HAL
open science

Knowledge-Based Categorization of Scientific Articles for Similarity Predictions

Nolwenn Bernard, Jonathan Weber, Germain Forestier, Michel Hassenforder, Bastien Latard

To cite this version:

Nolwenn Bernard, Jonathan Weber, Germain Forestier, Michel Hassenforder, Bastien Latard. Knowledge-Based Categorization of Scientific Articles for Similarity Predictions. International Conference on Theory and Practice of Digital Libraries (TPDL), Aug 2020, Lyon, France. pp.147-160, 10.1007/978-3-030-54956-5 . hal-03716931

HAL Id: hal-03716931

https://hal.science/hal-03716931v1

Submitted on 12 Sep 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Knowledge-Based Categorization of Scientific Articles for Similarity Predictions

Nolwenn Bernard*, Jonathan Weber†, Germain Forestier†, Michel Hassenforder†, and Bastien Latard*,†

* MDPI, Basel, Switzerland
{lastname}@mdpi.com
† Universit de Haute-Alsace, IRIMAS, Mulhouse, France

Abstract. Staying aware of new approaches emerging within specific areas can be challenging for researchers who have to follow many feeds such as journals articles, authors' papers, and other basic keyword-based matching algorithms. Hence, this paper proposes an information retrieval process for scientific articles aiming to suggest semantically related articles using exclusively a knowledge base. The first step categorizes articles by the disambiguation of their keywords by identifying common categories within the knowledge base. Then, similar articles are identified using the information extracted from the categorization, such as synonyms. The experimental evaluation shows that the proposed approach significantly outperforms the well known cosine similarity measure of vectors angles inherited from word2vec embeddings. Indeed, there is a difference of 30% for P@k ($k \in [1, 100]$) in favor of the proposed approach.

Keywords: Information retrieval · Categorization · Scientific literature · Document similarity

1 Introduction

Nowadays, the number of scientific articles available in digital format has exploded. Their processing is time-consuming and hence, automatic tools are widely used by researchers to stay up-to-date, as stated by Pain [25]. Improving bibliographic searches could have a positive impact on the scientific literature [30]. The major challenge of this process is its scalability; indeed, these days, databases such as arXiv[1] and Scilit[2] freely propose millions of articles. Thus, text mining is necessary for suggesting relevant documents regarding a topic. Text mining is commonly defined as the process of extracting interesting and nontrivial patterns or knowledge from unstructured text documents. This process involves different fields, such as information retrieval, text analysis, and categorization.

[1] https://arxiv.org/
[2] https://www.scilit.net/

This is the author's version of an article published at TPDL 2020. The final authenticated version is available online at: https://doi.org/10.1007/978-3-030-54956-5_11

The purpose here is to suggest semantically related scientific articles based on a categorization method and similarity measure. The approach proposed in this article aims to improve and extend upon the previous work presented by Latard et al. [18], henceforth referred to as the original approach. They stated an approach which categorizes articles using keyword disambiguation, which provides good results in terms of precision, yet with the drawback of low coverage (i.e., less than 50% of articles are categorized). Therefore, the objective of the proposed approach is to have a higher coverage than the original one in terms of categorized articles in order to compete with the probabilistic approaches broadly used in text mining. To achieve this, category assignation was transformed to be more permissive than the original approach and lemmatization was included in the categorization process.

As far as we know, a fully automated cross-domain information retrieval process for scientific articles exclusively based on an knowledge base does not exist. Using semantics in this type of process permits extension beyond the scope of possibilities [1,17] and the introduction of word sense disambiguation. Indeed, the use of synonyms, hypernyms, etc. is able to complete the query in comparison to the use of only keywords.

This paper starts by a brief overview of related works (Section 2). Then, the proposed method is explained (Section 3). Next, Section 4 presents a comparison with an approach using word embeddings and cosine similarity measure. Finally, the results will be discussed (Section 5).

2 Related works

In 2018, the number of new scientific papers published per year was estimated to be over 3 million [11]; therefore, providing the most relevant suggestions has become a challenge. This has led to a growing interest in information retrieval and information extraction. Information retrieval aims to retrieve the most relevant documents from a corpus based on a given query. Therefore, it often combines text mining techniques [8] and similarity measures in order to retrieve the closest documents. Text mining also embraces information extraction whose purpose is to extract meaningful information from unstructured documents.

