Hajra, Arben; Tochtermann, Klaus

Enriching Scientific Publications from LOD Repositories Through Word Embeddings Approach

Suggested Citation: Hajra, Arben; Tochtermann, Klaus (2016) : Enriching Scientific Publications from LOD Repositories Through Word Embeddings Approach, In: Garoufallou, E. Subirats Coll, I. Stellato, A. Greenberg, J. (Ed.): Metadata and Semantics Research. MTSR 2016, ISBN 978-3-319-49157-8, Springer, Cham, pp. 278-290, https://doi.org/10.1007/978-3-319-49157-8_24

Enriching Scientific Publications from LOD Repositories through Word Embeddings Approach

Arben Hajra[1], Klaus Tochtermann[2]

[1] South East European University (SEEU), Tetovo/Skopje, R. of Macedonia
{a.hajra}@seeu.edu.mk
[2] Leibniz Information Centre for Economics (ZBW), Kiel/Hamburg, Germany
{k.tochtermann}@zbw.eu

Abstract. The era of digitalization is increasingly emphasizing the role of Digital Libraries (DL), by increasing requirements and expectations of services provided by them. The interoperability among repositories and other resources continues to be a subject of research in the field. Retrieving publications related to a particular topic from different DLs, especially from diverse domains, require several clicks and online visits of many different points of access. However, achieving interoperability by cross-linking publications, authors and other related data would facilitate the scholarly communication in general. Starting from a single point, a scholar would be able to find resources i.e., publications and authors, previously enriched with several other information from different repositories. Repositories available as semantic web content, such as bibliographic Linked Open Data (LOD) datasets are the focus of this study. Primarily, we consider existing alignments among concepts between repositories. Improvements regarding the semantic measurements of relatedness of different resources are possible by the application of text-mining techniques. The paper introduces preliminary experiments conducted by vector space models through the application of TF-IDF and Cosine Similarity (CS). Additionally, the paper discusses experiments of applying a word embedding approach, with which we are focusing mainly on the context by distributed word representations, instead of word frequency, weighting and string matching. We apply the contemporary Word2Vec model as a similar deep learning approach to model semantic word representations.

Keywords: digital libraries, linked open data, semantic web, word embeddings, data mining, recommended systems

1 Introduction

Traditionally, libraries provide the basic information infrastructures for scholarly communication. The era of digitalization emphasized their role in this process, but at the same time, requirements and expectations of services provided by them increased. In this situation, Digital Libraries (DL) successfully managed to adapt to these challenges by improving the utilization of resources [14]. Nonetheless, there is still a gap between the demand and offer. Interoperability among resources continues to be the subject in this field even today [15], [16], [17], [18].

Often, scientific digital libraries are specialized in specific domains such as: economics, social sciences, computer sciences, agronomics, etc. Retrieving similar publications within the same DL is a common practice in most of DLs. However, recommending semantically similar publications from two or more different repositories is still an open field of research. Today, retrieving publications related to a particular topic, from different DLs and especially from different domains, is still very heuristic, and often require step-wise or as far as possible simultaneous navigations through the affected DLs. The current practice of Google Scholar gives an idea for such recommendations, however there are much more resources which are not made visible by services like this. To this end, achieving the interoperability among DLs by cross linking publications, authors and other related data would facilitate the scholarly communication in general. The idea is as follows: Starting from a single point of access, a scholar would be able to find resources i.e., publications and authors, previously enriched with several other information from different repositories. And when a scholar fetches a publication in a DL, the system will offer the scholar a list of semantically related publications from other repositories, an extended list of co-authors, and other related data corresponding to that publication.

Repositories available as semantic web content, such as bibliographic Linked Open Data (LOD) repositories [15], [29], are in the focus of this study. Primarily, we consider the existing alignments among concepts between repositories, by exploring best practices for consuming them. After that, we investigate the role of thesauri, including descriptors with the corresponding narrowed, broadened and extended concepts through Simple Knowledge Organization System Reference - SKOS[1] vocabulary. Improvements regarding the semantic measurements between resources are achieved by evaluating several text-mining techniques. In this study, we present preliminary experiments conducted by vector space models through the application of TF-IDF and Cosine Similarity (CS). Additionally, we extend the experiments by applying a word embedding approach, in which we are focusing mainly on the context by distributed word representations, instead of words frequency, weighting and string matching. The contemporary Word2Vec[2] model is applied as a similar deep learning approach to model semantic word representations.

