Journal Recommender: Recommendation Based on the Extension of BrCris' VIVO Ontology and OpenAlex

Ingrid Q. Pacheco[1], Giseli Rabello Lopes[1] and João Luiz Rebelo Moreira[2]

[1]*Graduate Program in Informatics – Computing Institute, Federal University of Rio de Janeiro (UFRJ), Rio de Janeiro, RJ, Brazil*

[2]*Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, PO Box 217, Enschede 7500, AE, The Netherlands*

Abstract

As a daily task, plenty of researchers submit articles to events and magazines hoping that their studies will reach more people. However, as the quantity of venues grows each day, it may be hard to know which are the best options. In the present work, an ontology-based recommender system is proposed using VIVO ontology and OpenAlex data. It discusses possible methods and validation procedures to ensure great quantitative measurements and to provide the best suggestions according to the available data.

Keywords

Recommender System, Published Articles, OpenAlex, VIVO Ontology, BrCris, Clustering, SBERT

1. Introduction

One of the most important steps in every scientist, academic, student, or researcher's academic life is publishing a scientific article [1]. The article gives more importance and meaning to the projects created considering it can benefit other people by aggregating more knowledge on their research.

By understanding that their work already had enough arguments to answer their questions and the problems identified have enough basement to be solved, the researcher detects the perfect moment to publish their article. However, considering the impact they want to have, there is a massive concern regarding which would be the ideal event to submit their article to, trying to minimize their chance of denial.

Regardless of the vehicle, the publication is part of the research. After all, it does not matter how relevant the results of the research are, if no one knows about them, they will not be impactful [2]. As for where to publish, the greater the recognition of the platform, the bigger the chance of people finding it and reading it, increasing its impact.

Proceedings of the XVI Seminar on Ontology Research in Brazil (ONTOBRAS 2023) and VII Doctoral and Masters Consortium on Ontologies (WTDO 2023), Brasilia, Brazil, August 28 - September 01, 2023.

ingridpacheco@dcc.ufrj.br (I. Q. Pacheco); giseli@ic.ufrj.br (G. R. Lopes); j.luizrebelomoreira@utwente.nl (J. L. R. Moreira)

CEUR Workshop Proceedings (CEUR-WS.org)

1.1. Problem Definition

Considering the existence of more than 9585 conferences and 4152 journals of Computer Science, it is hard for the authors to know which should receive submissions, even more considering that submitting an article for the wrong conference usually can cause rejection, delay, or less reading work [3]. In fact, this is such a relevant point that several publishers currently have their recommendation system tools, such as *IEEE*[1], *Springer*[2] and *Elsevier*[3].

Even though there are plenty of options, each has its limitations. All the mentioned above are limited by publication vehicles, that only bring conferences that were published by them, excluding diversity. In addition, there are systems that limit by research area, such as *Content-based Journals & Conferences Recommender System for CCF*[4], that focuses on one specific area. Moreover, limitations by location do not consider the reality the researcher is inserted in and retrieves results considering the global scope, with only best-graded conferences or the ones chosen by some organizations, usually all being from the same country. Finally, most of the research works cover only partial cold start conditions on new users, while an ontology-based approach solves it by providing initial knowledge [4].

Therefore, creating a recommender system that surpasses the mentioned limitations is essential to provide a more targeted way for users to find the appropriate conferences and journals.

1.2. Research Proposal

The present work aims to study and develop the most suitable ontology-based recommender system considering VIVO ontology and OpenAlex data. One of the main benefits of this approach is that the semantics of the ontology may help with the interconnections of works and authors and the evaluation of the research, and also solve the cold start problem, in which a recommendation has to be made for a new user or item [5].

To recommend something, data is necessary to know what the user usually likes and find similar products. However, sometimes, there is no previous data available and that is when the cold start problem happens. Some systems ask the user to rate a list of items or to list preferences, but given an ontology, previous data is available, and it is possible to be guided through the ontology to find the most accurate suggestions.

Hence, the proposal tends to not only surpass the cold start barrier and previously mentioned limitations but also outperform similar systems by using ontology to find similar articles to user input and their correlations to authors and co-authors, expanding the possibilities of the research. Moreover, content-based filtering will provide the top recommendations for the user.

To sum up, the objectives of the research proposal are:

1. To improve the VIVO ontology provided by BrCris by extending it using other ontologies to satisfy OpenAlex's needs.
2. To develop a recommender system for conferences and journals based on the VIVO ontology, considering existing data and user input.

[1] https://publication-recommender.ieee.org/home

[2] https://journalsuggester.springer.com/

[3] https://journalfinder.elsevier.com/

[4] http://www.keaml.cn:5500/prs

2. OpenAlex data

One of the most important steps when building a recommender system is to know which data will be used and what they have available. Considering the context of recommending a list of conferences based on an input of a possible article, a database containing published articles is necessary, to facilitate comparison.

