ZBW Publikationsarchiv

Publikationen von Beschäftigten der ZBW – Leibniz-Informationszentrum Wirtschaft
Publications by ZBW – Leibniz Information Centre for Economics staff members

Hajra, Arben; Tochtermann, Klaus

Article — Accepted Manuscript (Postprint)

Linking science: approaches for linking scientific publications across different LOD repositories

International Journal of Metadata, Semantics and Ontologies

Suggested Citation: Hajra, Arben; Tochtermann, Klaus (2017) : Linking science: approaches for linking scientific publications across different LOD repositories, International Journal of Metadata, Semantics and Ontologies, ISSN 1744-263X, Inderscience, Olney, Bucks, Vol. 12, Iss. 2/3, pp. 124-141,
https://doi.org/10.1504/IJMSO.2017.10011833

This Version is available at:
http://hdl.handle.net/11108/356

Kontakt/Contact
ZBW – Leibniz-Informationszentrum Wirtschaft/Leibniz Information Centre for Economics
Düsternbrooker Weg 120
24105 Kiel (Germany)
E-Mail: info@zbw.eu
https://www.zbw.eu/de/ueber-uns/profil-der-zbw/veroeffentlichungen-zbw

Standard-Nutzungsbedingungen:
Dieses Dokument darf zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden. Sie dürfen dieses Dokument nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen. Sofern für das Dokument eine Open-Content-Lizenz verwendet wurde, so gelten abweichend von diesen Nutzungsbedingungen die in der Lizenz gewährten Nutzungsrechte.

Terms of use:
This document may be saved and copied for your personal and scholarly purposes. You are not to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public. If the document is made available under a Creative Commons Licence you may exercise further usage rights as specified in the licence.

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Mitglied der Leibniz-Gemeinschaft

Linking science: approaches for linking scientific publications across different LOD repositories

Arben Hajra*

South East European University (SEEU),
Tetovo/Skopje, Macedonia
and
Leibniz Information Centre for Economics (ZBW),
Kiel/Hamburg, Germany
Email: a.hajra@seeu.edu.mk
*Corresponding author

Klaus Tochtermann

Leibniz Information Centre for Economics (ZBW),
Kiel/Hamburg, Germany
Email: k.tochtermann@zbw.eu

Abstract: Enriching the content of a digital library (DL) with additional information from other DLs and domains would facilitate the scholarly communication, scientific findings, and knowledge distribution. The implementation of semantic technologies by interlinking resources results in a new vision for interoperability among different DLs. Therefore, this research explores bibliographic Linked Open Data (LOD) repositories by investigating alignments among them. The application of global unigrams frequency is applied for determining the importance of terms on the set of metadata. The semantic relatedness of the retrieved publications is measured by comparing two main approaches with one another: Vector Space Model through TF-IDF and Cosine Similarity, versus a Deep Learning approach through Word2Vec implementation of Word Embeddings. In summary, they are performing with 40.5% difference, concerning the outcome of relevant retrieved publications. In addition to the given metadata, word embeddings achieve a better performance for short texts, such as publications titles.

Keywords: digital libraries; linked open data; semantic web; word embeddings; data mining; recommender systems.

Reference to this paper should be made as follows: Hajra, A. and Tochtermann, K. (2017) 'Linking science: approaches for linking scientific publications across different LOD repositories', *Int. J. Metadata, Semantics and Ontologies*, Vol. 12, Nos. 2/3, pp.124–141.

Biographical notes: Arben Hajra is a PhD Student at the Faculty of Engineering at Christian-Albrechts-Universität (CAU) in Kiel, Germany. Graduated Computer Sciences at South East European University, Macedonia, and since 2005 is engaged as Teaching Assistant at the same department. Starting from 2012, he is collaborating as Visiting Researcher at German National Library of Economics – Leibniz Information Centre for Economics (ZBW). His research interests include data science, semantic web, linked open data, data mining, and databases.

Klaus Tochtermann has been Professor for digital information infrastructures at Kiel University and also the director of the ZBW – Leibniz Information Centre for Economics, Germany, since 2010. He received his PhD in Computer Science in 1995. In 2002 he passed postdoctoral thesis (habilitation) entitled 'Personalisation in the Context of Digital Libraries and Knowledge Management'. In 1996 he spent his postdoc at the Centre for the Studies of Digital Libraries at the A&M University in Texas (USA). From 1997 to 2000 he was Department Head for Environmental Information Systems at the, FAW Ulm (Germany). Since October 2000 he has been the director of Austria's first industry-based research institute on knowledge management, Know-Centre. In 2004 he was appointed as Professor for Knowledge Management at Graz University of Technology. His current research includes research data management, open science and science 2.0.

Copyright © 2017 Inderscience Enterprises Ltd.

1 Introduction

Libraries present the first and the foremost source for scholarly communication. Traditionally, they provide the basic information infrastructures for sharing and discovering knowledge. During the era of digitalisation, libraries have become even more crucial in this process, by increasing and simplifying the accessibility of resources (Kling and McKim, 1999). At the same time, the need and wants of scholars have also changed fundamentally. Thus, libraries are now considered not only as a place for finding a particular piece of information, but also as a place where the required information would be enriched with various data from different places and domains. Therefore, instead of navigating for interlinking relevant information, a library would provide automated services for that purpose. In such a case, Digital Libraries (DL) successfully managed to adapt to these challenges by improving the utilisation of resources from different perspectives, such as quality of services, system performance and user experience (Xie, 2006; Garibay et al., 2010; Heradio et al., 2012). Even so, there is still an evident gap between demand and supply (Thanos, 2016).

By triggering a publication in a particular DL, apart from the standard metadata used for describing that publication, the system can offer some metrics such as downloads, views, citations and a list of related publications stored in that repository. However, do scholars need more? What about other related publications stored or indexed in different libraries, new author's correlations and other important information for enriching that resource? The interoperability among resources has been identified as a problem in different studies since many years ago (Paepcke et al., 1998; Sheth, 1999; Besser, 2002; Borgman, 2002) and continues to be the subject of research until today (Agosti et al., 2016). The achievement of interoperability among DLs by cross-linking publications, authors and other related data would facilitate the scholarly communication, scientific findings, knowledge retrieving and representation (Thanos, 2014). Starting from a single point of access, a scholar would be able to find resources, i.e. publications and authors, previously enriched with additional information from different (disconnected) repositories.

Identifying potential sources from which the data could be retrieved is a first step towards achieving this purpose. Repositories available as semantic web content, such as bibliographic Linked Open Data (LOD) repositories (Paepcke et al., 1998; Latif et al., 2016), build the focus of our research. Currently, a large number of libraries have exposed their data as RDF statements in the LOD cloud. Examples include the German National Library (DNB), Swedish National Library (LIBRIS), British National Bibliography (BNB), Europeana Digital Library, Library of Congress (LC), Leibniz Information Centre for Economics (ZBW), Food and Agriculture Organisation of the United Nations (FAO), DBLP Bibliography Database, etc. In principle, our interlinking process is relying on existing alignments among concepts used in different repositories and by exploring best practices for consuming these mappings. The role of thesauri used for indexing the data stored in repositories is investigated with particular attention. Thus, for the alignments, we include the descriptors with the corresponding narrowed, broadened and extended concepts through the Simple Knowledge Organisation System Reference – SKOS modelling scheme. Improvements regarding the semantic measurements between resources are achieved by evaluating text-mining techniques.

In this paper, we present experiments conducted by Vector Space Models (Salton et al., 1975) 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 of distributed word representations, instead of words frequency, weighting and string matching. The contemporary Word2Vec implementation is applied as a similar Deep Learning approach to model semantic word representations (Mikolov et al., 2013).

The main objectives of our work are to find a novel and automatic approach for cross-linking 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.

This paper is an extension of work originally reported in Metadata and Semantics Research Conference (Hajra and Tochtermann, 2016). It begins by highlighting the motivation and problem statement about the interoperability of resources between Digital Libraries. In Section 3, we continue by exploring the related work in the context of LOD consumption and recommender systems. Our approach is presented in Section 4, including the publications' centred metadata, repositories and the SKOS alignments between them. Sections 5 and 6 show the implementation of Vector Space Model, through TF-IDF and Cosine Similarity, and the Word Embedding approach, through Word2Vec, respectively. Each approach is elaborated through evaluations by highlighting the weaknesses regarding the generated results. The main results and findings are exposed in Section 7. This paper closed with a summary in Section 8.

2 Motivation and problem statement

The main aim of our work is to enrich the content of a DL with additional information from other DLs, which is closely related to a given publication. Assume we have found a publication and bibliographic information in one DL, we want to harvest other DLs for correlations to other publications and for additional bibliographic information. Thus, 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.

Typically, specialised DLs hold domain-specific information (e.g. economics) which makes it difficult to search across

different domains. For example, would a scholar need literature from economics and agriculture he or she would have to access two different DLs. This happens because scientific digital libraries are specialised in specific domains such as economics, social sciences, computer sciences, agronomics, etc. Recommending semantically similar publications within the same DL is a common practice in most of DLs. However, achieving interoperability by cross-linking authors and/or publications from 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 requires step-wise or as far as possible simultaneous navigations through the affected DLs. The current practice of Google Scholar, BASE (Bielefeld Academic Search Engine), Mendeley or Semantic Scholar from AI2 gives an idea for such recommendations. However, there are much more resources, which are not made visible by this kind of services, especially the interconnection between different domains (Jacsó, 2005; Dorsch, 2017).

Today, 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 an assured degree of semantic similarity, certainly present a complex process of lexical or string matching, mostly due to the diversity of ontologies and metadata vocabularies used for describing resources (Joshi et al., 2012). The usage of LOD, i.e. the aligned concepts between repositories, can be seen as hope for breaking down this heterogeneity.

3 Related work

The implementation of semantic technologies and the approach of interlinking resources known as Linked Data have given a new vision to the interoperability among information (Berners-Lee et al., 2001). Since the conceptualisation of Linked Data principles in 2006, as a set of best practices for publishing and interlinking structured data on the Web, the intention of them has been increased rapidly (Auer et al., 2016). The RDF data model appears to be a widely accepted model for data integration, knowledge representation, and interconnections. Owing to this, Digital Libraries often prefer to publish their indexes or even entire catalogues as RDF serialisations. This intention does not rely only on publishing; applying and consuming Linked Data principles in real applications is now a common practice. Among several examples, a remarkable one is Europeana: an aggregator and single access point to millions of books, paintings, films and museum objects (Doerr et al., 2010). Alignments of concepts, i.e. SKOS mappings among repositories/thesauri, can play a crucial role in the process of interoperability and interdisciplinary. The ARIADNE project highlights the importance of vocabulary linked data for integration of archaeological records (Binding and Tudhope, 2016). Several other projects put the focus on querying and retrieving information from LOD based on these alignments (Fernández et al., 2011; Joshi et al., 2012).

