Using Word Embeddings for Recommending Datasets based on Scientific Publications

Narges Tavakolpoursaleh[1], Johann Schaible[1], and Stefan Dietze[1,2]

[1] GESIS – Leibniz Institute for the Social Sciences, Cologne, Germany
[2] Institute for Computer Science, Heinrich-Heine University Duesseldorf, Germany
firstname.lastname@gesis.org

Abstract. In scholarly search systems, computing recommendations of the same type, for example, additional publications when reading a particular publication, is a well-approached problem. However, suggesting items from another type, e.g., research data when reading a publication, is rarely covered in scholarly recommendations. In this position paper, we employ word embeddings to approach the problem of such cross-domain recommendations in scientific search systems, more specifically, recommending research data based on publications. Besides various metadata, publication and research dataset entries comprise textual metadata (e.g. title, abstract), which allows to detect similar entries using word embeddings. We illustrate first results, major problems and possible solutions when using word embeddings for recommending datasets based on publications.

Keywords: Dataset Retrieval and Recommendations · Cross-Domain Recommendations · Word Embeddings.

CEUR-WS.org/Vol-2454/paper_59.pdf

1 Introduction

In digital libraries, such as arXiv[3], typically, a scientific search system aids users in finding literature covering a topic of interest [17]. To additionally alleviate the users' situation in finding appropriate literature, scientific search systems may also comprise recommender systems, which provide suggestions for items – sometimes previously unknown items – that are most likely of interest to a user [15]. One prominent use case in scholarly recommendations is suggesting literature which is similar to a publication the user is currently viewing. This resembles recommending items of the same type (i.e. the *domain* of the item) and is a well-approached problem in scientific search systems. However, there is another important use case that exploits recommendations from different types, i.e., *cross-domain recommendations*. We focus on the following prominent and more and more emerging example, which is rarely covered in scientific search systems: recommending research data when viewing a scientific publication.

Some information systems provide a search over various types of information in a given field of interest. For example, besides publications, the *GESIS-wide Search*[4]

Copyright ©2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

[3] https://arxiv.org, Accessed June 2019
[4] https://www.gesis.org/en/home/, Accessed June 2019

(GWS), comprises research datasets, questions as well as variables, and further information in the field of Social Sciences. Such entries of different types enable scholarly recommendation systems to provide the desired cross-domain suggestions.

Why is retrieving research data important? Research data is an important facilitator of scientific progress. Making it publicly available is crucial towards enabling open science, i.e., towards replicating and/or reproducing research outcomes as well as validating newly developed methods and insights [3]. When research data is archived in a digital form, the problem how to retrieve it, is mainly covered by *dataset search and retrieval*. Typically, retrieval systems return relevant datasets for explicitly formulated user queries [9]. Recommendations can further alleviate finding suitable research data. However, most recommendation approaches target rather dataset interlinking by using semantic technologies to match datasets with other datasets that overlap in their content. Recommending datasets based on publications can pose problems using this approach, as publications might use general datasets, e.g., statistics on a country's demographics, for rather specific topics, e.g., mobility of youth towards large cities. Utilizing the content description, such as the abstract of publications and datasets, is likely to be more promising, as both might contain needed information to detect similarities.

In this paper, we present our on-going work on using word embeddings for research dataset recommendations in the GESIS-wide Search based on scientific publication that a user is currently viewing. Word embeddings seem promising in detecting appropriate recommendations based on the textual metadata of both a publication and a dataset. We focus on the specific use case in which we define a recommended dataset as relevant, if that dataset has been subject to the publication, i.e., the publication cites that dataset. The main task of the recommender is thereby defined as: the recommended dataset should/could be used and/or cited if the user intends to build her research upon the currently viewed publication. We illustrate that our word embedding model, unfortunately, does not achieve promising results, provide possible reasons, as well as give a first outlook on possible solutions how to improve the recommendations.

