Automatic role discovery of scientific articles from a dynamic citation network enriched by the citation context *

Gérome Ferrand[1],
supervised by: Rémy Cazabet[1,2], and Marc Bertin[1,2]

[1] Claude Bernard Lyon 1 University, France
[2] IXXI Institute, 46 allée d'Italie, 69007 LYON, France
http://www.ixxi.fr

Abstract. Many studies propose to examine the scientific literature through citation networks. Among these studies, we can find the classification of scientific publications, which can be based on the analysis of the content of the articles or on their citations. Our contribution is essentially focused on the citations and the semantic complements provided by the contexts of these citations (such as the -IMRaD- sections and sentences containing the citations or the verbs used). In order to identify a typology of scientific articles, we embed the citation contexts in a vector space using Shallow Networks learning. In a second step, we express the articles (i.e. nodes) by a vector computed on the basis of the citation embedding vectors (i.e. links). Then, we perform a clustering on the nodes' vectors in order to group the articles having a semantic proximity, thus forming article classes. We performed this method on a subset of a scientific article dataset from the Public Library of Science (PLoS) open access database. This subset is composed of xx xxx articles and our method proposes x classes of scientific publications. The validation method shows that our procedure is able to exclude notions related to the field of application of the study (neuroscience, biochemistry, etc.) and favours more the intrinsic contribution to the type of the article (meta-analysis, survey, etc.).

Keywords: Scientometrics · Clustering · Embedding.

Abstract. De nombreux travaux proposent d'étudier la littérature scientifique par le biais de réseaux de citations. Parmi ces études, nous retrouvons la classification des publications scientifique, qui peut se faire sur la base de l'analyse du contenu des articles ou sur leurs citations. Notre contribution se fonde essentiellement sur les citations et les compléments sémantiques apportés par les contextes de ces citations (tel que les sections -IMRaD- et les phrases qui contiennent les citations ou encore les verbes utilisés). Afin d'identifier une typologie d'articles scientifique, nous intégrons les contextes de citation dans un espace vectoriel grâce

* Supported by organization IXXI

à l'apprentissage de Shallow Networks. Dans un second temps, nous exprimons les articles (ie: noeuds) par un vecteur calculé sur la base des vecteurs de citations (ie: liens). Ensuite, nous procedons à un clustering sur les vecteurs des noeuds afin de regrouper les articles ayant une proximité sémantiques, formant ainsi des classes d'article. Nous avons effectué cette méthode sur un sous-ensemble d'un jeu de donnée d'article scientifique issue de la base de données en open access de la Public Library of Science (PLoS). Ce sous-ensemble est composé de xx xxx articles et notre méthode propose x classes de publications scientifiques. La méthode de validation montre que notre procédure parvient à exclure les notions relatives au domaine d'application de l'étude (neuroscience, biochimie, etc.) et favorise davantage l'apport intrinsèque au type de l'article (méta-analyse, survey, etc.).

Keywords: Scientométrie · Clustering · Embedding.

1 Introduction

1.1 Background and Motivations

More and more scientific articles are published digitally each year. According to the Web of Science Core Collection database, 3,739,636 articles published in 2021 are available, compared to 3,100,263 articles in 2016. This represents an increase of a factor of 1.2 over 5 years. In addition to the development of the scientific community and its productions, this can also be explained by an increase in the indexing of scientific products, which makes it difficult to automatically evaluate this ever-growing network of information on a large scale. Article classification is one of the types of studies that have attracted a lot of attention in recent years, with the use of various methods. Many of them refer to network-based modelling ([TR21], [Saj+21]).

The bibliographic reference system is at the heart of the network structure of scientific productions. Indeed, any scientific publication is part of a set of publications and connects between them through citations to form a citation network. At the micro level, this makes it possible to trace the source and obtain additional information on the facts put forward. At the meso-scopic level, we can quantitatively evaluate the impact of an article according to its number of citations (i.e. incoming degree) in the network and by comparing this value with other articles in its discipline [Gar55]. Finally, from a macro point of view, these references make it possible to evaluate the evolution and structure of Science as proposed by many studies, such as [For+18], [CC13], [MJF15].

Most of the studies (and in particular those proposed above) propose to analyse science in several dimensions; with data directly from the text of the article, from the links that unite them via bibliographic references, with the temporal aspect provided by the dates of publication, the authors, the journals, etc. However, few of them use data relating to citation contexts, which also offer a qualitative contribution to the relations between articles, thus making

it possible to supplement the quantitative approaches that are naturally more frequent in the literature on science metrics.

The citation contexts available to us correspond to the sections in which the citations appear, the sentences containing the citations, or the verbs used in the citation sentences. The labelling of the links (i.e. the citations) by these citation contexts thus offers a semantic complement on the reasons which pushed an author to cite an article rather than another. Indeed, our contribution aims at proposing a tool to strip away the semantics related to the theme of the article in favour of the influence and role of the article at a local level.

We construct a citation graph, with the sentences or verbs of the citations as labels. We then embed these word vectors in a shallow network in order to obtain real number vectors. These vectors are then used to determine vectors defining the research papers. Finally a clustering is applied to these clusters in order to provide an article topology.

In order to present our approach in more detail, the rest of this report is organised as follows: to conclude this first section, we develop a state of the art in order to present related work. The method on automatic scientific article role discovery is summarised in section 2. The results and the evaluation of this method will be presented in section 3. Finally, a concluding discussion is proposed in section 4.

1.2 State-of-the-art

Citation Proximity Analysis In 2009, Gipp et al. proposed a method called co-citation proximity analysis (CPA) whose aim is to determine a measure of semantic proximity between scientific publications on the basis of the position of their citations in the text [GB09].

This similarity measure is based on the assumption that, in the full text of a document, documents cited close to each other tend to be more strongly related than documents cited further away.

In the preliminary studies for our method, we propose a slightly different approach according to which 2 cited documents are semantically close if they are cited within the same IMRaD section.

TF-IDF and the frequency analysis methods In the field of frequency analysis, 2 particularly used methods are opposed. The bag-of-word method consists in representing a text by the frequency of the words that compose it. The N-grams method consists of evaluating the probability of occurrence of the sequences of words in the text.

Another method, called TF-IDF, represents the text based on the principle that the importance of a word increases proportionally to its frequency of occurrence, but is compensated by the frequency of occurrence of the word in the text corpus [SPA72].

Doc2Vec There are several models based on shallow networks (i.e. neural networks with only one hidden layer). The self-supervised method word2vec is the original one which allowed the development of many other derived self-supervised methods [Mik+13]. This model can use the Continuous Bag Of Word (CBOW) or skip-gram artchitectures for learning (see Fig. 1). The CBOW architecture predicts a word based on the contextual words given as input (window). The skip-gram architecture predicts the most frequent context words based on an input word. These networks can be concluded with a Softmax activation function for classification. Once trained, the model learns to project words into a vector space, where nearby words have semantic proximity.

The Doc2Vec model derives from the Word2Vec model, but takes as input an identifier of the target document (an article in our case) as an additional word in the contextual window [LM14]. Doc2Vec can therefore create vectors representing words but also vectors of scientific articles.

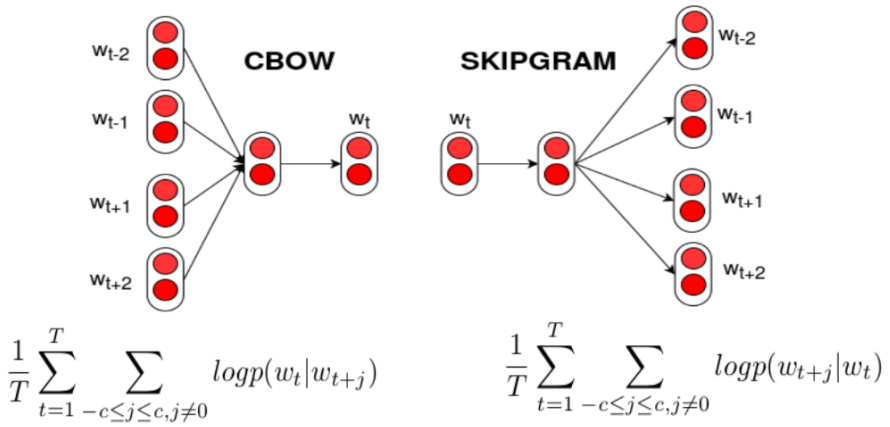


Fig. 1: CBOW and Skip-Gram architectures.

2 Method

2.1 Dataset

To produce this method, we based ourselves on the open access articles published by the Public Library of Science (PLoS)[3]. This corpus of articles gathered 263 548 files (i.e. 26.7 GB) encoded according to the *Journal Article Tag Suite* (JATS) technical standard[4], which is a specific XML format allowing to describe scientific literature published online.

[3] PLoS Open Access data: https://plos.org/text-and-data-mining/.
[4] JATS specifications available here: https://jats.nlm.nih.gov/.

Prior to our work, a first step was to process this data to store it in a SQL database in order to extract the different entities such as articles, journals, authors, bibliographic references, abstracts, titles, sections, etc.

A second step was to query the SQL server to reconstruct the articles we are interested in - see subset part - on the computing machine in JSON format. This step is not essential, but it allows us to reduce the access time to the data and to keep only the information strictly necessary for our objective, which is to build the citation network.

In this section we describe the datasets and subsets of data that we have generated and then used during our work. These explanations allow us to grasp certain details in the rest of this report, but also to identify possible bias in case these datasets could be reused in future work.