The main intention of our work is to find a novel and automatic approach for crosslinking scientific publications from different repositories. In our view, the implementation of deep learning approach for language processing is proposed as the most comprehensive approach for this purpose. To this end, we show how we can automatically determine the semantic similarity between publications, even if only a small set of metadata is available.

2 Motivation and Problem Statement

Recommender systems are applied in several fields, therefore it is inevitable to explore their application in scholarly communication, particularly in digital libraries

[1] https://www.w3.org/TR/swbp-skos-core-spec
[2] https://code.google.com/p/word2vec/

[11], [12] [13]. However, the common implementation of recommending systems in DLs is mainly a practice used within the same repository. Recommending and interlinking publications by cross-linking relevant information from several repositories still remains a challenge [19], [20]. At the moment, repositories are considered as isolated silos, which make it difficult to process matching similar resources by using the same query string in different repositories. Cross-linking resources, i.e., scientific publications with assured degree of semantic similarity, certainly presents a complex process of lexical or string matching, mostly due the diversity of ontologies and metadata vocabularies used for describing resources.

3 Proposed Approach and Related Work

Recommender systems for scientific publications are generally grounded on content analysis, user profiles and collaborative filtering with incontestable role of social data [21], [22], [23], [24]. However, in this work we are following a different strategy for initiating and retrieving the list of recommended relevant resources. In essence, the user triggers the search and selects a paper from a DL that best fits his or her requirements. In a next step, the selected publication is enriched with closely related publications, authors and similar information found in other repositories.

The interoperability is initiated from one repository by considering all existing metadata for a single publication, such as: title, authors, abstract and keywords. Using this information, we are connecting to other external repositories to search for possible semantically related publications and other related information (e.g. author details) to the initial publication.

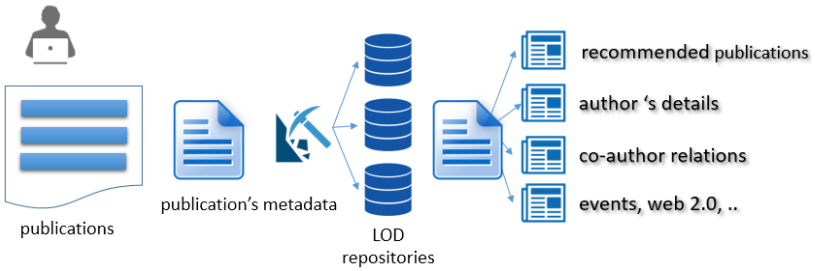


Fig. 1. Enriching a scientific publication with information from other repositories.

In order to achieve this, we leverage already available contents on the semantic web, such as Linked Open Data (LOD) repositories, as one of the most promising data sources [30]. As such, the existing alignments among concepts between repositories are considered with the corresponding narrowed, broadened and extended concepts through the SKOS vocabulary. At the same time, the deployment of several data mining techniques is crucial in this process [21]. In our work we apply two approaches the vector space model and word embedding approach.

This work was evaluated with the content of the EconStor[3] repository, which is a leading Open Access repository in Germany. Through EconStor, the German National Library of Economics - Leibniz Information Centre for Economics (ZBW) offers a platform for Open Access publishing to researchers in economics. ZBW also maintains the Standard Thesaurus Wirtschaft (STW)[4], which is the Thesaurus for Economics used for description and indexing purposes.

3.1 Aligned Concept between Repositories and Thesauruses

Most of LOD repositories as part of LOD cloud[5], offer a number of incoming/ongoing links to other datasets for mapping several resources or concepts that have the same meaning. EconStor, through the STW thesaurus, has numerous mappings to other thesauri and vocabularies. For instance, for Agrovoc (Multilingual Agricultural Thesaurus)[6] 1,027 *skos:exactMatch* alignments exist, while for TheSoz[7] (Thesaurus Social Sciences) 3,022 *skos:exactMatch* and 1,397 *skos:narrowMatch* are available. According to this, the initial experiments are done between EconStor and OpenAgris[8] based on structural similarity between these two repositories. Both of them offer an open catalog as part of LOD cloud with available SPARQL endpoints and RDF dump files, as well as thesauri on both sides, STW and Agrovoc respectively.

