Blending Topic-based Embeddings and Cosine Similarity for Open Data Discovery

Maria Helena Franciscatto[1], Marcos Didonet Del Fabro[1], Luis Carlos Erpen de Bona[1],
Celio Trois[2] and Hegler Tissot[3]

[1]*Department of Informatics, Federal University of Paraná, Curitiba, Brazil*
[2]*Technology Center, Federal University of Santa Maria, Santa Maria, Brazil*
[3]*Drexel University, Philadelphia, U.S.A.*

Keywords: Source Discovery, Open Data, LDA, Word2Vec, Cosine Similarity, Machine Learning.

Abstract: Source discovery aims to facilitate the search for specific information, whose access can be complex and dependent on several distributed data sources. These challenges are often observed in Open Data, where users experience lack of support and difficulty in finding what they need. In this context, Source Discovery tasks could enable the retrieval of a data source most likely to contain the desired information, facilitating Open Data access and transparency. This work presents an approach that blends Latent Dirichlet Allocation (LDA), Word2Vec, and Cosine Similarity for discovering the best open data source given a user query, supported by joint union of the methods' semantic and syntactic capabilities. Our approach was evaluated on its ability to discover, among eight candidates, the right source for a set of queries. Three rounds of experiments were conducted, alternating the number of data sources and test questions. In all rounds, our approach showed superior results when compared with the baseline methods separately, reaching a classification accuracy above 93%, even when all candidate sources had similar content.

1 INTRODUCTION

Open data repositories makes information public to all citizens, promoting the monitoring and evaluation of government actions, data reuse, and improvement of services provided to the population. Such initiatives empower citizens, not only by making them more informed, but allowing them to transform data into something else, which is the true value of Open Data (Nikiforova and McBride, 2021).

However, the global increase of Open Data has led to the need to maintain numerous databases for storing important information, making the manipulation of this data a non-trivial task (Zhang and Yue, 2016). In other words, retrieving the information the user expects requires accessing structured and unstructured data, which lose significance if they are not presented clearly and meaningfully (Beniwal et al., 2018). With respect to Open Data, the information is usually available through data tables or CSV files, and when the amount and diversity of data are elevated, the visualization becomes confusing, affecting the users capability of performing comparisons and evaluations (Porreca et al., 2017). Thus, existing open data portals are considered complex by non-technical users, either by the format in which the data are presented, or the difficulty in finding the desired information (Osagie et al., 2017; Attard et al., 2015).

In fact, users usually can not easily analyze Open Data without expert assistance, and when they do, the required information may be scattered over several data tables or data sources (Djiroun et al., 2019). As a consequence, it becomes their responsibility to spend time finding, downloading, and evaluating datasets without the proper support from open data portals and platforms (Helal et al., 2021). The difficulty in finding the desired information not only affects the general user experience, but impacts scientific cooperation regarding the analysis and integration of heterogeneous data sources, which can be applied for solving complex problems in several areas (Sowe and Zettsu, 2015). These issues could be alleviated if the data source that best fits the user need was retrieved at first-hand, decreasing the manual work when searching for the right information; this task is known as *Source Discovery* (Abelló et al., 2014).

Source Discovery tasks have potential to leverage Open Data access and provide support to the

Franciscatto, M., Fabro, M., Erpen de Bona, L., Trois, C. and Tissot, H.
Blending Topic-based Embeddings and Cosine Similarity for Open Data Discovery.
DOI: 10.5220/0011040700003179
In *Proceedings of the 24th International Conference on Enterprise Information Systems (ICEIS 2022) - Volume 1*, pages 163-170
ISBN: 978-989-758-569-2; ISSN: 2184-4992

end user, by identifying, among several candidates sources, which one is most likely to contain the information needed. Several discovery methods have been proposed to measure, e.g., table relatedness (Nargesian et al., 2018), or find similarity joins between data collections (Xu et al., 2019). Although these studies focus on finding similarities between pairs of datasets, data tables or attribute columns, user-centered approaches that consider the query context for dataset recommendation are sparse, especially concerning Open Data (Dawes et al., 2016). With Open Data Discovery, the overall user experience when querying portals could be improved, favoring the access to public data and consequently their use in human-oriented applications. Therefore, there is a need to investigate approaches focused on reducing the complexity in discovering relevant open data sets.

