

Using an Ensemble of Features for Personalized Recommendations of Scientific Publications

Discussion Paper

Paolo Tenti¹, James Thomas², Rafael Peñaloza¹ and Gabriella Pasi¹

¹*IKR3 Lab, University of Milano-Bicocca, Milan, Italy*

²*EPPI Centre, UCL Social Research Institute, University College London*

Abstract

Maintaining reviews of scientific publications as soon as new relevant publications are available is a typical challenge to many research communities. We address this challenge as a content-based recommendation problem, where the publications already selected for a review drive the recommendation of the new publications. In addition, resources such as domain databases, ontologies and academic graphs provide structured information about publications (e.g., authors, journals, conferences). Our experiments show that a simple model based on that structured information to represent publications achieve high precision and recall, and outperform models that use more sophisticated representations based on embeddings.

Keywords

Content-based recommendation, Scientific papers recommendations, Text classification

Introduction

A common issue to many research communities is building and maintaining meaningful collections of references to scientific publications related to a specific research topic; to this aim reviews are compiled, which report such references in an organized way. In this context, a first challenge is discovering the initial collection of publications to be considered for compiling a review; this can be done by searching across several scientific databases and journals and going through several passages of filtering and refinement. A second challenge is maintaining reviews; that is, capturing new relevant publications as soon as they are available after a review has been constructed, to the aim of keeping the review updated. Both processes are partially manual, long-running, and error-prone. We argue that the problem of maintaining existing reviews can entrain a content-based recommendation problem: publications in existing reviews can be used to automatically find and recommend new publications to the reviews' owners.

IIR 2021 – 11th Italian Information Retrieval Workshop, September 13–15, 2021, Bari, Italy

✉ p.tenti1@campus.unimib.it (P. Tenti); james.thomas@ucl.ac.uk (J. Thomas); rafael.penalozan@unimib.it (R. Peñaloza); gabriella.pasi@unimib.it (G. Pasi)

🌐 <https://ikr3.disco.unimib.it/people/paolo-tenti/> (P. Tenti);

<https://iris.ucl.ac.uk/iris/browse/profile?upi=JTHOA32> (J. Thomas); <https://rpenalozan.github.io/> (R. Peñaloza);

<https://ikr3.disco.unimib.it/people/gabriella-pasi/> (G. Pasi)

🆔 0000-0003-2432-3018 (P. Tenti); 0000-0003-4805-4190 (J. Thomas); 0000-0002-2693-5790 (R. Peñaloza);

0000-0002-6080-8170 (G. Pasi)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

Problem statement

Let \wp be the domain of publications, and \mathfrak{R} the domain of reviews, such that a review $R \in \mathfrak{R}$ is a finite set of scientific publications, i.e., $R = \{p_1, \dots, p_k \mid p_i \in \wp\}$.

A set $\overline{\mathfrak{R}} \subseteq \mathfrak{R}$ of reviews and a set of brand-new publications $\overline{\wp} \subseteq \wp$ are given. For any review $R \in \overline{\mathfrak{R}}$, the problem is to retrieve a set $\overline{\wp}_R \subseteq \overline{\wp}$ of publications that are relevant to R .

To define the notion of relevance, we first define a similarity function $\Phi : \mathfrak{R} \times \wp \rightarrow [0, 1]$ that given a review $R \in \mathfrak{R}$ and a publication $p \in \wp$, returns their similarity score. Relevance can be modeled in this context as a binary function $\Xi_\lambda : \mathfrak{R} \times \wp \rightarrow \{True, False\}$ that given a threshold λ returns true if and only if $\Phi(p, R) > \lambda$. For any review $R \in \overline{\mathfrak{R}}$, the set of relevant publications $\overline{\wp}_R = \{p \mid p \in \overline{\wp} \wedge \Xi_\lambda(p, R)\}$ will be recommended.

Methodological framework

Resources such as scientific databases (e.g., PubMed ¹), ontologies (e.g., Unified Medical Language System ², PICO ³) and academic graphs (e.g., Microsoft Academic Graph [1]), provide relational information about publications, such as title, abstract, authors, citations, references, journals, and conferences. That information can be used to enrich publications with descriptive features. Hence, we study the opportunity of using these features to construct multiple vector representations of publications, to address the problem of recommending scientific publications as described above.

Given a set of publications features $\mathcal{F} = \{\pi_1, \dots, \pi_n\}$, for a publication $p \in \wp$ we define $v(p) = \{v^{\pi_1}(p), \dots, v^{\pi_n}(p) \mid \pi_n \in \mathcal{F}\}$ as the set of vectors that represent p , where $v^\pi(p)$ stands for the vector representation of p with respect to the feature π .

We define a methodological framework for using the available features to compute a compound similarity function between a publication and a review. This framework is simple yet extensible.

First, we define a method to construct the vectors $v^\pi(p)$ for a given publication p and feature π . In addition, we define the representation of a review $R \in \mathfrak{R}$ with respect to a feature π as $v^\pi(R) = v^*(\{v^\pi(p) \mid p \in R\})$, where v^* is an aggregation function over the set of publications representations with respect to the same feature π .