2 Related Work

Whereas finding information in scientific search systems that satisfies a user's information need is a well-elaborated topic in classical information retrieval [17,16], specifically targeting the goal to retrieve research data is still a growing field [6]. To this day, still most research data repositories use the same approaches to retrieve research data as for publications, since there are only a few studies (including user behavior studies) which seem to be more promising than the established document retrieval methods [6]. For recommender systems in scientific search systems, according to [1], content-based filtering is the most common recommendation approach (55%), followed by collaborative filtering (18%) and graph-based recommendation approaches (16%), while the remaining recommender systems use rather hybrid approaches. A major reason for this, is that collaborative filtering requires a large collection and investigation of user profiles and graph-based approaches require a well-designed knowledge graph describing and linking the data in a repository [8]. Content-based approaches merely use the entries' metadata, which especially in digital libraries is rather rich.

Prior works on the general problem of dataset recommendation focus on particular scenarios, for instance, recommendation of datasets for interlinking (dataset-dataset-recommendation). Ellefi et al. [5] use clustering and established schema-matching metrics to recommend datasets with overlapping schemata, i.e., overlapping content. Lopes et al. [12] considers the link graph among datasets to recommend datasets which link to the same or similar resources. Given the lack of reliable and exhaustive metadata for research datasets, prior work in the field of dataset retrieval and dataset recommendation relies on techniques for dataset profiling [2], for instance, in order extract and represent dataset metadata capturing various dimensions of relevance. Thus, we restrict ourselves to first utilize only the textual metadata of publications and research datasets.

Word embedding techniques like Latent Semantic Indexing or word2vec can be utilized to capture the contents' metadata and provide semantics to the content [13]. Recent works on unsupervised representation learning have the intent to embed context to predict the words in a sentence [10] or the nodes in a graph [14]. Learning the vector space representations of words have facilitated obtaining distributional semantics of words [10] and have been shown to perform well in many natural language processing tasks of understanding the word-context [11]. Determining the semantic similarity between items is also a related problem in the application of recommending datasets based on publications. Therefore as the first experiment, we applied Mikolov's Doc2Vec [10] which is as an extension to Word2Vec for learning document-level embeddings.

3 Data and Approach

GESIS-wide Search: In this paper, we exploit the contents of the integrated search system GESIS-wide Search [7] for recommending datasets based on publications that a user is currently viewing. The GESIS-wide search comprises publications (ca. 95k), research data (ca. 84k), questions and variables (ca. 12.7k), as well as instruments and tools (370) in the field of Social Sciences, and thus allows for such cross-domain recommendations between these four types of data. The publications are mostly in English and German language and are annotated with further textual metadata like title, abstract, topic, persons, and other. Metadata on research data comprises (among others) a title, topics, datatype, abstract, collection method, universe, primary investigator, as well as contributor in English and/or German.

Recommendation Task: When recommending items to a user, the following question arises: what is the general task of the recommendation? This means, is the recommended item supposed to complement, be as similar as possible to, or even contradict the item viewed or downloaded by the user? Additionally, other parameters of a recommendation, such as novelty and the impact on the domain, can be quite important to satisfy the users' information needs. In scientific search systems, all these dimensions might play a role when defining a *relevant* recommendation. However, this also makes it quite difficult to design and evaluate (cross-domain) recommendations, as with all these parameters, there are different definitions of the relevance and/or the usefulness of a recommended item. In the task of recommending datasets based on a publication, it might be desirable to recommend datasets which support the publication, complement the publication's findings, are cited in the paper, or are related in some other way,

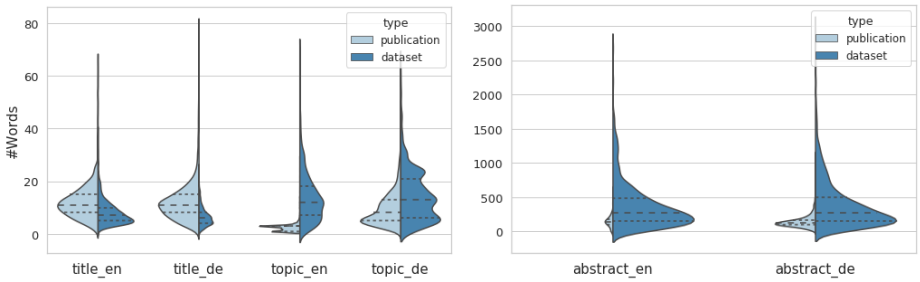

Fig. 1: Number of English (en) and German (de) words in title, topic and abstract in metadata of publications and datasets

e.g., topic, domain, temporal and geographical coverage. For our prototype, we focus on a first simple use case. Datasets that are cited by a publication the user is currently viewing are considered relevant, i.e., the ground truth. This resembles the following scenario: which datasets should/could be used and/or cited if the user intends to build her research upon the currently viewed publication.