OpenAlex is an open catalog of the global research system. They automatically index journals and articles from Crossref and other sources as MAG, ORCID, ROR, DOAJ and others [6]. It was chosen as the main data source due to the possibilities the API provides, such as how data is structured, their correlations, the capabilities of querying them, and how its data is also provided as a snapshot.

The API entities include works (published works), authors (people who write the works), sources (where works are hosted), institutions (institutions to which authors claim affiliations), publishers (companies and organizations that distribute works), funders (organizations that fund research) and geo (locations from where entities come from).

Although the API is compelling, it has some limitations that let the need of hosting their snapshot data on AWS, aiming to make it easier and faster to fetch and process their data.

3. Methods

Considering the chosen data source, what and how it was going to be used was the natural following procedure. Although OpenAlex's structure is very diverse, there was no need for all data to be used. The first step was identifying which part of it was necessary, reducing it to the properties *ID* (the ID of the work), *doi* (the DOI of the work), *title* (the title of the work), *primary_location.source* (the source it was published on), *concepts* (the concepts extracted for the work) and *abtract_inverted_index* (the abstract of the work, as an inverted index).

As the work data has, alone, more than 1 Terabyte, they were added to a Virtual Machine where their properties could be searched and later, processed. The virtual machine is hosted on AWS by Laguna, and it also includes other structures contained on OpenAlex. However, to provide more meaning and be able to explore semantic interconnections, it had to be suited to the VIVO ontology, used by BRCris to provide tools to the Brazilian academic community.

3.1. Improving VIVO ontology

The VIVO ontology aims to provide an easier way to search and browse data about researchers and their works, allowing applications to emerge faster with all gathered data. Nevertheless, its core doesn't satisfy all the needs BrCris has to represent the Brazilian academic research, and consequently, they improved the main ontology by adding *Graduate Program*, *Referee Role*, and *Community* as extensions for the Brazilian context.

Thence, OpenAlex came forth as another data source to make BrCris more complete and powerful, as it has data that is not included in their database yet. Even so, some improvements to the ontology must be done.

As mentioned before, the main properties used from OpenAlex's works are *doi*, *title*, *primary_location.source*, *concepts* and *abtract_inverted_index*. Some of them have a direct correlation to properties mapped on the ontology, such as dc:title for title, skos:Concept for concepts, bibo:doi for doi and bibo:Journal for primary_location.source. When it comes to the abstract, each term had to be joined in a single string before mapping it to vivo:Abstract.

After the mapping was structured, a script in Python was created to pass through every work on OpenAlex and add their mappings to BrCris' solution. When this process was over, it was time to receive the complementary data provided by the user.

3.2. Receiving user input

To match the existing work data and receive recommendations of conferences, some data must be inputted to make the comparisons possible. Among plenty of possible inputs, two were considered relevant to the system, title, and abstract.

To be able to receive some input from the user, a web application was created, with Javascript (with React library) for the frontend and Python (with FastAPI) for the backend. The project is public and open on Github[5].

The Search page takes the user where they input their work data and receive the recommendation. Furthermore, it has some filtering that is completely optional for the user. They can choose which kind of recommendation they want to receive (only conferences, only journals, or both of them) and the countries the options come from (in case they want to find conferences from a specific location). In the future, the filtering options are intended to grow and also contemplate other particularities.

With all data inputted by the user, some processing must be done, and for this, the project OpenAlex Concept Tagging will be highlighted.

3.3. OpenAlex Concept Tagging

OpenAlex has a public repository with the code used to extract concepts for each work in the catalog, the OpenAlex Concept Tagging[6]. Besides the open code, it also has documentation explaining the methodology, data used, and how it was conceptualized.

The document explains the process of eliminating properties and knowing which could be used to extract concepts. At the beginning, they had the *Paper title*, *Document type*, *Journal Title*, *Author*, *Affiliation*, *Publication date* and *Abstract*.

However, not all data had to be used. Publication date, authors, and affiliations were discarded due to not having anything to do with the tagging, large number of appearances and only being on 25% of works, respectively. Journal title only remained because the quantity of document types that are journals is more than 50%, showing the importance of the feature on the model.

It is also important to highlight that the model is trying to replicate the MAG model, which was discontinued by Microsoft at the end of 2021. This means that the model will not assign new concepts, as it was trained using MAG historical tagging data.

The first step of the recommender system is to receive the data inputted by the user and run the model on them, extracting their concepts. Even though the project is open on Github, their

[5] https://github.com/projetos-codic-ibict/JournalRecommendation

[6] https://github.com/ourresearch/openalex-concept-tagging/tree/main

data on s3 is forbidden, which led to the creation of a project on BrCris' AWS with a copy of their code, to make this step available for the system.