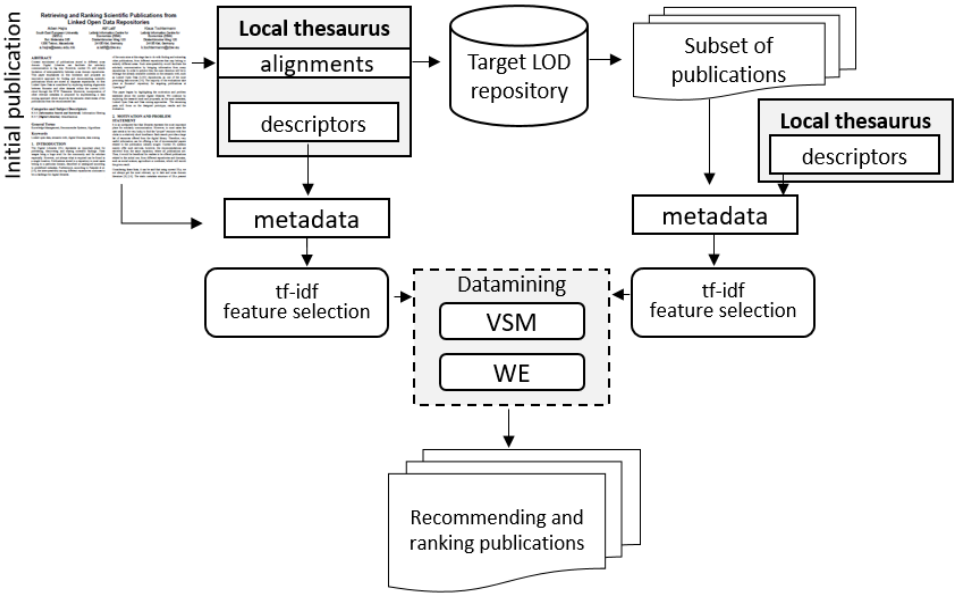
Retrieving information relying on the linked data knows to generate very high recall (Cuzzocrea et al., 2015). Usually, the result is dominated by the information that can be so different from what the user is interested, i.e. not relevant to the user, or any relevant information cannot be displayed. Providing the user with the desired information, several parameters must be considered, such as the previously selected item or any other kind of preferences (Di Noia et al., 2012). Such that, it is inevitable to explore the application of recommender systems in scholarly communication, particularly in digital libraries (Mooney and Roy, 2000; Huang et al., 2002; Smeaton and Callan, 2005). The common implementation of recommending systems in DLs is mainly a practice used within the same repository. Therefore, recommending and interlinking publications by cross-linking relevant information from several repositories remains a challenge (Horava, 2010; Passant, 2010; Dietze et al., 2013). The systems for retrieving and recommending scientific publications are generally grounded on content analysis, user profiles and collaborative filtering with the incontestable role of social data (Sugiyama and Kan, 2010; Lops et al., 2011; Park et al., 2012; Bobadilla et al., 2013). Hence, in this work, we are following a different strategy for initiating and retrieving the list of recommended relevant resources.

4 The approach

The approach followed in this work is entirely based on the set of metadata that are used to describe a paper in a repository, rather than an input query from the user. In essence, the user triggers the search and selects a paper from a DL that best fits her or his requirements. In a next step, the selected publication is enriched with closely related publications, authors and similar information. In addition, 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.

In order to achieve this, we leverage already available contents on the semantic web, such as LOD repositories, as one of the most promising data sources (Berners-Lee et al., 2001). The interest in consuming these data is growing rapidly day-by-day. Such as, the existing alignments among concepts between repositories are considered with the corresponding narrowed, broadened and extended concepts through the SKOS modelling scheme. In order to retrieve a semantically similar set of publications with the initial publication, the deployment of Recommender Systems, i.e. data mining techniques, is applied. In our work, we investigate two approaches, the Vector Space Model and Word Embedding approach. An ultimate overview of this approach is represented in Figure 1.

Figure 1 Approaches for cross-linking scientific publications



4.1 Selected repositories

This work was evaluated with the content of the EconStor repository, which is a leading Open Access repository in Germany (Latif et al., 2014). 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. Its repository metadata are published as more than 40,000 bibliographic records as RDF triples. ZBW also maintains the Standard Thesaurus Wirtschaft (STW), which is the Thesaurus for Economics used for description and indexing purposes (Neubert, 2009).

As target repository, at this phase, we are pointing to OpenAgris, the multilingual bibliographic database for agricultural science (Anibaldi et al., 2015). Its records are enhanced with AGROVOC (Multilingual Agricultural Thesaurus), maintained by Food and Agriculture Organisation of the United Nations (FAO) (Caracciolo and Keizer, 2011; Caracciolo et al., 2012). The thesaurus covers several areas including food, nutrition, agriculture, fisheries, forestry, and the environment.

Thus, the initial experiments are done between EconStor and OpenAgris based on the structural similarity between these two repositories. Both of them offer an open catalogue as part of LOD cloud with available SPARQL endpoints and RDF dump files, as well as thesauri on both sides, STW and AGROVOC, respectively. The main reason has to do with the idea of interlinking repositories from different fields, for achieving an interdisciplinary connection.

The experiments are performed on locally stored dump files from the explored data sets, i.e. EconStor, STW Thesaurus, AGROVOC, and OpenAgris. Regarding OpenAgris, the version with updates of the year 2013 is used with 201,038,257 statements. Data sets are stored in the Ontotext GraphDB database.

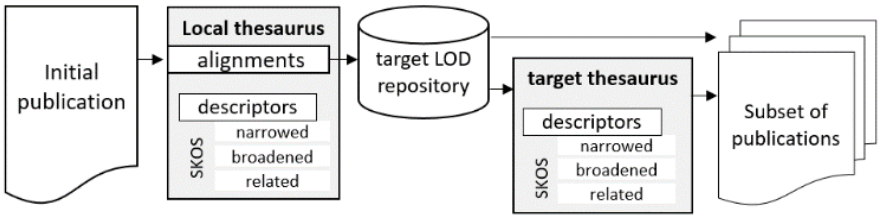
4.2 Publication's metadata

As worked out in the previous sections, the interoperability is initiated from one repository, where in addition to the aligned concepts we are 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 information related to the initial publication. Thus, for a particular paper from EconStor, denoted as $\boldsymbol{p}$, a set of publications D is retrieved, where $D = \{d_1, d_2, d_3, \ldots, d_n\}$ is a subset of that repository. Normally, for each selected paper from EconStor different set of publications will be recommended, such as $\{p_1, D_1\}, \{p_2, D_2\}, \ldots, \{p_n, D_n\}$.

Therefore, from the EconStor publications we are considering the title(p_t), abstract(p_{abs}), keywords(p_k) and descriptors(p_d), while from the targeted repository, i.e. OpenAgris, the title(d_t), abstract(d_{abs}) and descriptors(d_d), since keywords are not provided.

The presence of thesauri on both sides allows the possibility of having an extension of these metadata. That way, each descriptor used for a paper in EconStor or OpenAgris can be narrowed, broadened or presented with related terms, through the SKOS modelling scheme. Let us take an example by selecting a publication from EconStor titled '*Do inflation and high taxes increase bank leverage?*' The main descriptors, in this case, are *Inflation*, *Corporate taxation*, *Equity capital*, *Bank*, *Banking history* and *Sweden*. From here, based on STW thesaurus, *inflation* is narrowed to *Stagflation*, *Hyperinflation* and *Core inflation*, broadened to *Price level*, while related to *Anti-inflation policy*, *Inflation theory*, *Inflation rate* and *Wage-price spiral*.

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



4.3 Aligned concepts between repositories/ thesauruses

Most of LOD repositories as part of LOD cloud offer a number of incoming/outgoing links to other data sets 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 1027 skos:exactMatch alignments exist, while for TheSoz (Thesaurus Social Sciences) 3022 skos:exactMatch and 1397 skos:narrowMatch are available. In fact, alignments in this situation are mappings among concepts between these two thesauruses. For example, the descriptor 'Inflation' is aligned to AGROVOC with absolutely the same label 'inflation' using the URI http://aims.fao.org/aos/agrovoc /c_3857. However, this does not have always to be so; sometimes the mapped concepts can be in plural (ex. Biofuel to biofuels) or a completely different word. In addition, the narrowed and broadened concepts of a particular descriptor may differ in other thesauruses comparing to STW.

Given this, the first step for cross-linking a publication from one repository with other publications from different repositories is done based on the existing aligned descriptors. It is worth mentioning that from the set of descriptors of an EconStor paper, not all of them are aligned to other repositories. However, in the experiments, we are considering publications that have at least one descriptor with the outgoing link.

Based on our previous evaluation conducted using 112 publications, the list of retrieved publications according to the aligned concepts between repositories was extremely wide (Hajra et al., 2014). For example, in order to deliver more details, the concept '*inflation*' is used for describing 2754 documents in OpenAgris catalogue, while '*income*' in 21,838. Since a publication can have several such aligned descriptors, the insertion of all of them through a union is resulting in even a broader outcome. Meanwhile, the attempt to find publications in the target repository, with the same set of descriptors as in the initial one, does not return any publication. The hierarchical navigation between concepts with the use of knowledge organisation 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 offering a shorter list of recommended publications and the opportunity to be ranked.

Therefore, we use alignments between repositories or thesauruses for retrieving an initial set of publications, especially for reformulating a search query from one vocabulary to another (Joshi et al., 2012; Hajra et al., 2014; Binding and Tudhope, 2016). The importance of these descriptors, as well as the alignments among them, is considered as undisputed since experts in relevant fields set them manually. 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.

5 Vector space model

In such a situation, when using aligned concepts generates a wider range of results, we need further processing to narrow this subset. For this purpose, 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.

5.1 Determining the most important terms of a publication

We use the TF-IDF to represent the terms of a paper in a common vector space. The terms of the paper from the initial repository are represented in the vector $\vec{V}(p)$, while each paper in the target repository will embody a particular vector, $\vec{V}(d_i), i = 1, n$. The selection of terms for populating these vectors has a direct impact on the generated results, elaborated later in this section. Additionally, the frequency of a term in the vector can shadow the importance of an important term, with lower frequency. Thus, the importance of each word from the selected metadata is weighted by applying the TF-IDF algorithm (Ramos, 2003; Manning et al., 2008).

Based on the Google Books Ngrams, there is a data set of n-grams consisted of unigrams to 5-grams corpus (Brants and Franz, 2006; Evert, 2010; Norvig, 2013). In this work, we focus on unigrams, i.e. individually words and their frequency. Thus, locally we have saved a data set consisting of 319,999 words in English language and their frequency of usage. Table 1 gives a short overview of some words and their frequency in that data set. As expected, the word '*the*' is the most used with 0.0393 frequency.

