Paragraph-based intra and inter-document similarity using neural vector paragraph embedding

Bart Thijs

▶ **To cite this version:**

Bart Thijs. Paragraph-based intra and inter-document similarity using neural vector paragraph embedding. 2019. hal-02015772

HAL Id: hal-02015772
https://hal.science/hal-02015772

Preprint submitted on 12 Feb 2019

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.

KU LEUVEN

FACULTEIT ECONOMIE EN BEDRIJFSWETENSCHAPPEN

Management, Strategy and Innovation (MSI)

Paragraph-based intra- and inter- document similarity using neural vector paragraph embedding

Bart Thijs

MSI_1901

Paragraph-based intra- and inter- document similarity using neural vector paragraph embeddings

Bart Thijs[1]

[1] *bart.thijs@kuleuven.be*
KU Leuven, ECOOM, FEB, Naamsestraat 61, 3000 Leuven (Belgium)
Univ Grenoble Alpes, LIG, SIGMA, 38000 Grenoble (France)

Abstract

Science mapping using document networks is based on the assumption that scientific papers are indivisible units with unique links to neighbour documents. Research on proximity in co-citation analysis and the study of lexical properties of sections and citation contexts indicate that this assumption is questionable. Moreover, the meaning of words and co-words depends on the context in which they appear. This study proposes the use of a neural network architecture for word and paragraph embeddings (Doc2Vec) for the measurement of similarity among those smaller units of analysis. It is shown that paragraphs in the 'Introduction' and the 'Discussion' section are more similar to the abstract, that the similarity among paragraphs is related to -but not linearly- the distance between the paragraphs. The 'Methodology' section is least similar to the other sections. Abstracts of citing-cited documents are more similar than random pairs and the context in which a reference appears is most similar to the abstract of the cited document. This novel approach with higher granularity can be used for bibliometric aided retrieval and to assist in measuring interdisciplinarity through the application of network-based centrality measures.

Introduction

Document networks with weighted edges based on similarities using either citation links (Small, 1994), lexical similarity (Wang & Koopman, 2017) or combinations (eg. Ahlgren & Colliander, 2009; Thijs & Glänzel, 2018) and unweighted approaches using direct citations links (eg. Boyack, 2017) have been used for science mapping exercises. In these studies, documents were assumed to be indivisible units, each with unique links to neighbouring nodes holding a single value indicating the strength of the similarity in the case of a weighted variant. Then, these networks were subject of community detection or clustering approaches which resulted in hard clustering assigning documents to single groups or clusters of papers. However, it becomes more and more clear that the basic underlying assumption in these models is questionable. A document is often not to be reduced to such a single point or entry in the knowledge space.

A recent paper (Thijs & Glänzel, 2018) using full texts from the journal Scientometrics identified papers that shifted easily from one cluster to another after slight changes in the weighting parameter in the combined citation-lexical approach. A set of papers on institutional performance was split into two groups with the first focusing on university ranking and name disambiguation problems associated with this topic and the latter one related to institutional performance in social sciences. The first group was merged into the topic of '*Research Assessment*' while the latter one becomes part of the set of papers labelled as '*Field and Regional Studies*'. Both these clusters were not specifically labelled as dealing with institutional research. It is impossible to indicate whether one or the other grouping was better. Both had similar quality scores.

A similar observation was made by Boyack (2017) when he compared local cluster solutions in the field of Astronomy with the topics identified in his global science map. Several typical Astronomy topics from the global map did contain publications that were not in the initial data set due to their lack of compliance with the retrieval strategy. Other papers from the initial set were to be found in topics that were clearly not primarily on Astronomy. These papers had a large portion of their links to non-Astro papers. Working on the same data set, I identified

papers studying the effect of absence of gravity on the growth of plants connecting both agriculture with astronomy (see Kiss et al., 2014 as an example).
This leads to the proposition that a more fine-grained approach should be applied in science mapping where the unit is no longer the document but at lower levels like sections, paragraphs or even sentences. This has already been alluded to by several studies using an enhanced co-citation analysis which incorporates also the proximity of the cited references. Several studies indicate the co-cited publications are more similar if the distance of their in-text citation is smaller (see Gipp & Beel, 2009). Complementary to these co-citation-based findings, others have reported different lexical properties of the subsequent sections in scientific papers. Bertin et al. (2016) reported different use of verbs and rhetorical structures surrounding references across different sections.