PTP subset All of this work was carried out on a (personal) machine which is accessible to the public and therefore has limited computing power. Consequently, the number of entities in our dataset had to be reduced.

We have kept only PLoS articles that cite PLoS articles and those that are cited by PLoS articles. We call this subset *Plos To Plos*, with the acronym *PTP* (see Fig. 2 for a better understanding.). In other words, not all scientific articles are present in our subset. Articles from different journals than PLoS are not included in our subset. Any article that is not directly related to at least one other PLoS article is excluded from the PTP subset.

This reduction in the volume of articles allowed us to have a denser underlying network than with the initially available set of articles. Indeed, the presence of an article within the PTP subset implies that it is at least connected to another node of the network, therefore this dataset excludes all isolated nodes.

Furthermore, PTP maximises the information about each node, as all the data of a node in the network is (in theory) present in the database. This avoids having DOIs without information. A DOI[5] (Digital Object Identifier) is a unique identifier for each article and therefore for each node. Each article thus has a file in JSON format with its metadata according to the schema in Fig. 3. The JSON format allows faster access to the necessary data; each file corresponds to a node and therefore follows the intrinsic structure of our network modelling, unlike the SQL database, which is unstructured with respect to our modelling. This storage method provides a good compromise between accessibility and flexibility of the schema at the scale of our dataset, while maintaining granularity and scalability.

PTP composition and network topology The PTP subset has 172,877 nodes, and 317,069 links. But the articles in this dataset have a total of 6,641,892 outcomming degree. Therefore, we can claim that the PTP dataset represents at most 4.77% of the real network (by the ratio between the 2 previous node values), noting that this represents an upper bound on the true ratio. Indeed,

[5] DOI specifications available here: https://www.doi.org/doi_handbook/pub_agreements/DOICoreSpecificationv1.pdf.

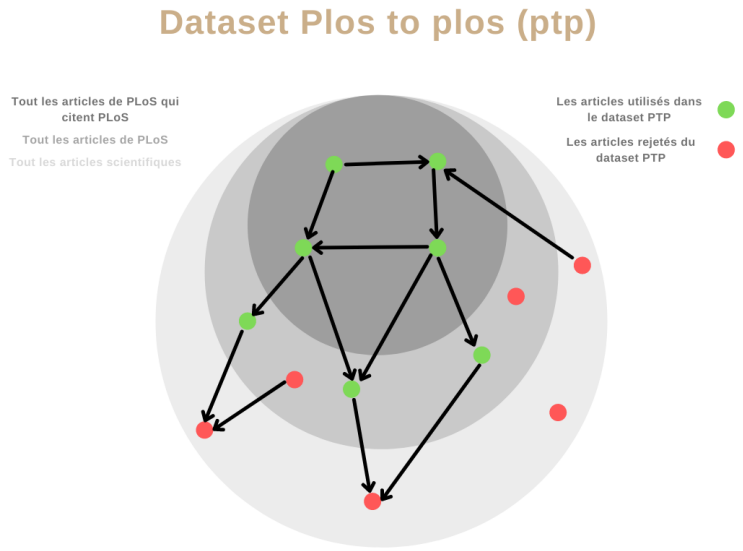


Fig. 2: A simple diagram to illustrate which articles are present within the PTP subset.

we should also count the nodes that are isolated from the network, and thus not cited by the PLoS papers, etc.

The articles in this dataset are distributed over a period from 2003 to 2019 (see Fig. 4). A slight anomaly can be noticed during the year 2015: the volume of articles seems abnormally low compared to the other years. But we will ignore this phenomenon.

Articles are published in the main PLoS journals but much more frequently in the *PLoS One* journal (see Fig. 5).

The articles deal mainly with biology-related topics (see Fig. 6), although some of them do not provide any information on topics, and are therefore not classified in this way.

Finally, concerning the citations, we notice that most of the citations are found in the introduction and discussion sections, followed by the method section, to the detriment of the result section, which remains a part in which there are the least citations (see Fig. 7).

Structurally, this network is a DAG, which means that it is directed and acyclic (only theoretically because we have observed in practice the presence of loops, and therefore of mutual citations). We observe an average degree of 2.42 in the undirected network, and thus 1.21 in/out-degree in the directed network. A density of the order of 1e-05, and 118,532 related components.

This network is supposed to be very close to what is called a scale free network [Sol65], satisfying the following equation:

```
{"doi": "10.1371/journal.pone.0155720",
    "type": "Research Article",
    "title": "The Effectiveness of Mindfulness-Based...",
    "abstract": "Perinatal mental health difficulties...",
    "journal": "plos one",
    "year": 2016,
    "subjects": ["Medicine and health sciences", ...],
    "nbRef": 75,
    "authors": [{
            "name": "Billie",
            "surname": "Lever Taylor"
        }, ...],
    "references": [{
            "doi": "10.1371/journal.pone.0096110",
            "imradLocs": ["I"],
            "sentences": ["There is also good empirical
                evidence..."],
            "ctxVerbs": [["be", ...]]
        }]
```

Fig. 3: Article schema stored in JSON format. Each article has a unique DOI to identify it, the type of the article, a title, an abstract, its journal (PLoS has different journals), its year of publication, the list of topics, the total number of references of the article (i.e. length of the articles cited in the bibliography), its author list and the list of references to PLoS articles. Each reference has the DOI for the corresponding PLoS article, the section where the citation appears, the sentence in which the citation is found and the infinitive verb extracted from the previous sentence.

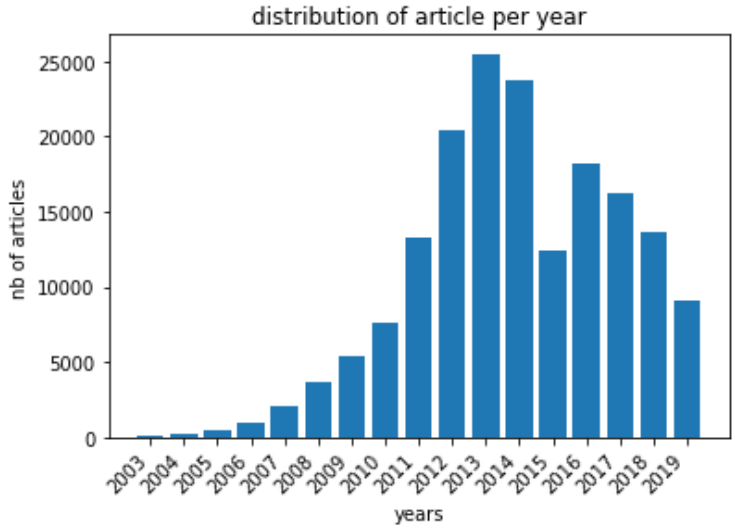


Fig. 4: Distribution of article in the PTP dataset per year.

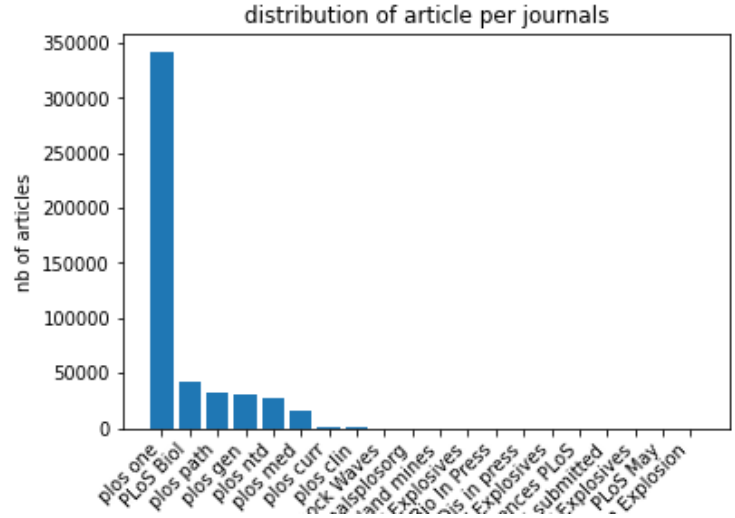


Fig. 5: Distribution of article in the PTP dataset per journal.

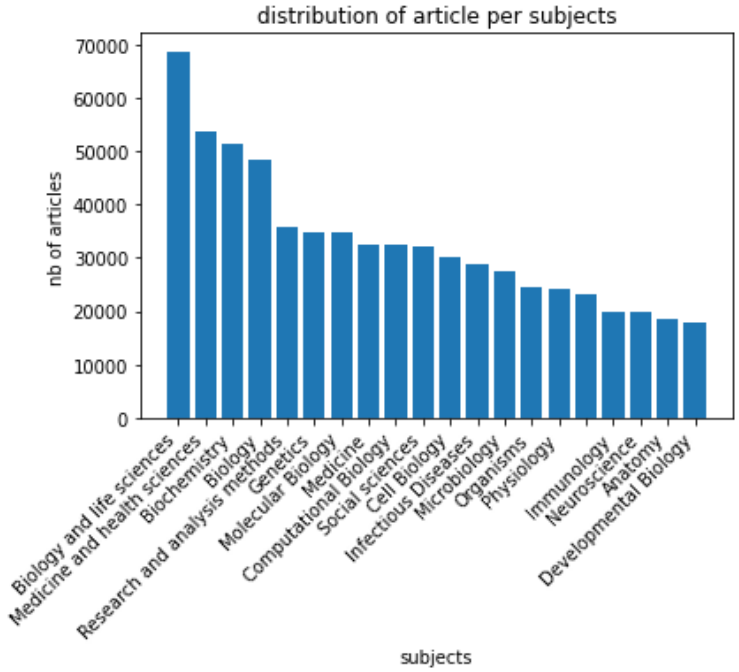


Fig. 6: Distribution of article in the PTP dataset per topic.