Based on our previous evaluation conducted using 112 publications, the list of retrieved publications according to the aligned concepts between repositories was extremely wide [8]. For example, in order to deliver more details, the concept "biofuel" from EconStor is aligned to Agrovoc as "biofuels", and is used for describing 7083 documents in OpenAgris catalog. By including all the existing aligned concepts describing a paper, the list will be even broader. Hierarchical navigation between concepts with the use of knowledge organization systems by broadening and narrowing the concepts, e.g., the notion of Germany broadened to Europe and narrowed to Berlin, helps to reduce complexity by narrowing down the number of results. However, the outcome is not satisfactory for measuring similarity among publications and offering a shorter list of recommended publications.

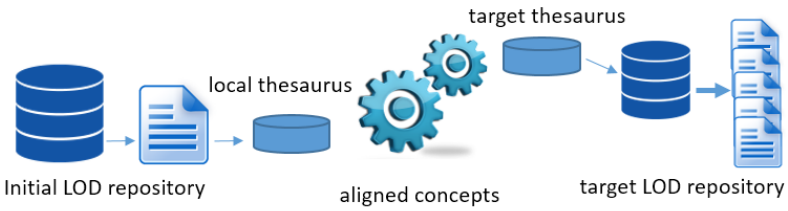


Fig. 2. Retrieving scientific publications from LOD repositories based on concepts' alignments

[3] http://www.econstor.eu/
[4] http://zbw.eu/stw
[5] http://linkeddata.org/
[6] http://aims.fao.org/agrovoc
[7] http://www.gesis.org/en/services/research/thesauri-und-klassifikationen/social-science-thesaurus/
[8] http://aims.fao.org/openagris

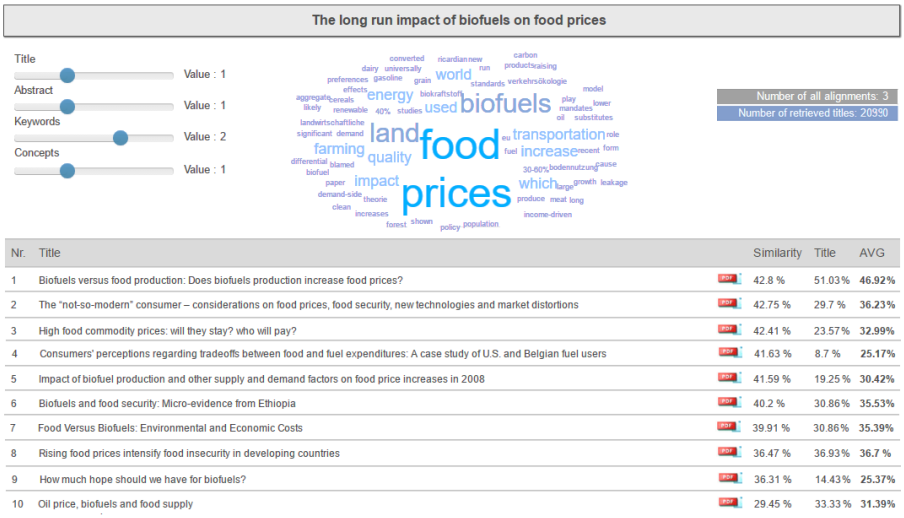
Therefore, we also use alignments between repositories/thesauruses for retrieving an initial set of publications, especially for reformulating a search query from one vocabulary to another [8]. The presence of thesauri in the primary and targeting repository can be useful for extending the corpus of metadata concepts, which, as we will show later, is very significant for further analyses.

3.2 Publications Metadata and Vector Space Model

In such a situation, the involvement of other metadata, such as title, abstract and keywords is mandatory. By including these elements in the implementation of data mining approaches among the set of metadata and thesauri concepts, the similarity between publications is calculated and used for ordering purposes.