This work presents a Source Discovery approach that blends topic-based embeddings and Cosine Similarity for inferring an optimal open data source, given an input query. Specifically, we apply a hybrid method named LDA-W2V, which uses LDA for detecting the most representative content of a data source, and Word2Vec for measuring how semantically close it is from a user query. Complementary, Cosine is applied as a syntactic similarity measure between source and query, so the best source is retrieved based on its structure and context.

Our approach was evaluated on its ability to discover the right source for a query among a set of eight candidates. For proving its consistency, we conducted three rounds of experiments, alternating the number of data sources and test questions, and comparing it with Cosine measure and LDA-W2V separately. The results showed that our approach was superior in all rounds of experiments, reaching a classification accuracy above 93%. This rate demonstrates that our approach based on LDA-W2V and Cosine Similarity is able to discover the most related data source for a user question, even when all candidate sources have similar content.

2 BACKGROUND

This section presents related work and concepts involved in the present study, including Source Discovery, Cosine Similarity, and LDA-W2V algorithm.

2.1 Source Discovery

A *data discovery problem* occurs when users and analysts spend more time looking for relevant data than analyzing it (Fernandez et al., 2018). So, Source Discovery is a process that aims to mitigate this obstacle, finding one or more relevant data sources (among many possible sources) suitable to a user query.

The concept has been widely investigated in several domains, especially in Business Intelligence (BI), assuming that data sources must be discovered on-the-fly for dealing with real time and situational queries (Abelló et al., 2013). Considering this need, many organizations have been encouraged to build a navigational data structure to support source discovery or to use tools for deriving insights from datasets (Helal et al., 2021). With respect to Open Data, Source Discovery is able to facilitate the access to publicly available datasets, designed to be reused for human benefit.

Several studies propose Source Discovery mechanisms to handle user queries. We can mention, e.g., the Aurum system (Fernandez et al., 2018), which allows people to flexibly find relevant data through properties of the datasets or syntactic relationships between them. The DataMed approach (Chen et al., 2018) also includes a Source Discovery task for finding relevant biomedical datasets from heterogeneous sources, making them searchable through a web-based interface.

Source Discovery implementation may be supported by several tasks such as schema discovery and query reformulation (Hamadou et al., 2018). Mostly, similarity methods are also used to determine the source that best relates to the user query. Some examples of these methods are presented in the following.

2.2 Cosine Similarity

Cosine Similarity measures similarity as the angle between two vectors being compared, assuming that each word in a text or document corresponds to a dimension in a multidimensional space (Gomaa et al., 2013). When measuring the angle of the documents, smaller the angle, higher the similarity. So, considering that cosine of 0° is 1, two vectors are said to be similar when Cosine Similarity is 1 (Gunawan et al., 2018). Cosine Similarity calculation is represented in Equation 1, where $\vec{A}$ and $\vec{B}$ are attribute vectors.

$$\cos(A,B) = \frac{\vec{A} \cdot \vec{B}}{|\vec{A}| \cdot |\vec{B}|} \tag{1}$$

Cosine Similarity is perhaps the most frequently applied proximity measure in information retrieval (Korenius et al., 2007). It has been widely studied in several application domains, such as medical diagnosis (Rafiq et al., 2019) and recommendation systems (Kotkov et al., 2018). In this study, it is investigated for finding related data sources in a discovery task. Concerning this goal, the next subsections present a hybrid approach based on LDA and Word2Vec.

2.3 Hybrid LDA-W2V

Latent Dirichlet Allocation (LDA) is a generative probabilistic model of a corpus, based on the idea that documents are represented as random mixtures over latent topics, and each topic is characterized by a distribution over words (Blei et al., 2003). Thus, given a document, paragraph, or sentence, LDA determines the most relevant topics, based on words in a topic that appear more often in the narrative compared to the words related to other topics (Bastani et al., 2019).

In (Jedrzejowicz and Zakrzewska, 2020), the authors propose a hybrid model based on LDA. Specifically, the model joins LDA with Word2Vec algorithm (Goldberg and Levy, 2014), considering that word embeddings allow to capture semantics when processing vast amounts of linguistic data. First, it assumes a set of documents, which are preprocessed for obtaining a list of representative words or *tokens*. Each document set of tokens is given as input to the LDA algorithm, which predicts the topics (i.e., words and their proportions) that best describe the document content. For exemplifying, a document containing information on universities could be represented by a topic containing the following words and proportions: *0.048*"high" + 0.047*"education" + 0.033*"students" + 0.033*"university" + 0.032*"course" + 0.032*"public" + 0.032*"federal" + 0.031*"private" + 0.031*"administrative."*