Word embedding covers techniques in natural language processing, where words of the vocabulary are mapped to real-valued vectors. Semantic similarities are identified based on the usage of a word in the corpus and its neighbors, as stated by Firth [2]: *"You shall know a word by the company it keeps!"*. In the literature, word embedding is generally associated with word2vec [21], which is a tool based on a multi-layer neural network. For example, word2vec generated models have been used by the Microsoft Academic search engine [12] and by the Computer Science Ontology classifier [27].

Vector space model is described by Salton et al. [28] as the representation of the corpus into a $m * n$, matrix where columns represent the corpus documents and rows embrace all terms of the entire corpus vocabulary. This vector representation is widely used because it provides a practical way to manipulate and

compare documents. Indeed, common similarity measures [9], such as Euclidean distance and cosine similarity, take advantage of this vector representation.

Word sense disambiguation [23] is the capacity to identify the sense of an ambiguous word regarding its usage context. In many text mining applications [20,29], this step of disambiguation is crucial. Knowledge sources are essential for word sense disambiguation; they can be structured, such as thesauri and ontologies, or unstructured, like sense-annotated corpora.

The categorization workflow [18] used as a base for this work takes advantage of keywords' metadata to find semantic relations between an article's keywords. In another work, Latard [16] uses these metadata to define a similarity metric to retrieve similar articles in a corpus.

3 Proposed approach

3.1 Categorization

The categorization of scientific articles is the first step of the proposed approach. Word sense disambiguation is important knowing that an article's keywords can be ambiguous; for example, *synthesis* can be understood as a logical reasoning in a mathematics context or as chemical compound production in chemistry. Therefore, the knowledge base BabelNet [24] is used to select the most consistent keyword's meanings depending on the article's context. It is a multilingual encyclopedic dictionary and a semantic network based on the integration of several semantic lexicons (WordNet [22], VerbNet) together with collaborative databases (Wikipedia[3] and other Wiki data). It can be seen as a dictionary where a single word has different meanings called synsets in the rest of this paper with different senses. Synsets contain several elements of information and for this system, only three of them are kept: categories (C), domains (D), and neighbor synsets (N). A synset s is defined as follows:

$$s = \{C, D, N\} \tag{1}$$

Categorization workflow. From their research work, Gil-Leiva and Alonso-Arroyo [5] highlighted that the keywords provided by authors bring relevant and meaningful information. In *Web Of Science*[4], keywords are provided by the authors or by the algorithm "KeyWords Plus" [4] or both. Given that we assume that an article's keywords are legitimate, they are used as the only input in this step of the process.

Fig. 1 represents the categorization's step workflow [16]. An article's keywords without any preprocessing are used as input to search for an exact match in BabelNet, but many keywords composed of several words are not indexed in BabelNet. Hence, in the case of a first stage without results, a second stage, in which keywords are split, is proposed. Then connected synsets are identified (e.g.,

[3] https://www.wikipedia.org/
[4] https://www.webofknowledge.com/

Fig. 2) and their related data, such as categories and domains, are extracted and considered as contextually related to this article.

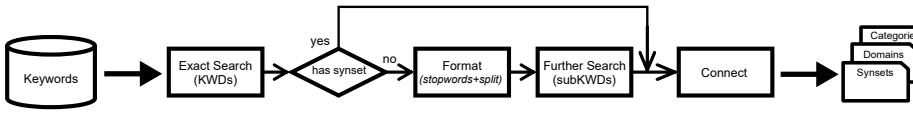


Fig. 1: Simplified workflow of the categorization stage

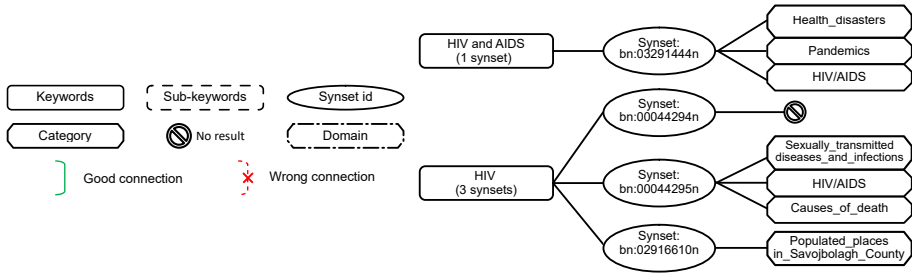


Fig. 2: Simple connection by categories

Improvement. In order to categorize more articles than the original approach, in other words, increasing the article coverage, less restrictions are applied during the synset connections. For this part, three points are taken into consideration: lemmatization, synsets connections by domains, and an article's categories.

