher or a group of researchers. However, due to their ever-increasing size, the analysis of academic social networks, on which a scientific article recommendation system is based, has become a complex task, and one of the strategies adopted in the scientific literature to overcome this difficulty is the partitioning of the initial network into smaller subgroups, for example, researchers sharing the same areas of interest will be grouped into thematic communities, a strategy that reduces processing time and storage space since the observations relate to a group of researchers interested in the same themes, hence the interest of the community concept, which provides a relevant analytical framework for Understanding the collective and organizational dynamics of the overall network [6].

The essential of the literature on recommending systems shows that there is a consensus on the Classification of Recommending Systems for the three categories: methods based on collaborative filtering, content-based methods and hybrid methods [8][9][10][11].

Our article is organized as follows. In section 2 we present the bulk of the literature on recommendation approaches based on theme modeling and thematic community schemes, in section 3 we present our recommendation approach that aims to mitigate the problem of cold start and data sparsity. In order to evaluate the effect of taking into account theme modeling in the identification of thematic communities in academic social networks we conducted experiments on a real data set that we have build around 13 researchers and 60 articles, the remaining part of our approach, i.e. recommendation generation, will be the subject of a forthcoming article.

A. Content-Based Methods (CBF)

The content-based filtering process essentially takes place in two phases. The learning of the profile vector of the active
researcher based on the history of his activity. There are different methods for this, including LDA (Latent Dirichlet Allocation) [12] (section 2.3). And the representation as a feature vector of the candidate article, defined by models such as TF-IDF which produces a weighted vector of terms [13] or sentence extraction that produces a description of the content of each candidate article in the form of a list of key words that reflect the essence of the topics addressed in each candidate article [14]. The second phase uses a similarity function that takes as input the profile vector of the active researcher and the vector representing the candidate article and provides a prediction score. Generally, the similarity function is the cosine of the angle formed by the two vectors in question, this last phase produces a ranked list of articles whose top N articles will be recommended to the target researcher [15]. In the field of recommending scientific articles, content-based filtering is the most widely used [16].

B. Methods based on collaborative filtering (CF)

Collaborative filtering is based on the sharing of opinions and evaluations between researchers on certain articles. The underlying idea is: if a researcher A evaluates or rates a U paper in the same way as another researcher B, if both researchers A and B have previously enjoyed other articles in a similar way. And unlike content-based filtering, collaborative filtering is independent of the content of candidate articles [17], [18]. Typically, researchers evaluations of articles are represented by a scoring matrix, where each line corresponds to a researcher's evaluation history. Recommendations produced in a collaborative filtering process are based on similarity between researchers.

According to [19][20] there are two classes for collaborative filtering, memory-based methods and model-based methods.

1) Memory-based methods

This method exploits the entire usage matrix (Fig) to generate recommendations. The term memory (neighborhood) refers to users as well as articles. Thus the algorithms of this method can be divided into two categories: user based or items based.

User based: introduced by GroupLens [21]. The principle of this method is to first determine which users are similar to the active user, which is equivalent to estimating the similarities between the row in the active user usage matrix with all other rows [22][23], then fills in the empty cells of the usage matrix with a prediction score. The calculation of similarity between the vectors representing the users is measured by the cosine, or Pearson's coefficient. The latter is the most widely used and the most efficient in terms of predictive accuracy [23].

Items based: The principle of this method consists in predicting the appreciation of the active user for a candidate article, based on the assessments of the active user for articles similar to the candidate article[24]. The determination of similar articles can be calculated by the cosine of the article attribute vectors or the Pearson's coefficient.

2) Model-based methods

These methods are based on learning machine techniques such as probabilistic models (naive Bayesian classifier) [25], clustering, and the most popular, latent factor models [26]. In order to address the shortcomings of memory-based collaborative filtering [27]. A model-based process learns to recognize complex patterns on offline training data so that it can generate predictions on test data or real-world online data.

A probabilistic model consists of calculating the probability $P(a|b,i)$ that the user $a$ assigns the score $b$ to the item $i$ knowing its previous scores. The prediction $\text{pred}(b; i)$ matches, either to the rating with the highest probability, or to the expected rating, as defined by the formula [26].

$$\text{pred}(a,i) = \sum_{b \in B} b \cdot P(b | a,i)$$

$B$ is the set of values that a note can take.

The clustering technique is often used as an intermediate step to bring together researchers sharing the same areas of interest or to group articles addressing the same topics into clusters, which will be exploited for further processing.

C. Hybrid Methods

A hybrid recommendation process combines both content-based and collaborative filtering techniques to increase recommendation performance [27].

II. RELATED WORK

Our work is linked to two main lines of research

- Modelling of researcher profiles based on the LDA scheme.
- The discovery of thematic communities in an academic social network, for the recommendation of scientific articles.

We review the relevant literature dealing with these two areas.