Word Embeddings: Le et al. [10] introduced an unsupervised algorithm that learns the vector representations from texts of different lengths to predict the surrounding words in a sample of a paragraph. This Paragraph Vector framework (Doc2Vec) maps every paragraph to a unique vector and concatenates it together with vectors of words, in order to predict the next word in a context. We used this framework for representing the context of research datasets and publications in a vector space. Subsequently, we computed the distances between the dataset representations and the representations of publications. Finally, we measured the semantic similarities and provided a list of recommendations ranked from most to least similar.

In more detail, datasets and publications in the GESIS-wide search are described with several general textual metadata like title, abstract, author or investigator, topic, and other type-specific metadata. We decided to utilize only titles and abstracts, as first both types have these labels, and second they are focused on the main topic of their contents. We concatenated the title and abstract of all items, i.e., we did not separate between dataset title/abstract and publication title/abstract but rather put them together, and trained the Doc2Vec model. Fig. 1 shows the number of words in titles, topics, and abstracts in publications and datasets. We set up a 300-dimensional vector space with a window size of five models for German and English language words and built the vocabulary of the entire corpus ($177k$ items). For computing the similarity between datasets and publications, we have compared the paragraph vectors in the vector space of items. We trained two models for English and German words. Subsequently, we measured the similarities between $98k$ items ($62k$ datasets and $36k$ publications with English metadata) using the English language model, and $78k$ items ($20k$ datasets and $57,884$ publications with German metadata) using the German language model.

Table 1: Publication-Dataset connections statistics in GWS corpus

Corpus	Statistic
$9,373$	# publications citing research data
$2,823$	# unique datasets
$22,201$	total # of dataset citations
2.368	avg. citations per publication

Retrieved top-1000 similar items (publication + dataset)	
$1,294$ (5.82%)	total # of relevant datasets retrieved
327	# of relevant datasets retrieved @10

4 Preliminary Results and Discussion

As mentioned, our recommendation task considers suggesting a cited dataset in a publication as *relevant* for the user who is currently viewing this publication. Thus, we computed "related dataset"-links of publications in GWS and considered them as relevant for dataset recommendation. As an example of those links, dataset with title *"Role of Government -ISSP 1985"* [5] is cited by the publication entitled: "*Police powers"* [6]. Table 1 illustrates our corpus as well as the number of correctly retrieved datasets.

When retrieving the recommendable datasets for each publication, we observed the rank of used/cited datasets in the retrieval results. The outcome was not as we expected, since we could retrieve only 5.82% (i.e., only $1,294$ out of $22,201$) of all used/cited datasets in the first $1,000$ results (and only 327 in the top-10).

Using only the abstract and the title of publications and datasets, we found that it is difficult to retrieve datasets which are utilized in publications. This can have various reasons, such as the insufficient amount of words in the title and abstract and the lack of consideration of other, potentially useful information, such as publication dates or the dataset citation context as representation of a dataset. In general, the amount of metadata per record in the GWS corpus is quite different. Some records have well and prosperous metadata whereas others are poorly described, e.g., restricted to a short title and authors name. Additionally, quite an amount of datasets did not even have an abstract describing the dataset, but rather some keywords and bullet points placed as abstract. Also, training over the mix of publications and datasets might cause a problem. One possible solution would be to train embeddings for datasets and documentations separately. Another possibility to improve the results is to include a publication's abstract in the datasets' descriptions which are cited by this publication. Among other reasons, as mentioned before, the actual relevance of a recommendation is difficult to assess, which indicates that offline evaluations might be inappropriate in recommendation scenarios, as they are limited in representing the users' interests. This means, a retrieved dataset in higher rank could still be semantically relevant to the currently viewed publication although it is not applied/cited in the publication.