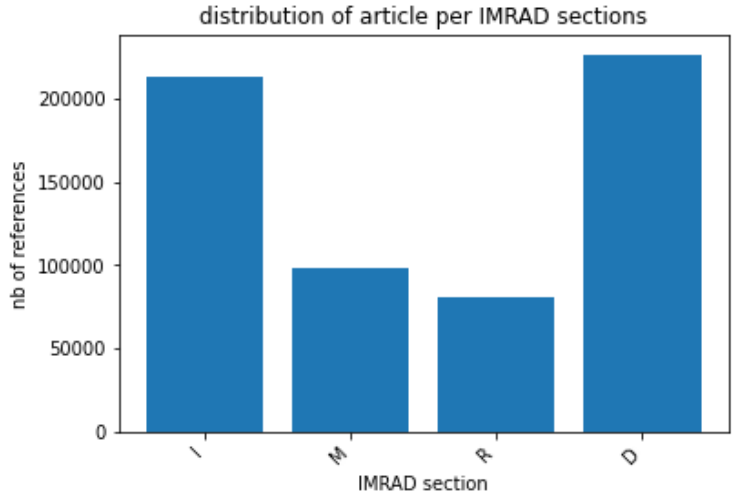


Fig. 7: Distribution of citations in the PTP dataset per section.

$$P(k) = k^{\gamma} \tag{1}$$

with gamma generally bounded between 2 and 3 [BB03] (see Fig. 8).

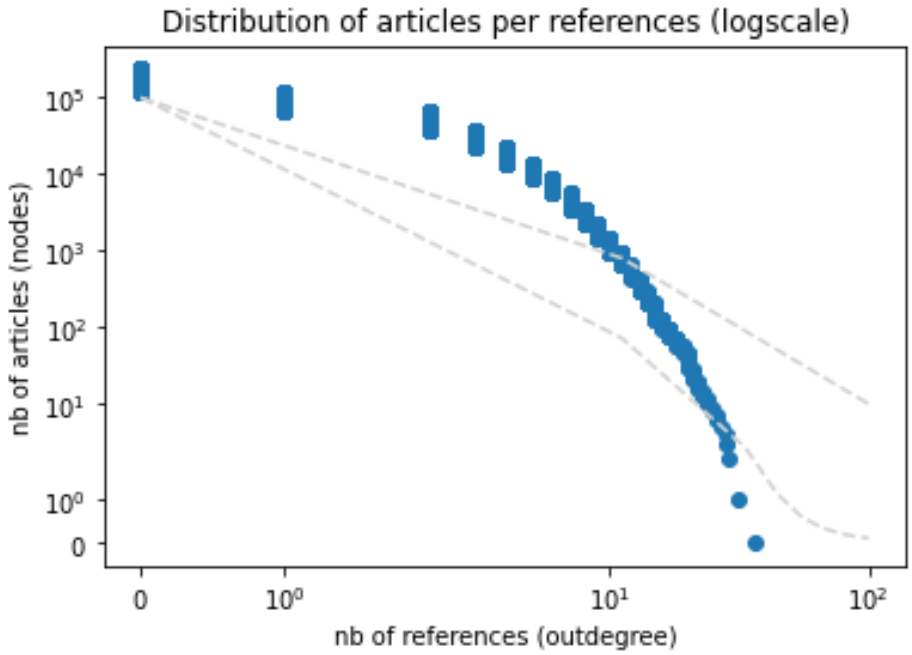


Fig. 8: Distribution of number of article in the PTP dataset per number of (directed) reference.

In practice, our distribution of the number of articles per number of references tends towards this limit without being perfectly matched. We suppose that this is due to the multiple points that we have previously mentioned about the modifications made to the dataset.

K-Core subsets Some preliminary studies have been carried out on a subset of the PTP dataset. This subset was selected on the basis of the K-cores of the PTP citation network. The k-core (core of order k) of a graph G(V, E) is the largest subgraph $SG(V^{SG}, E^{SG})$ such that all nodes have at least a degree k, i.e.

$$\forall u \in V^{SG}, k_u^{SG} \leq k, \tag{2}$$

with k_u^{SG} the degree of node u in subgraph SG.

Indeed, the selection of a k-Core allows to increase the density of our subgraph. This increase in density implies less volatile values and therefore more

pronounced trends. However, the selection of a k-core also causes changes in other parameters of the original network, so it is a question of finding a compromise among these new characteristics.

The 5-core presents a sufficient balance to increase the density and the clustering coefficient, while reducing the number of connected components and the number of nodes (it is nevertheless necessary to keep a certain number of nodes - i.e. a few thousand - in order to be able to continue the analysis). The features of the subnetwork included in the 5-core are as follows:

- number of vertices = 6624
- average degree (as undirected G) = 0.1709
- density = 0.00066
- number connected components: = 13
- Clustering Coefficient (of G) = 0.1435

2.2 Preliminary studies

These are the preliminary studies that led us to the method we propose in this report. Although these works have no immediate links with the proposed method, we believe that they allow us to better grasp the issues at stake in the initial problem and some of the phenomena highlighted in these analyses could be useful for anyone wishing to pursue or use our contribution. It is for these reasons that we wish to present these preliminary studies while remaining brief on each of them.

The preliminary studies started with clustering on topological features from the citation network. Then we tried to implement community detection methods in order to see how they can contribute to solve our problem. Co-citation networks were then experimented on. Finally, a parallel analysis to study the evolution of the citations of the most cited articles over time was performed in order to investigate to which degree the temporal aspect could determine a typology of scientific articles.

The core of the contribution discussed in this report can be found in the Pipeline section.

Clustering on network features In a first step, an exploratory study was carried out in order to find structurally recurrent patterns. Then, we built the citation network on the basis of the scientific articles present in the PTP dataset.

To do so, we started by doing feature selection by extracting features from the intrinsic metrics of Network science such as the degree normalized by section, the different centralities, etc. However, none of our studies led to conclusive results.

Community detection We also tried to use community detection methods. However, we realised that this type of methodology led to the identification of very fine themes such as the sequencing of the human genome with a certain type of method.

Co-citations networks We then tried to transform the general structure of the network. Indeed, the co-citations network allows us to have a new link between the different articles. This parallel analysis allowed us to highlight the particularity of the articles that are mostly cited in the method section. These articles obtain a particularly high betweeness score. In other words, articles that are highly cited in the method section tend to be cited more by articles of various themes.

Evolution of the citations of the most cited articles over time We also studied the contribution of the temporal aspect to find patterns allowing the discrimination of a topology within our dataset. However, despite interesting patterns, we came to the conclusion that we lacked data if we wanted to obtain reliable trends.

2.3 Pipeline

Overview We build a network of citations with for each link the sentence that contains the citation corresponding to that link. Each of these sentences is processed in order to keep only the lemmas of the words that compose the sentence: each citation sentence thus forms a tuple of lemma. A shallow network (based on the Doc2Vec model) then learns to embed each of these tuples into a vector space. This space represents the semantic similarity of 2 citation vectors by the distance between them. A first clustering can be performed in order to evaluate a (continuous) typology of citations, and more generally, a short intermediate analysis of the vector space allows a better semantic understanding of the citations.

In a second step, the vectors representing the citations are used to determine the vector of each scientific paper. Indeed, we assume that each paper can be defined by the set of citations citing it. In this way, the vectors of the papers correspond to the average of the vectors of its incoming citations (we can assume that a weighted average of the vectors of the incoming citations may be more relevant to define the vector of a paper). We can then cluster on the basis of these new vectors (i.e. paper vectors) to determine a (continuous) typology of scientific papers.

We have tried 2 different approaches based on this pipeline. These approaches diverge in the parameters that are used: the first approach uses the citation sentences while the second one focuses mainly on the verbs that are used in these same citation sentences. We will see in the results section how these two approaches differ in their outputs.

In the rest of this pipeline section, we will first discuss in detail how we build the citation network, then how we parse and embed the citation contexts in the vector space. We then move on to the computation and embedding of scientific papers. Finally we conclude this part by clustering the vectors of the nodes (i.e. the scientific papers).

Citation network with citation context As mentioned in the PTP composition and network topology section, the citation network thus formed is (in theory) a Directed Acyclic Graph (DAG), i.e. each link - representing a citation - is directed to a node - representing an article - and the temporal dimension implies that loops within this network are (in theory) not possible. Each node is identified using the DOI associated to the article it represents and each link is aggregated either by the verbs of the citation sentence or by the sentence itself.

Embedding on citation contexts Before starting the pipeline, it is necessary to exclude some links. Since we assume that the vectors of papers can be determined on the basis of the vectors of citations, we also assume that we need a certain number of incoming citations to qualify the destination node. Therefore, we have chosen to set this threshold to 3 citations ($\Theta_c = 3$), so if the sum of the number of incoming links of a node is less than 3, then these are removed. Furthermore, we can only keep citations with a certain minimum number of 3 words, because below this value we consider that the citation does not contribute enough semantically.

In the model with citation sentences as parameters, it is necessary in a second step to parse them using tools specific to Natural Langage Processing (NLP). This procedure aims to remove all words with a low semantic value, commonly called stop-words, and to keep only the lemmas of the remaining words. These lemmas allow us to match words with identical bases (e.g. the words "medically" and "medicalisation" have a common base which is "medical", so we can assume that these 2 words have a strong semantic proximity and therefore the documents containing them deal with identical subjects).