Other approaches that move away from the assumption of a document as an indivisible unit are probabilistic topic modelling techniques like LDA (Blei, 2012) where documents are linked to different topics with a weight relative to the probability that the topic is relevant to the document. With LDA, one has to set the number of topics prior to the analysis. Gal et al (2017) applied this technique to a set of publications from cardio-vascular research with an initial set of 200 topics where after expert validation only 166 remained relevant. Three issues remain unsolved when using topic modelling approaches. First, there is the initial decision on number of topics, next, the document remains the unit of analysis and the probability that a document is related to a particular topic is attributed to the document as a whole and finally, the use of bag-of-word approach in the learning phase neglects the differences in meaning a word can have depending on its context. Leydesdorff and Hellssten (2006) demonstrated how words and co-words retrieve their meaning from their presence in sentences and broader context.

This study proposes a new approach that moves the granularity towards smaller units of text namely the paragraphs in the different sections across a full paper and applies an analytical technique that tries to capture the meaning of words and phrases not only from its position relative to other words in its neighbourhood but also from the overall subject or topic covered by the paragraph. Vector word embeddings using neural networks architecture like GloVe (Pennington, Socher & Manning, 2014) and Word2Vec (Mikolov et al. 2013) are able to map words to a low dimensional space, with high performance. These word embeddings have however a single representation in the vector space neglecting the different meaning a word can have across different contexts. This is solved by adding an additional paragraph or document layer to the learned model which holds this context information (Quoc & Mikolov, 2014).

This study will use the Doc2Vec implementation in GenSim (Rehurek & Sojka, 2010) for the calculation of the word and paragraph embeddings and for the calculation of the similarities between low level units or fragments of text extracted from all PlosOne publications up to december 2018. As such, the current research is the first to apply these techniques at such a large scale with the following objectives. First, I'll try to measure the intra document similarity between the different paragraphs and between the paragraphs and the abstract. It is assumed that paragraphs close to each other have higher similarity, that paragraphs in the 'Introduction' and the 'Discussion'-section have higher similarity with the abstract and each other and that paragraphs in the 'Methodology' are least similar to all other paragraphs Next, the research focusses on documents linked through a citation. It is assumed that the context surrounding the reference is most similar to the content of the cited paper. In a last section of the study, the location of the reference in the citing document is mapped with the different paragraphs in the cited document in order to retrieve the cited information relevant to the citing document. The

results from this study can have applications like bibliometric aided information retrieval or can assist in the identification of interdisciplinary research.

Data & Methodology

Data

This study uses publications from PlosOne downloaded from PubMed Central ftp://ftp.ncbi.nlm.nih.gov/pub/pmc. The downloaded set contains publications indexed until December 7th, 2018 and it holds 204,846 documents from 2006 onwards. The papers are provided in XML-format following the '*Journal Article Tagging Suite*' (JATS) standard. This schema divides the information in three main elements: <front>, <body> and <back> with an underlying structure of elements and attributes and complies with ANSI standard Z39.96-2012 (ANSI, 2012). This format, provided as an XML-schema, is then converted by the *Java Architecture for XML Binding* library (JAXB) into generated Java source code. This generated Java library serves then as a unmarshalling toolbox which can convert any XML-document compliant with the JATS-schema into a set of Java Objects (POJOs). This toolbox incorporates parts of the *CorpusHandling* library developed by *CyCorp* and available under Apache license (version 2.0) from GitHub (see https://github.com/cycorp/CorpusHandling/). Each XML-document is unmarshalled into a Java object and parsed in order to extract:

- Bibliographic information like title, article number, publication year
- Sections and paragraphs holding the actual text fragments of the paper
- In-text references identified by the <Xref>-tag
- References at the end of the paper

It is assumed that papers published in PlosOne adhere to the IMRaD structure (Introduction, Methodology, Results and Discussion) or a variation where the Methodology section is at the end of the paper (IRDaM) following the description of the distribution of sections across PlosOne publications (Bertin, et al. 2013). The title heading each section is used to classify the text fragment to one of the following classes:

I. Introduction; Background
II. Data; Material; Methodology; Design
III. Results
IV. Discussion; Summary; Conclusion
V. Other sections

Paragraphs are identified by XML-element tags <*sec*> and <*p*>, extracted and given a sequence number. Sentences are extracted and numbered within each paragraph. Figure 1 presents a paragraph from the first paper published in PlsOne (Harris et al, 2006).

Early investigations focused on the role of neurons in subcortical stations and primary somatosensory cortex (SI) in coding low frequency “flutter” vibrations (below 50 Hz) **[1]-[3]**, while more recent work has emphasized the role of cortical areas “downstream” from SI, such as the second somatosensory cortex (SII) and regions of frontal cortex **[4]**, **[5]**. Which of these different areas, and which features of the neural activity within these areas, are essential components in forming the percept of a vibration? A series of psychophysical experiments with humans provided evidence that neural processes in SI contribute to frequency discriminations. In a task designed to resemble that performed by monkeys in the aforementioned neurophysiological studies, subjects compared two sequential vibrations and reported which had the higher frequency.