We use the vector space model, in which we weight each concept from the selected metadata by applying the TF-IDF algorithm. The similarity among publications, i.e., vectors of concepts, is measured as deviation of angles between each document vector, by using the CS. Thus, iteratively we measure the similarity between metadata of our initial publication with the metadata of publications from the target repository.

For this purpose, we conducted heuristic evaluations when analyzing the impact of each element. As shown in figure 3, the developed prototype makes it possible to adjust the relevance of each metadata set by weighting the title, abstract, keywords, and considering all the aligned concepts (including narrowed, broadened and related terms). The example in figure 3 shows that for the selected publication with current adjustment of the metadata, the words “food” and “price” become crucial. This results in retrieving also publications semantically distant with the initial publication, which are related to *food* and *agriculture* rather than *economy*.



Nr.	Title	Similarity	Title	AVG
1	Biofuels versus food production: Does biofuels production increase food prices?	42.8 %	51.03 %	**46.92%**
2	The “not-so-modern” consumer – considerations on food prices, food security, new technologies and market distortions	42.75 %	29.7 %	**36.23%**
3	High food commodity prices: will they stay? who will pay?	42.41 %	23.57 %	**32.99%**
4	Consumers' perceptions regarding tradeoffs between food and fuel expenditures: A case study of U.S. and Belgian fuel users	41.63 %	8.7 %	**25.17%**
5	Impact of biofuel production and other supply and demand factors on food price increases in 2008	41.59 %	19.25 %	**30.42%**
6	Biofuels and food security: Micro-evidence from Ethiopia	40.2 %	30.86 %	**35.53%**
7	Food Versus Biofuels: Environmental and Economic Costs	39.91 %	30.86 %	**35.39%**
8	Rising food prices intensify food insecurity in developing countries	36.47 %	36.93 %	**36.7 %**
9	How much hope should we have for biofuels?	36.31 %	14.43 %	**25.37%**
10	Oil price, biofuels and food supply	29.45 %	33.33 %	**31.39%**

Fig. 3. Combination of metadata components from a scientific paper for retrieving recommended publications from other repositories.

The combination among the metadata is crucial for determining the semantic relativeness among the initial and retrieved publications. Different combinations among these parameters would result in different list of retrieved publications from the targeted repository. The impact can also be seen in the generated results. In this case, the first ranked publication is semantically very near to the initial publication.

Concerning this, in our previous work we have achieved very significant results by enriching author profiles with additional information from different digital libraries [25]. In another study [8], considering different cases, different combinations of these metadata also led to good results. Based on the evaluations done with 112 publications, the count-based approach with TF-IDF and Cosine Similarity [8], repeatedly shows that irrelevant terms are highly ranked. This results in compromising outcomes, i.e., recommending semantically distant publications to a particular publication. Therefore, the right combination of metadata terms for this purpose is very experimental. The above mentioned data mining techniques, TF-IDF and CS do not offer much to achieve a completely automated process [27].

4 Deep Learning Approach

Determining the semantic similarity between two texts represent a complex and challenging process. In general, there are several approaches introduced based on lexical matching, handcrafted patterns, term-weighting and syntactic parse trees [9], [10]. Indeed, lexical features, like string matching and frequency of words in a text, do not capture semantic similarity in a satisfied level [9], [27]. Hence, the deep learning approach for language processing based on neural network language models outperforms traditional count-based distributing models on word similarity [2]. Current trends for determining word similarities, i.e., semantic similarities among texts, rely on vector representations of words by using neural networks, known as *word embeddings* or word representations [1], [2], [3], [4], [5], [6], [7], [9], [26], [27], [28].

4.1 Word Embeddings

In deep learning, word embeddings currently represent the most outstanding field. It is the main discussed subject in almost every publication regarding the semantic representation of words in a low-dimensional vector [1], [2], [3], [4], [5], [6], [7], [9], [26], [27], [28]. Their presence is evident in many areas, such as in Natural Language Processing (NLP), Information Retrieval (IR) and generating search query strings. Word embeddings insert the complete vocabulary into a low-dimensional linear space. The embedded word-vectors are trained over large collections of text corpuses through neural networking models. Thus, words are embedded in a continuous vector space where semantically similar words are mapped to vectors. Learning the word embeddings is totally unsupervised method computed on a predefined text corpus.