More concretely, if we consider the sentence below:

"*This assembly into miRISC has been implicated in miRNA functions* [19−24]."

the processing gives the following vector of lemmas:

$$['thi','assembl','mirisc','implic','mirna','function']$$

Once each link in the network has these new lemma vectors, we can proceed to their embedding. To do this, we first need to label each lemma vector with an identifier. We can then set the Doc2Vec model with the following parameters:

The Doc2Vec method (implemented by the Gensim library) allows to use 2 different architectures. The one chosen in our approach is the so-called distributed bag of words (PV-DBOW). This one grants an increased computation speed and its architecture corresponds more to what we wish to set up (in comparison to PV-DM). Another parameter makes it possible to choose whether the model should learn on the words at first or directly on the documents. We opted for increased accuracy by choosing the option to learn on words before learning on documents. We also opted to add noise to the citation contexts by setting the associated parameter to 15. According to the size of the citation contexts which

can be very variable, the window size is set to 2 and on another hand, words having less than 2 occurrences in the citation list are ignored. Only 20 epochs are needed for this type of model, as they only need a small number of passes to learn. Finally, the number of worker threads, allowing to increase the speed of computation, can vary according to the machine used.

Once the learning process is complete, the output vectors have a size of 100 in order to capture a maximum of information on the combinations of input lemmas. These vectors, formed of 100 real numbers, can then be studied for a better intermediate understanding of the final results.

Node vector computation On the basis of the vectors included in the 100-dimensional space mentioned above, we can now compute the vectors of the nodes. This means performing an average which can be weighted by a weight on the corresponding link. The idea behind this method is that if we do not know the nature of something or someone, we can refer to the set of elements that are directly connected to it and thus determine the nature of the object or individual by the union of the definition that the elements have of it.

The results of our approach are based exclusively on a simple average (i.e. the weights are 1). Each node then obtains a vector of real numbers of size 100.

Node clustering As a last step, we propose to perform a clustering on all the vectors in order to group the articles by their distance. We opted for the k-means method because of the large number of points in the space and therefore the fast computation of this tool. Moreover, this method offers an ease of implementation and a guarantee of convergence: once again, these characteristics are not negligible given the number of data.

3 Results

3.1 Comparative studies

Intermediary results on the vectors of citation contexts Concerning the intermediate clustering that can be performed on the vectors of the citation contexts, we can first observe the results with the citation sentences by mapping the Fig. 9 with Fig. 10.

The 2D representation in the Fig. 9 was obtained following a PCA, although the clustering was well done on the 100 dimensions. The semantic proximity between the different clusters can be observed here. The number of clusters was obtained by averaging the inertia and the distortion, favouring the value closer to the inertia in the case of a non-integer average (this method is inspired by the elbow method [Tho53]).

One can easily imagine the links that can exist between the different neighbouring clusters of this space thanks to the lexical field in Fig. 10. These groups of 5 words, on 2 lines, represent the words having a certain relevance in each cluster according to their position in the list. The first line corresponds to the

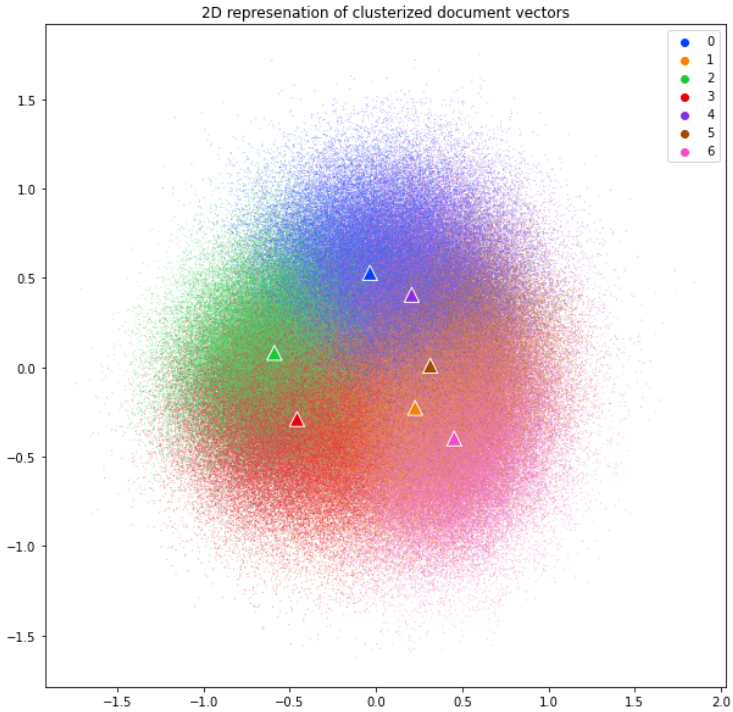


Fig. 9: Representation of the citation vectors from the initial 100D citation sentences in a 2D space (using a PCA)

```
###- cluster 2 -###
['kinas', 'repress', 'actin', 'cytoplasm', 'phosphoryl']
['protein', 'activ', 'cell', 'regul', 'gene']


###- cluster 4 -###
['snp', 'annot', 'haplotyp', 'genom', 'sequenc']
['sequenc', 'gene', 'genom', 'use', 'identifi']


###- cluster 0 -###
['kit', 'prepar', 'stain', 'mm', 'incub']
['describ', 'previous', 'use', 'perform', 'cell']


###- cluster 3 -###
['inflamm', 'inflammatori', 'carcinoma', 'cytokin', 'endotheli']
['cell', 'mous', 'express', 'infect', 'cancer']


###- cluster 5 -###
['task', 'learn', 'spike', 'nois', 'stimulu']
['model', 'use', 'network', 'studi', 'method']


###- cluster 6 -###
['art', 'hospit', 'educ', 'men', 'woman']
['studi', 'patient', 'risk', 'report', 'health']


###- cluster 1 -###
['habitat', 'reef', 'predat', 'coral', 'marin']
['speci', 'popul', 'studi', 'may', 'host']
```

Fig. 10: list of the 5 most characteristic words of the clusters of the context sentences vectors. the first line corresponds to the score named *custom score*, the second line comes from the score named *tf-idf like score*.

score resulting from the first equation below, which we generically call '*custom score*'. The second line follows the second equation below and we call it '*tf-idf like score*'.

$$CustomScore(w) = \frac{F_w^c}{|W_c|} \times \frac{F_w^T}{|T|} \tag{3}$$

$$TfIdfLikeScore(w) = \frac{F_w^c}{|W_c|} \times \log_{10}(\frac{|c|}{F_w^T}) \tag{4}$$

With:
w a word belonging to the set of all words W,
c a cluster belonging to the set of all clusters T,
F_w^c the frequency of a word w in a cluster c,
F_w^T the frequency of a word w in all clusters T,
W_c the set of all words in a cluster c.

These 2 measures offer different words in general, which complement each other to get an idea of the theme treated by the given cluster.

We can then compare this results with Fig. 11 and Fig. 12 with the model taking as parameters the vectors coming from the verbs of the citations.

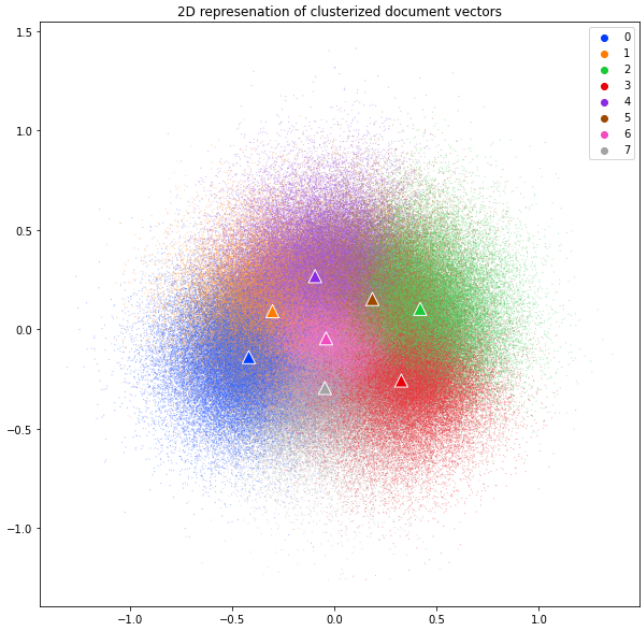


Fig. 11: Representation of the citation vectors from the initial 100D citation verbs in a 2D space (using a PCA)

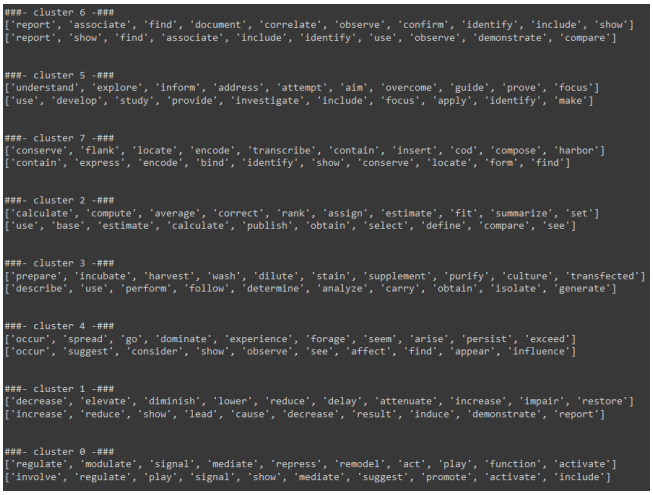


Fig. 12: list of the 10 most characteristic words of each clusters of the context verbs vectors. the first line corresponds to the *custom score*, the second line comes from the *tf-idf like score*.