The choice of using the exact model used by OpenAlex was to not have the possibility of the found concept being outside the scope of OpenAlex, therefore, increasing the chance of finding similar articles using the concepts. Following this, the process of matching them was possible.

3.4. Initial method experimentation

The initial method of experimentation was simpler and aimed to understand how data could be used in order to provide recommendations for conferences and journals.

How is it possible to find the best journals when only having minimal data about a work? This was the key question to find the starting point of the system. It's not necessary to only think about the user's input, but also similar works to them and where they were published on.

Even so, it is definitely a simpler way to solve the problem, as a lot more features could be taken into consideration to provide a better outcome, such as the impact scores of the vehicles, the country the researcher lives in, where they submitted before as well as their co-authors (if available). All these possibilities were not discarded, but as the starting point, finding similarity between existing works and user input was the best choice.

First, the title and abstract inputted were used by the OpenAlex Concept Tagging model to extract concepts. With them two paths were followed, clustering and Sentence-BERT (SentenceTransformers) with cosine similarity[7]. For each path two features were used each time, Concepts and Abstract_Inverted_Index, resulting in four possible results.

Sentence-BERT was chosen due to its considerable capability of finding semantic textual similarity and semantic search. It can be used to compute sentence / text embeddings for more than 100 languages that can be compared using cosine-similarity to find sentences with similar meanings [7]. Its biggest limitation is the sentence length, whereas the provided methods use a limit of 128 word pieces, with longer inputs being truncated [8]. As the concepts are smaller than the limit, it is not a problem.

On the other hand, K-Means is a classic clustering method based on analysis and comparisons among numerical values to classify the entry. The unsupervised choice was due to not having a specific classification of the works considering the chosen properties, letting the process become automatic. Its limitations are requiring to specify the number of clusters, being sensitive towards outliers, and that it forms spherical clusters only. Other methods were not discarded, but these were the most interesting to begin with.

In the clustering path, the library Sklearn[8] was used along with its *K-Means* method. The entries were either the concepts or abstracts of existing works and user input. For the abstracts, the terms were joined in a single string.

The entries were pre-precessed, removing stop words, tokenizing, and so on. The functions *TfidfVectorizer* and *fit_transform* were called, which converts a collection of raw documents to a matrix of document-term. The clusters were later created using *K-Means* and *fit*. The journals that published the works contained in each cluster were saved in a dictionary, and the cluster

[7] https://www.sbert.net/

[8] https://scikit-learn.org/stable/index.html

the user entry belongs to was predicted using the function *predict*. Finally, for the returned cluster, the journals were ranked due to the number of times they appeared.

On the other hand, there's the SBERT path. The system embeds both entries from existing works and from the user using the function *encode* from the Model *SentenceTransformer*. Latter, it compares both of them using cosine similarity and returns the value for each pair (one entry from existing work and one from the user). As the user entry is unique, it brings its similarity to each work. With the numbers, it is possible to find the works most similar to what the user is writing and, therefore, get their journals to create a ranking.

3.5. Methodology Improvements

All the mentioned methods are suitable alternatives to be tested, but in order to provide more meaningful data, the use of VIVO ontology is desired. As discussed on 3.1, OpenAlex's data will be mapped into the VIVO ontology and added to BRCris' solution.

As an ontology-based approach, it will solve the cold-start problem by providing initial knowledge about articles, their authors, and the relationship between them. Moreover, it will also be possible to find information about their preferences and co-authors, which allows further use of collaborative filtering.

The improvements in the methodology refer to the use of the ontology. The primary steps are the same, as the user's inputs runs the OpenAlex Concept Tagging model, extracting concepts. Then, they are queried on the ontology to find works that have them (are similar). Subsequently, it runs similarity techniques on these articles and ranks their journals to return the final listing. A diagram of the mentioned system is illustrated on figure 1.

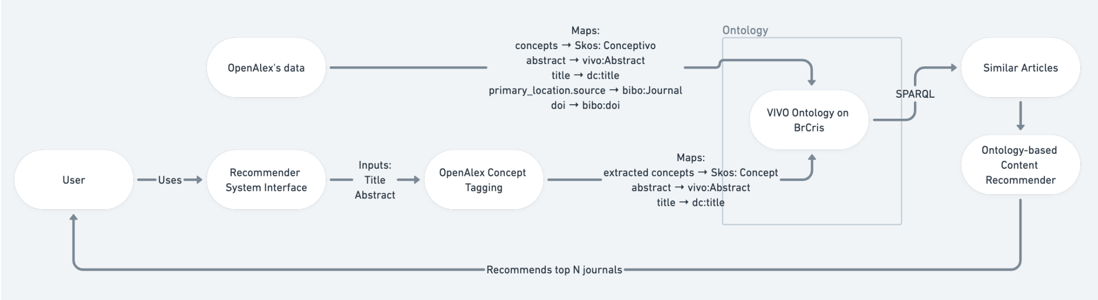

Figure 1: System Architecture

The usage of ontology not only improves the performance of the system by permitting to find similar concepts but also allows future improvements as it has personal data from researchers and their co-authors, to make it more personalized.