The next step of the approach performs a Word2Vec (W2V) Extension, which aims to extend words in sources topics by using similar words acquired from Word2Vec model. The similarity is measured by Cosine Similarity (see Subsection 2.2): Supposing a topic word W_n, the W2V model is traversed to find similar words $[w2vWord_1, ..., w2vWord_n]$. Then, the similarity score between a pair $[W_n, w2vWord_n]$ is multiplied by the LDA proportion score for W_n, obtaining a derived proportion DP_n. Each similar word found represents an extended word EW_n. Following the previous topic example, the W2V model retrieves the word *institution* as similar to the topic word *university*, with a similarity score 0.71. This score is multiplied by LDA proportion score for *university*, 0.033, thus obtaining DP_n=0.023. *institution* becomes an extended word EW_n, and generates a tuple $[W_n, DP_n, EW_n]$, e.g., [*university, 0.023, institution*]. After word extension, the approach tries to classify an input sentence (the test set) into a topic, based on the probabilities (DP_n) within all possible tuples. In short, the input sentence is assigned to the topic that contains the highest probabilities for each sentence word.

LDA has become the most popular topic modeling algorithm used, due to its applicability in several contexts and ability to analyze large documents (Gottfried et al., 2021). We leveraged the benefits from both LDA and Word2Vec, applying this hybrid solution for source discovery involving open datasets.

3 SOURCE DISCOVERY APPROACH BASED ON COSINE AND LDA-W2V

Accessing Open Data often represents an obstacle for regular users, due to the amount of data made available, its format and diversity. A Source Discovery mechanism could replace users manual work when querying, by retrieving desirable data and improving the overall experience. Thus, this section presents a Source Discovery approach based on LDA-W2V (Jedrzejowicz and Zakrzewska, 2020) and Cosine Similarity for recommending the open data source that best fits a user question.

Cosine Similarity was chosen due to its simplicity, as it only requires term-frequency vectors from two sets being compared to calculate similarity. Also, it is commonly applied to measure similarity beteween a query and an item, based on common features between them (Ristoski et al., 2014; Di Noia et al., 2012). However, when context information is not available, this measure may fail in determining similarity (Orkphol and Yang, 2019). Thus, our approach is complemented with a hybrid LDA approach based on (Jedrzejowicz and Zakrzewska, 2020), which combines semantic capabilities from both LDA and Word2Vec for classification tasks. The LDA model has clear internal structure that allows efficient inference, and it is independent of the training documents number, thus being suitable for handling large scale corpus (Liu et al., 2011). The combination LDA-W2V + Cosine allows to apply both syntactic and semantic capabilities for discovering which data source is more adequate to answer an input question[1]. Our approach is demonstrated in Figure 1.

The approach takes as input a set of candidate sources (represented by CS1 to CS4 in the figure) and a user input sentence. First, we extract from each can-

[1] Other syntactic and semantic methods could be joined for similar purposes.

didate source its schema information[2], preprocessing it for removing special characters and stopwords from the text. The preprocessing stage results in a set of tokens for each source (T_{CSn}) and for the input sentence (T_{IS}). The sources tokens are given as input to our LDAW2V-based implementation, where a Topic Detection module outputs several topics for each source. In other words, after LDA topic detection, each source is represented by several topics composed by words (W_n) and their proportions (P_n), that determine word relevance in the analyzed text. The word extension through Word2Vec occurs as mentioned in Subsection 2.3: words that are similar to a topic word W_n are captured along with their similarity score, in order to form tuples containing extended words EW_n and probabilities DP_n derived from P_n multiplication.

After W2V Extension, we obtain tuples $[W_n, DP_n, EW_n]$ for each candidate source. So, the Matching step[3] of our LDA-W2V algorithm receives and traverses all tuples, aiming to find the best correlations for an input sentence, i.e., a query. In other words, the query tokens (T_{IS}, extracted after preprocessing), are used to search for equivalent W_n or EW_n with higher DP_n. When a match is found, a *probability array* PA_{CSn} is created for each candidate source, containing one DP_n for each query token T_{IS}. Otherwise, if a query token is not found in the extension, a default probability (0.000001) is inserted in PA_{CSn}.