Lemmatization is the process of reducing the different forms of a word to one single form, commonly called a lemma. For example, play is the lemma of "playing, played, plays". Obtaining generic keywords which may have a greater chance of being indexed in BabelNet is a good solution. Qazanfari et al. [26] showed that using lemmas improved the precision and accuracy of their recommendation system. That is why lemmatization is included in both search approaches (i.e., exact and further). Indeed, the initial keyword set has the first chance to give results and lemmatization is attempted in a second chance in the case of an empty result set. Let us focus on the keyword "Software development processes" to demonstrate the benefit of this feature, as it is not indexed in BabelNet and therefore, nothing is returned using the exact search. At the this point, the original method launches the further search with the sub-keywords "Software development", "Software processes" and "development processes" and gives the set of categories C_O. This approach lemmatizes the input before trying the exact search again with "Software development process" which retrieves the categories C_L. Even if the retrieved categories are close, a deviation is noticeable

with the ones from C_O and, indeed, *Marketing* digresses from the main topic.
$C_O = \{Software_project_management, Software_development, Project_management, Product_development, Marketing, Computer_occupations\}$
$C_L = \{Software_development_process, Formal_methods, Methodology, Software_engineering\}$

In the original method, only categories were involved in the synset connection process because the focus was on precision. Yet, the generality of the domains [5] can be used as an advantage because it is more probable that articles share common domains. This implies that there are more potential articles with data. Let us focus on an example, an article with the keywords *"kerosene reforming, novel combustion technologies, hydrogen assisted combustion"* does not have connected synsets by categories. If domains are taken in consideration (Fig. 3), there are connections, thus, some synsets are validated (e.g., bn:00020144n, bn:00014024n) and added to the set A_S, which regroups all the synsets of the article A.

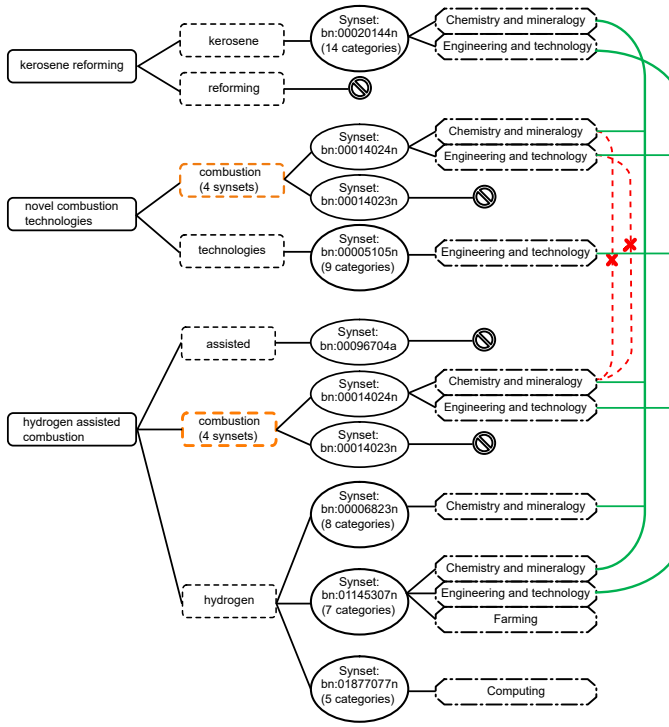


Fig. 3: Domains connection

[5] BabelNet has 34 domains (e.g., "Computing", "Astronomy") and many categories that are mostly inherited from Wikipedia.

The last transformation of the original method concerns the category attribution. An article's categories have an important impact on the data augmentation phase. For the purpose of increasing an article's coverage, it was decided to assign all categories from connected synsets to the article (Eq. 2). For example, the categories "Sexually_transmitted_diseases_and_infections", "Health_disasters", "Pandemics", "Causes_of_death" will be added to "HIV/AIDS" for the article in Fig. 2.

$$\begin{gathered} \text{Let } s_1, s_2 \in A_S \\ A_C = \{c \mid c \in (s_1.C \cup s_2.C)\} \end{gathered} \tag{2}$$

where A_C is all the categories of the article A.

3.2 Related articles

Finding related articles is the second step of the proposed process, where data inherited from categorization are exploited. It is divided into two steps: 1) data augmentation and 2) similarity measurement.

Data augmentation. Data augmentation is necessary because disambiguated words might be very specific and thus, rare in the corpus. Therefore, neighbors such as synonyms and hypernyms are extracted from BabelNet's knowledge base to expand matching possibilities with other articles. Yet, to avoid bringing unrelated neighbors, they are only selected if they share at least one category with the article.