Similar to the previous results, we have in Fig. 11 the vector space based on the citation verbs and in Fig. 10 the 10 most relevant words for each cluster. The method is strictly identical, except for the input parameters. Naturally, given the input vectors, the most relevant words are essentially verbs. They are therefore the most characteristic verbs of each cluster.

Complementary to this semantic analysis, we conducted a study on the distribution of the sections associated with the citations in the space. This study, available in Figs. 21, 22, 23, 24, 25, 26, allows us to see the proximity between the citations occurring in the Introduction and in the Discussion, in particular, but also the most characteristic verbs for each section.

Results on the vectors of scientific papers Concerning the final results on the clustering of the vectors of the scientific articles, we can first observe the raw result given in fig. 13 for the vectors computed on the citation sentences and in fig. 14 for the vectors computed on the citation verbs.

At first sight, we can notice a lesser presence of points in opposition to the intermediate results carried out directly on the citation vectors. This is because an article gathers several citations. Moreover, considering the elements previously presented, we can affirm that there are at least 3 times less points when comparing the 2 vector spaces. We can also observe a reduction in the amplitude of the dispersion of the points in the space in comparison with fig. 9. This can have several origins, in particular attributable to the method of computation used (i.e. averaging), indeed, the average of points tends to determine an output value inscribed in the points used in input. Consequently, the dispersion of

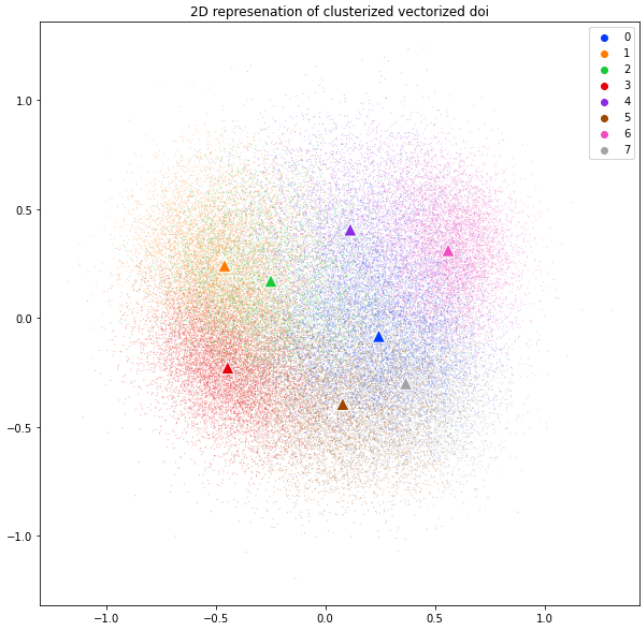


Fig. 13: repartition of the papers vectors by clusters (vectors based on citation sentences)

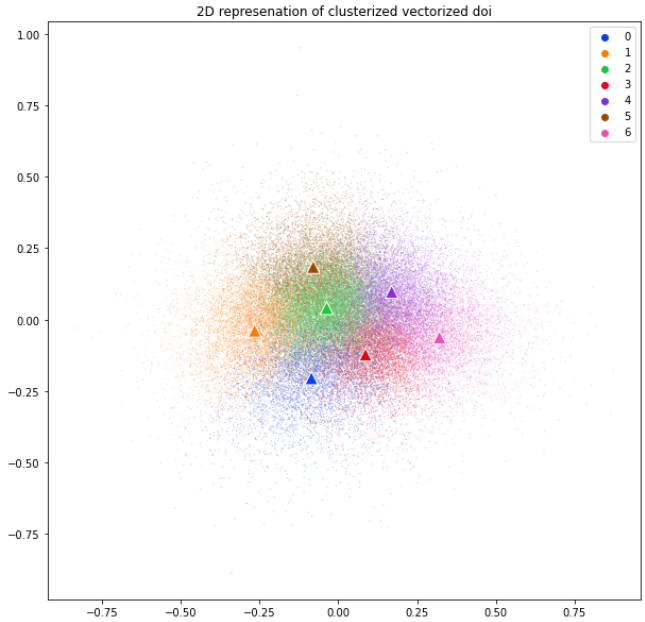


Fig. 14: repartition of the papers vectors by clusters (vectors based on citation verbs)

the averages will be inscribed in the dispersion of the initial points. This phenomenon can also be attributed to a reduction in the diversity of information contained in the vectors and consequently a reduction in the gaps between the points.

The figs. 15, 16, 18 allows to consult the most characteristic terms of each cluster among the topics filled in by the authors, the titles as well as the abstracts.

The ranking of topics was performed with the custom score method, while the analyses on titles and abstracts were performed with the tf-idf like score method.

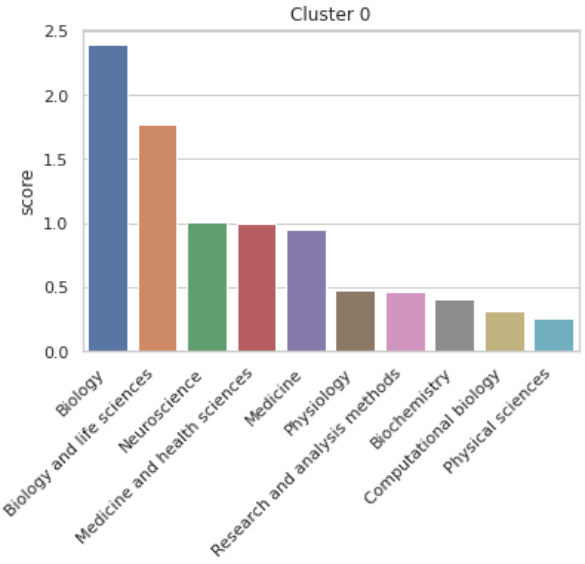
The articles in this dataset are mainly from the field of biology, therefore the terms Biology and Biology and life science are present in almost all of their keywords. It is for this reason that we observe these two terms at the top of each of the topic rankings. In fact, it is interesting to observe them from the third term onwards in order to get an idea of the field studied in each of the articles in each of the clusters.

Moreover, each of the 3 analysis can be put in relation to each other in order to refine the general topic of each cluster. For example, for the space of vectors calculated from the citation sentences, we can determine that cluster 0 has the words (all analyses combined) "neuroscience", "brain", "network", "learn", "human", "neuron", "connect", etc. We can then assume that cluster 0 deals with topics inherent to neuroscience, around human brains. We can also note the relationship between cluster 0 and cluster 1 of the same vector space. Indeed, cluster 1 is characterised by words close to cluster 0 (to a lesser extent) such as "biochemistry", "Genetics", "molecular cell Biology", "model organisms", "neuroscience", "cell", "cancer", "stimul", "activ", "tumor", "neuron", "tissue", etc. We can naturally observe similar relationships such as 0 and 1, between other clusters and other vector spaces.

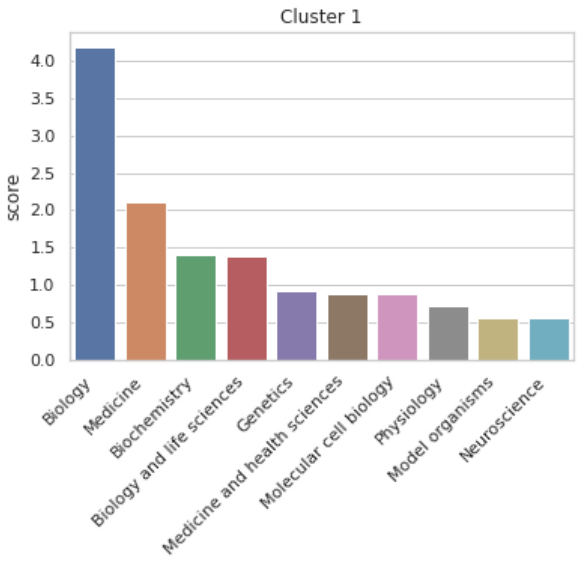
However, the initial idea of our contribution is to find a method able to determine a topology not on the thematic level but according to the methodology structuring the article and thus the knowledge contained in it. To do so, we have tried to study the distribution of certain types of articles in space. The aim was to see how the regions of the space are able to capture these types of articles(see [Fig. x] for an example). On the other hand, we have tried to find thematic articles in order to compare how these are distributed in this space.

We can distinguish 2 types of words, the methodological words and the thematic words. In fact, what we would ideally like is that the articles with methodological words in their title, indicated at the top of the figure, tend to be placed in a region and not to be distributed uniformly in the space. In contrast, we would ideally like to see articles with thematic words in their title spread relatively evenly across space. This would mean that embedding makes it easier to discriminate articles on the basis of their methodology rather than their theme.