Figure 1. Text fragment taken from Harris et al (2006)
In-text references are marked in bold and underlined.

Next, in-text references (<*xref*>) are linked with the complete reference at the end of the paper and available identifiers like PMID, PMCID or DOI of cited papers are retrieved. This enables the linking of individual paragraphs to the cited paper. The position of the in-text reference with respect to extracted sentences in the paragraph is recorded. The in-text references in Figure 1 are marked in bold and underlined. The first <xref>-element is linked to the first three entries in the reference list at the end of the paper as it indicates the range between reference 1 and 3. The next two elements refer to the fourth and fifth entry. Each element in the text is replaced by the corresponding PMCID, PMID or DOI depending on the available data in the reference list.

Methodology

After extraction of the data from the XML-file, a set of processing steps is applied in order to obtain the vector word and paragraph embeddings. A pre-processing procedure as described in Thijs et al (2017) and Glänzel & Thijs (2017) based on the Stanford Natural Language Processing library (Chen & Manning, 2014) and the Lucene text search engine is used for the extraction of sentences, application of Part of Speech tagging, stemming, removal of stop-words and selection of noun phrases. Document identifiers like PMCID of the cited references are processed as noun phrases and retained at the original position within each sentence. Table 1 presents the results after pre-processing of the text fragment in figure 1. A list of all cited documents is added at the end of each paragraph as an additional ‘sentence’. The choice to include the cited references in the final paragraph embeddings is not without consequences. It adds a bibliographic-coupling-like component to the embeddings.

Table 1. Parsed content of paragraph in Harris et al (2006) per section, paragraph and sentence.

Section	Paragraph	Sentence	Parsed Content
I	5	0	earli investig role neuron subcort station primari somatosensori cortex si low frequenc flutter vibrat 50 hz pmc2118947 pmc4959494 pmc4977839 recent work role cortic area si second somatosensori cortex sii region frontal cortex pmc12368806 pmc10884334
I	5	1	differ area featur neural activ area essenti compon percept vibrat
I	5	2	seri psychophys experi human neural process si frequenc discrimin

I	5	3	task monkey aforement neurophysiolog subject two sequenti vibrat higher frequenc
…	…	…	…
I	5	15	pmc2118947 pmc4959494 pmc4977839 pmc12368806 pmc10884334 …

The mathematical representation of the text fragments or paragraphs is based on vector representations built by a neural network architecture in an unsupervised machine learning algorithm. The applied methodology was first developed for distributed word embeddings at Google by Mikolov et al (2013) as a more complex substitute for simple vector-based representations like N-gram models. These embeddings are used to predict a word given the surrounding words in its context. The context is a sliding window with a fixed word length. The context is also applied to the identifiers of the cited references. Quoc & Mikolov (2014) extended the model for the inclusion of document or paragraph embeddings to outperform traditional bag-of-word approaches. Just like the original word embeddings model, the paragraph is represented by a vector in the same space as the words. It complements each fixed word length context used for the prediction of the words in the paragraph. The vector representation is unique for each paragraph and it is not shared among paragraphs and can be thought of as the representation of the topic the paragraph is dealing with.

The neural network used for training this model is a single layered architecture with a fixed dimensionality. In contrast to the LDA approach, these dimensions are not linked to topics and no external validation of the validity of the dimensions is required.

The Python implementation included in the Gensim library (Rehurek & Sojka, 2010) is used in this study. The algorithm is named '*Doc2Vec*' and takes the paragraph as a list of words as input with an additional tag identifying paragraph. This tuple is called a '*TaggedDocument*' in the library. The abstracts are tagged by the PMCID and paragraphs with tags containing the PMCID, section classification and sequence number of the paragraph. A cosine is calculated between the vector embeddings to measure the similarity between the text fragments.

The first set of analyses focusses on the intra-document similarity between paragraphs and abstract and among paragraphs. Figure 1 provides a schematic overview of the different analytical steps in this study. The intra-document similarity is indicated at the left-hand site. Within paper A, the abstract is compared with each paragraph and each paragraph with all subsequent paragraphs within the same section and across sections. The sequence number of the paragraphs are used to indicate the distance between the paragraphs in the text.