Similarity measurement. To determine how similar two articles are, a similarity equation (Eq. 3) was defined [16] based on the three different ways to connect articles. A similarity measurement between sets is computed with weighted Jaccard indexes.

$$\begin{aligned} sim(A_i, A_j) = \frac{1}{\alpha+\beta+\gamma} * \Bigg(& \alpha\ jac(K_i, K_j) + \frac{\beta}{2}\ jacKN(K_i, N_j, K_j) \\ & + \frac{\beta}{2}\ jacKN(K_j, N_i, K_i) + \gamma\ jacNN(N_i, N_j, K_i, K_j) \Bigg) \end{aligned} \tag{3}$$

where:
- K_x is the set of keywords' synsets of the article A_x
- N_x is the set of neighbors' synsets of the article A_x
- $jac()$, $jacKN()$ and $jacNN()$ are three Jaccard index variants, respectively defined in Eq. 4, Eq. 5 and Eq. 6

There are three different ways to connect articles together which use keywords' synsets extracted from both categorization and data augmentation:

1. *Keyword intersection*: Articles share the same keywords' synsets, and this is the obvious and most reliable connection.

$$jac(K_i, K_j) = \frac{|K_i \cap K_j|}{|K_i \cup K_j|} \tag{4}$$

2. *Keyword-Neighbor intersection*: Keywords' synsets of the first article belong to neighbors of the second article and vice-versa. This connection is considered as moderately reliable.

$$jacKN(K_i, N_j, K_j) = \frac{|K_i \cap N_j|}{|K_i \cup N_j| - |K_i \cap K_j|} \tag{5}$$

3. *Neighbor intersection*: Articles share the same neighbors; this is the farthest and least reliable connection.

$$jacNN(N_i, N_j, K_i, K_j) = \frac{|N_i \cap N_j|}{|N_i \cup N_j| - (|K_i \cap N_j| + |N_i \cap K_j|)} \tag{6}$$

An heuristic analysis of the Eq. 3 was realized to evaluate the impact of the coefficients α, β, and γ. This showed that as the three ways to connect articles together do not have the same confidence, maximizing α will provide more accurate results than the maximization of β and γ because the keyword intersection is the most reliable. Knowing that, the values of 4, 2 and 1 were, respectively, chosen for α, β, and γ for the rest of this article [16].

4 Evaluation

4.1 Dataset

The analysis presented in this article was performed using *Web of Science Dataset WOS-46985* from Kowsari et al. [14], which has been specifically used for text classification [15,7]. This dataset contains 46,985 articles from *Web of Science* divided into seven domains and 134 categories. For the rest of the evaluation, only domains were taken into consideration because the proximity of the dataset's categories might lead to high overlapping.

4.2 Metric

Eye tracking studies on user behavior regarding ranked results of a web search [6,10] showed that the higher the rank, the less attention is paid by the user to consult this suggestion. Hence, given the number of scientific articles in the literature, precision at k (P@k) is a suitable metric to evaluate this method, as the focus is on the first k elements which are considered as the most similar. Indeed, it is improbable that a user wants to read all similar articles retrieved in the literature. P@k is defined as follows:

$$\text{P@k} = \frac{\#Relevant\ articles\ in\ top_k}{k} \tag{7}$$

Articles sharing the same domain were called relevant articles; the maximum value of k was experimentally set to 100. Yet, this metric has a weak point: the dependency on k [19]. Let us say that k equals 10 and the method proposes only six related articles even if these six articles are relevant, the precision would not be 1 but 0.6. In order to minimize this, a customized top_k, called "linked top_k" was created. The aim is to increase the number of relevant articles retrieved using their top_k (Fig. 4).

$$top_k = \{w_i, ..., w_k\} \quad 1 \leq i \leq k$$
$$\text{where } w_i \text{ is the weight of the } i^{th} \text{ related article} \tag{8}$$