Word embeddings currently have two well-known models of implementation: the Word2Vec algorithms proposed by Mikalov et al. for Google [7] and GloVe model

from Pennington et al. at Stanford [28]. Our experiments and evaluations are based on Word2vec due to the performance and computational cost.

Word2Vec Embeddings

As noted before, Word2Vec is a novel word embeddings approach, which learns a vector representation for each word using neural network language model [7]. Two implementations of Word2Vec can be found, continuous bag-of-words (CBOW) and Skip-gram. CBOW predicts a word from the context of input text (surrounding words), while Skip-gram predicts the input words from the target context (surrounding words are predicted from one input word).

Word2Vec uses the hierarchical softmax training algorithm, which best fits for infrequent words while negative sampling better for frequent words and low dimensional vectors. Based on the previous analyses in [7], [26], [27], the skip-gram model with the use of hierarchical softmax algorithm is particularly efficient regarding the computational cost and performance. CBOW is recommended as more suitable for larger datasets. As such, the model can be trained on conventional personal machines with billions of words, achieving the ability to learn complex word relationships [7], [26].

Currently, there are several implementations of Word2Vec in different environments. The native proposed code is optimized in the C programming language. However, Deeplearning4j[9] implements a distributed form of Word2Vec for Java and Scala, while Gensim[10] and TensorFlow[11] offer a python implementation of Word2Vec.

4.2 Training and Building the Model

The experiments in this work are based on the Gensim package, which is a python implementation of Word2Vec model. Gensim provides very significant optimization regarding the computational speed, which overpasses even the native C implementation. Currently, there are several pre-trained models on different datasets, such as Google News, DBpedia and Freebase. However, considering the specificity of the domain, we prefer to train our own word vectors for deploying the experiments.

The model is trained in a text corpus for generating a set of vectors, which are word representations of words in that corpus. Thus, through a SPARQL query we retrieve all the titles, abstracts and keywords of 37,917 publications from EconStor. Since Gensim's Word2Vec expects a sequence of sentences as input, several preprocessing steps are performed at the corpus, such as conversion to utf8 unicode, lowercasing, removing numbers and punctuations. Finally, the model is trained in corpus of 12,329,307 raw words and 683,937 sentences. Before training the process, several parameters are determined that affect the training speed and performance. Based on our dataset size, only words that appear more than two times in the corpus are considered. The dimensionality space of the words inside a vector is set to 100, which means that each word is represented with 100 most similar words in that vector. More words

[9] http://deeplearning4j.org/word2vec.html
[10] https://radimrehurek.com/gensim/models/word2vec.html
[11] https://tensorflow.org/versions/r0.8/tutorials/word2vec/index.html

in a vector means better quality, although bigger dataset must be used. The model has been trained in the hierarchical skip-gram architecture in a laptop with i5 CPU 1.7 GHz, 8 GB RAM memory. Surprisingly, the time it took was very fast, 109.5 sec, and thus far beyond our expectations.

4.3 Analyzing the Model

This section presents the investigation of the learned model. We performed several analyses on top of trained model in section 4.2. One of the most interesting analyses regarding the word representation approach is about finding the set of closest words based on a particular entered word. For instance, regarding the economic domain of the trained corpus, we are interested to see what the model learned about the economic concept "money" and a more general one, "food". Table 1, lists ten nearest terms that Word2Vec has calculated for these words.

The generated results are very impressive. For example, the word "liquidity" is ranked as the most similar to "money" with a degree of similarity .764 out of 1, and all others are intuitively very close to it. Moreover, a word is represented in a relationship to hundred words like this, as defined at the training parameters. To our knowledge it is almost impossible to generate such a result through dictionaries or thesauruses. Thus, if we are referring to the STW thesaurus described in section 3, the concept "money" is not represented with many meaningful terms, regarding the SKOS vocabulary. Even the usage of other external resources, such as WordNet synonyms, does not offer such an impressive set of related terms.