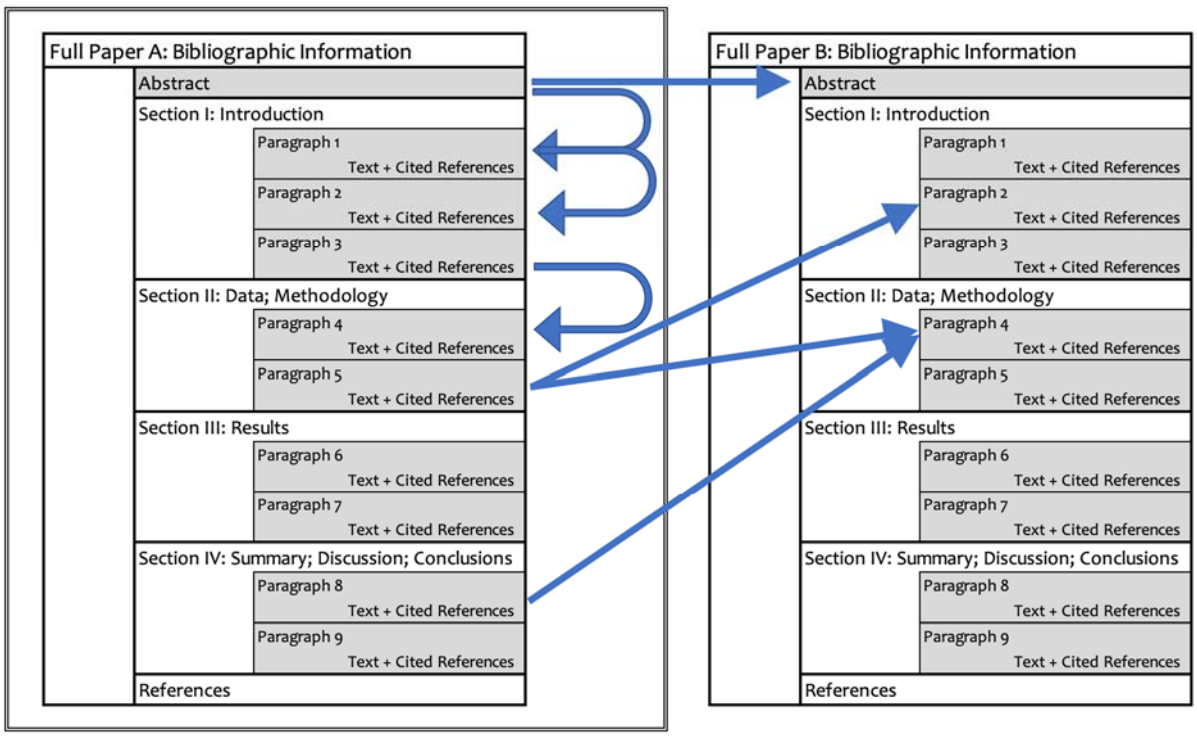

Figure 2. Schematic overview of the different comparisons.

The second set of analyses focusses on the similarity between citing and cited pairs of documents. In fig 2. there is a citation from paper A to paper B. The similarity of both abstracts in a citing-cited document pair is compared to the similarity in a randomly selected document pair. As the paragraph holding the in-text reference to the cited paper can be located, a next analysis compares the similarity between the citing paragraph and cited abstract and paragraphs across all sections within the citing-cited document pair.

Results

Descriptive Statistics

204,846 PlosOne publications have been downloaded and processed. 99% of these are recorded as '*research article*' in the Web of Science database and the remaining 1% as '*review*'. Table 2 provides the distribution of papers over publication years and the average number of paragraphs, together with the share of documents with the IMRaD sections in any order. Almost all documents have an introduction and discussion section.

Table 2. Descriptive statistics for downloaded PlosOne papers per year. Sections are classified based on the header of the section.

Publication Year	Publications	Average number of paragraphs	Introduction (I)	Methodology (II)	Results (III)	Discussion (IV)
2006	137	29.32	100.0%	99.3%	84.7%	100.0%
2007	1230	30.97	100.0%	98.5%	86.8%	99.4%
2008	2820	31.14	96.3%	95.7%	86.6%	96.0%
2009	4537	31.90	97.0%	96.3%	87.8%	96.8%
2010	6925	32.17	97.5%	96.9%	88.6%	97.3%
2011	14043	32.17	98.2%	97.3%	89.7%	97.9%

2012	24102	32.28	97.3%	96.0%	89.0%	97.0%
2013	32973	31.68	95.6%	93.4%	86.0%	95.3%
2014	30467	32.66	98.6%	96.1%	87.8%	98.4%
2015	28126	33.60	99.8%	96.8%	87.5%	99.6%
2016	22092	34.10	99.8%	96.7%	88.0%	99.6%
2017	20499	34.32	99.5%	96.2%	87.3%	99.2%
2018	16895	34.01	94.9%	91.2%	83.0%	94.6%
Total	**204846**	**32.92**	**97.9%**	**95.5%**	**87.3%**	**97.7%**

The publications contain on average 32.92 paragraphs and 3.86 different sections. This is below the values reported by Bertin et al (2013). This probably due to differences in parsing and extraction of the XML-elements. Subsections indicated by <sec>-elements as a child from another <sec>-element are not considered as separate sections and obtain their classification from their parent element. Section and paragraph elements without text as value were not considered as separate paragraphs.