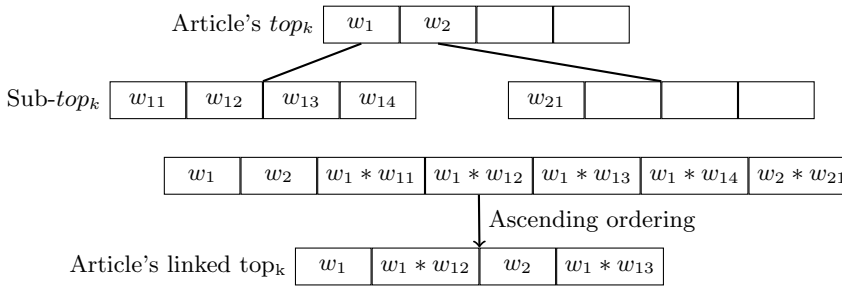


Fig. 4: Article's linked top_k, with $k = 4$

4.3 Word2vec and cosine similarity (w2v-cos)

Mikolov et al. [21] introduced word2vec as a machine learning approach using a multi-layer neural network to learn semantic word proximity. The models trained by word2vec learning techniques represent the syntactic probabilities of words' co-occurrence and can be used to predict the next words in a sequence. For the experiment, the model trained[6] on a part of Google News dataset (100 billion words) was employed. It contains 300-dimensional vectors for 3 million words and phrases. These vectors are exploited to compute the similarity between two words.

In this case, each split keyword is passed to the network and the 300-dimension vectors, in return, are averaged to obtain a mean vector and then stored. To determine the similarity between two articles the cosine similarity measure (Eq. 9) was chosen.

$$sim_c = \frac{V_i \cdot V_j}{|V_i| \times |V_j|}$$
$$\text{where } V_x \text{ is the mean vector of the article } A_x. \tag{9}$$

[6] https://code.google.com/archive/p/word2vec/

The cosine is defined in the interval $[-1, 1]$ but only the positive space $[0, 1]$ is used in this experiment because all components of article vectors are non-negative [13]. Two articles with a cosine similarity of 1 are considered as the same while a similarity of 0 means no correlation between them.

4.4 Analysis

The first notable point is that w2v-cos is largely outperformed by the other approaches. Indeed, it stagnates around a precision of 19% for k between 1 and 100 (Fig. 5a) while the original approach has an average precision of 25% and the proposed approach reaches 49%. The article coverage as well as the permissiveness of the approach during the categorization step can justify the wide difference between the original and proposed approaches.

As explained earlier, P@k has limits, especially its dependency on the k value, and thus, the original approach is penalized more than the proposed approach due to the coverage (46.8% against 88.2% for the proposed approach). The w2v-cos approach is less impacted; indeed, there are only 11 articles without vector representation. In order to minimize this dependency, a second version of P@k was created (Eq.10), which permits the precision to be evaluated on real proposed articles. With this new metric, the precision is not biased by the difference between k and the number of articles in top_k, as is the case with P@k.

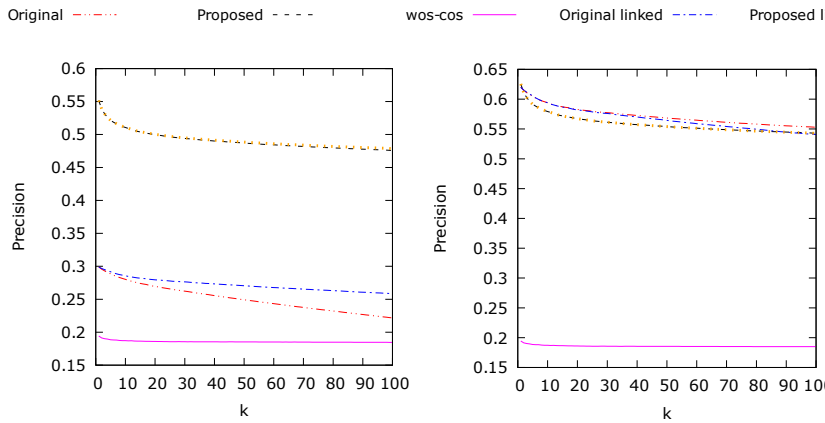
$$\text{P'@k} = \frac{\#\textit{Relevant articles in } top_k}{\text{Number of articles in } top_k} \tag{10}$$

Fig. 5b illustrates that the usage of this new metric, the w2v-cos curve (i.e., green) is the same as in Fig. 5a because this approach always proposes the maximum number of articles, except for the 11 articles without vector representation. Concerning the two others, Fig. 5b shows a slightly advantage for the original approach; yet, it is necessary to nuance with the total number of pairs retrieved (Table 1). In fact, P'@100 is nearly the same but the proposed approach retrieved 4,108,432 pairs, which is more than twice the number retrieved by the original approach.