We further define $\Phi^\pi(p, R)$ as the function to calculate the similarity between a publication p and a review R , with respect to a feature π . Finally, we define the function to calculate the similarity between p and R as $\Phi(p, R) = \Phi^*(\{\Phi^\pi(p, R) \mid \pi \in \mathcal{F}\})$ where Φ^* is an aggregation function over feature-specific similarity scores.

Evaluation

We conducted a series of experiments on a manually labeled dataset of domain-specific reviews and a few tens of thousands of publications. Note that the research domain is homogeneous,

¹<https://pubmed.ncbi.nlm.nih.gov>

²<https://www.nlm.nih.gov/research/umls/index.html>

³<https://linkedata.cochrane.org/pico-ontology>

and thus the reviews are quite similar to each other.

We evaluated our method as a multi-class, multi-label classification problem. Specifically, classes are the reviews, and each publication might be relevant to multiple reviews. Thus, the problem is to predict all relevant labels (i.e., reviews) for a certain, previously unseen, publication.

We considered title, abstract, citation network, authors, journals, conferences, topics and ontological categories as features, and their representations as either Tf-Idf vectors or binary vectors. For topics we considered the Field of Studies extracted from the Microsoft Academic Graph. We extracted ontological categories from PiCO, a domain specific ontology.

The best performing model showed precision of 97.7% with recall of 99.2%. Similar results were confirmed by experiments on a different dataset. We observed the following:

- Ensembles of features over-performed textual features alone, either with Tf-Idf based representations or with embeddings.
- For text-based models, Tf-Idf based representations considerably beat embeddings on precision. Our interpretation was that, in a context where reviews come from the same domain, key phrases are better suited to capture the differences between them.
- Titles generally performed better than abstracts when using ensembles of features and are computationally more efficient.
- Representing publications by means of simple binary or Tf-Idf based vectors suffices to achieve good performance, in contrast to more sophisticated solutions, such as representations based on embeddings.
- To implement v^* many approaches are possible. Our experiments suggest that the best performing ones capture the most representative properties of reviews, rather than any single property regardless of their importance.
- To implement Φ^* , our experiments show that simple mathematical operators suffice to achieve good results and keep the model computationally efficient and highly explainable. Among the benefits of explainability, note that for any given review it is easy to capture the relevant features to achieve good recommendation performance.

Contributions and open challenges

Several Web resources (i.e., academic graphs, domain specific scientific databases) provide structured information about scientific publications, such as title, abstract, authors, citations and potentially ontological categories. Our work shows a methodological framework that can make use of such structured information to achieve high-precision and high-recall in the down-stream task of domain-specific, personalized recommendation of scientific publications.

In addition, our work shows that representing publications with ensembles of features outperforms representations based on embedding vectors [2, 3, 4]. Our interpretation is that to discriminate publications' membership to reviews that belong to the same domain, the signals coming from key phrases and ontological categories are more relevant than those from more general semantic representations, like the ones obtained with text embeddings.

Constructing more sophisticated embeddings to represent publications that capture both their content and relational properties [5] might achieve comparable or better performance in

a more domain independent and task agnostic way. However, we argue that using a simple similarity mathematical model based on easy to capture features is computationally inexpensive, easier to train, more interpretable and still highly generalizable.

Finally, we believe that there are still open challenges to address. Our experiments show the importance of some features like topical and ontological categories. However, the availability of such features might be domain-dependent, or hard to extract. On the one hand it would be worth studying the impact of using text embeddings over titles and abstracts in synergy with standard features (i.e., n-grams over titles and abstracts, citation network and co-authoring), to see if they could compensate for more sophisticated and domain-dependent features (i.e., ontological categories, fields of study). On the other hand, it would be worth studying generalizable methods for extracting ontological features from publications [6].

References

- [1] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-J. P. Hsu, K. Wang, An overview of microsoft academic service (mas) and applications, in: Proceedings of the 24th International Conference on World Wide Web, 2015, pp. 243–246.
- [2] D. Cer, Y. Yang, S. yi Kong, N. Hua, N. L. U. Limtiaco, R. S. John, N. Constant, M. Guajardo-Céspedes, S. Yuan, C. Tar, Y. hsuan Sung, B. Strope, R. Kurzweil, Universal sentence encoder, arXiv preprint arXiv:1803.11175 (2018).
- [3] C. W. Schmidt, Improving a tf-idf weighted document vector embedding., arXiv preprint arXiv:1902.09875 (2019).
- [4] S. Arora, Y. Liang, T. Ma, A simple but tough-to-beat baseline for sentence embeddings, in: ICLR 2017 : International Conference on Learning Representations 2017, 2017.
- [5] D. Nozza, E. Fersini, E. Messina, Cage: Constrained deep attributed graph embedding, Information Sciences 518 (2020) 56–70.
- [6] V. Gutiérrez-Basulto, S. Schockaert, From knowledge graph embedding to ontology embedding? an analysis of the compatibility between vector space representations and rules, in: KR, 2018, pp. 379–388.