The Neural Network Model

The final *Doc2Vec*-model is trained on 6.95 million text fragments from abstracts and paragraph texts. The neural network contains 400 nodes and training is done over ten iterations. Before training, a vocabulary was created with 3.90 million unique words. The total number of words included was 440 million. A sliding window of 7 words was used to establish the context for each word. It took about 15 hours to train this model on an average server requiring not more than 27Gb of RAM.

Intra-document similarity

First, the analysis focuses on intra-document similarity. Figure 3. shows the distribution of the cosine similarity of the abstract with distinct paragraphs across the four identified sections. The '*Introduction*' is most similar to the abstract, while the '*Methodology*' section is least similar. The inclusions of in-text references in the final paragraph embeddings and the absence of references in abstracts can act as a damping factor for the similarity between abstract and actual text fragments. However, the higher amount of references in the introduction and discussion (Bertin et al 2013) does not prevent the higher similarity between these sections and the abstract.

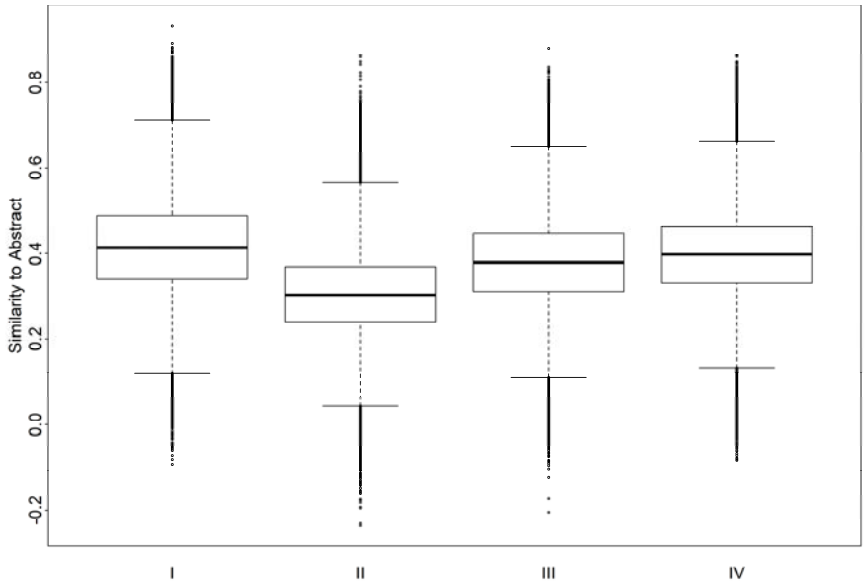

Figure 3. Distribution of similarity between Abstract and different sections in PlosOne papers.

Next, the similarity is calculated between paragraphs within sections and across section in each paper. The average similarities are presented in table 3. The intra-section similarity ranges from 0.32 for the '*Introduction*' to 0.29 for the '*Methodology*'. Looking at similarities across section, it can be observed that '*Introduction*' and '*Discussion*' are more similar, and '*Methodology*'-paragraphs are least similar to paragraphs in other sections.

Table 3. Average similarity of paragraphs across sections

	I	**II**	**III**	**IV**
I: Introduction	**0.32**	0.26	0.27	0.31
II: Methodology	0.26	**0.29**	0.27	0.26
III: Result	0.27	0.27	**0.31**	0.28
IV: Discussion	0.31	0.26	0.28	**0.30**

It is worthwhile to complement this analysis by adding the distance between paragraphs to the analysis. Figure 4 plots the average similarity between two paragraphs against the distance between them in the text. As each paragraph gets a sequence number in the processing phase it is easy to calculate the distance between them. The plot distinguishes between two groups, namely the distance between paragraphs inside one section opposed to the distance across sections. The solid line indicates the within section similarity and starts with the highest value. It rapidly declines with an increasing distance. The similarity between paragraphs across sections stars much lower and takes an increase and slow decline afterwards.