As expected, using linked top_k increases the number of proposed pairs and slightly increases P@k (Fig. 5a). The influence of linked top_k is more remarkable for the original approach; indeed, it lowered the number of missing pairs by

Approach	P@100	P'@100	# proposed pairs	# correct pairs	# missing pairs
w2v-cos	0.185	0.185	**4,697,400**	867,377	**1,100**
Original	0.222	**0.553**	1,831,027	1,041,567	2,867,473
Original-linked	0.259	0.541	2,248,478	1,215,297	2,450,022
Proposed	0.475	0.543	4,108,432	2,234,326	590,068
Proposed-linked	**0.478**	0.542	4,142,957	**2,246,657**	555,543

Table 1: Comparison of retrieved pairs with k=100, best values are in bold



(a) P@k (b) P'@k

Fig. 5: Curves comparing the different approaches, suffix linked for the approaches using linked top_k

14.6 points. Missing pairs represent the difference between the expected number of pairs (i.e., # articles * k) and the number of pairs the approach is able to propose. Yet, this linking tends to bring more wrong pairs than good ones, as demonstrated with a subtle decrease of P'@100: 1.2 points for the original approach and 0.2 points for the proposed one. In fact, using top_k of the relevant articles as an intermediary for the construction of linked top_k (Fig. 4) introduces a loss of confidence in the proposed articles. Therefore, if the focus is on quantity, applying this linking can be a good compromise because the loss of precision is negligible.

As presented in Fig. 5b, the precision of the different approaches can be improved. The first thing to notice is the proximity of certain domains, such as *Medical* and *Psychology*. This proximity is reflected in Table 2, which presents the first seven combinations most found for k between 1 and 100. Indeed, for each approach, the combination Medical/Psychology appeared in the first three positions. Let us exemplify this with a concrete case using the original approach, where K_i is in the *Medical* domain and K_j is in *Psychology*:

- $K_i = \{$***Alzheimer's disease***, *cerebrovascular disease,* ***dementia****, estrogen, menopause, prevention*$\}$

- $K_j = \{$***Alzheimer's disease****, nonverbal communication, emotional prosody, behavioral and psychological symptoms of* ***dementia*** *(BPSD)*$\}$

Despite the fact that these articles have distinct domains, their keywords (K_i, K_j) are quite similar. Hence, after the categorization step, they have the same synsets and are considered as similar, which makes sense in reality. On the contrary, certain associations seem illogical, like the first set of keywords K_i in *Psychology* and the second one K_j in *MAE*:

- $K_i = \{$*Facial,* ***Emotion****, Lateralisation, Stroke,* ***Perception***$\}$

- K_j = {*Battery electrical vehicles, Repertory grid technique, Comparison of modes of transport, Subjective* ***perception***, ***Emotions***}

The categorization step for these sets in the original approach finds the same synsets coming from perception and emotions and thus, the similarity is 1. In that case, using the proposed approach, or w2v-cos, which is more permissive, does not assign a similarity of 1; indeed, more keywords have data and so the disambiguation is better. Hence, the notion of dissimilarity between these sets is introduced, which makes sense.

Moreover, the domain overlapping of w2v-cos shows that this approach has trouble distinguishing *Medical* from all other domains. This can explain such a gap in terms of the precision between this approach and the other two. For example, the two following sets of keywords (K_i, K_j) related to the fields of medicine and electrical engineering are considered as being highly similar even though there is no obvious relation. Moreover, no relation is found using the other two approaches, thus qualifying these sets as uncorrelated.

- K_i = {*HCV, Flaviviridae, epidemiology, Saudi Arabia*}

- K_j = {*Electric machines, Machine control, Magnetic losses, Multilevel systems, Physics-based modeling, Power system simulation, System analysis*}

This example highlights a limit of the w2v-cos approach; indeed, the model assigns at least one vector for 99.9% of the articles using a probabilistic method [21]. This type of method can be seen as a black box; thus, the method to build vectors is unclear, which complicates the understanding of such incongruous associations.

Original		Proposed		w2v-cos	
Medical/Medical	20,074,800	Medical/Medical	46,084,918	Medical/Medical	**2**2,149,123
Psychology/Psychology	11,383,939	Psychology/Medical	18,407,072	CS/Medical	19,929,028
Psychology/Medical	9,754,968	Psychology/Psychology	17,936,117	Medical/Psychology	19,578,757
Biochemistry/Medical	7,119,827	Medical/Biochemistry	16,363,304	Medical/ECE	18,017,840
ECE/ECE	6,368,952	CS/CS	15,058,490	Biochemistry/Medical	17,761,059
Biochemistry/Biochemistry	6,173,909	ECE/ECE	12,009,851	Medical/Civil	13,602,961
CS/CS	5,639,581	Biochemistry/Biochemistry	11,268,296	Medical/MAE	10,052,781