Table 1. Top ten most similar words based on the words "money" and "food", generated through word2vec from our text corpus.

a. for the word "*money*"

Word	**Similarity**
liquidity	.764
credit	.723
loan	.709
debt	.654
lending	.644
borrowing	.643
asset	.642
short-term	.634
bank	.633
bond	.632

b. for the word "*food*"

Word	**Similarity**
energy	.789
agricultural	.786
water	.767
land	.756
crop	.701
fuel	.694
transport	.694
agriculture	.691
electricity	.690
milk	.684

The trained model can be used for several other semantic language processing. Accordingly, there is a possibility to retrieve a list of most similar words by subtracting words from a given set of words. Thus, from a set of metadata we have the possibility to include and exclude several concepts. For example, from the set of metadata concepts defined for a publication, we want to consider the terms "bank", "oil" and "price" by excluding the term "food". Therefore, based on this formula [(bank + oil + price) – (food)], the trained model offers the term *currency* with .764 similarity, *li-*

quidity with .734 and *spreads* with .695. This means that the retrieved publications according to this expression are semantically related to these terms.

5 Results and Discussions

This study presents several approaches with regard to the initial purpose to enrich scientific publications of a DL with other relevant information from other repositories. However, the main challenge is the determination of semantic relatedness between the initial and retrieved publications.

As emphasized in section 3, the implementation of count-based approach through TF-IDF and Cosine Similarity, requires a large set of metadata from the publications, to measure the similarity degree. Moreover, the right combination of metadata elements is crucial. Hence, in several cases the presence of a more general concept used in these metadata had negative impact on the result. For example, regarding the publication titled "*The long run impact of biofuels on food prices*", the word "*food*" has been determinant in the similarity measurements when only the title has been considered for the calculation. Thus, the retrieved publications have been related to "*agriculture"* and "*food diets"*, which semantically are not that close to the initial publication. By including the abstract and keywords, improvements were evident. However, this applies heuristic involvements in the evaluation of results. Moreover, the count-based approach shows significant weakness in recognizing relationships among terms, even in the cases when the presence of thesauri is evident.

Based on the developed prototype, we have evaluated randomly 37 publications from EconStor. For each selected publication, the prototype retrieves and orders the most semantically similar publications from OpenAgris. The process is the same as in figure 3, however the similarity is calculated through Word2Vec instead of TF-IDF and CS. The top ten retrieved publications are manually analyzed in order to determine the semantic relevance with the initial publication. In a situation like this, the implementation of word embeddings approach shows outstanding results, even with smaller amount of metadata and combinations among them. In 100% of the cases, the Word2Vec embedding approach overcome TF-IDF with Cosine Similarity.

Table 2 depicts the results of one from the 37 evaluated publications, by comparing the results generated in both approaches with two different sets of metadata. Firstly, the similarity degree between publications A and B is calculated only on titles (T), such as $sim(T_a, T_b)$. As such, for the first retrieved publication on that list, Word2Vec has generated **.804** similarities with the EconStor publication titled "*The long run impact of biofuels on food prices*". The count-based implementation of Cosine Similarly gives **.5103** similarities between the same titles. In the same example, analyses are extended by including other metadata terms in the similarity calculations. Hence, from the EconStor publications the **title(T_a)**, **abstract(A_a)**, **keywords(K_a)** and **descriptors(D_a)** are considered, while from the OpenAgris publications the **title(T_b)**, **abstract(A_b)** and **descripts(D_b)**. The last two columns of table 2 show the similarity among these metadata comparatively, $sim(T_aA_aK_aD_a, T_bA_bD_b)$. By considering the first publication from table 2, TF-IDF with CS generates **.428** similarity degree among

them, while Word2Vec gives **.962**. The row number ten emphasizes even more the discrepancy between the generated results. In that case, we realized that the retrieved publication is closely related to the EconStor publication, however the first method has generated only .2757 similarity degree compared to .9343 from Word2Vec.

Table 2. The similarity degree between a particular EconStor publication with OpenAgris publications, calculated with TF-IDF & CS versus Word2Vec.