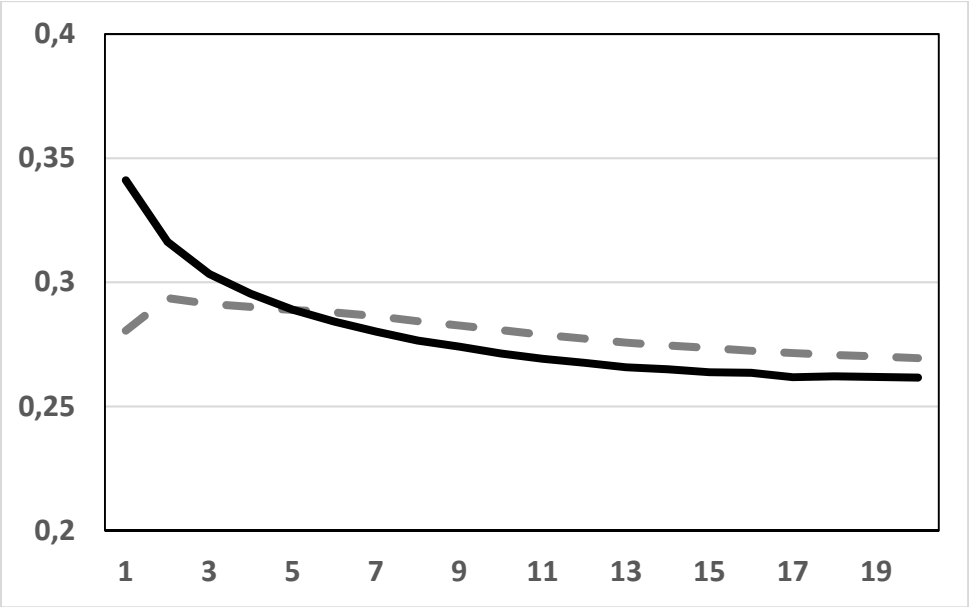

Figure 4. Average similarity between paragraphs related to the distance within the document. (Solid line: within one section, Dashed: across sections)

The overall image in figure 4 can easily be explained by the low similarity between the '*Methodology*' and '*Results*' sections with the two other sections. The main structure of PlosOne papers is either IMRaD or IRDaM with the '*Methodology*' or '*Result*' section in between '*Introduction*' and '*Discussion*' creating higher distance between these sections with higher similarity. Remarkable is the crossing of the two lines near a distance of 5 between paragraph. From then on, paragraphs from different sections are more similar than within sections. Probably, topics or themes already raised in a previous section are retaken in the light of the obtained results or applied methodology.

Between document similarity

The analyses in this next section will all focus on similarity across documents.

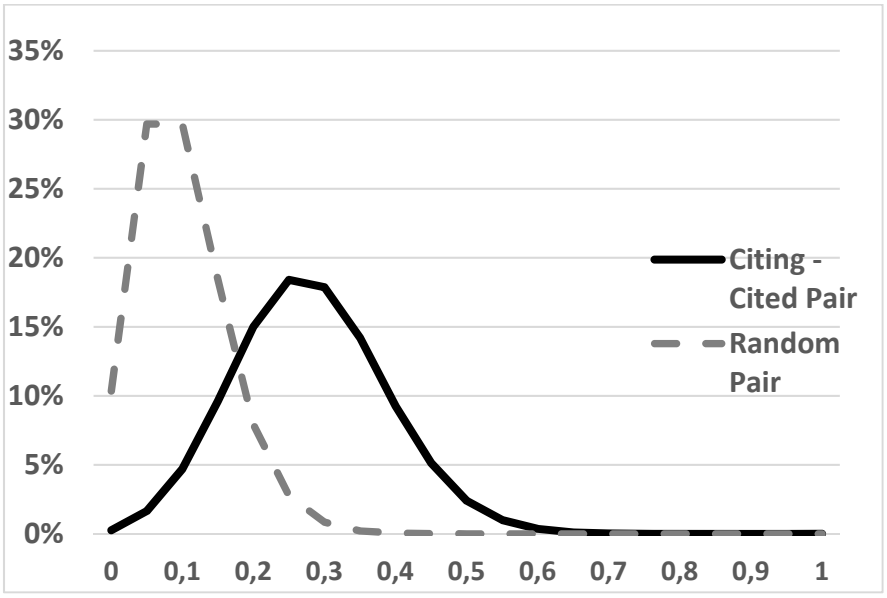

Figure 5. Distribution of similarity as measured through vector document embeddings of abstracts of document pairs (solid line: Citing-Cited document pair, dashed line: random pairs).

In order to have a baseline or reference point, the similarity between abstracts in citing-cited document pairs is gauged against the similarity between two randomly selected abstracts. The distribution of both sets of similarities have been plotted in figure 5.