A. Modelling of researcher profiles based on the LDA scheme

The subject modeling is a powerful and practical tool for semantic exploration and subject extraction. One of the methods that has been the subject of many scientific publications in the field of recommendation systems, in particular the recommendation of scientific articles, is the LDA generative probabilistic model. Table .1 (see annex) presents a selection of relevant publications.

B. The discovery of thematic communities in an academic social network, for the recommendation of scientific articles

The detection of communities of interest consists in identifying the best possible graphical partitioning of a network. In the context of our study this translates into the identification of subsets of researchers grouped together on the basis of similarity of thematic interest without any explicit social interaction between them. Recommendation approaches based on community structures can significantly limit the number of users in the process of calculating the prediction and the results are more relevant given that the size of the data is limited to community members only [28]. In
Considering an article \( d \) composed of \( N \) words:

\[ d_i = \{ w_1, w_2, \ldots, w_N \} \]

There is then a probabilistic relationship between the words \( w_i \), the topics noted \( z_k (k \in \mathbb{N}) \) and the article in question \( d \):

- \( P(z_k|d) \): the probability of topic \( z_k \) in the document \( d \)
- \( P(w_i|z_k) \): the probability of word \( w_i \) in topic \( z_k \)

#### C. LDA model for the representation of a researcher’s profile

A researcher’s profile reflects his thematic interests, it is generated from its corpus, by automatic modeling of themes, a widely used technique for semantic exploration and theme extraction in large volumes of textual documents. In this perspective we apply the LDA model, An unsupervised learning technique that treats an article as a vector of words to identify unobservable themes. In the last decade many articles have addressed the use of the LDA model in recommending scientific articles[34][35][36][37].

Considering an article \( d \) composed of \( N \) words:

\[ d_i = \{ w_1, w_2, \ldots, w_N \} \]

The latent variables:

- \( \theta \) document topic distribution.
- \( z \) word topic assignment.
- \( w \) observed word.

The model has two parameters to infer from the observed data, which are the distributions of the latent variables \( \theta \) (document-topic) and \( z \) (subject-word). By determining these two distributions, it is possible to obtain the topics of interest on which researchers write.

### III. PROPOSED APPROACH

In this section, we propose our approach to recommending scientific articles in order to solve the cold start problem that occurs in recommendation systems due to the lack of information on the one hand about a new user who has not yet interacted with the scientific articles, such as publishing an article, downloading, sharing,...and whose corpus is considered to be empty and on the other hand about a new article that has not been the subject of interest from researchers.

Our approach, unlike the approaches described in section 3.1, is based on the modeling of subjects to build very precise profiles taking into account only the thematic area of interest of the researchers without taking into account other types of information that do not bring precision to the profile and may overload it. The problem of sparcity, particularly in the rating matrices, is dealt with by reducing the size of the space (researchers, articles) to a very small space (researchers, topics), and from the latter further reduces the size of the space to communities of thematic interest, which restricts the storage space and the processing time of the recommendation process to the neighbours of the active researcher only.

#### A. Description of the approach

We propose to carry out our approach with the following steps.

- **Step 1**: acquisition of data on the researcher by the web crawler technique. Fig. 2 Indeed a crawling tool extracts all the information relating to the articles in the browsing history, the annotated tags that contain summaries of interesting articles, ... this step produces a schema (researcher, corpus), will follow a pre-processing for the construction of a dataset for the experimentation phase.

- **Step 2**: we apply the LDA model learning on the corpus of each researcher, to extract the different themes that constitute the thematic
areas of interest for each researcher. Thus for each researcher we have his profile in the form of a probability vector on the subjects. We prepare for the next step, a matrix whose lines are the profile vectors of the researchers.

**Step 3:** Researchers who have a high probability for the same themes will form thematic clusters, by applying a flexible clustering algorithm. (fig.6).

**Step 4:** An algorithm is applied to detect thematic communities of interest, starting from the graph of researchers weighted by a similarity between the profiles of the corresponding researchers. Thus, the neighborhood of a target researcher will be identified among similar researchers in his community. (fig.7).

**Step 5:** for a candidate article for recommendation on the basis of a prediction score, a correlation function will be applied between the profile of the target researcher and the vector representing the article in question in the form of a subject probability vector. for a set of candidate articles for recommendation, a ranking of the correlation scores in descending order.

**IV. EXPERIENCE ET ANALYSE**

We have represented our network of researchers by a graph weighted by the cosine similarity between the different entities. (fig.5).

Each node represents a researcher and his identifier.
The communities obtained after applying Blondel's algorithm [40] with the Gephi tool produced the following classes : fig.8.

Results :

- Modularity : 0.366
- Modularity with resolution : 0.366
- Number of communities : 4

The communities discovered :

Community_0 = \{researcher_1, researcher_2, researcher_3, researcher_12\}.