		$sim(T_a, T_b)$		$sim(T_a A_a K_a D_a, T_b A_b D_b)$	
	Title	**TF-IDF with CS**	**Word2Vec**	**TF-IDF with CS**	**Word2Vec**
1	Biofuels versus food production: Does biofuels production increase food prices?	.5103	**.8040**	.4280	**.9620**
2	The "not-so-modern" consumer – considerations on food prices, food security, new technologies and market distortions	.2970	**.7780**	.4275	**.8904**
3	High food commodity prices: will they stay? who will pay?	.2357	**.7740**	.4241	**.9204**
4	Consumers' perceptions regarding tradeoffs between food and fuel expenditures: A case study of U.S. and Belgian fuel users	.0871	**.6521**	.4163	**.9368**
5	Impact of biofuel production and other supply and demand factors on food price increases in 2008	.1925	**.8660**	.4159	**.9594**
6	Biofuels and food security: Micro-evidence from Ethiopia	.3086	**.6592**	.4023	**.8903**
7	Food Versus Biofuels: Environmental and Economic Costs	.3087	**.6723**	.3991	**.8043**
8	Rising food prices intensify food insecurity in developing countries	.3693	**.7710**	.3647	**.9461**
9	How much hope should we have for biofuels?	.1443	**.4420**	.3631	**.9318**
10	Oil price, biofuels and food supply	.3086	**.8320**	.2757	**.9343**

The word embeddings approach evidently overcome the count-based and text-matching approach. The results generated here are significantly better even with smaller amount of concepts included in similarity calculation. The similarity calculated by Word2Vec shows outstanding performance, even when only titles are compared. The presence of other metadata, such as the abstract and keywords, improves the calculation of semantic similarity between publications. By considering the performed evaluations, the word embeddings approach evidently contribute for enriching a scientific publication with semantically related information, such as other publications from different repositories.

6 Summary

The main intention of this work was to emphasize the advantages resulting from an improved interoperability among different Digital Libraries and to investigate different algorithms to achieve this interoperability. Thus, by cross-linking data from different places, a particular resource would be enriched with several other information. This results in a significant enhancement of scholarly communication in general, regarding time consuming and quality of the required information. The idea is to perform a single query in a single place (e.g. their favorite DL) and still to offer scholars information from different repositories, based upon this single query. Ultimately, a

selected publication in a DL, will be enriched with a list of recommended publications from other DLs, such as, additional information about authors, conferences, etc.

In order to achieve this, we needed to find this information and then determine its relevance i.e., semantic similarity between two different resources. For this purpose, bibliographic Linked Open Data repositories are considered by investigating the alignments among them. We applied several data mining techniques, such as TF-IDF and Cosine Similarity, among the publications metadata. The generated results, showed that the traditional count-based and text-matching approach require a heuristic way to determine a satisfactory level of semantic similarity among publications. Given this, we also followed the deep learning approach to model semantic word representations. The implementation of a contemporary Word2Vec model results in an outstanding outcome. This is achieved by simplifying the combination process between the metadata, and even more, by performing it on a smaller set of metadata, such as title's concepts only. However, significant improvements are evident by extending the set of metadata with concepts from the abstract and keywords. More detailed analysis with these sets of metadata, and the expansion of the evaluations range will be investigated in our future work.

References

1. Lebret, R., Collobert, R.: Rehabilitation of Count-based Models for Word Vector Representations. In Computational Linguistics and Intelligent Text Processing. pp. 417-429. Springer International Publishing (2015)
2. Levy, O., Goldberg, Y., Dagan, I.: Improving distributional similarity with lessons learned from word embeddings. In Transactions of the Association for Computational Linguistics. 3, 211-225 (2015)
3. Bengio, Y., Schwenk, H., Senécal, J. S., Morin, F., Gauvain, J. L.: Neural probabilistic language models. In Innovations in Machine Learning. pp. 137-186. Springer Berlin Heidelberg. (2006)
4. Collobert, R., Weston, J.: A unified architecture for natural language processing: Deep neural networks with multitask learning. In Proceedings of the 25th international conference on Machine learning. pp. 160-167. ACM (2008)
5. Mnih, A., Hinton, G. E.: A scalable hierarchical distributed language model. In Advances in neural information processing systems. pp. 1081-1088 (2009)
6. Turian, J., Ratinov, L., Bengio, Y.: Word representations: a simple and general method for semi-supervised learning. In Proceedings of the 48th annual meeting of the association for computational linguistics. pp. 384-394. Association for Computational Linguistics. (2010)
7. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In arXiv preprint arXiv. 1301.3781 (2013)
8. Hajra, A., Latif, A., Tochtermann, K.: Retrieving and ranking scientific publications from linked open data repositories. In Proceedings of the 14th International Conference on Knowledge Technologies and Data-driven Business. p. 29. ACM (2014)
9. Kenter, T., de Rijke, M.: Short text similarity with word embeddings. In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. pp. 1411-1420. ACM (2015)
10. Robertson, S., Zaragoza, H.: The probabilistic relevance framework: BM25 and beyond. In Information Retrieval Vol. 3, No. 4, 333–389. Now Publishers Inc (2009)