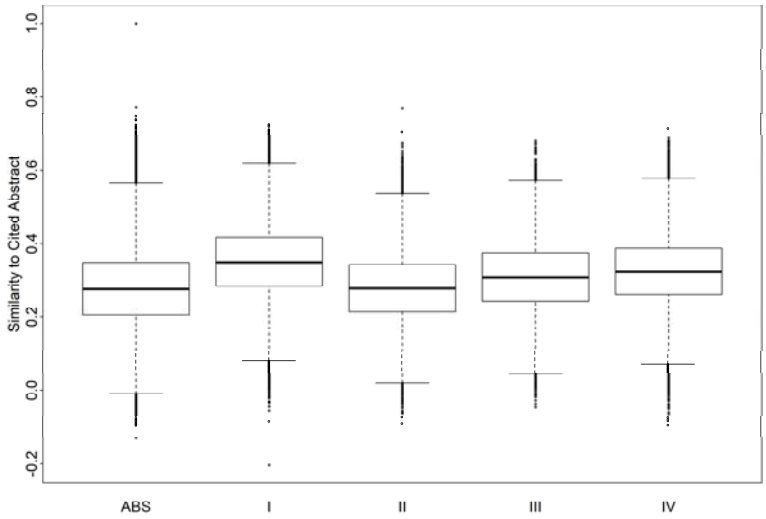

Figure 6. Distribution of similarity between abstract or citing section and abstract of cited document.

The similarity between random selected abstracts is just below 0.09, while the average for the citing-cited pair of documents is 0.28. The distribution of similarities of abstracts of citing and cited pairs of documents is also in figure 6. Here it is contrasted by the distribution of similarity of the paragraph in which the reference appears and the abstract of the cited document, grouped by citing section. Once more, paragraphs in the '*Introduction*' show the highest similarity with the cited abstract. The median in the second box is highest while the first box (abstract to abstract) has the lowest median. This shows clearly that the information in these individual paragraphs bear different content or information than the abstract.

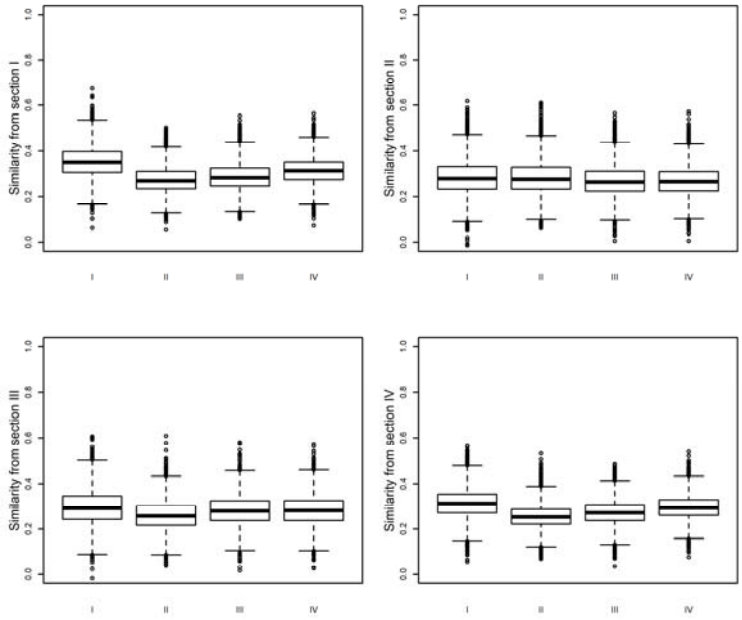

Figure 7. Distribution of similarity between citing section and different sections in cited document.

For the last analysis, the similarity is calculated between the citing paragraph and all paragraphs in the cited document. A citation does not contain -it exceptionally does- a reference to the exact location of the relevant concept or topic in the cited document. This last analysis selects only those PlosOne papers cited at least 5 times by other PlosOne papers. Figure 7. plots the average similarity between citing paragraphs and cited paragraphs across different sections. A plot for each section at the citing side is given. Each plot contains a box per section in the cited document. Paragraphs from the '*Introduction*' and '*Discussion*' section are most similar with the '*Introduction*' in the cited document with 'Discussion' ranked second. This pattern changes when looking at the citing paragraphs in the '*Methodology*'-section. Here cited '*Methodology*' and '*Introduction*' score equally.

Discussion and Conclusion

The results obtained in this study support the statement that a more fine-grained approach using paragraphs is applicable in science mapping and that it will provide additional insights in the topic structure underlying scientific papers. As earlier observed (see Bertin et al. 2013), each section in a publication serves different purposes with distinct reference distribution. Here it is shown that there is also a textual difference between sections but also within sections. The further paragraphs are separated from each other in a section the less similar they are. The use of vector word and paragraph embeddings can be useful for several applications in quantitative science studies. In the following section, applications of intra- and inter document similarity are presented.