Table 2: Most found domain combinations with the number of occurrences for each approach

5 Discussion

As stated previously in Section 4.4, the proposed approach outperforms w2v-cos used as a baseline; yet, the precision can be improved. A possible solution to increase the precision would be to introduce a threshold. The similarity is defined as between 0 and 1, with 0 corresponding to no correlation. Knowing that, the study of the distance[7] distribution of top_k (Fig. 6) can bring relevant information towards finding a critical distance. As shown in Fig. 6, as the distance

[7] distance = 1 - similarity

increases, the proportion of wrong pairs increases while the precision decreases. Thus, finding the distance where the wrong predictions start to represent more than 50% of the proposed pairs and filtering the pairs with a higher distance in articles' top_k might improve the precision. However, this solution implies a loss of proposed articles.

In this case, the proposed approach has an average of 55.8% for P'@k ($k \in [1, 100]$). The critical distance is reached at 0.94, where good predictions represent 51.4% of the proposed pairs. At this point, filtering will decrease the number of proposed pairs by 20.4 points. This drop goes along with an augmentation of P'@k average to 56.6%. In this context, the difference between the proposed approach and its filtered version is marginal and it shows that introducing a threshold impacts the precision. Given this, more variants of the threshold selection need to be tested such that there will be more information regarding its influence and the user can choose whether to accept this compromise.

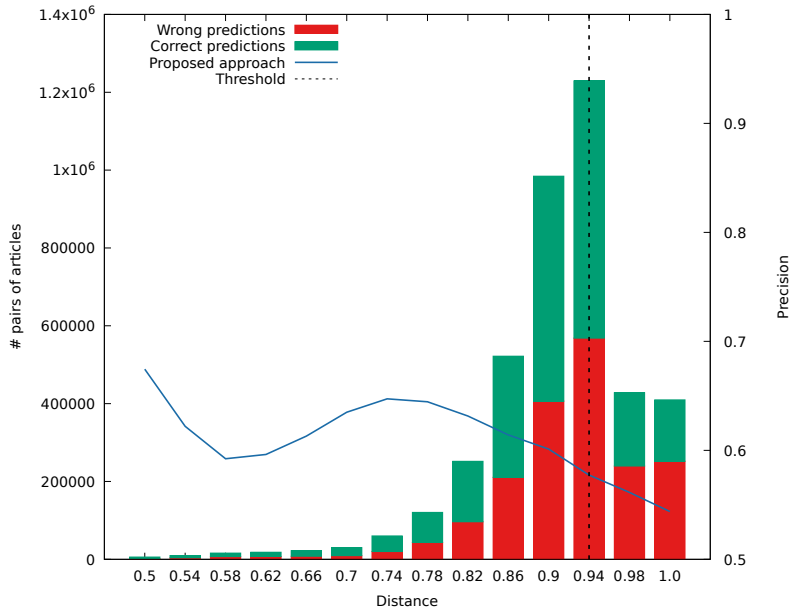


Fig. 6: Distance distribution and precision of top_k using the proposed approach.

6 Conclusion

The aim of this paper was to present a process of suggesting related scientific articles. This process is composed of two major steps: the categorization and the suggestion using a similarity measure. The objective of increasing the coverage

compared to the original approach was completed. Indeed, the coverage reaches 88.2% compared to 48.6% for the original approach. However, on this point, w2v-cos is still superior but it is outperformed by the proposed approach in terms of precision.

The experiments permitted us to establish that the proposed approach can compete against probabilistic methods such as baseline w2v-cos. The analysis highlights that word sense disambiguation is more efficient in the proposed approach, leading to a much better precision than w2v-cos.

The proposed approach provides promising results and improves upon the original one. In the future, the reproducibility of this approach could be evaluated using another dataset. Moreover, to support the previous assumption, a comparison with other probabilistic approaches such as the binary independence retrieval model [3] will be done in future work.