3.6. Validation

One of the most important parts of a recommender system is to understand if it works and the precision and accuracy of its results. Therefore, a method of validating its outcome is necessary. For validating the system, the process will be divided into two parts, an offline and an online validation.

For the offline validation, a part of the database will be used to assess the accuracy of the system. Their abstracts will be used as input along with the titles. Therefore, the outcome should be the journal or conference they were published in or some similar to it. As the accuracy might not be 100%, it is also important to consider the possibility of recommending a similar option, and the result will be assessed due to the key concepts related to it. Considering the use of an ontology, it is easier to know if a journal is similar/correlated to another, as well as concepts.

For the online validation, a set of works will be used in the assessment. In this case, a web application will be created showcasing the abstract and title of a work and 5 options of journals/conferences, being 3 of them options the model considered as a recommendation and 2 it didn't. The user must select the options they would submit the work to. The process will be repeated for a couple of works to provide a better outcome.

The number of works to be assessed will be defined by how many people will participate in the assessment, considering it will most likely involve researchers, students, and professors from a wide range of areas. In addition, due to the need to involve third parties, it will only happen when the first validation is over and the accuracy is how it was expected.

From the mentioned approach, it will be possible to calculate plenty of quantitative measurements, such as precision and accuracy of the results. For example, if there are 5 options, the system chose 1, 3 and 5 as the suggestion and the final user chose 2, 3 and 5, the precision would be 100% as the quantity is 3 for both of them, but the accuracy would be 60% as it mistakenly labeled 1 and 3 as the opposite of what the user did.

Another important validation is between a system that uses ontology and one that doesn't. To prove the better performance of the first, the offline validation will run for both (the initial methods and improved methodology) and the online validation will provide some options from the first and some from the second, in order to understand which the user prefers.

4. Conclusion

In order to optimize time and effort, a recommender system for conferences and journals based on the article the user is writing is demanded. However, to avoid the cold start problem and consider future user data, if available, an ontological approach is beneficial, improving the recommendation capability. Therefore, this project aimed to explore how VIVO ontology and OpenAlex data could be used to ensure the best accuracy and precision the model could have.

As future work, the recommender system must be extended to consider impact scores from vehicles, available user data on Lattes, and conferences co-authors have presented at, evolving it to a hybrid methodology with both collaborative filtering and content-based approaches. All this data is available on VIVO ontology, but it will require some implementation both on the backend to suit the model to consider user data, and also on the frontend with a login, to know which researcher is using the system. Furthermore, it would be great to provide those who don't have Lattes, to add their personal information, in order to make the recommendation more suitable to their preferences.

References

[1] D. V. Cuong, D. H. Nguyen, S. Huynh, P. Huynh, C. Gurrin, M.-S. Dao, D.-T. Dang-Nguyen, B. T. Nguyen, A framework for paper submission recommendation system, in: Proceedings of the 2020 International Conference on Multimedia Retrieval, ICMR '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 393–396. URL: https://doi.org/10.1145/3372278.3391929. doi:10.1145/3372278.3391929.

[2] J. Robens, The importance of academic publishing and the open access evolution, 2018. URL: https://www.aje.com/arc/the-importance-of-academic-publishing-and-the-open-access-evolution/.

[3] D. Wang, Y. Liang, D. Xu, X. Feng, R. Guan, A content-based recommender system for computer science publications, Knowledge-Based Systems 157 (2018) 1–9. URL: https://www.sciencedirect.com/science/article/pii/S0950705118302107. doi:https://doi.org/10.1016/j.knosys.2018.05.001.

[4] J. Joy, N. S. Raj, R. V. G., Ontology-based e-learning content recommender system for addressing the pure cold-start problem, J. Data and Information Quality 13 (2021). URL: https://doi.org/10.1145/3429251. doi:10.1145/3429251.

[5] M. Jacobs, Decision support system afor programming language selection: a literature review (2021).

[6] J. Priem, H. Piwowar, R. Orr, Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts, 2022. arXiv:2205.01833.

[7] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019. URL: https://arxiv.org/abs/1908.10084.

[8] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, CoRR abs/1908.10084 (2019). URL: http://arxiv.org/abs/1908.10084. arXiv:1908.10084.