Table 1 The list of unigrams, the word and their frequency based on Google Books Ngrams

Word (w)	*Frequency (f_w)*
The	0.03933837507090550000
Of	0.02236252533830050000
And	0.02210015761953700000
To	0.02063676420967820000
High	0.00058731326672819600
Money	0.00032340969045539800
Food	0.00030630268711382300
Bank	0.00015568028973689900
Taxes	0.00005725775413448000
Inflation	0.00001456795810462590
Leverage	0.00000687497277142300

Before populating the vector $\vec{V}(p)$ with terms from the set of publications' metadata, several pre-processing steps are performed, such as removing punctuations, lowercase and encoding the data to Unicode character encoding (UTF-8). Additionally, the list of 'stopwords' is applied for avoiding the iteration of Table 1 for very frequent words. After that, each word that becomes part of the vector is weighted by considering a very naive method. In the case when the word (w) is listed in the frequency data set, the weight of that word is determined as $w_{weight} = \log 10\left(1+\frac{tf}{n}\right)*(1-f_w)$. Otherwise, if the word is not part of that list, the weight will be calculated based on the metadata distribution, such as $w_{weight} = \log 10\left(1+\frac{tf}{n}\right)$. We are applying a global unigrams frequency of words, instead of generating corpus-based frequency, which is a common practice in TF-IDF implementation. The only reason for this approach relies on avoiding the domain influence over the generated frequencies, since we are aiming to cross-link interdomain information.

Consider the title of the publication '*Do inflation and high taxes increase bank leverage?*' After the pre-processing steps, the vector $\vec{V}(p)$ will contain the words *inflation*, *high*, *taxes*, *increase*, *bank*, and *leverage*. In this case, the overall number of words in the vector is denoted as n, n = 6, while the frequency of the words in the document (i.e. title) is denoted as *tf*.

As shown in Figure 3, for a given paper from the initial repository, the developed prototype makes it possible to adjust the relevance of each metadata component: the value can be increased or decreased by weighting the title(p_t), abstract(p_{abs}), keywords(p_k) and descriptors(p_d).

The example shows that if we only consider the title of the selected publication, the words '*leverage*' and '*inflation*' are more crucial, whereas '*high*' is less important. This is because in general '*high*' occurs very often (based on Table 1).

In the second adjustment, when all the metadata components are taken, the word '*bank*' is assigned as the most essential term, followed by '*inflation*' and '*capital*'. The top-ten most important terms regarding this adjustment for this publication are listed in Table 2. Besides the fact that the term 'high' appears seven times in these metadata, it is ranked as the sixth most important one. A better visual interpretation of these weights is shown in Figure 3b.

Figure 3 Adjusting the relevance of the metadata components for the initial publication

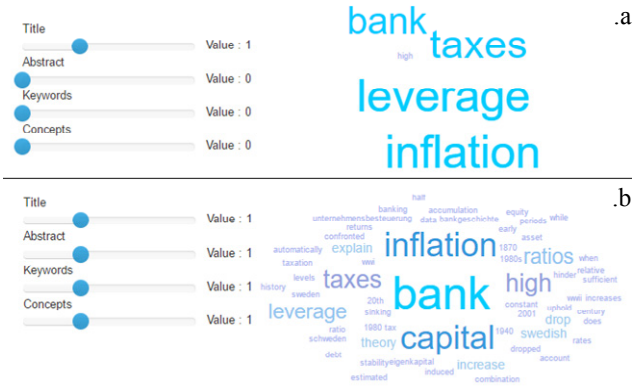


Table 2 The top-ten most important terms of a publication metadata based on a specific adjustment among them

Rank	*Word*	*tf*	*Weight*
1	Bank	10	0.065918249938192
2	Inflation	7	0.047178390392642
3	Capital	7	0.047173543186314
4	Corporate	6	0.040739099875212
5	Taxes	5	0.034212141508546
6	High	7	0.034194006135951
7	Leverage	4	0.027583331996128
8	Ratios	4	0.027583262848340
9	Swedish	1	0.014010412785021
10	Explain	1	0.014010119258181

5.2 Measuring the similarity of publications

The similarity of publications, i.e. vectors of concepts, is measured as the deviation of angles between each document vector, by using the Cosine Similarity. Thus, iteratively we measure the similarity between metadata of our initial publication with the metadata of publications from the target repository, $sim(p,d_i)$, for $i=1,n$. As shown in Figure 3, the combination among the metadata is crucial for determining the weight of the terms in the initial publication. The proper selection can be seen as the right bait for a successful 'fishing'. Different combinations among these parameters would result in different lists of retrieved publications from the targeted repository. The impact can also be seen in the generated results.

Concerning this, in our previous work, we have achieved significant results by enriching author profiles with additional information from different digital libraries (Hajra et al., 2015). In another study (Hajra et al., 2014), considering different cases, different combinations of these metadata also led to good results. For this purpose, we conducted heuristic evaluations

when analysing the impact of each element. In the absence of any golden rule, as the most determinant combination we have perceived the combination of all of them by doubling the title ($2p_t$, p_{abs}, p_k, p_d). The title is most representative, as author tends to include the key terms regarding the subject.

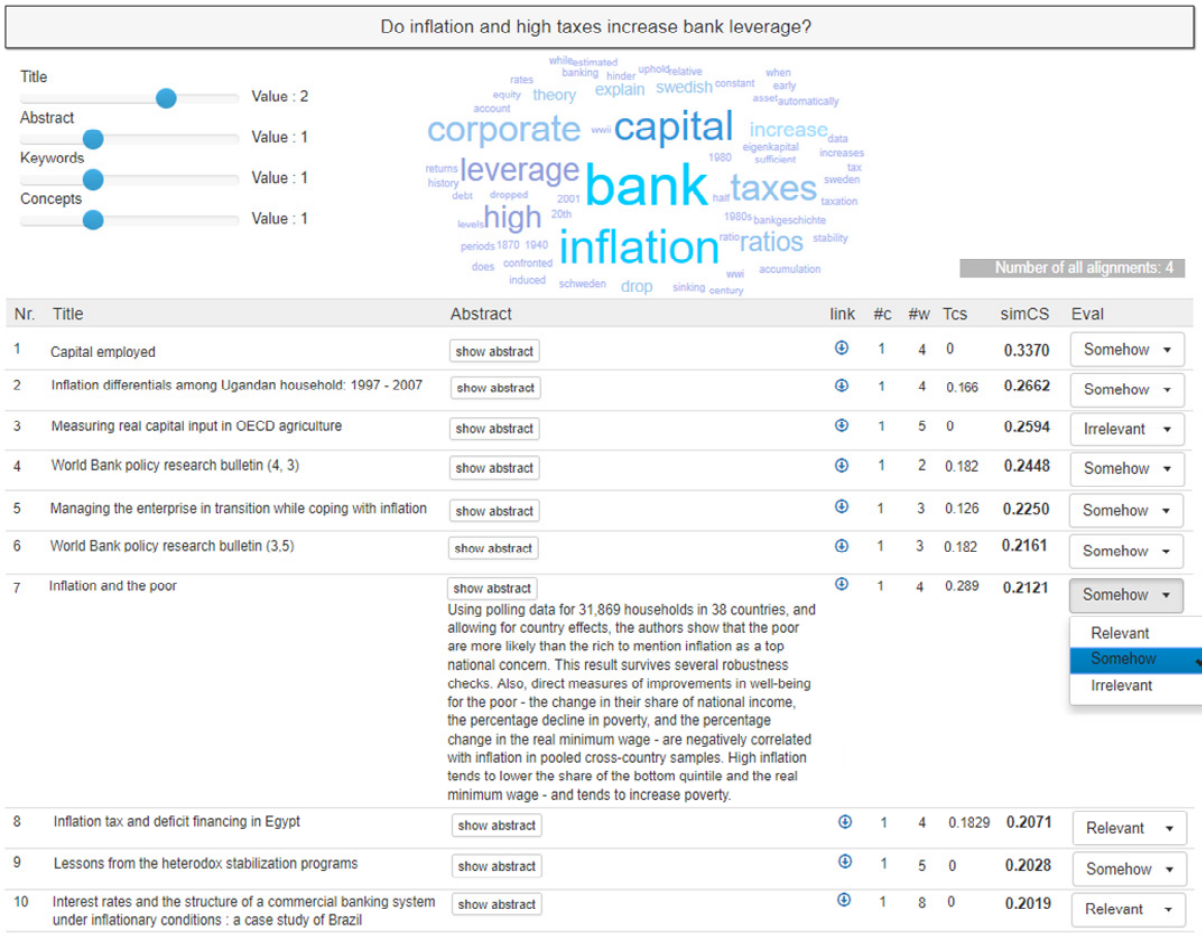
5.3 *Evaluation of VSM approach*

We have evaluated 57 EconStor publications, with the developed prototype. After triggering a title from EconStor, the system retrieves an ordered list of most similar publications from other repositories, in this case, OpenAgris. As can be seen in Figure 4, the prototype generated values for several parameters: *#c* represents the number of common descriptors in both sites. *#w* is the number of common words in these publications. *Tcs* represents the Cosine Similarity measured only on titles, while *simCS* the Cosine Similarity measured with all the defined components, i.e. *sim*[($2p_t$, p_{abs}, p_k, p_d), ($2d_t$, d_{abs}, d_d)]. From the generated results in Figure 4, the prototype shows that the first retrieved publication has 0.3370 Cosine Similarity with our publication, while the similarity between the titles is zero since there are no common words. The value of eight in the parameter *#c* represents the intersection of common words in the two publications. The prototype indicates that the average number of tokens from the initial publication is about 72, while it is 79 for the target publication.

However, the frequency of tokens from the metadata is crucial for scoring results. The word '*inflation*' appears eight times in our paper and 20 times in the first ranked target paper. Conversely, the number four has only two words in common, i.e. the words '*inflation*' and '*bank*', and a small number of other noisy words. The number of the equivalent descriptors used for describing a paper in both repositories generally is one; except in a few cases, the publication is described by two or more descriptors in both repositories.

In order to determine the relevance of the retrieved target publications, human evaluations are done for the top-ten ranked results. These evaluations are done by analysing and comparing the titles (Resnick, 1961) and continuing with the abstract using the possibility for full-text reading. For each publication in the top-ten, a value *i* is assigned for *irrelevant*, *s* for *somehow relevant* or *r* as *relevant*. Considering the example in Figure 4, the publication number three is evaluated as irrelevant and seven others as somehow relevant, while only two of them are depicted identified as closely related to the initial publication. However, in another example, '*Food prices and political instability*' in the top-ten, five are identified as irrelevant, four somehow and only one as relevant.

Figure 4 The combination of metadata components from a scientific paper for retrieving recommended publications from other repositories



The precision (i.e. the list of the relevant document) is improved by factorising the title in the scoring and ordering. Therefore, a simple experimentation, by performing the ranking as the average of *title* and *all other metadata*, i.e. $avg(Tcs, SimCS)$, shows significant improvement regarding the number of relevant publications in the top-ten. However, this has negative implications for the relevant publications with a not meaningful title. In that case, several relevant publications will not be ranked highly. The tenth ranked publication from Figure 4 that is evaluated as relevant will not be in the top-ten because of the zero value in *Tcs*. As a result, the role of an abstract and other metadata components such as keywords or descriptions is crucial, when the title is not subject representative (of the type '*What next?*' or '*Lessons learned*'). When the title does not contain significant common terms with the triggered publican's metadata, ex. '*Capital employed*' in Figure 4, VSM fails to calculate any similarity.