While the LDA-W2V process occurs, the sources tokens T_{CSn} and the input sentence tokens (T_{IS}) are sent to the Cosine Similarity module, responsible for vectorizing them for application in the Cosine formula (see Equation 1). At this step, each source, as well as the input sentence, are converted to term-frequency vectors V_{IS} and V_{CSn}, so the Cosine Similarity cos_{CSn} is the distance measured between these vectors. The output from this module is the Cosine Similarity for each candidate source, which is inserted in the probability array PA_{CSn}, along with the LDA-W2V derived probabilities. After completing the LDA-W2V and Cosine tasks, each candidate source is represented by an array of joint probabilities, and the average probability for each array is calculated. Finally, the source most likely to meet a user question is selected by considering the array with highest average.

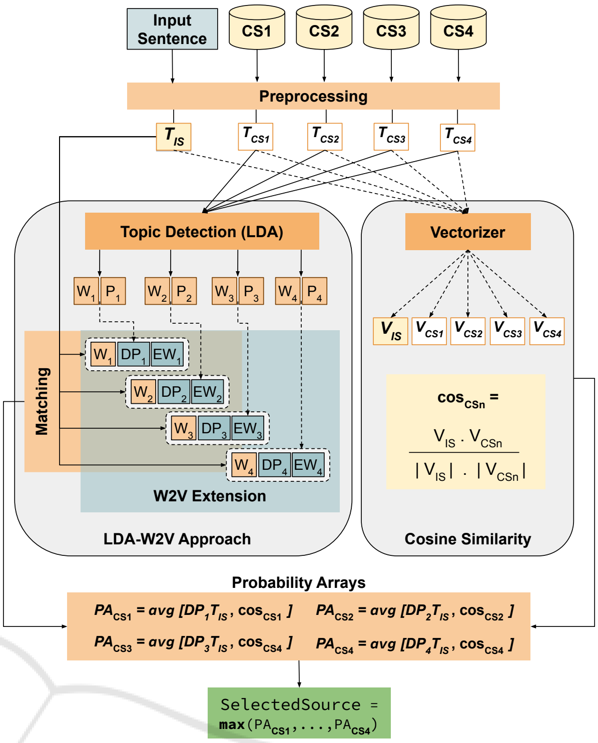

Figure 1: Overview of the source discovery approach based on Cosine Similarity and LDA-W2V.

Practical Example: Suppose two candidate sources, CS1 and CS2, and a user input sentence *IS* "How many rural schools have computers?". After preprocessing, the query tokens T_{IS} are [*rural, school, computer*]. After W2V Extension, each candidate source is represented by several tuples $[W_n, DP_n, EW_n]$, from which T_{IS} will be searched. If all query tokens are found in the CS1 extension, the array PA_{CS1} for this source could be, e.g., [0.0024, 0.032, 0.0096]. The same query tokens for CS2 could originate an array PA_{CS2} [0.0012, 0.0018, 0.000001], considering that "computer" token was not found in the extension. Thus, each source array will contain different probabilities, one for each query token. After Cosine Similarity calculation, cos_{CS1} is *0.85* for CS1, whereas cos_{CS2} is *0.7* for CS2. Both values are included in their respective probability arrays, resulting in $PA_{CS1} = [0.0024, 0.032, 0.0096, 0.85]$ and $PA_{CS2} = [0.0012, 0.0018, 0.000001, 0.7]$. After the arrays are arranged, the probabilities average is calculated for each source, so the source with highest average is chosen as the most likely one to meet the input query. In this example, CS1 would be the selected one.

[2] We consider the schema information all metadata contained in a CSV file, or text in a data dictionary file, which summarize the source content or its purpose.

[3] Since the original LDA-W2V approach was focused on classifying texts into different topics, we implemented an adapted version of the Matching step, where the text (i.e., the input sentence) is classified into one source based on a probability array.

4 EXPERIMENTS

In this section we present how our approach was implemented to deal with candidate open data sources and input queries.

We applied our approach to discover the data source that best matches an input question, considering eight possible datasets, i.e., candidate sources, named FIES[4], INEP[5], PROUNI[6], CadUnico[7], School Census[8], IBGE[9], DataSUS[10], and Ibama[11]. All candidate sources were extracted from Brazilian Open Data datasets. Half of them (INEP, FIES, PROUNI, and School Census) contain educational data from different contexts. The other four datasets are very distinct: CadUnico contains socioeconomic information on low-income citizens and families, IBGE aggregates social, economic and environmental indicators from Brazilian cities and states, DataSUS manages health information from Brazilian healthcare networks, and Ibama dataset contains information related to environment actions and use of natural resources.