The fig. 19 allows us to summarise all the points concerned by the previous analysis by averaging all the points for each keyword. The methodological words

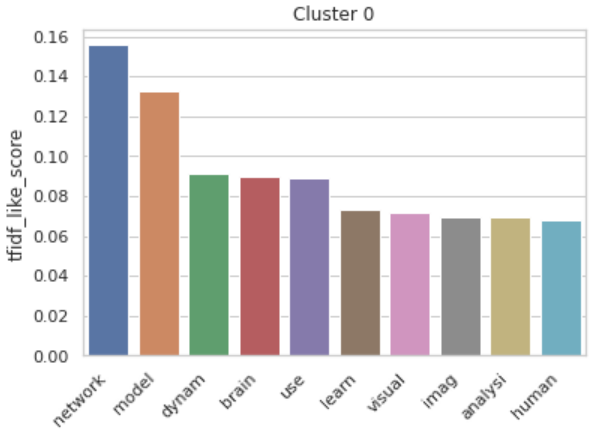


(a) most relevant topics in cluster 0

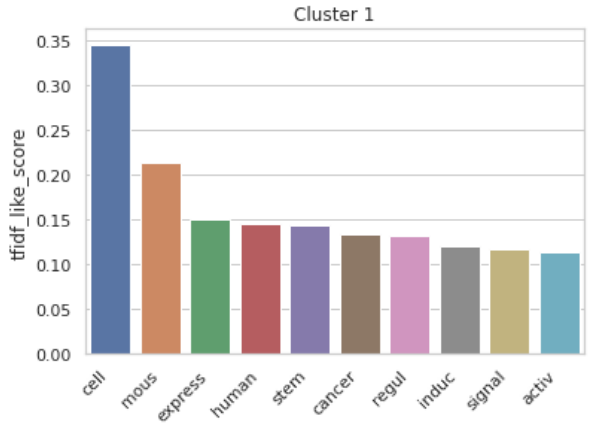


(b) most relevant topics in cluster 1

Fig. 15: information retrieval over topics in cluster 0 and 1 (vectors based on citation sentences)

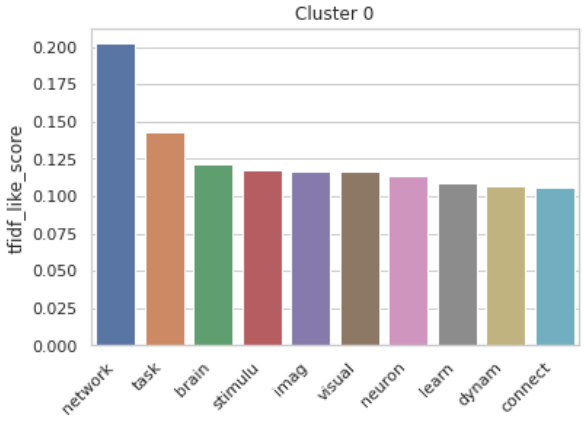


(a) most relevant words in title in cluster 0

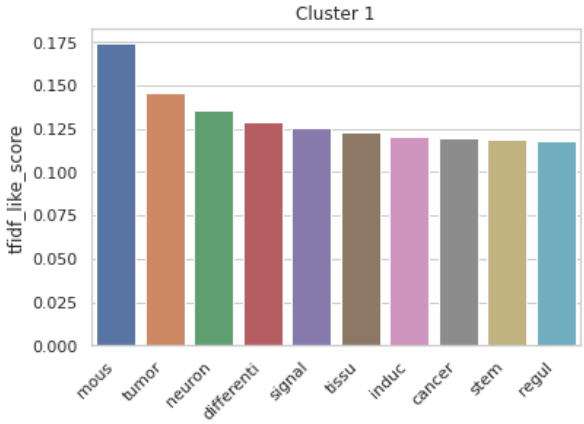


(b) most relevant words in title in cluster 1

Fig. 16: information retrieval over titles in cluster 0 and 1 (vectors based on citation sentences)

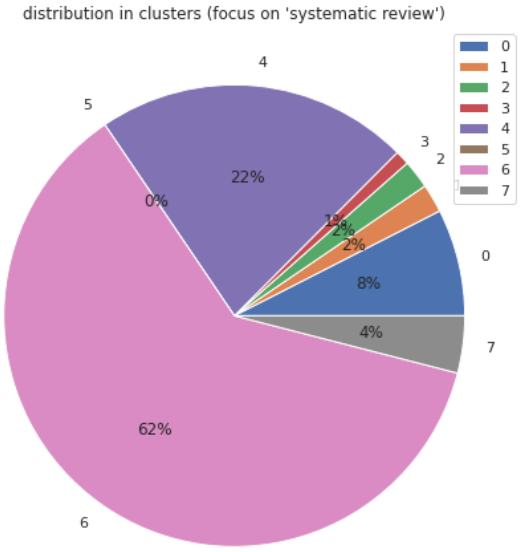


(a) most relevant words in abstracts in cluster 0

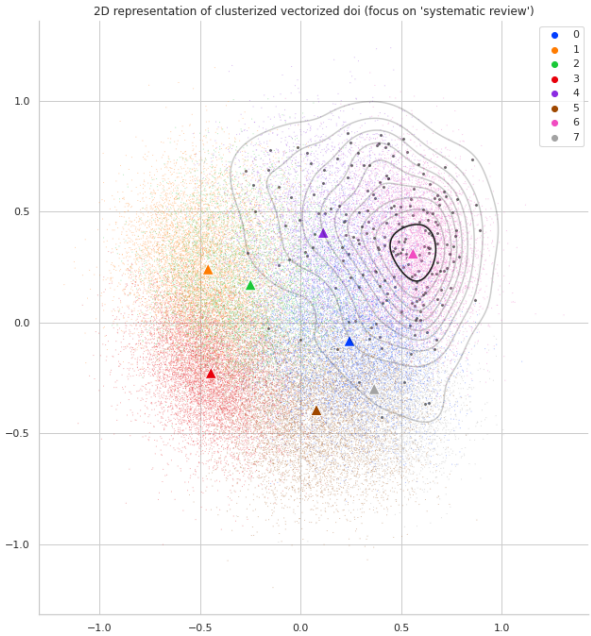


(b) most relevant words in abstracts in cluster 1

Fig. 17: information retrieval over abstracts in cluster 0 and 1 (vectors based on citation sentences)



(a) Distribution of the cluster over the subset of article having *systematic review* in title (vectors based on sentences)



(b) Repartition in the space of the subset of article having *systematic review* in title (vectors based on sentences)

Fig. 18: Analysis of the subset of article having *systematic review* in title (vectors based on citation sentences)

are in blue while the thematic words are in orange. It can be assumed that the more central a word is, the more uniformly it is distributed in space.

The fig. 20 allows us to compare the distances to the origin of thematic articles and methodological articles.

Ideally, we would like the orange points representing the thematic articles to be as close to 0 as possible, while the blue points representing the methodological articles are as far away from 0 as possible. However, we notice that the difference in distance is slightly more pronounced in the approach taking the citation verbs as input parameters. This may lead to the idea that the use of verbs tends to better discriminate articles according to their methodology.

4 Discussion

In line with other studies, we were able to observe that citations within sections are not done in the same way. We have also seen that citations in the Introduction and in the Discussion have a strong semantic similarity, both in terms of the wording used and the verbs used.

We construct a citation graph, with the sentences or verbs of the citations as labels. We then embed these word vectors in a shallow network in order to obtain real number vectors. These vectors are then used to determine vectors defining the research papers. Finally a clustering is applied to these clusters in order to provide an article topology.

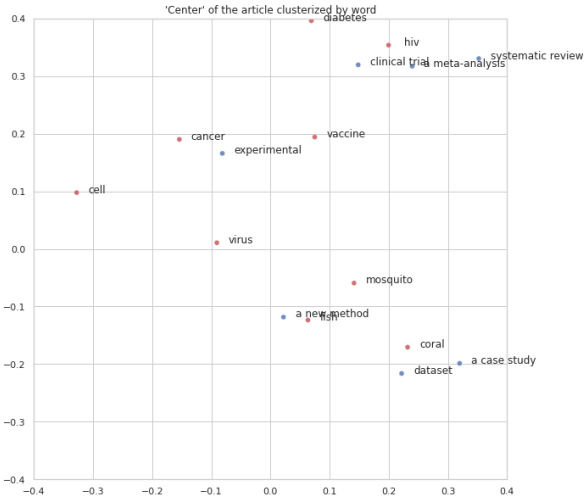
We have found that citations and their contexts are difficult to categorize the structure of the articles, although we have managed to get less than optimal results. The use of verbs in citation sentences seems to perform better than whole citation sentences for this purpose. The results are not yet conclusive, but eventually with further work we could hope to optimise them.

4.1 Areas of improvements

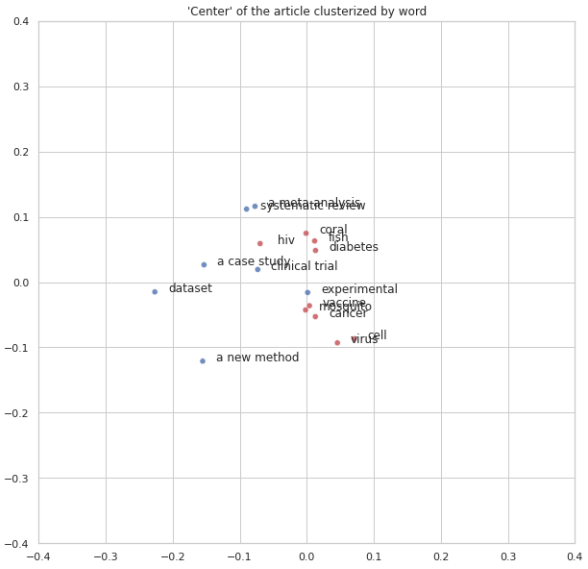
There are several areas of improvement that can be imagined in order to optimise our results.