The count-based approach with TF-IDF and Cosine Similarity generates satisfactory results for retrieving relevant publications from other repositories if a satisfactory amount of metadata is provided. This is particularly true if the intersection between the compared documents results in common words (Hajra et al., 2014). Despite that, we have identified several weaknesses showed with this approach.

5.4 Limitations of VSM

The main issue with this approach is that it is strictly related to the intersection of common words among compared documents. Such that, a simple morphological variation between words delivers the result. The attempt for achieving uniform words, i.e. converting to singular, or by applying stemming or lemmatisation, shows improvements. However, we need to be very careful with this process, since the evaluations show that in several cases the stemming or lemmatisation can be so 'aggressive' by changing a word significantly. In Figure 4, our title with the title of publication number ten initially generates zero similarity, after stemming the words '*bank*' will be matched, since '*inflationary*' stemmed as '*inflationari*'.

The semantic interconnection between words or the context of use is not taken into account, as we cannot find any similarity between words such as '*bank*' and '*credit*'. This implies that a large number of relevant publications might not be among the top. The application of external vocabularies such as WordNet, for the availability of synonyms about the given words, complicates the process further. The variety of synonyms for a single word broadens the result by making it too far from the initial publication.

This approach repeatedly shows those irrelevant terms to be highly ranked. For example, take the publication titled '*Food prices and political instability*', based on the combination ($2p_t$, p_{abs}, p_k, p_d), the word '*food*' becomes dominant. This results in compromising outcomes, i.e. recommending semantically distant publications to that publication. In this case, as the first ranked publication we retrieve '*Food Security in Older Australians from Different Cultural Backgrounds*'. Therefore, the right combination of metadata terms for this purpose is very experimental.

Another aspect worth mentioning is that this approach shows poor results when measuring the similarity of vectors with only a few terms in them, such as the similarity between titles of publications. As one of many examples that show the weakness of these approaches when relying on short texts is the similarity between these two titles '*Do inflation and high taxes increase bank leverage?*' and '*Lessons from heterodox stabilization programs*', which results with zero.

Based on this, in general, TF-IDF and CS do not offer much for achieving a completely automated process for measuring the semantic relativeness among the initial and retrieved publications (Baroni et al., 2014). Based on these insights, we have explored several other approaches for finding an optimal solution that includes the semantic component for similarity measurement and ranking. The Latent Semantic Analysis (LSA) (Deerwester et al., 1990) or Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is an option in this direction. However, based on the evaluations in several studies, these approaches do not offer the best solution for our cases (Baroni et al., 2014; Pennington et al., 2014; Kim et al., 2016). According to this, we are focused on the neural word embedding as one of the most promising approaches in the NLP.

6 Deep learning approach

Determining the semantic similarity between two texts is a complex and challenging process. In general, there are several approaches introduced based on lexical matching, handcrafted patterns, term-weighting and syntactic parse trees (Robertson and Zaragoza, 2010; Kenter and de Rijke, 2015). Indeed, lexical features, like string matching and frequency of words in a text, do not capture semantic similarity on a satisfactory level (Baroni et al., 2014; Kenter and de Rijke, 2015). Hence, the deep learning approach for language processing based on neural network language models outperforms traditional count-based distributing models on word similarity (Levy et al., 2015). 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 (Bengio et al., 2006; Collobert and Weston, 2008; Mnih and Hinton, 2009; Turian et al., 2010; Mikolov et al., 2013; Baroni et al., 2014; Pennington et al., 2014; Kenter and de Rijke, 2015; Kusner et al., 2015; Lebret and Collobert, 2015; Levy et al., 2015).

6.1 Word embeddings

In deep learning, word embeddings currently represent the most outstanding field. It is the most discussed subject in almost every publication regarding the semantic representation of words in a low-dimensional vector (Bengio et al., 2006; Collobert and Weston, 2008; Mnih and Hinton, 2009; Turian

et al., 2010; Mikolov et al., 2013; Baroni et al., 2014; Pennington et al., 2014; Kenter and de Rijke, 2015; Kusner et al., 2015; Lebret and Collobert, 2015; Levy et al., 2015). 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 network models. Thus, words are embedded in a continuous vector space where semantically similar words are mapped to close vectors. Learning the word embeddings is totally unsupervised and computed on a predefined text corpus.

Out of the word embedding techniques two are the most prominent: the Word2Vec algorithms proposed by Mikolov et al. (2013) for Google and GloVe model from Pennington et al. (2014) at Stanford. Our experiments and evaluations are based on Word2Vec due to the performance and computational cost.

6.1.1 Word2Vec embeddings

Word2Vec is a novel word embeddings approach, which learns a vector representation for each word using neural network language model (Mikolov et al., 2013). 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 Mikolov et al. (2013), Baroni et al. (2014) and Kusner et al. (2015), 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 data sets. As such, the model can be trained on conventional personal machines with billions of words, achieving the ability to learn complex word relationships (Mikolov et al., 2013; Kusner et al., 2015).

Currently, there are several implementations of Word2Vec in different frameworks. The native proposed code is optimised in the C programming language. However, Deeplearning4j implements a distributed form of Word2Vec for Java and Scala, while Gensim and TensorFlow offer a Python implementation of Word2Vec.

6.2 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 significant optimisation regarding the computational speed, which overpasses even the native C implementation. Currently, there are several pretrained models on different data sets, 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 on 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 pre-processing steps are performed at the corpus, such as conversion to utf8 Unicode, lowercasing, removing numbers and punctuations. Finally, the model is trained on the corpus of 12,329,307 raw words and 683,937 sentences. Before training the process, several hyper-parameters are determined that affect the training speed and performance. Based on our data set size, every word in the corpus is considered with a window value of five. The dimensionality space of the words inside a vector is set to 300, which means that each word is represented with 300 most similar words in that vector. More words in a vector may increase the quality of the model, although bigger data set must be used.

The hierarchical Skip-gram architecture is used for training the model in a laptop with i5 CPU 1.7 GHz, 8 GB RAM memory. Surprisingly, the time it took was very fast, 129.7 sec, far beyond our expectations.

6.3 Analysing the model

This section presents the investigation of the learned model. We performed several analyses on top of the trained model in Section 6. One of the most interesting analyses regarding the word representation approach is about finding the set of related 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 concept '*inflation*', as a purely economic concept, and the concept '*food*', as a general concept. Table 3 lists the ten nearest terms that Word2Vec has calculated for these words.

The generated results are very impressive. For example, the word '*output*', '*nominal*' and '*volatility*' are ranked as the most similar to '*inflation*' with a degree of similarity 0.644, 0.611 and 0.604 out of 1. In fact, that value is more accurately to be denoted as the degree of relatedness among these concepts, rather than the similarity (Faruqui et al., 2016). In general, all the listed words are intuitively very close to it. Moreover, a word is represented in relatedness to 300 words, as defined by 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 thesauri described in Section 4, the concept '*inflation*' 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.

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 or 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 0.764 similarity, '*liquidity*' with 0.734 and '*spreads*' with 0.695. Such implementation can be productive in the first steps of the selection of terms from the metadata set for populating the vector $\vec{V}(p)$. Especially if there are, present a human interaction in the selection process. However, in our evaluations we do not practise such an approach. Thus, the choice of concepts is completely automatic, as described in Section 5.1.

Table 3 Top-ten most similar words based on the words '*inflation*' and '*food*', generated through Word2Vec from our text corpus

(a) For the word 'inflation'		*(b) For the word 'food'*	
Word	*Similarity*	*Word*	*Similarity*
Output	0.644	Energy	0.789
Nominal	0.611	Agricultural	0.786
Volatility	0.604	Water	0.767
gdp	0.590	Land	0.756
Aggregate	0.570	Crop	0.701
Persistence	0.561	Fuel	0.694
Macroeconomic	0.543	Transport	0.694
Price	0.535	Agriculture	0.691
Inflationary	0.532	Electricity	0.690
Forecast	0.531	Milk	0.684

6.4 Evaluation of word embedding approach

Based on the developed prototype, we have evaluated exactly the same 57 EconStor publications, used in Section 5.3. For each selected publication, the prototype retrieves and ranks the most semantically similar publications from OpenAgris. The process is the same as in Section 5.3; however, as can be noted from Figure 5, in this approach we have introduced two more measurement components: *Tw2v* which denotes the Word2Vec similarity among titles, and *simW2V* that is the Word2Vec similarity among all the publication's metadata, i.e. $sim[(2p_t, p_{abs}, p_k, p_d), (2d_t, d_{abs}, d_d)]$. The ordering is performed according to *simW2V* scoring.

As expected, the implementation of word embeddings approach shows a different list of retrieved publications, compared to Cosine Similarity (CS) in Figure 4. The results from Figure 5 make it obvious that the values generated through Word2Vec overcome those generated by CS. Figure 5 represents one of the depicted results from the evaluated publication, which is the same as in Figure 4, '*Do inflation and high taxes increase bank leverage?*' The results are shown in both approaches with two different sets of metadata.

Firstly, the similarity degree between publication p and d_i is calculated only by using titles, such as $sim(p_t, d_{ti})$. As such, for the first retrieved publication on that list Word2Vec has generated a similarity of 0.5680, shown in *Tw2v* column. The count-based implementation of CS gives 0 score between the same titles, shown in *Tcs*. This is one of many examples that prove the ability of the word embedding approach to work even with small amounts of metadata.

In the same example, analyses are extended by including other metadata terms in the similarity calculations. Hence, from the EconStor publications the title(p_t), abstract(p_{abs}), keywords(p_k) and descriptors(p_d) are considered, while from the OpenAgris publications the title(d_t), abstract(d_{abs}) and descriptors(d_d). The last two columns of Figure 5 compare the similarity between these metadata in both approaches: *simCS* and *simW2V*. By considering the first publication from Figure 5, TF-IDF with CS generates 0.2019 similarity degree among them, while Word2Vec gives 0.8733. The differences of generated results in both approaches are very obvious. The fact that word embeddings reach to rank top publications which was not possible with the previous approach is most important. Therefore, the third ranked publication through Word2Vec, manually judged as relevant (see Figure 5), does not appear in top-ten retrieved publications in the first approach (see Figure 4).