In the next subsection, we describe implementation details concerning LDA-W2V and Cosine Similarity methods included in our approach.

4.1 Implementation Details

First of all, each candidate source schema information was manually extracted either from a CSV or a data dictionary in its respective open data portal. The extracted information was placed in an auxiliar CSV file that summarizes metadata from all sources, so that each row contains information from a different source. Each source content within this file was preprocessed (see Figure 1) through lowercasing, removal of special characters, removal of stopwords, and stemming, resulting in a set of meaningful tokens (represented by T_{CSn} in the Figure). LDA-W2V and Cosine Similarity were implemented in Python[12], taking the CSV file as input, along with test questions (the input sentences). Each test question was preprocessed the same was as the sources content, resulting in a list of tokens T_{IS}. Examples of test questions are shown in Table 1.

Table 1: Examples of Test Questions (English Translation)

Query	Source
How many rural schools have computers?	School Census
How many families in CadUnico receive income greater than R$1000?	CadUnico
How many students that benefited from racial quotas dropped out of a course in 2018?	INEP
What is the territory size of the Amazonas state?	IBGE

For the LDA-W2V module of our approach, the *Gensim* implementation[13] was used, and each T_{CSn} was sent as a corpus to the LDA model. Multiple runs were performed with LDA, alternating the *num_topics* parameter in the model, which determines the number of latent topics to be extracted from each corpus. The parameter value ranged from 8 to 10, based on the coherence measure[14] for each source, which evaluates how coherent the produced topics are, by capturing their interpretability on the LDA distribution.

Since all candidate sources are from Brazilian open data portals, we implemented W2V Extension (see Figure 1) with a Portuguese pre-trained model from *FastText*[15]. The model receives the LDA topic distribution (words and proportions) from each source, and searches for similar words for the extension. For measuring the similarity between a topic word W_n and a model word, we used Word2Vec's *similar_by_vector* method that finds the top-N similar words given a word vector. The N parameter was set as 50, and the retrieval of model words similar to topic words was based on a similarity threshold, represented as K, which assumed value 0.45. Only similarity scores above K were multiplied with the W_n proportion score to derive the proportion DP_n (see Subsection 2.3).

For measuring Cosine Similarity between a query and a candidate source, we extracted the term-frequency vectors for each query and source. For this task, Python Counter tool[16] was applied in the preprocessed query and source (T_{IS} and T_{CSn}). The Counter maps tokens in T_{IS} and T_{CSn} as dictionary keys, and their counts as dictionary values, thus obtaining term-frequency vectors V_{IS} and V_{CSn} in the format $\{$*"word": 4, "otherword": 2*$\}$. The vectors are sent to the Cosine function, which performs the calculation shown in Equation 1.

[4] http://dadosabertos.mec.gov.br/fies

[5] https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/censo-da-educacao-superior

[6] https://dados.gov.br/dataset/mec-prouni

[7] https://dados.gov.br/dataset/microdados-amostrais-do-cadastro-unico

[8] https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/censo-escolar

[9] https://www.ibge.gov.br/

[10] https://opendatasus.saude.gov.br/dataset

[11] http://www.ibama.gov.br/dados-abertos

[12] https://github.com/mariahelenaaf/OpenDataDiscovery

[13] https://radimrehurek.com/gensim/models/ldamulticore.html

[14] https://radimrehurek.com/gensim/models/coherencemodel.html

[15] https://fasttext.cc/

[16] https://docs.python.org/3/library/collections.html

After executing both LDA-W2V and Cosine modules, we obtain a probability array PA_{CSn} for each candidate source, as demonstrated in Figure 1. Next, we present the results of our approach performing Source Discovery tasks.

5 RESULTS AND DISCUSSION

In this section we describe the results obtained with our approach performing Source Discovery tasks, evaluating its capability to choose the most likely open dataset to answer an input query. The evaluation involved eight candidate datasets (see Section 4) and test sets containing 48 and 74 questions, respectively, to which a source should be inferred. All questions and their correct answers (i.e., sources names) were defined manually, based on indicators available on the data monitoring platforms SIMOPE[17] and LDE[18]. Examples of the test questions are shown in Table 1.

The evaluation was performed in three rounds, alternating the number of datasets and questions for investigating possible changes in classification results. Also, in each evaluation round, our approach was tested against Cosine and LDA-W2V individually for observing accuracy variation. The comparison between all methods accuracy is shown in Table 2.