Perhaps more data could positively change our approach. Indeed, the exclusivity of the PLoS articles may cause a lack of caraterization of the articles when computing the vectors of the scientific papers based on the citation vectors. On the other hand, increasing the minimum number of citations could also positively change our approach. Among the other parameters to be modified, it is possible that by further optimising (at the expense of the computational capacity) the shallow network used, this could possibly offer better results.

Another limitation is that the clusters formed are not distinctly separated, directly impacting the quality of the topology generated. Firstly because the visualisation is enabled by a PCA which transforms the data into 2D and therefore reduces the information. Secondly, because the embedding is based on the words present in the set of citations and naturally not those that are not present.

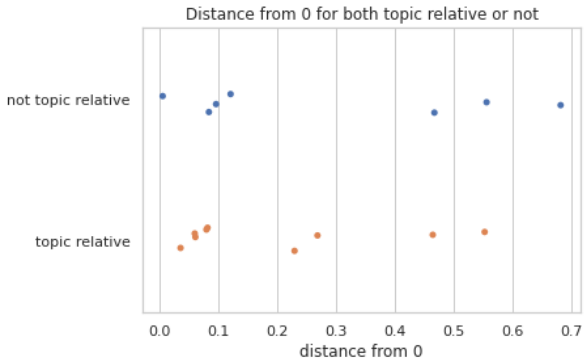


(a) Repartition of the points corresponding to the average of each article having the given words present in title (vectors based on sentences)

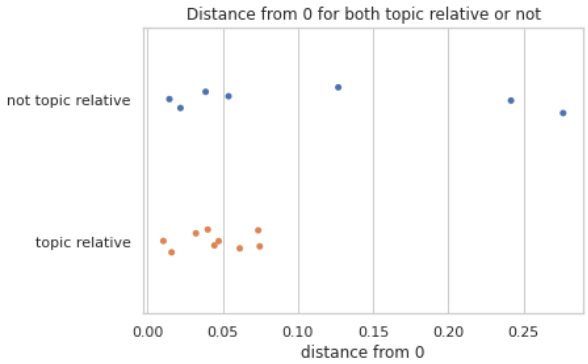


(b) Repartition of the points corresponding to the average of each article having the given words present in title (vectors based on verbs)

Fig. 19: Comparison of the repartition of the points corresponding to the average of each article having the given words present in title, between vectors based on sentences and verbs.



(a) Distance from 0 of the points corresponding to the average of each article having the given words present in title (vectors based on sentences)



(b) Distance from 0 of the points corresponding to the average of each article having the given words present in title (vectors based on verbs)

Fig. 20: Comparison of the distance from 0 of the point corresponding to the average of each article having the given words present in title, between vectors based on sentences and verbs.

Therefore, in reality, a void is supposed to appear in the space where the citations not present in the corpus could be, but in fine the points in the space are distributed in a relatively continuous way and thus the inter-cluster limits are also continuous.

Concerning the methods used, we have suggested that a clustering by cos-sin similarity would probably be more advantageous than a clustering by classical distance. Furthermore, reducing the number of vector dimensions or performing a PCA could also provide better clustering as the k-means method does not scale well on high dimensional vectors.

Another limitation to be improved concerns the evaluation method we propose. Indeed, it is obvious that this one does not answer adequately to check the quality of our approach. Therefore, the improvement of the latter may be crucial in the elaboration of our results.

Finally, we believe that an approach based on the recent Graph Neural Network (GNN) methods could bring a new way of perceiving the typology we have tried to highlight.

References

[BB03] ALBERT-LÁSZLÓ BARABÁSI and ERIC BONABEAU. "Scale-Free Networks". In: *Scientific American* 288.5 (2003), pp. 60–69. ISSN: 00368733, 19467087. URL: http://www.jstor.org/stable/26060284 (visited on 08/17/2022).

[CC13] David Chavalarias and Jean-Philippe Cointet. "Phylomemetic Patterns in Science Evolution—The Rise and Fall of Scientific Fields". In: *PLOS ONE* 8.2 (Feb. 2013), pp. 1–11. DOI: 10.1371/journal.pone.0054847. URL: https://doi.org/10.1371/journal.pone.0054847.

[For+18] Santo Fortunato et al. "Science of science". In: *Science* 359.6379 (2018), eaao0185. DOI: 10.1126/science.aao0185. eprint: https://www.science.org/doi/pdf/10.1126/science.aao0185. URL: https://www.science.org/doi/abs/10.1126/science.aao0185.

[Gar55] Eugene Garfield. "Citation Indexes for Science". In: *Science* 122.3159 (1955), pp. 108–111. DOI: 10.1126/science.122.3159.108. eprint: https://www.science.org/doi/pdf/10.1126/science.122.3159.108. URL: https://www.science.org/doi/abs/10.1126/science.122.3159.108.

[GB09] Bela Gipp and Joeran Beel. "Citation Proximity Analysis (CPA) - A new approach for identifying related work based on Co-Citation Analysis". In: vol. 2. July 2009.

[LM14] Quoc Le and Tomas Mikolov. "Distributed Representations of Sentences and Documents". In: *Proceedings of the 31st International Conference on Machine Learning*. Ed. by Eric P. Xing and Tony Jebara. Vol. 32. Proceedings of Machine Learning Research 2. Bejing, China: PMLR, 22–24 Jun 2014, pp. 1188–1196. URL: https://proceedings.mlr.press/v32/le14.html.

[Mik+13] Tomas Mikolov et al. *Efficient Estimation of Word Representations in Vector Space*. 2013. DOI: 10.48550/ARXIV.1301.3781. URL: https://arxiv.org/abs/1301.3781.

[MJF15] Matteo Morini, Pablo Jensen, and Patrick Flandrin. "Temporal evolution of communities based on scientometrics data". In: *Sciences des données et humanités numeriques*. Projet ARESOS (Reconstruction, Analyse et Accès aux Données dans les Grands Réseaux Socio-Sémantiques), ISC-PIF. Paris, France, Nov. 2015. URL: https://hal.inria.fr/hal-01282805.

[Saj+21] Naseer Sajid et al. "Exploiting Papers' Reference's Section for Multi-Label Computer Science Research Papers' Classification". In: *Journal of Information and Knowledge Management* 20 (Mar. 2021), p. 2150004. DOI: 10.1142/S0219649221500040.

[Sol65] Derek J. de Solla Price. "Networks of Scientific Papers". In: *Science* 149.3683 (1965), pp. 510–515. DOI: 10.1126/science.149.3683.510. eprint: https://www.science.org/doi/pdf/10.1126/science.149.3683.510. URL: https://www.science.org/doi/abs/10.1126/science.149.3683.510.

[SPA72] KAREN SPARCK JONES. “A STATISTICAL INTERPRETATION OF TERM SPECIFICITY AND ITS APPLICATION IN RETRIEVAL”. In: *Journal of Documentation* 28.1 (Jan. 1972), pp. 11–21. ISSN: 0022-0418. DOI: 10.1108/eb026526. URL: https://doi.org/10.1108/eb026526.

[Tho53] Robert L. Thorndike. “Who belongs in the family?” In: *Psychometrika* 18.4 (Dec. 1953), pp. 267–276. ISSN: 1860-0980. DOI: 10.1007/BF02289263. URL: https://doi.org/10.1007/BF02289263.

[TR21] Mauro Dalle Lucca Tosi and Julio Cesar dos Reis. “SciKGraph: A knowledge graph approach to structure a scientific field”. In: *Journal of Informetrics* 15.1 (2021). DOI: 10.1016/j.joi.2020.101109. URL: https://ideas.repec.org/a/eee/infome/v15y2021i1s175115772030626x.html.

A Annexes

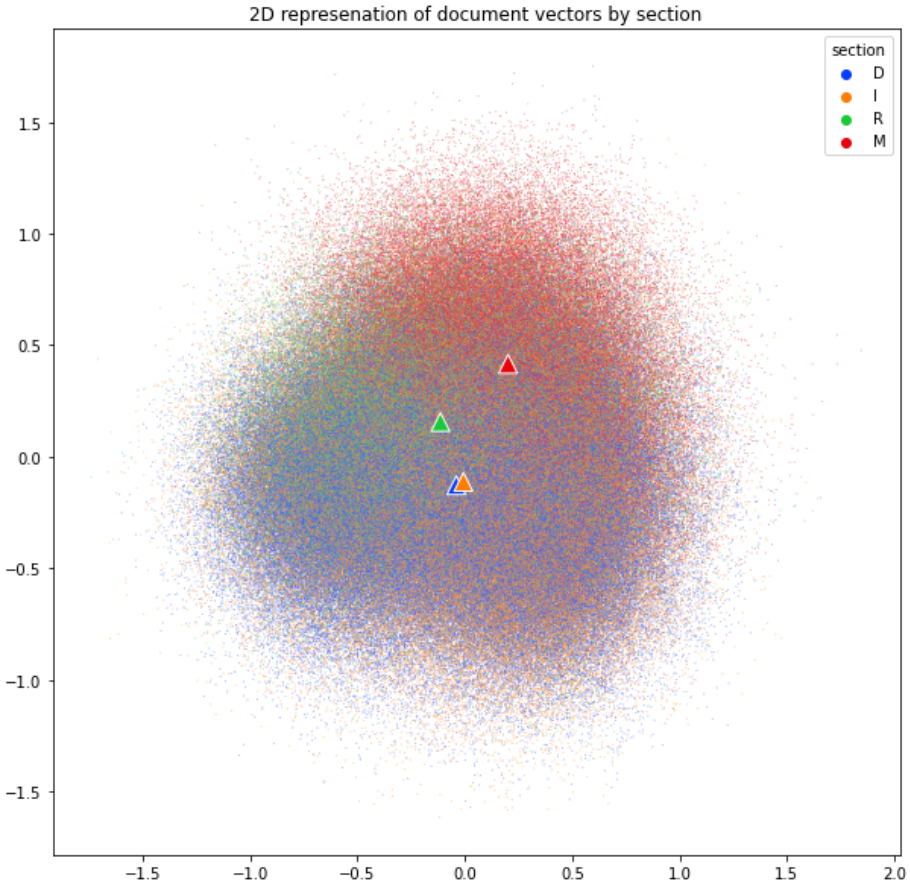


Fig. 21: Repartition of citations by sections in space (vectors based on citations sentences).

```
###- cluster D -###
['find', 'agreement', 'futur', 'might', 'could']
['studi', 'report', 'cell', 'also', 'previou']


###- cluster I -###
['million', 'decad', 'understand', 'world', 'attent']
['studi', 'recent', 'cell', 'gene', 'use']


###- cluster R -###
['figur', 'fig', 'tabl', 'p', 'mutant']
['fig', 'cell', 'gene', 'express', 'studi']


###- cluster M -###
['primer', 'http', 'softwar', 'mm', 'describ']
['describ', 'use', 'previous', 'perform', 'data']
```

Fig. 22: Most relevant words by section (vectors based on citations sentences).