Figure 5 The similarity measurement is scored with Cosine Similarity and Word2Vec. The results are ordered based on Word2Vec similarity score. The relevance of the retrieved publications is evaluated manually

Nr.	Title	Abstract	link	#c	#w	Tcs	Tw2v	simCS	simW2V	Eval
1	Interest rates and the structure of a commercial banking system under inflationary conditions : a case study of Brazil	show abstract		1	8	0	0.568	0.2019	**0.8733**	Relevant
2	Inflation tax and deficit financing in Egypt	show abstract		1	4	0.1829	0.7481	0.2071	**0.8644**	Relevant
3	Inflation and the rule-of-thumb method of adjusting the discount rate for income taxes	show abstract		3	8	0.1873	0.7125	0.1978	**0.8565**	Relevant
4	Capital rising in the Baltic States: lessons learned and future prospects	show abstract		1	8	0	0.5081	0.1561	**0.8523**	Somehow
5	Effects of tax incentives on long-run capital formation and total factor productivity growth in the Canadian sawmilling industry	show abstract		1	6	0	0.6921	0.165	**0.8478**	Somehow
6	Future capital requirements should be studied	show abstract		1	6	0	0.6133	0.1589	**0.8437**	Irrelevant
7	Returns, interest rates, and inflation: how they explain changes in farmland values	show abstract		1	5	0.1179	0.7174	0.1679	**0.8335**	Somehow
8	Macroeconomic factors influencing lending rates	show abstract		1	6	0	0.5844	0.1268	**0.8326**	Somehow
9	Dollarization and exchange rate fluctuations	show abstract		1	6	0	0.6729	0.1506	**0.8299**	Somehow
10	Liberalising foreign investments by pension funds: positive and normative aspects	show abstract		1	9	0	0.5768	0.0571	**0.8226**	Irrelevant

The fact that word embedding overcomes CS, regarding the score value, cannot be adopted with automatism as the ultimate approach. Since the scores are used for ranking purposes, we have extended the human evaluation in both approaches, comparatively. Accordingly, same as in the first approach, the top-ten retrieved publications are manually analysed in order to determine the semantic relevance with the initial publication.

6.5 Limitations of word embeddings approach

In the case when the word embedding model is trained on the corpus of one data set, then the vocabulary of that corpus is embedded in word arrays. The usage of the model for measuring semantic similarity between two texts from different data sets is facing in a large set of 'unknown' words. In our case the model is trained from the EconStor data, thus the Word2Vec has detected several missing words from the OpenAgris when similarity measurement is calculated. We have ignored all the words that are not part of the trained model. However, this has implications in the generated results, i.e. the result to be generated on a few terms that cannot be representative for the publication from the non-trained corpus.

Using a model trained on a wider range of covered vocabulary, such as Google News, decreases the number of missing words. However, the application of this model does not make evident any improvements regarding the relevance of the top retrieved publications. Building a model on top of the experimented data sets, the initial and the targeted repository is resulting in different distributions of semantically related words in arrays. Therefore, considering the combination of EconStor and OpenAgris for building the model, Word2Vec gives more general context to a particular word, instead of closely related economic correlations. Thus, in this situation the most semantically similar words to '*food*' are listed, *seafood* 0.71, *foodstuff* 0.69, *grocery* 0.66, *restaurant* 0.651, *consumer* 0.642, *menu* 0620, etc. As shown, there is a huge difference compared to the same word in Table 3b. By applying this kind of model, we are facing a decreased performance in the task to determine the semantic similarity between two publications, according to human judgements. The embeddings trained on specific domain corpora generate better results compared to a more general model such as Wikipedia or Google News, for specific related tasks (Ye et al., 2016). In different scenarios, the combination of local and global context corpora in the learning process is fruitful for a more general word representation (Huang et al., 2012).

Word embedding is an unsupervised process, such that the selected data set for training the model is crucial for the quality of the model. Therefore, the absence of terms in the training phase, word frequency and neighbourhoods can be determining factors. Even the predefined hyper-parameters like the dimensions of the distributed words on arrays, the window size, negative samples or the minimum count can play a crucial role. Based on the performed experiment, we conclude that word embedding is very sensitive to these tuning parameters. Similar conclusions, regarding the tuned parameters, are noted in related work, i.e. Huang et al. (2012), Levy et al. (2015), Schnabel et al. (2015), Faruqui et al. (2016), Ye et al. (2016), Zamani and Croft (2016).

Recent trends are placing the emphasis on the combination of word embedding with other traditional approaches (ex. LSA, BM25, TF-IDF), or with other different word representation methods (ex. Mikolov, Glove) (Niraula et al., 2015; Kim et al., 2016). As claimed, such a combination is resulting in a better performance regarding the measurement of semantic words relatedness or even semantic texts similarity. In Kim et al. (2016) a combination of word embeddings with BM25 or Word Mover's Distance (WMD) (Kusner et al., 2015) for measuring a similarity between a query and a document is proposed. While De Boom et al. (2016) show an attempt to combine word embeddings with TF-IDF information. The use of weighted centroids of word embeddings with WMD to re-rank the retrieved documents is evident in Brokos et al. (2016).

7 Results and discussions

This study presents and evaluates several approaches dealing with the enrichment of scientific publications of a DL with other relevant information from other repositories. As the most important input parameter for our solutions, we take the closely related publications indexed in these data sets. Therefore, the main challenge is the determination of semantic relatedness between the initial and retrieved target publications. Starting from the aligned concepts between the LOD repositories, we extended our research with two additional approaches for measuring and determining that semantic relatedness.

As emphasised in Section 5, 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 frequency of a more general concept in these metadata had a negative impact on the result. For example, regarding the publication titled '*Food Prices and Political Instability*', the word '*food*' has been determinant in the similarity measurements. Thus, the retrieved publications have been related to '*agriculture*', '*food security*' or '*health*' rather than '*food prices*' or '*politics*', which semantically are not close to the initial publication. Different adjustment among the metadata components results in improvements of the retrieved results. However, this applies heuristic involvements in the evaluation of results. Moreover, the count-based approach shows significant weakness in recognising relationships among terms, even in the cases when the presence of thesauri is evident. Therefore, its performance is strictly related to the presence of the same words among the compared texts. In order to overcome such limitations, we have investigated the word embedding, as most comprehensive and promising. The evaluations are done comparatively, in both approaches at the same time, on concrete data sets. The generated results of the top-ten retrieved publications are judged by humans regarding their relevance to the triggered publication.

Nowadays, there are several publications mainly addressing the evaluation of word relatedness, i.e. semantic similarity among words, based on word embedding approach. Almost all of these evaluations take place in already human annotated data sets such as WordSim353 (Finkelstein et al., 2001) or SimLex-999 data set (Hill et al., 2016). Another set of publications are focusing on IR, by evaluating the binomial query-retrieved documents, or question–answer. Even in these cases, there are several humanly annotated data sets, such as TREC (Hersh et al., 2007) or PubMed (Lin and Wilbur, 2007), with already predefined thresholds. However, even our case represents a common IR task; we find it more appropriate for evaluating the proposed approaches on tangible cross-domain repositories.

The main task in our case relies on the semantic relatedness among documents, i.e. publications from different domain repositories. Therefore, there is an obvious difference in how the retrieval is initiated. We are starting by considering all the metadata of a publication, rather than a user-entered query. When a user makes a query, it consists of carefully chosen appropriate terms, without 'noisy' terms in it. While in the publications' metadata, the importance of metadata components, i.e. title, abstract, keywords, should be determined additionally. Except that, the weight of the terms inside these components plays a crucial role. Thus, different combinations among these metadata lead to different results. This is one of the reasons for performing our evaluations on these types of data sets.

7.1 The results

As mentioned before, in total 57 publications have been evaluated. The process is described in detail in Section 5.3, regarding Vector Space Model, and Section 6.4 concerning Word Embeddings approach. Figures 4 and 5 also provide more details. For each of these 57 EconStor publications, the prototype has retrieved 300 publications from the target repository, i.e. OpenAgris. Iteratively we have evaluated the top-ten retrieved publications, ordered on both approaches, with two different sets of metadata (all the metadata versus titles). Thus, for each of these EconStor publications p_i, a set of publication D_i is retrieved, where $D = \{d_1, d_2, d_3, \ldots, d_{300}\}$ is a subset of OpenAgris repository.

Table 4 depicts an example of two such evaluations. By default as a reference, the ordering is done based on CS score, denoted as *top10CS*. After that, for each EconStor publication, i.e. *publication1*, *publication2*, the retrieved results are ordered by word embeddings, similarity score, denoted as *topW2V*. Therefore, the relevance of the retrieved publications is assessed and labelled manually with *i*, *s* and *r*. For clarifying this, let us take a closer look at Table 4. Considering the EconStor *publication 1*, the first retrieved result based on CS is evaluated as irrelevant (*i*), while the first ranked result based on Word2Vec is judged as relevant (*r*). Thus, at the end of each column, the evaluation results of both approaches are shown cumulatively, concerning the relevance. The generated results make evident the discrepancies between the applied approaches.

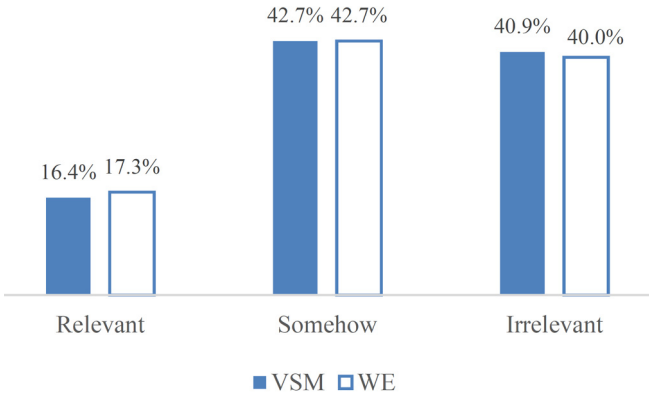
Table 4 An example of the top-ten retrieved and evaluated publications for two EconStor publications, ordered in both approaches by all the metadata and titles only

	publication1 (p_1)								*publication2 (p_2)*							
	All metadata				*Only titles*				*All metadata*				*Only titles*			
rank	*top10CS*	*relCS*	*topW2V*	*relW2V*	*titleCS*	*relTcs*	*titleW2V*	*relTw2v*	*top10CS*	*relCS*	*topW2V*	*relW2V*	*titleCS*	*relTcs*	*titleW2V*	*relTw2v*
1	d_1	*i*	d_5	*r*	d_8	*r*	d_{22}	*s*	d_1	*r*	d_1	*r*	d_8	*s*	d_{12}	*s*
2	d_2	*r*	d_{59}	*r*	d_1	*i*	d_2	*r*	d_2	*s*	d_{33}	*r*	d_7	*s*	d_8	*s*
3	d_3	*r*	d_{28}	*i*	d_6	*i*	d_3	*r*	d_3	*r*	d_{13}	*r*	d_{23}	*i*	d_5	*s*
4	d_4	*i*	d_{57}	*s*	d_4	*i*	d_7	*s*	d_4	*s*	d_{41}	*s*	d_{12}	*s*	d_{23}	*i*
5	d_5	*r*	d_{39}	*s*	d_5	*r*	d_6	*i*	d_5	*s*	d_{27}	*s*	d_4	*s*	d_{19}	*i*
6	d_6	*i*	d_{60}	*i*	d_2	*r*	d_{59}	*r*	d_6	*r*	d_3	*r*	d_{14}	*s*	d_1	*r*
7	d_7	*s*	d_{42}	*i*	d_{13}	*i*	d_{14}	*r*	d_7	*s*	d_5	*s*	d_{10}	*s*	d_{28}	*i*
8	d_8	*r*	d_{34}	*s*	d_{23}	*i*	d_{42}	*i*	d_8	*s*	d_{20}	*r*	d_1	*r*	d_7	*s*
9	d_9	*i*	d_{66}	*i*	d_{46}	*i*	d_{39}	*s*	d_9	*s*	d_{36}	*s*	d_6	*r*	d_{22}	*s*
10	d_{10}	*i*	d_3	*r*	d_3	*r*	d_{60}	*i*	d_{10}	*s*	d_{15}	*r*	d_{17}	*s*	d_{27}	*s*
r		4		3		4		4		3		6		2		1
s		1		3		0		3		7		4		7		6
i		5		4		6		3		0		0		1		3

Thus, Word2Vec ranks referring again to *publication1* (p_1) in Table 4, the 59th retrieved publication according to CS (d_{59}), as fifth (d_5). At the same time, there are several cases in which Word2Vec has re-ranked a top-ten publication, which has been ordered below 100 by CS.