In the first evaluation round, we conducted the evaluation with all data sources (i.e., Prouni, School Census, FIES, INEP, CadUnico, IBGE, Ibama, and DataSus) and the test set containing 48 questions. Considering Cosine individually, the right source was chosen for 87.5% of the test questions, whereas for LDA-W2V, the accuracy rate reached 79.17%. We can observe, in Table 2, that our approach blending the two methods improved the classification accuracy considerably, reaching 93.75% (an increase of 6.25% in the individual Cosine result, and 14.28% in the LDA-W2V result).

For assessing our approach consistence, we conducted a second round of evaluations, using the test set containing 74 questions and the same eight open data sources. Although a small improvement had been observed for LDA-W2V (79.73%), Cosine accuracy was impacted negatively (83.78%), hence impacting our model's accuracy (87.74%). This can be explained by the structure of some test questions, whose content were less specific. E.g., if a question is mostly composed by words that are common in many candidate sources, there might be a misclassification. The accuracy reduction can also be due to the stemming method applied in preprocessing stage, since the reduction of some Portuguese words may originate ambiguous tokens, thus confusing the classifier. Despite that, our approach had superior results than the isolated methods: the right source was chosen for 87.74% of the questions, meaning 4.05% of increase when compared to Cosine, and 8.11% when compared to LDA-W2V.

Table 2: Classification accuracy in three evaluation rounds with different sources (S) and questions (Q).

S	Q	Cosine Sim.	LDA-W2V	**Our Approach**
8	48	87.5	79.17	**93.75**
8	74	83.78	79.73	**87.74**
5	48	87.5	81.25	**93.75**

In the third evaluation round, the consistency of our approach was estimated from another perspective: we removed three of the most distinctive data sources from the pool of candidate sources, leaving only the datasets Prouni, School Census, FIES, INEP, and CadUnico. From these, CadUnico is the only source that does not contain education data. Our objective was to verify whether the classification accuracy would be satisfactory with data sources containing very similar content (i.e., metadata). The results are summarized in Table 2 and Figure 2.

Cosine and LDA-W2V reached 87.5% and 81.25% of accuracy, respectively, whereas our approach reached 93.75%. The rates had little variation when compared to the ones in the first evaluation round, which is a very promising result. Indeed, four out of five datasets contained information on Education, representing a more complex classification task compared to predicting distinct classes; yet, our approach accuracy rate was over 93%. Moreover, as we can see in Figure 2, there was a significant prediction improvement for each of the five datasets in comparison to Cosine and LDA-W2V predictions, especially for School Census and Cadunico datasets, where no incorrect classification occurred.

We observed that LDA-W2V was the method with the lowest accuracy (around 80%) in all executions. This can be explained by the high variability caused by the probabilities assigned in W2V Extension step (see Subsection 2.3). For exemplifying, let us consider the input query "*Number of students who have housing assistance at federal universities*". Although the source that best answers this question is INEP, FIES source was the predicted one, since its probability array contained higher values for some tokens such as *assistance* and *federal*. This specific behavior may have impacted our model accuracy, especially considering that half of the databases used contained educational information.

[17] https://seppirhomologa.c3sl.ufpr.br/

[18] https://dadoseducacionais.c3sl.ufpr.br

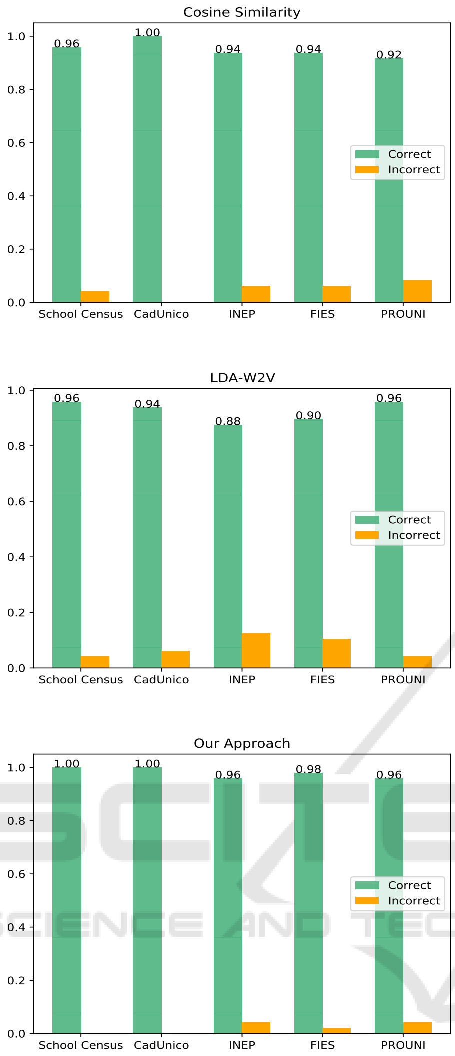

Figure 2: Classification accuracies in the third evaluation round.