References

1. Ensan, F., Bagheri, E.: Document retrieval model through semantic linking. In: WSDM. pp. 181–190. ACM (2017)
2. Firth, J.G.: A synopsis of linguistic theory 1930-1955 in studies in linguistic analysis. Oxford (1962)
3. Fuhr, N.: Probabilistic models in information retrieval. The computer journal **35**(3), 243–255 (1992)
4. Garfield, E.: current eamments. Current contents **32**, 3–7 (1990)
5. Gil-Leiva, I., Alonso-Arroyo, A.: Keywords given by authors of scientific articles in database descriptors. Journal of the American society for information science and technology **58**(8), 1175–1187 (2007)
6. Guan, Z., Cutrell, E.: An eye tracking study of the effect of target rank on web search. In: SIGCHI. pp. 417–420. ACM (2007)
7. Heidarysafa, M., Kowsari, K., Brown, D.E., Meimandi, K.J., Barnes, L.E.: An improvement of data classification using random multimodel deep learning (rmdl). International Journal of Machine Learning and Computing **8**(4) (2018)
8. Hotho, A., Nürnberger, A., Paass, G.: A brief survey of text mining. LDV Forum **20**, 19–62 (2005)
9. Huang, A.: Similarity measures for text document clustering. In: NZCSRSC. vol. 4, pp. 9–56 (2008)
10. Joachims, T., Granka, L.A., Pan, B., Hembrooke, H., Gay, G.: Accurately interpreting clickthrough data as implicit feedback. In: SIGIR. vol. 5, pp. 154–161 (2005)
11. Johnson, R., Watkinson, A., Mabe, M.: The STM Report: An overview of scientific and scholarly publishing (2018), https://www.stm-assoc.org/2018_10_04_STM_Report_2018.pdf
12. Kanakia, A., Shen, Z., Eide, D., Wang, K.: A scalable hybrid research paper recommender system for microsoft academic. In: The World Wide Web Conference. pp. 2893–2899. ACM (2019)
13. Korenius, T., Laurikkala, J., Juhola, M.: On principal component analysis, cosine and euclidean measures in information retrieval. Information Sciences **177**(22), 4893–4905 (2007)
14. Kowsari, K., Brown, D.E., Heidarysafa, M., Meimandi, K.J., Gerber, M.S., Barnes, L.E.: Hdltex: Hierarchical deep learning for text classification. In: ICMLA. pp. 364–371. IEEE (2017)

15. Kowsari, K., Heidarysafa, M., Brown, D.E., Meimandi, K.J., Barnes, L.E.: Rmdl: Random multimodel deep learning for classification. In: ICISDM. pp. 19–28. ACM (2018)
16. Latard, B.: Scientific Search Engines: From the Categorization to the Information Retrieval. Ph.D. thesis, Universit de Haute-Alsace (2019)
17. Latard, B., Weber, J., Forestier, G., Hassenforder, M.: Towards a Semantic Search Engine for Scientific Articles. In: TPDL. pp. 608–611. Springer (2017)
18. Latard, B., Weber, J., Forestier, G., Hassenforder, M.: Using semantic relations between keywords to categorize articles from scientific literature. In: ICTAI. pp. 260–264. IEEE (2017)
19. Manning, C., Raghavan, P., Schütze, H.: Introduction to information retrieval. Natural Language Engineering **16**(1), 100–103 (2010)
20. Menaka, S., Radha, N.: Text classification using keyword extraction technique. International Journal of Advanced Research in Computer Science and Software Engineering **3**(12), 734–740 (2013)
21. Mikolov, T., Chen, K., Corrado, G.S., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
22. Miller, G.A.: Wordnet: a lexical database for english. Communications of the ACM **38**(11), 39–41 (1995)
23. Navigli, R.: Word sense disambiguation: A survey. ACM Comput. Surv. **41**, 10:1–10:69 (2009)
24. Navigli, R., Ponzetto, S.P.: Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence **193**, 217–250 (2012)
25. Pain, E.: How to keep up with the scientific literature (2016), https://www.sciencemag.org/careers/2016/11/how-keep-scientific-literature
26. Qazanfari, K., Youssef, A., Keane, K., Nelson, J.: A novel recommendation system to match college events and groups to students. IOP Conference Series: Materials Science and Engineering **261**(1), 1–15 (2017)
27. Salatino, A.A., Osborne, F., Thanapalasingam, T., Motta, E.: The CSO Classifier: Ontology-Driven Detection of Research Topics in Scholarly Articles. In: TPDL. vol. 11799, p. 296. Springer (2019)
28. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Commun. ACM **18**, 613–620 (1975)
29. Shehata, S.: A wordnet-based semantic model for enhancing text clustering. In: ICDM. pp. 477–482. IEEE (2009)
30. Shemilt, I., Simon, A., Hollands, G.J., Marteau, T.M., Ogilvie, D., O'Mara-Eves, A., Kelly, M.P., Thomas, J.: Pinpointing needles in giant haystacks: use of text mining to reduce impractical screening workload in extremely large scoping reviews. Research Synthesis Methods **5**(1), 31–49 (2014)