Regarding the top-ten retrieved publications, based on all metadata, the word embedding approach yields 70.9% completely different documents in the top-ten, compared to the Vector Space Model. Thus, only 29.1% of the same retrieved publications appear in the top-ten, by both approaches. These cases are shown in Table 4 (with highlighted background). For the entire set of evaluations, with all metadata, the Vector Space Model, i.e. TF-IDF with CS, yields 16.4% relevant publications in the top-ten, 42.7% somehow relevant and 40.9% irrelevant. In the same set, word embeddings, i.e. Word2Vec, yields 17.3% relevant publications, 42.7% somehow relevant and 40% irrelevant. A better graphical representation of these data is depicted in Figure 6.

Figure 6 Humanly evaluation of the top-ten retrieved publications based on the Vector Space model and the Word Embedding approach

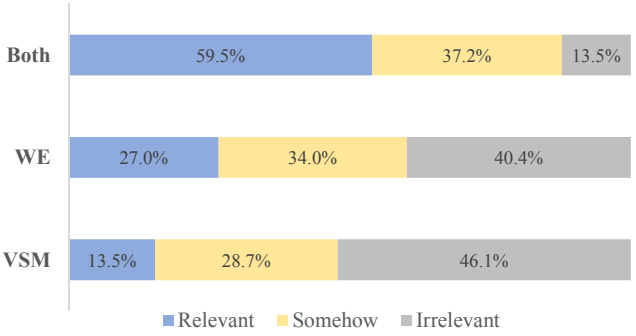


At a first glance, it seems that the two approaches have is a minor difference in the generated results, according to the relevance of the top-ten retrieved publications. However, a more detailed analysis shows quite interesting occurrences. Moreover, concerning only the list of relevant publications inside the top-ten, Word2Vec catches 27% of all relevant publications, while CS with TF-IDF catches 13.5%. Thus, both of them show 59.5% of the same relevant publications in the top-ten. Regarding the irrelevant documents, WE gives 40.4% versus 46.1% of VSM, in that list. Figure 7 highlights more details about these proportions.

We also note that WE is able to generate better results, as far as relevance is concerned. It also reaches to 'seize' publications that even have little or no similar concepts among themselves. This is because of WE's ability to present correlations between words.

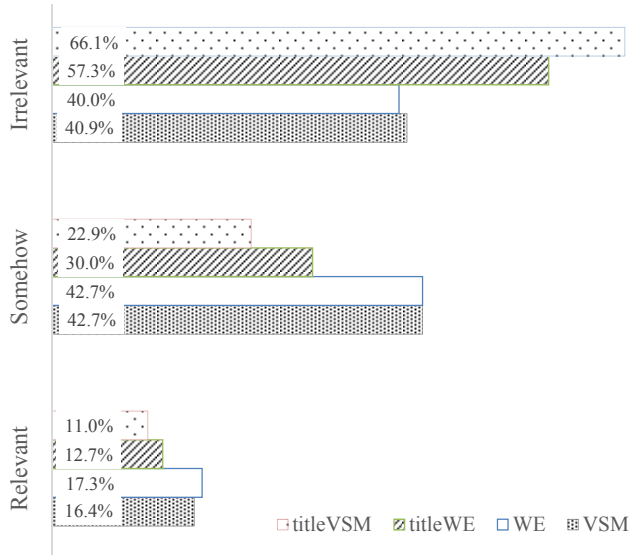
The number of irrelevant results is in the frame of expectations, taking into account the different domains between the repositories where the evaluations take place. In the case when the selected publications are purely economic, such as '*Taxes, wages and working hours*', both approaches give zero relevant recommendations, and four somehow relevant. Conversely, for inter-domain publications such as '*Politics, globalization, and food crisis discourse*', or '*Public policies against global warming*' the system manages to retrieve four very relevant publications. The other reason is related to the limited number of records for each search at target repository. For evaluation purposes, the prototype processes only 300 publications, for every EconStor paper at that repository, i.e. OpenAgris. Increasing that number means increasing the possibility for more relevant publications, but at the same time increasing the cost of processing.

Figure 7 The relevance of the retrieved result based on Vector Space Model and Word Embedding approaches, separately by including the common results



In addition to the given metadata, word embedding achieves a good performance in smaller texts also (Kenter and de Rijke, 2015; Galke et al., 2017). In parallel, we have analysed and evaluated the relevance of the retrieved documents if only the similarity between titles is used as ordering score. For example, between the titles '*Do inflation and high taxes increase bank leverage?*' And '*Are government regulations pushing food prices higher?*' The Word2Vec has scored 0.7223 similarity degree, compared to zero of the CS score.

Figure 8 The relevance of the retrieved result based on VSM and WE approach, generated in all metadata versus titles



The results presented in Figure 8 point out the slight domination of WE in terms of performance only in titles. Therefore, WE achieved to retrieve 12.7% relevant publication versus 11% of VSM. In addition, VSM retrieves 8.8% more irrelevant documents than WE.

The score generated as the combination of all the metadata, i.e. $sim[(2p_t, p_{abs}, p_k, p_d), (2d_t, d_{abs}, d_d)]$ achieves to catch 17.3% more relevant or somehow relevant publications rather than the score calculated on titles, $sim(p_t, d_t)$, referring to the Word Embeddings approach. Furthermore, only 28.2% of publications in the top-ten are the same in both ordering scores.

7.2 Discounted cumulative gain (DCG) metrics

A formal way of presenting the results is done by applying the Discounted Cumulative Gain (DCG) measure, as the most notable metric for quantifying the performance of ranking high relevant documents (Järvelin and Kekäläinen, 2000). The formulation of DCG is defined as below, where the main inputs are the relevance value of the retrieved documents (rel_i) with the corresponding ranked positions (i). While, n represents the number of evaluated documents, which in our case is continually 10.

$$DCG_n = \sum_{i=1}^{n} \frac{rel_i}{log_2(i+1)} = rel_i + \sum_{i=2}^{n} \frac{rel_i}{log_2(i+1)}$$

The application of DCG to our evaluated data requires a translation of the relevance values from literals to numbers. Such that, *r* that stands for relevant is denoted with 2, *s* of somehow as 1, and *i* for irrelevant as 0. Thus, in total there are three relevance values, $rel_i \in \{0,1,2\}$. Table 5 embodies exactly the *publication 1* from Table 4, after including these translations. As can be noted in Table 5, the DCG score is calculated for both approaches, i.e. CS and W2V on all metadata and titles comparatively. Therefore, four ranking strategies are shown. The end of each column shows the sum of these values as stated in the formulation. Therefore, considering the same example, the DCG_{10} score for Cosine Similarity on all metadata is 4.0 while the DCG_{10} score of Word2Vec on the same metadata is 4.973.

However, the DCG score is not the best solution for measuring the performance of several approaches with a different set of metadata, regarding the ranking of relevant documents (Järvelin and Kekäläinen, 2002; Wang et al., 2013). For that purpose, several other modifications of DCG can be applied in different contexts. In our case, since we are operating with the fixed number of evaluated documents over all approaches, i.e. ten, the normalised discounted cumulative gain (nDCG) is applied. For this purpose, the normalisation of the results, based on the relevance order, is performed. For each of the columns in Table 5 (*relCS*, *relW2V*, *relTcs*, *relTw2v*), the DCG is recalculated after sorting the retrieved documents in decreasing order of relevance. For example, *relCS* now will be ordered such as $(2_1 2_2 2_3 2_4 1_5 0_6 0_7 0_8 0_9 0_{10})$. The DCG value calculated in this way is known as the Ideal DCG (IDCG). Hence, *relCS* have $IDCG_{10}$ of 5.510. The normalised discounted cumulative gain (nDCG) represents the fraction of DCG with ideal DCG. In this case, for the *relCS* example in Table 5, we have $nDCG_{10} = 4.0/5.51 = 0.726$.

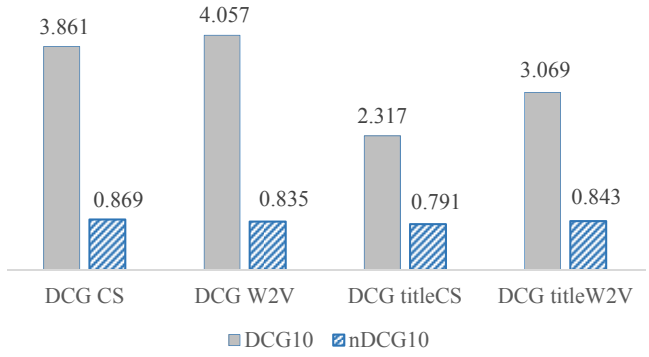
Table 5 An example of generating the DCG_{10} score on the top-ten retrieved publications for one EconStor publication

publication1 (p_1)								
	All metadata		*Only titles*		*All metadata*		*Only titles*	
position (i)	*relCS*	*relW2V*	*relTcs*	*relTw2v*	*DCG CS*	*DCG W2V*	*DCG titleCS*	*DCG titleW2V*
1	0	2	2	1	0.000	2.000	2.000	1.000
2	2	2	0	2	1.262	1.262	0.000	1.262
3	2	0	0	2	1.000	0.000	0.000	1.000
4	0	1	0	1	0.000	0.431	0.000	0.431
5	2	1	2	0	0.774	0.387	0.774	0.000
6	0	0	2	2	0.000	0.000	0.712	0.712
7	1	0	0	2	0.333	0.000	0.000	0.667
8	2	1	0	0	0.631	0.315	0.000	0.000
9	0	0	0	1	0.000	0.000	0.000	0.301
10	0	2	2	0	0.000	0.578	0.578	0.000
			DCG_{10}		4.000	4.973	4.064	5.373
			$IDCG_{10}$		5.510	5.436	5.123	6.200
			$nDCG_{10}$		0.726	0.828	0.793	0.867

The interpretation of scores can lead us to a better understanding of the performance of the proposed approaches. The computed DCG_{10} and $nDCG_{10}$ scores over all 57 evaluated EconStor publications are visualised in Figure 9. Therefore, from the same figure we can conclude that in both metadata sets DCG value shows a better performance of W2V compared to CS. When all the metadata are considered, the DCG_{10} of W2V is 4.057 while CS is 3.861. This insight shows that W2V achieved to catch in top-ten more documents that are relevant compared to CS. The discrepancy is even more notable when only titles are considered, i.e. 3.069 versus 2.317 in favour of W2V.