Despite that, the experiments involving Open Data Discovery tasks with our approach have shown promising results: the accuracy rates in all evaluation rounds were above 87%, where half of the open datasets used were of very similar subject (i.e., Education). From these results, we infer that possible weaknesses of each method individually (Cosine or LDA-W2V) were overcome by their combination, allowing to explore both semantic and syntactic features for discovery tasks. It is important to highlight that, in our blended model, only metadata and/or sources descriptions were used to classify the test questions, which reinforces the quality of the results and their potential for Open Data Discovery domain.

6 CONCLUSIONS

This paper presented a Source Discovery approach that joins topic-based embeddings from the LDA-W2V algorithm and Cosine Similarity for inferring an open data source that is most likely to answer an input query. By blending both measures, we leverage syntactic and semantic advantages for detecting meaningful content. We evaluated the approach by conducting three rounds of Source Discovery experiments involving eight candidate open datasets and alternated test sets: for each test question, our approach has to infer which of the candidate sources was the most suitable. The classification results showed an accuracy above 93%, i.e., a superior rate when compared with LDA-W2V and Cosine separately. From these experiments, we conclude that our hybrid approach is able to discover the most related data source for a user question. The encouraging results also demonstrate that our approach has potential to improve Open Data transparency and user support.

As future work, we aim to implement a case-based recommendation strategy allied to Source Discovery, since it can leverage user feedback on the suggested source to improve future classifications. Also, our objective is to investigate other approaches and voting mechanisms that can be applied to Source Discovery tasks, for building a unified solution that combines different advantages. Finally, we aim to evaluate a Source Discovery solution with real users, in order to promote Open Data access through a good querying experience.

ACKNOWLEDGMENTS

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brazil (CAPES), Finance Code 001, and Fundo Nacional de Desenvolvimento da Educação (FNDE), through research project “Pesquisa e Inovação na Integração e Visualização de Dados do Programa Nacional do Livro Didático”.

REFERENCES

Abelló, A., Darmont, J., Etcheverry, L., Golfarelli, M., Mazón, J.-N., Naumann, F., Pedersen, T., Rizzi, S. B., Trujillo, J., Vassiliadis, P., et al. (2013). Fusion cubes: Towards self-service business intelligence. *International Journal of Data Warehousing and Mining (IJDWM)*, 9(2):66–88.

Abelló, A., Romero, O., Pedersen, T. B., Berlanga, R., Nebot, V., Aramburu, M. J., and Simitsis, A. (2014).

Using semantic web technologies for exploratory olap: a survey. *IEEE transactions on knowledge and data engineering*, 27(2):571–588.

Attard, J., Orlandi, F., Scerri, S., and Auer, S. (2015). A systematic review of open government data initiatives. *Government information quarterly*, 32(4):399–418.

Bastani, K., Namavari, H., and Shaffer, J. (2019). Latent dirichlet allocation (lda) for topic modeling of the cfpb consumer complaints. *Expert Systems with Applications*, 127:256–271.

Beniwal, R., Gupta, V., Rawat, M., and Aggarwal, R. (2018). Data mining with linked data: Past, present, and future. In *2018 Second International Conference on Computing Methodologies and Communication (ICCMC)*, pages 1031–1035. IEEE.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Chen, X., Gururaj, A. E., Ozyurt, B., Liu, R., Soysal, E., Cohen, T., Tiryaki, F., Li, Y., Zong, N., Jiang, M., et al. (2018). Datamed–an open source discovery index for finding biomedical datasets. *Journal of the American Medical Informatics Association*, 25(3):300–308.

Dawes, S. S., Vidiasova, L., and Parkhimovich, O. (2016). Planning and designing open government data programs: An ecosystem approach. *Government Information Quarterly*, 33(1):15–27.

Di Noia, T., Mirizzi, R., Ostuni, V. C., Romito, D., and Zanker, M. (2012). Linked open data to support content-based recommender systems. In *Proceedings of the 8th international conference on semantic systems*.