Community_1 = \{researcher_4, researcher_5, researcher_6, researcher_13\}.

Community_2 = \{researcher_9, researcher_10, researcher_11\}.

Community_3 = \{researcher_7, researcher_8\}.

V. CONCLUSION AND FUTURE WORK

In this article we have proposed a scientific article recommendation approach based on the discovery of thematic communities of interest in the context of an academic social network. Taking into account only the subjects to model the profiles of the researchers made it possible to partition the network on the basis of similarity between researchers sharing a high probability for the same subject. Then a community detection algorithm is applied on all the clusters formed during the previous step to identify ultimate communities around thematic interests. The results we have obtained up to step 4 have shown that the inclusion of subjects in the modeling of researcher profiles with abstraction of other social information has provided very promising results so we plan to apply a process of recommending scientific papers for a targeted researcher and for these community neighbours.

REFERENCES

[1] Amami, M., Pasi, G., Stella, F., & Faiz, R. (2016). An LDA-based approach to scientific paper recommendation. Paper presented at the International Conference on Applications of Natural Language to Information Systems.

[2] Sugiyama, K., & Kan, M.-Y. (2015). A comprehensive evaluation of scholarly paper recommendation using potential citation papers. International Journal on Digital Libraries, 16(2), 91-109.

[3] Beel, J., Gipp, B., Langer, S., & Breitinger, C. (2016). Research-paper recommender systems: A literature survey. International Journal on Digital Libraries, 17(4), 305-338.

[4] X. Bai, M. Wang, I. Lee, Z. Yang, X. Kong and F. Xia, “Scientific Paper Recommendation: A Survey,” in IEEE Access, vol. 7, pp. 9324-9339, 2019.

[5] Cohendet, P., Creplet, F., Dupouët, O. 2000.«Communities of practice and epistemic communities: a renewed approach of organizational learning within the firm ». Proceedings of the 5th Workshop on Economies with Heterogeneous Interacting Agents (WEHIA). Marseille. 15-17 June.

[6] Luc Brunet et André Savoie « la face cachee de l’organisation » edition

[7] Presses de l’Université de Montréal 2003.

[8] 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.

[9] Burke, R. (2000). Knowledge-Based Recommender Systems. Dans Kent, A. (Éd.), Encyclopedia of Library and Information Systems (Vol. 69): Marcel Dekker.

[10] Resnick, P., & Varian, H. R. (1997). Recommender systems. Commun. ACM, 40(3), 56-58.ReutersThomson.

[11] Schafer, J. B., Konstan, J., & Riedl, J. (1999). Recommender systems
TABLE I: SELECTION OF RELEVANT ARTICLES ON LDA AND RECOMMENDATION OF ARTICLES

| Authors | Year | Articles |
|---------|------|----------|
| Jeledar et al | 2019 | Natural Language Processing via LDA Topic Model in Recommendation Systems. |
| Kim, et al | 2019 | Research paper classification systems based on TF-IDF and LDA schemes |
| Amami et al | 2017 | A graph based approach to scientific paper recommendation. |
| Sahijwani et al | 2017 | User Profile Based Research Paper Recommendation. |
| Dai et al | 2017 | Explore semantic topics and author communities for citation recommendation in bipartite bibliographic network. |
| Younus et al | 2014 | Utilizing Microblog Data in a Topic Modelling Framework for Scientific Articles’ Recommendation. |
| Kim et al | 2013 | TWITOB: A Recommendation System for Twitter Using Probabilistic Modeling |
| Li et al | 2013 | Scientific articles recommendation. |
| Wang et al | 2011 | Collaborative Topic Modeling for Recommending Scientific Articles |
| Chenguang et al | 2010 | Research Paper Recommendation with Topic Analysis |
| Author           | year | Articles                                                                 |
|------------------|------|--------------------------------------------------------------------------|
| Alfarhood et al  | 2019 | Collaborative Attentive Autoencoder for Scientific Article Recommendation |
| Xia et al        | 2019 | Scientific Article Recommendation: Exploiting Common Author Relations and Historical Preferences |
| Sun et al        | 2018 | A hybrid approach for article recommendation in research social networks |
| Wan et al        | 2018 | HAR-SI: A novel hybrid article recommendation approach integrating with social information in scientific social network |
| Mao et al        | 2018 | Academic Social Network Scholars Recommendation Model Based on Community Division |
| Wang             | 2016 | Academic Paper Recommendation Based on Community Detection in Citation-Collaboration Networks |
| Reihanian et al  | 2016 | Topic-oriented Community Detection of Rating-based Social Networks |
| Sobharani et al  | 2013 | Application of clustering to analyze academic social networks |
| Zhao et al       | 2012 | Topic oriented community detection through social objects and link analysis in social networks |
| Ding et al       | 2011 | Community detection: topological vs. topical |