However, an interesting sighting shows the analysis of $nDCG_{10}$ score. The value of 0.869 at CS comparing to 0.835 at W2V let to know that CS manages to perform a better ranking of the relevant documents. Thus, although W2V attains to catch more relevant or somehow relevant documents in the top-ten, CS performs a better ranking. Nonetheless, this is not the case when the comparison is done on titles only, where the dominance of W2V is evident in every aspect, emphasising, even more, its outperforming capability in short texts.

Figure 9 The average Discounted Cumulative Gain (DCG) and Normalised DCG (nDCG) score for VSM and WE approach, generated on two different metadata sets



8 Summary

The main objective of our research was to emphasise the advantages resulting from an improved interoperability among different Digital Libraries and to investigate different algorithms to achieve this interoperability. By cross-linking data from different repositories, a given resource can be enriched with additional information in the form of similar publications. This results in a significant enhancement of scholarly communication in general, regarding time consumption and quality of the required information. The idea is to perform a single query in a single repository (e.g. the favourite DL) and to offer scholars information from different repositories, based upon this single query. Ultimately, a selected publication from one DL can be enriched with a list of recommended publications from other DLs, additional information about authors, conferences, etc.

In order to achieve this, we needed to find this information and to determine its relevance, i.e. semantic similarity between two different resources. For this purpose, bibliographic LOD repositories are considered to investigate the alignments among them. The evaluated results show that the list of retrieved publications according to each aligned concept between repositories was extremely heterogeneous. While the attempt to find publications in the target repository, with the same set of descriptors as in the initial one, did not return any further publications. Therefore, we use alignments between repositories for retrieving an initial set of publications, especially as an important component for determining the weight of the terms in the metadata set.

The semantic relatedness of the retrieved publications with the triggered publication is measured by applying two main approaches comparatively. The generated results show that the traditional count-based and text-matching approach through TF-IDF and CS are satisfactory. However, it relies on heuristics to determine a higher level of semantic similarity among publications. Its performance is closely related to the common words among the compared publications. The disability for determining the words relatedness appears to be the main weakness, even in the cases when the presence of thesauri is evident.

Given this, we applied the deep learning approach to model semantic word representations. The implementation of contemporary Word2Vec implementation is an important 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 the concept appearing in the title concepts only. Substantial improvements are evident by extending the set of metadata with concepts from the abstract and keywords. The results show that the implementation of the word embedding approach managed to retrieve those top-ranked relevant recommended publications, which the previous approached has ranked far below the top positions. Therefore, 27% of all relevant publications are caught by Word2Vec only, while 13.5% by CS with TF-IDF. Thus, they are performing with 40.5% difference concerning the outcome of relevant retrieved publications. A proper interlacement between these approaches leads to further promising improvements.

In addition, the results are presented by applying the Discounted Cumulative Gain (DCG) measure and Normalised Discounted Gain (nDCG). These scores prove a light dominance of Word2Vec as it shows more relevant documents in the top-ten than CS. The discrepancy is even more notable when only titles are considered, regarding the DCG_{10} score of 3.069 for Word2Vec, versus 2.317 of CS. However, although W2V attains to catch more relevant or somehow relevant documents in the top-ten, the $nDCG_{10}$ value indicates that CS manages to perform a better ranking when it performs on all the metadata set.

In conclusion, as a result of the applied approaches, publications stored in a particular data set, i.e. digital library, are enriched with closely related semantic recommendations from other LOD repositories. This will enhance the visibility of publications from a single place by sparing the scholar for further navigations in other digital libraries. The research can be extended with several other combinations of the proposed approaches and metadata. At the same time, new methods can be introduced. However, in any case, a human assessment of the relevance of the retrieved results is necessary, knowing that it is expensive and inconsistent.

References

Agosti, M., Ferro, N. and Silvello, G. (2016) 'Digital library interoperability at high level of abstraction', *Future Generation Computer Systems*, Vol. 55, pp.129–146.

Anibaldi, S., Jaques, Y., Celli, F., Stellato, A. and Keizer, J. (2015) 'Migrating bibliographic datasets to the semantic web: the AGRIS case', *Semantic Web*, Vol. 6, No. 2, pp.113–120.

Auer, S., Heath, T., Bizer, C. and Berners-Lee, T. (2016) 'LDOW2016', *Proceedings of the 25th International Conference Companion on World Wide Web – WWW '16 Companion*, ACM Press, New York, NY, USA, pp.1039–1040.

Baroni, M., Dinu, G. and Kruszewski, G. (2014) 'Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors', *Proceedings of Association for Computational Linguistics*, Vol. 1, pp.238–247.

Bengio, Y., Schwenk, H., Senécal, J-S., Morin, F. and Gauvain, J-L. (2006) 'Neural probabilistic language models', *Innovations in Machine Learning*, Springer-Verlag, Berlin/Heidelberg, pp.137–186.

Berners-Lee, T., Hendler, J. and Lassila, O. (2001) 'The semantic web', *Scientific American*, Vol. 284, No. 5, pp.28–37.

Besser, H. (2002) 'The next stage: moving from isolated digital collections to interoperable digital libraries', *First Monday*, Vol. 7, No. 6. Available online at: https://www.learntechlib.org/p/95119/

Binding, C. and Tudhope, D. (2016) 'Improving interoperability using vocabulary linked data', *International Journal on Digital Libraries*, Vol. 17, No. 1, pp.5–21.

Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003) 'Latent dirichlet allocation', *Journal of Machine Learning Research*, Vol. 3, pp.993–1022.

Bobadilla, J., Ortega, F., Hernando, A. and Gutiérrez, A. (2013) 'Recommender systems survey', *Knowledge-Based Systems*, Vol. 46, pp.109–132.

Borgman, C.L. (2002) 'Challenges in building digital libraries for the 21st century', *Proceedings of 5th International Conference on Asian Digital Libraries*, Springer, Berlin, Heidelberg, pp.1–13.

Brants, T. and Franz, A. (2006) *Web 1T 5-gram Corpus Version 1.1*, Google Inc., Linguistic Data Consortium, Philadelphia, PA.

Brokos, G-I., Malakasiotis, P. and Androutsopoulos, I. (2016) *Using Centroids of Word Embeddings and Word Mover's Distance for Biomedical Document Retrieval in Question Answering*. Available online at: https://scirate.com/arxiv/1608.03905

Caracciolo, C. and Keizer, J. (2011) 'Thesaurus alignment for linked data publishing', *Proceedings of the International Conference on Dublin Core and Metadata Applications 2011 Thesaurus*, The Hague, The Netherlands, pp.37–46.

Caracciolo, C., Morshed, A., Stellato, A., Johannsen, G., Jaques, Y. and Keizer, J. (2012) 'Thesaurus maintenance, alignment and publication as linked data: the AGROVOC use case', *International Journal of Metadata, Semantics and Ontologies*, Vol. 7, No. 1, pp.65–75.

Collobert, R. and Weston, J. (2008) 'A unified architecture for natural language processing', *Proceedings of the 25th International Conference on Machine Learning – ICML '08*, ACM Press, New York, NY, USA, pp.160–167.

Cuzzocrea, A., Lee, W. and Leung, C.K. (2015) 'High-recall information retrieval from linked big data', *2015 IEEE 39th Annual Computer Software and Applications Conference*, IEEE, Taichung, Taiwan, pp.712–717.

De Boom, C., Van Canneyt, S., Bohez, S., Demeester, T. and Dhoedt, B. (2016) 'Learning semantic similarity for very short texts', *Proceedings – 15th IEEE International Conference on Data Mining Workshop, ICDMW 2015*, doi:10.1109/ICDMW.2015.86.

Deerwester, S., Dumais, S.T., Landauer, T.K., Furnas, G.W. and Harshman, R.A. (1990) 'Indexing by latent semantic analysis', *Journal of the American Society for Information Science*, Vol. 41, No. 6, pp.391–407.

Dietze, S., Sanchez-Alonso, S., Ebner, H., Yu, H.Q., Giordano, D., Marenzi, I. and Nunes, B.P. (2013) 'Interlinking educational resources and the web of data', *Program*, Vol. 47, No. 1, pp.60–91.

Di Noia, T., Mirizzi, R., Ostuni, V.C., Romito, D. and Zanker, M. (2012) 'Linked open data to support content-based recommender systems', *Proceedings of the 8th International Conference on Semantic Systems – I-SEMANTICS '12*, ACM Press, New York, NY, USA, pp.1–8.

Doerr, M., Gradmann, S., Hennicke, S., Isaac, A., Meghini, C. and Van de Sompel, H. (2010) 'The europeana data model (EDM)', *World Library and Information Congress: 76th IFLA General Conference and Assembly*, Gothenburg, Sweden, pp.10–15.

Dorsch, I. (2017) 'Relative visibility of authors' publications in different information services', *Scientometrics*, Vol. 112, No. 2, pp.917–925.

Evert, S. (2010) 'Google web 1T 5-grams made easy (but not for the computer)', *Proceedings of the NAACL HLT 2010 Sixth Web as Corpus Workshop*, Association for Computational Linguistics, Los Angeles, CA, USA, pp.32–40.

Faruqui, M., Tsvetkov, Y., Rastogi, P. and Dyer, C. (2016) *Problems With Evaluation of Word Embeddings Using Word Similarity Tasks*. Available online at: http://arxiv.org/abs/1605.02276

Fernández, M., Cantador, I., Lopez, V., Vallet, D., Castells, P. and Motta, E. (2011) 'Semantically enhanced information retrieval: an ontology-based approach', *Web Semantics: Science, Services and Agents on the World Wide Web*, Vol. 9, No. 4, pp.434–452.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G. and Ruppin, E. (2001) 'Placing search in context: the concept revisited', *Proceedings of the 10th International Conference on World Wide Web*, ACM, Hong Kong, China, pp.406–414.

Galke, L., Mai, F., Schelten, A., Brunsch, D. and Scherp, A. (2017) 'Comparing titles vs. full-text for multi-label classification of scientific papers and news articles', arXiv preprint arXiv:1705.05311.