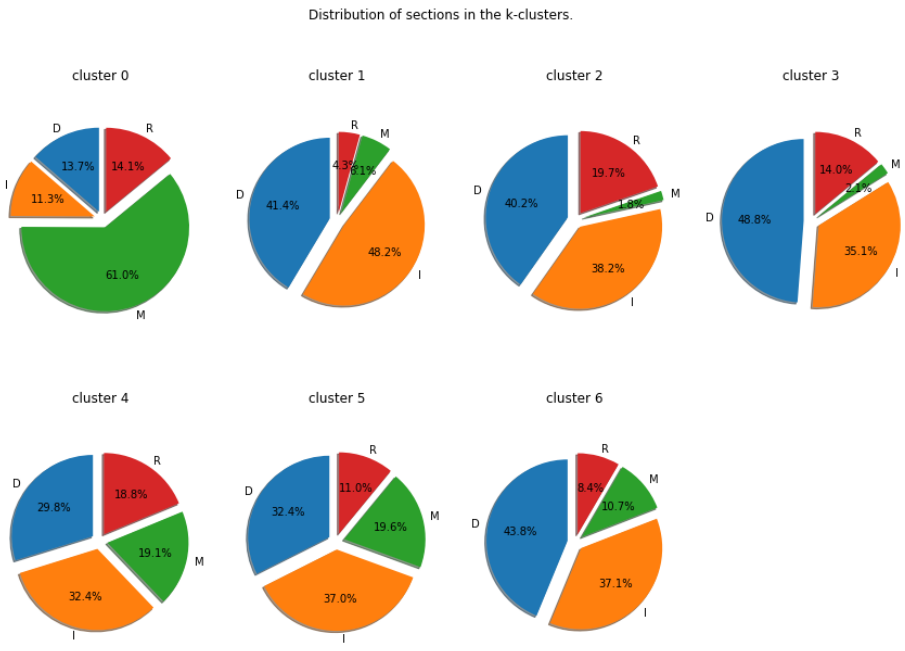


Fig. 23: Distribution of sections by clusters (vectors based on citations sentences).

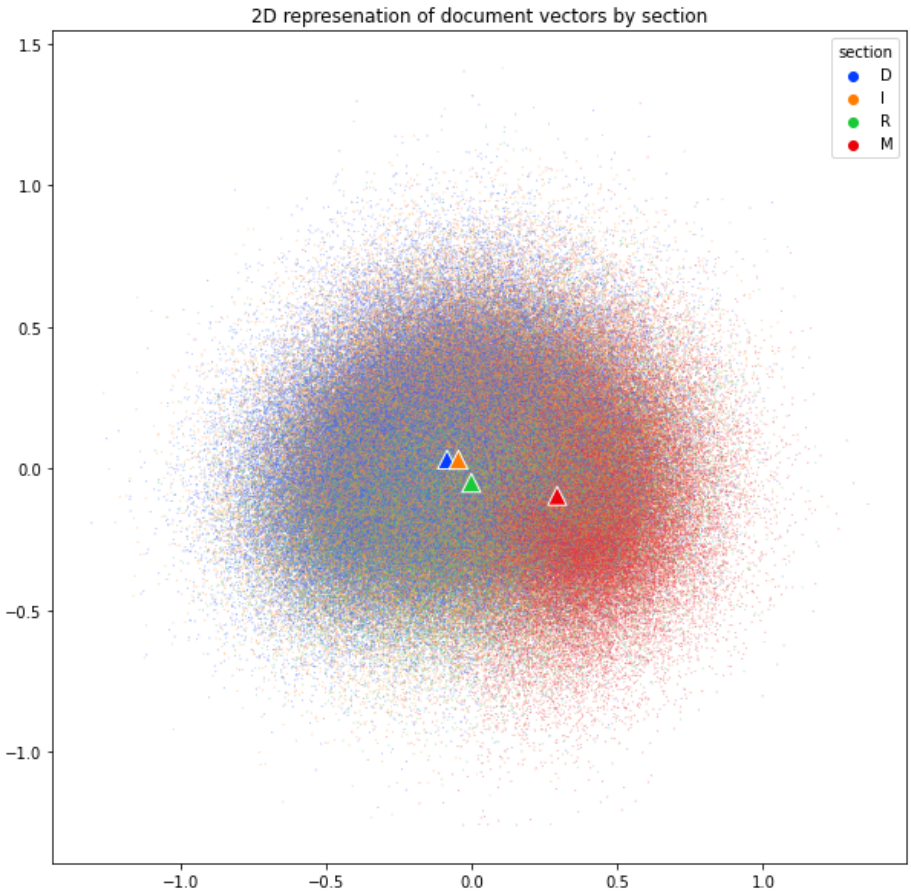


Fig. 24: Repartition of citations by sections in space (vectors based on citations verbs).

```
###- cluster D -###
Top 10 in custom interesting words:
['interest', 'speculate', 'corroborate', 'agree', 'contrast', 'seem', 'note', 'support', 'explain', 'emphasize']
Top 10 in tfidf-like interesting words:
['show', 'report', 'find', 'suggest', 'use', 'associate', 'observe', 'increase', 'include', 'demonstrate']


###- cluster I -###
Top 10 in custom interesting words:
['understand', 'transmit', 'term', 'discover', 'diabetes', 'emerge', 'name', 'elucidate', 'focus', 'live']
Top 10 in tfidf-like interesting words:
['include', 'show', 'use', 'identify', 'associate', 'report', 'increase', 'find', 'suggest', 'demonstrate']


###- cluster R -###
Top 10 in custom interesting words:
['=', 'enrich', 'localize', 'exclude', 'correspond', 'examine', 'expect', 'encode', 'bind', 'express']
Top 10 in tfidf-like interesting words:
['show', 'use', 'report', 'include', 'identify', 'observe', 'find', 'compare', 'express', 'see']


###- cluster M -###
Top 10 in custom interesting words:
['prepare', 'incubate', 'wash', 'supplement', 'calculate', 'genotyped', 'normalize', 'describe', 'accord', 'amplify']
Top 10 in tfidf-like interesting words:
['describe', 'use', 'perform', 'base', 'follow', 'accord', 'obtain', 'determine', 'calculate', 'include']
```

Fig. 25: Most relevant words by section (vectors based on citations verbs).

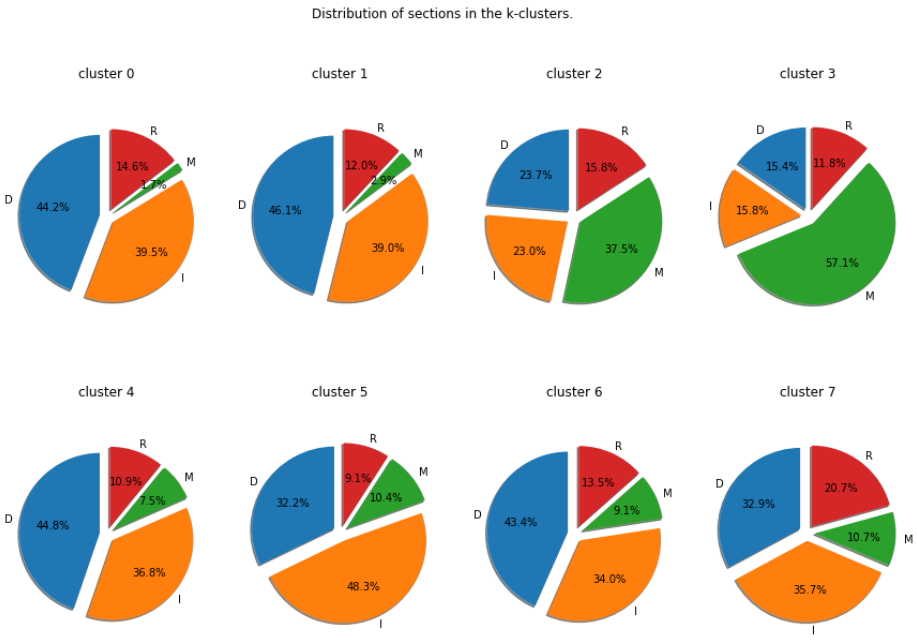


Fig. 26: Distribution of sections by clusters (vectors based on citations verbs).