11. Mooney, R. J., Roy, L.: Content-based book recommending using learning for text categorization. In Proceedings of the fifth ACM conference on Digital libraries. pp. 195-204. ACM (2000)
12. Huang, Z., Chung, W., Ong, T. H., Chen, H.: A graph-based recommender system for digital library. In Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries. pp. 65-73. ACM (2002)
13. Smeaton, A. F., Callan, J.: Personalisation and recommender systems in digital libraries. In International Journal on Digital Libraries. 5(4), 299-308 (2005)
14. Kling, R., McKim, G.: Scholarly communication and the continuum of electronic publishing. In arXiv preprint cs/9903015 (1999)
15. Paepcke, A., Chang, C. C. K., Winograd, T., García-Molina, H.: Interoperability for digital libraries worldwide. In Communications of the ACM. 41(4), 33-42 (1998)
16. Borgman, C. L.: Challenges in building digital libraries for the 21st century. In Digital Libraries: People, Knowledge, and Technology. pp. 1-13. Springer Berlin Heidelberg (2002)
17. Besser, H.: The next stage: Moving from isolated digital collections to interoperable digital libraries. First Monday, 7(6) (2002)
18. Sheth, A. P.: Changing focus on interoperability in information systems: from system, syntax, structure to semantics. In Interoperating geographic information systems. pp. 5-29. Springer US (1999)
19. Dietze, S., Sanchez-Alonso, S., Ebner, H., Qing, Y. H., Giordano, D., Marenzi, I., Pereira N. B.: Interlinking educational resources and the web of data: A survey of challenges and approaches. Program, 47(1), 60-91 (2013)
20. Horava, T.: Challenges and possibilities for collection management in a digital age. In Library Resources & Technical Services, 54(3), 142-152 (2011)
21. Park, D. H., Kim, H. K., Choi, I. Y., Kim, J. K.: A literature review and classification of recommender systems research. In Expert Systems with Applications, 39(11), 10059-10072 (2012)
22. Bobadilla, J., Ortega, F., Hernando, A., Gutiérrez, A.: Recommender systems survey. In Knowledge-Based Systems. 46 109-132 (2013)
23. Lops, P., De Gemmis, M., Semeraro, G.: Content-based recommender systems: State of the art and trends. In Recommender systems handbook. pp. 73-105. Springer US (2011)
24. Sugiyama, K., Kan, M. Y.: Scholarly paper recommendation via user's recent research interests. In Proceedings of the 10th annual joint conference on Digital libraries. pp. 29-38. ACM (2010)
25. Hajra, A., Radevski, V., Tochtermann, K.: Author Profile Enrichment for Cross-Linking Digital Libraries. In Research and Advanced Technology for Digital Libraries. pp. 124-136. Springer International Publishing (2015)
26. Kusner, M., Sun, Y., Kolkin, N., Weinberger, K.Q.: From Word Embeddings To Document Distances. In Proceedings of the 32nd International Conference on Machine Learning (ICML-15). pp. 957-966 (2015)
27. Baroni, M., Dinu, G., Kruszewski, G.: Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In ACL (1). pp. 238-247 (2014)
28. Pennington, J., Socher, R., Manning, C. D.: Glove: Global Vectors for Word Representation. In EMNLP. Vol. 14, pp. 1532-1543 (2014)
29. Latif, A., Scherp, A., Tochtermann, K.: LOD for Library Science: Benefits of Applying Linked Open Data in the Digital Library Setting. In KI-Künstliche Intelligenz. 1-9 (2015)
30. Berners- Lee, T., Hendler, J., Lassila, O.: The semantic web. In Scientific American. 284(5), 28-37 (2001)