Garibay, C., Gutiérrez, H. and Figueroa, A. (2010) 'Evaluation of a digital library by means of quality function deployment (QFD) and the Kano model', *The Journal of Academic Librarianship*, Vol. 36, No. 2, pp.125–132.

Hajra, A., Latif, A. and Tochtermann, K. (2014) 'Retrieving and ranking scientific publications from linked open data repositories', *Proceedings of the 14th International Conference on Knowledge Technologies and Data-Driven Business – i-KNOW '14*, ACM Press, New York, NY, USA, pp.1–4.

Hajra, A., Radevski, V. and Tochtermann, K. (2015) 'Author profile enrichment for cross-linking digital libraries', *International Conference on Theory and Practice of Digital Libraries*, Springer, Cham, Switzerland, pp.124–136.

Hajra, A. and Tochtermann, K. (2016) 'Enriching scientific publications from LOD repositories through word embeddings approach', *Research Conference on Metadata and Semantics Research*, Springer International Publishing, Cham, Switzerland, pp.278–290.

Heradio, R., Fernandez-Amoros, D., Cabrerizo, F.J. and Herrera-Viedma, E. (2012) 'A review of quality evaluation of digital libraries based on users' perceptions', *Journal of Information Science*, Vol. 38, No. 3, pp.269–283.

Hersh, W., Cohen, A.M. and Roberts, P. (2007) 'TREC 2007 genomics track overview', *The Sixteenth Text Retrieval Conference (TREC 2007)*, National Institute for Standards & Technology, Gaithersburg, MD, USA.

Hill, F., Reichart, R. and Korhonen, A. (2016) 'Simlex-999: evaluating semantic models with (genuine) similarity estimation', *Computational Linguistics*, MIT Press, Cambridge, MA, USA.

Horava, T. (2010) 'Challenges and possibilities for collection management in a digital age', *Library Resources & Technical Services*, Vol. 54, No. 3, pp.142–152.

Huang, Z., Chung, W., Ong, T-H. and Chen, H. (2002) 'A graph-based recommender system for digital library', *Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries – JCDL '02*, ACM Press, New York, NY, USA, pp.65–73.

Huang, E.H., Socher, R., Manning, CH. and Ng, A.Y. (2012) 'Improving word representations via global context and multiple word prototypes', *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers – Volume 1*, Association for Computational Linguistics (ACL '12), Stroudsburg, PA, USA, pp.873–882. Available online at: http://dl.acm.org/citation.cfm?id=2390524.2390645

Jacsó, P. (2005) 'Google scholar: the pros and the cons', *Online Information Review*, Vol. 29, No. 2, pp.208–214.

Järvelin, K. and Kekäläinen, J. (2000) 'IR evaluation methods for retrieving highly relevant documents', *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval – SIGIR '00*, ACM Press, New York, NY, USA, pp.41–48.

Järvelin, K. and Kekäläinen, J. (2002) 'Cumulated gain-based evaluation of IR techniques', *ACM Transactions on Information Systems*, Vol. 20, No. 4, pp.422–446.

Joshi, A.K., Jain, P., Hitzler, P., Yeh, P.Z., Verma, K., Sheth, A.P. and Damova, M. (2012) 'Alignment-based querying of linked open data', *OTM Confederated International Conferences 'on the Move to Meaningful Internet Systems'*, Springer, Berlin, Heidelberg, pp.807–824.

Kenter, T. and de Rijke, M. (2015) 'Short text similarity with word embeddings', *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management – CIKM '15*, ACM Press, New York, NY, USA, pp.1411–1420.

Kim, S., Wilbur, W.J. and Lu, Z. (2016) 'Bridging the gap: a semantic similarity measure between queries and documents', arXiv preprint arXiv:1608.01972.

Kling, R. and McKim, G. (1999) 'Scholarly communication and the continuum of electronic publishing', arXiv preprint cs/9903015. Available online at: http://arxiv.org/abs/cs/9903015

Kusner, M.J., Sun, Y., Kolkin, N.L. and Weinberger, K.Q. (2015) 'From word embeddings to document distances', *Proceedings of the 32nd International Conference on International Conference on Machine Learning – Volume 37*, JMLR.org, pp.957–966. Available online at: http://dl.acm.org/citation.cfm?id=3045221

Latif, A., Borst, T. and Tochtermann, K. (2014) 'Exposing data from an open access repository for economics as linked data', *D-Lib Magazine*, Vol. 20, No. 9, p.2.

Latif, A., Scherp, A. and Tochtermann, K. (2016) 'LOD for library science: benefits of applying linked open data in the digital library setting', *KI – Künstliche Intelligenz*, Vol. 30, No. 2, pp.149–157.

Lebret, R. and Collobert, R. (2015) 'Rehabilitation of count-based models for word vector representations', *International Conference on Intelligent Text Processing and Computational Linguistics*, Springer, Cham, Switzerland, pp.417–429.

Levy, O., Goldberg, Y. and Dagan, I. (2015) 'Improving distributional similarity with lessons learned from word embeddings', *Transactions of the Association for Computational Linguistics*, Vol. 3, pp.211–225.

Lin, J. and Wilbur, W.J. (2007) 'PubMed related articles: a probabilistic topic-based model for content similarity', *BMC Bioinformatics*, Vol. 8, No. 1, p.423.

Lops, P., de Gemmis, M. and Semeraro, G. (2011) 'Content-based recommender systems: state of the art and trends', *Recommender Systems Handbook*, Springer, Boston, MA, USA, pp.73–105.

Manning, C.D., Raghavan, P. and Schütze, H. (2008) *Introduction to Information Retrieval*, Cambridge University Press, New York, NY.

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013) 'Efficient estimation of word representations in vector space', arXiv preprint arXiv:1301.3781.

Mnih, A. and Hinton, G.E. (2009) 'A scalable hierarchical distributed language model', *Advances in Neural Information Processing Systems*, ACM, Vancouver, British Columbia, Canada, pp.1081–1088.

Mooney, R.J. and Roy, L. (2000) 'Content-based book recommending using learning for text categorization', *Proceedings of the Fifth ACM Conference on Digital Libraries – DL '00*, ACM Press, New York, NY, USA, pp.195–204.

Neubert, J. (2009) 'Bringing the "thesaurus for economics" on to the web of linked data', *Proceedings of the WWW Workshop on Linked Data on the Web (LDOW 2009)*, 20 April 2009, Madrid, Spain.

Niraula, N.B., Gautam, D., Banjade, R., Maharjan, N. and Rus, V. (2015) 'Combining word representations for measuring word relatedness and similarity', *Proceedings of the 28th International Florida Artificial Intelligence Research Society Conference*, AAAI Press, Hollywood, FL, USA, pp.199–204.

Norvig, P. (2013) *English Letter Frequency Counts: Mayzner Revisited or ETAOIN SRHLDCU*, Norvig.com. Available online at: http://norvig.com/mayzner.html

Paepcke, A., Chang, C-C.K., Winograd, T. and García-Molina, H. (1998) 'Interoperability for digital libraries worldwide', *Communications of the ACM*, Vol. 41, No. 4, pp.33–42.

Park, D.H., Kim, H.K., Choi, I.Y. and Kim, J.K. (2012) 'A literature review and classification of recommender systems research', *Expert Systems with Applications*, Vol. 39, No. 11, pp.10059–10072.

Passant, A. (2010) 'Measuring semantic distance on linking data and using it for resources recommendations', *AAAI Spring Symposium: Linked Data Meets Artificial Intelligence*, AAAI Press, Stanford, CA, USA, p.123.

Pennington, J., Socher, R. and Manning, C.D. (2014) 'Glove: global vectors for word representation', *EMNLP*, Association for Computational Linguistics, Doha, Qatar, pp.1532–1543.

Ramos, J. (2003) 'Using TF-IDF to determine word relevance in document queries', *Proceedings of the First Instructional Conference on Machine Learning*, Vol. 242, pp.133–142.

Resnick, A. (1961) 'Relative effectiveness of document titles and abstracts for determining relevance of documents', *Science*, Vol. 134, No. 3484, pp.1004–1006.

Robertson, S. and Zaragoza, H. (2010) 'The probabilistic relevance framework: BM25 and beyond', *Foundations and Trends® in Information Retrieval*, Vol. 3, No. 4, pp.333–389.

Salton, G., Wong, A. and Yang, C.S. (1975) 'A vector space model for automatic indexing', *Communications of the ACM*, Vol. 18, No. 11, pp.613–620.

Schnabel, T., Labutov, I., Mimno, D.M. and Joachims, T. (2015) 'Evaluation methods for unsupervised word embeddings', *EMNLP*, Association for Computational Linguistics, Lisbon, Portugal, pp.298–307.

Sheth, A.P. (1999) 'Changing focus on interoperability in information systems:from system, syntax, structure to semantics', *Interoperating Geographic Information Systems*, Springer, Cham, Switzerland, pp.5–29.

Smeaton, A.F. and Callan, J. (2005) 'Personalisation and recommender systems in digital libraries', *International Journal on Digital Libraries*, Vol. 5, No. 4, pp.299–308.

Sugiyama, K. and Kan, M-Y. (2010) 'Scholarly paper recommendation via user's recent research interests', *Proceedings of the 10th Annual Joint Conference on Digital Libraries – JCDL '10*, ACM Press, New York, NY, USA, p.29.

Thanos, C. (2014) 'The future of digital scholarship', *Procedia Computer Science*, Vol. 38, pp.22–27.

Thanos, C. (2016) 'A vision for open cyber-scholarly infrastructures', *Publications*, Vol. 4, No. 2, p.13.

Turian, J., Ratinov, L. and Bengio, Y. (2010) 'Word representations: a simple and general method for semi-supervised learning', *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics (ACL '10), Stroudsburg, PA, USA, pp.384–394. Available online at: http://dl.acm.org/citation.cfm?id=1858681.1858721

Wang, Y., Wang, L., Li, Y., He, D., Chen, W. and Liu, T-Y (2013) 'A theoretical analysis of NDCG ranking measures', *Proceedings of the 26th Annual Conference on Learning Theory (COLT 2013)*, Princeton, NJ, USA, pp.1–30.

Xie, H. (2006) 'Evaluation of digital libraries: criteria and problems from users' perspectives', *Library & Information Science Research*, Vol. 28, No. 3, pp.433–452.

Ye, X., Shen, H., Ma, X., Bunescu, R. and Liu, C. (2016) 'From word embeddings to document similarities for improved information retrieval in software engineering', *Proceedings of the 38th International Conference on Software Engineering – ICSE '16*, ACM Press, New York, NY, USA, pp.404–415.

Zamani, H. and Croft, W.B. (2016) 'Estimating embedding vectors for queries', *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, ACM Press, New York, NY, USA, pp.123–132.

Websites

https://www.w3.org/2004/02/skos/

https://scholar.google.com

https://www.base-search.net

https://www.mendeley.com

https://www.semanticscholar.org/

http://www.europeana.eu

http://www.ariadne-infrastructure.eu/

http://graphdb.ontotext.com