[5] `https://search.gesis.org/research_data/ZA1490`
[6] `https://search.gesis.org/publication/gesis-bib-24288`

In the next steps, we intend to improve our model by using a pre-trained vector space where the representation of the known words are determined, or refine the model by assigning a weight to each word (e.g., a simple TF-IDF or attention layer). Additionally, one can represent the GWS datasets, publications, and their relationships within a graph. This can serve a lot of applications such as node recommendation and link prediction [18]. Considering more metadata for each item, such as authors or publication years, can also improve the result. Finally, we intend to perform an online evaluation of our approaches using a Living Lab [4] and compare them to the default "more-like-this"-baseline SOLR offers out of the box by analyzing click-through-rates and similar.

References

1. Beel, J., Gipp, B., Langer, S., Breitinger, C.: Paper recommender systems: a literature survey. International Journal on Digital Libraries **17**(4), 305–338 (2016)
2. Ben Ellefi, M., Bellahsene, Z., John, B., Demidova, E., Dietze, S., Szymanski, J., Todorov, K.: RDF Dataset Profiling - a Survey of Features, Methods, Vocabularies and Applications. Semantic Web Journal Accepted in August 2017, to appear.
3. Borgman, C.L.: The conundrum of sharing research data. Journal of the American Society for Information Science and Technology (6), 1059–1078 (jun)
4. Breuer, T., Schaer, P., Tavakolpoursaleh, N., Schaible, J., Wolff, B., Mueller, B.: STELLA: Towards a Framework for the Reproducibility ofOnline Search Experiments. In: Proceedings of the Open-Source IR Replicability Challenge (OSIRRC) (accepted) (2019)
5. Ellefi, M.B., Bellahsene, Z., Dietze, S., Todorov, K.: Dataset recommendation for data linking: An intensional approach. In: ESWC. Springer (2016)
6. Gregory, K., Groth, P., Cousijn, H., Scharnhorst, A., Wyatt, S.: Searching data: A review of observational data retrieval practices in selected disciplines. Journal of the Association for Information Science and Technology (2019)
7. Hienert, D., Kern, D., Boland, K., Zapilko, B., Mutschke, P.: A digital library for research data and related information in the social sciences. In: ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL) (forthcoming) (2019)
8. Koren, Y., Bell, R.: Advances in Collaborative Filtering (2011)
9. Kunze, S.R., Auer, S.: Dataset Retrieval. In: 2013 IEEE Seventh International Conference on Semantic Computing. pp. 1–8. IEEE (sep)
10. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. pp. II–1188–II–1196. ICML'14 (2014)
11. Levy, O., Goldberg, Y.: Neural word embedding as implicit matrix factorization. In: Advances in neural information processing systems. pp. 2177–2185 (2014)
12. Lopes, G., Paes Leme, L.A., Nunes, B., Casanova, M., Dietze, S.: Two approaches to the dataset interlinking recommendation problem (2014)
13. Musto, C., Semeraro, G., De Gemmis, M., Lops, P.: Word embedding techniques for content-based recommender systems: An empirical evaluation. In: RecSys Posters (2015)
14. Narayanan, A., Chandramohan, M., Venkatesan, R., Chen, L., Liu, Y., Jaiswal, S.: Graph2Vec: Learning Distributed Representations of Graphs. CoRR (2017)
15. Ricci, F., Rokach, L., Shapira, B.: Introduction to Recommender Systems Handbook, pp. 1–35. Springer US, Boston, MA (2011)
16. White, R.: Interactions with search systems (2016)
17. Witten, I.H.I.H., Bainbridge, D., Nichols, D.M.: How to build a digital library. Morgan Kaufmann Publishers (2010)
18. Zhou, C., Liu, Y., Liu, X., Liu, Z., Gao, J.: Scalable graph embedding for asymmetric proximity. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)