Djiroun, R., Boukhalfa, K., and Alimazighi, Z. (2019). Designing data cubes in olap systems: a decision makers' requirements-based approach. *Cluster Computing*, 22(3).

Fernandez, R. C., Abedjan, Z., Koko, F., Yuan, G., Madden, S., and Stonebraker, M. (2018). Aurum: A data discovery system. In *2018 IEEE 34th International Conference on Data Engineering (ICDE)*, pages 1001–1012. IEEE.

Goldberg, Y. and Levy, O. (2014). word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*.

Gomaa, W. H., Fahmy, A. A., et al. (2013). A survey of text similarity approaches. *International Journal of Computer Applications*, 68(13):13–18.

Gottfried, A., Hartmann, C., and Yates, D. (2021). Mining open government data for business intelligence using data visualization: A two-industry case study. *Journal of Theoretical and Applied Electronic Commerce Research*, 16(4):1042–1065.

Gunawan, D., Sembiring, C., and Budiman, M. (2018). The implementation of cosine similarity to calculate text relevance between two documents. In *Journal of Physics: Conference Series*, volume 978, page 012120. IOP Publishing.

Hamadou, H. B., Ghozzi, F., Péninou, A., and Teste, O. (2018). Querying heterogeneous document stores. In *20th International Conference on Enterprise Information Systems (ICEIS 2018)*, volume 1, pages 58–68.

Helal, A., Helali, M., Ammar, K., and Mansour, E. (2021). A demonstration of kglac: a data discovery and enrichment platform for data science. *Proceedings of the VLDB Endowment*, 14(12):2675–2678.

Jedrzejowicz, J. and Zakrzewska, M. (2020). Text classification using lda-w2v hybrid algorithm. In *Intelligent Decision Technologies 2019*, pages 227–237. Springer.

Korenius, T., Laurikkala, J., and Juhola, M. (2007). On principal component analysis, cosine and euclidean measures in information retrieval. *Information Sciences*, 177(22):4893–4905.

Kotkov, D., Konstan, J. A., Zhao, Q., and Veijalainen, J. (2018). Investigating serendipity in recommender systems based on real user feedback. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, pages 1341–1350.

Liu, Z., Li, M., Liu, Y., and Ponraj, M. (2011). Performance evaluation of latent dirichlet allocation in text mining. In *2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, volume 4, pages 2695–2698. IEEE.

Nargesian, F., Zhu, E., Pu, K. Q., and Miller, R. J. (2018). Table union search on open data. *Proceedings of the VLDB Endowment*, 11(7):813–825.

Nikiforova, A. and McBride, K. (2021). Open government data portal usability: A user-centred usability analysis of 41 open government data portals. *Telematics and Informatics*, 58:101539.

Orkphol, K. and Yang, W. (2019). Word sense disambiguation using cosine similarity collaborates with word2vec and wordnet. *Future Internet*, 11(5):114.

Osagie, E., Waqar, M., Adebayo, S., Stasiewicz, A., Porwol, L., and Ojo, A. (2017). Usability evaluation of an open data platform. In *Proceedings of the 18th Annual International Conference on Digital Government Research*, pages 495–504.

Porreca, S., Leotta, F., Mecella, M., Vassos, S., and Catarci, T. (2017). Accessing government open data through chatbots. In *International Conference on Web Engineering*, pages 156–165. Springer.

Rafiq, M., Ashraf, S., Abdullah, S., Mahmood, T., and Muhammad, S. (2019). The cosine similarity measures of spherical fuzzy sets and their applications in decision making. *Journal of Intelligent & Fuzzy Systems*, 36(6):6059–6073.

Ristoski, P., Mencía, E. L., and Paulheim, H. (2014). A hybrid multi-strategy recommender system using linked open data. In *Semantic Web Evaluation Challenge*, pages 150–156. Springer.

Sowe, S. K. and Zettsu, K. (2015). Towards an open data development model for linking heterogeneous data sources. In *Knowledge and Systems Engineering (KSE), 2015 Seventh International Conference on*, pages 344–347. IEEE.

Xu, P., Lu, J., et al. (2019). Towards a unified framework for string similarity joins. *Proceedings of the VLDB Endowment*.

Zhang, C. and Yue, P. (2016). Spatial grid based open government data mining. In *Geoscience and Remote Sensing Symposium (IGARSS), 2016 IEEE International*, pages 192–193. IEEE.