Applications

The use of intra-document similarity between paragraphs can extend the study of interdisciplinarity. Currently two main approaches are applied for the study of interdisciplinarity of scientific publications namely the use of subject classifications of cited references (Leydesdorff & Rafols, 2011, Wang et al 2015) and the disciplinary profile of the researchers involved (Abramo et al., 2012). Using the lexical information embedded in the distinct paragraphs and sections combined with the similarity to cited documents can provide a novel third approach. Network-based statistics like node distance, centrality and modularity are appropriate measures for central concepts in the study of interdisciplinarity like disparity, balance and variety (Wang et al 2015).

Another application of this fine-grained approach is in information retrieval. Context based word embeddings provide enhancements at both *needle* and *haystack* side. Key words in search strategies can be complemented by their specific context which defines their meaning and the same model is used to characterize the paragraphs at the haystack side. Moreover, other applications of word embeddings (see eg. Mikolov et al., 2013) have shown that mathematical operations on vectors like subtractions are possible and retain their topical characterization. This allows the creation of search strategies starting from a set of keywords without the need to list all possible alternatives or variations but also to provide a set of keywords or papers that are irrelevant to the search and should be excluded.

Limitations.

The novel approach presented in this study does not come without limitations. At first, there is the need for open access to the full paper in the required JATS-format. The procedure could be rewritten to be applicable on HMTL data or even on parsed PDF. However, the main advantage

of JATS is that it specially targeted towards scientific journal articles and parsing is less prone to errors.
The scoring of additional documents not in the original dataset is possible through the neural network algorithm. The obtained model can even be trained to incorporate these additional documents, but the procedure initially starts with the creation of a word vocabulary which cannot be updated. This puts a burden on the extensibility of the model as the proposed approach also takes the publication identifiers of the cited documents. These identifiers are merged into vocabulary as if they were words. The set of cited documents in the additional data set will thus be limited to the original set of cited documents. Other approaches for word embeddings like LSH or random projects suffer also from this limitation.
The model reduces the document space from extreme sparse with hyper dimensionality into a dense matrix with limited predefined number of dimensions. Using such a dense matrix for the creation of document networks results in a near complete network where a similarity can be calculated for nearly any given pair of documents. It is very hard to use these near complete weighted networks as a basis for clustering techniques or community detection. Only the application of thresholds on the similarity can solve this issue which comes with computational constraints as the similarity of each pair of documents has to be calculated prior to the application of the threshold. Hashing algorithms like LSH can be used to solve this issue.

The creation of the neural network model involves tuning several parameters like number of underlying nodes or dimensions, learning rate, learning iterations, minimum threshold for rare words, down-sampling rate for frequent words, sliding window length for the word context. With the last option, frequent words are removed with a probability relative to the inverse of their frequency which results in actual larger windows. The Word2Vec also provides two different learning approaches. Each of these parameters can have an influence on the final obtained model. More research is required to study the effect of this hyperparameter tuning on the final validity of the model and resulting vector embeddings.

Conclusion

Vector word and paragraph embeddings provide a novel approach for the calculation of within and between document similarities. The technique is used to create neural network based mathematical representations of text fragments of smaller size like paragraphs. Within such a vector space, the cosine of the angle between the vectors can be used to indicate the similarity between the underlying text fragments. The Word2Vec and Doc2Vec implementations provide an easy to use library for the creation of the word embeddings and similarity calculations. The application of the technique shows that the paragraphs in the '*Introduction*' and '*Discussion*' section are most similar to the abstract but that the '*Methodology*' has a much lower similarity with abstract. Combined with lower number of references in this section, the paragraphs are less presented in document-based approaches using abstracts and citations for the creation of document networks. When looking at citing-cited pairs of documents, the paragraph containing the actual reference to the cited paper shows a higher similarity with the abstract of the cited paper. This is especially the case with paragraphs from the introduction. The novel approach can have several applications in quantitative science studies like the study of interdisciplinarity or bibliometric aided information retrieval, but the technique suffers still from limitations which can damper the validity of the obtained results.

References

Abramo, G., D'Angelo, C. A., Costa, F. D. (2012). Identifying interdisciplinarity through the disciplinary classification of coauthors of scientific publications. Journal of the Association for Information Science & Technology, 63(11), 2206–2222.

Ahlgren, P., Collinader, C., (2009), Document-document similarity approaches and science mapping: Experimental comparison of five approaches. *Journal of Informetrics*, 3 (1), 49-63.

American National Standards Institute, (2012). JATS: Journal Article Tag Suite. ANSI/NISO Z39.96-2012, National Information Standards Organization.

Bertin M., Atanassova I., Larivière V., Gingras, Y. ,(2013) The distribution of References in Scientific Papers: an Analysis of the IMRaD Structure. In: *Proceedings of the 14th International Conference of the International Society for Scientometrics and Informetrics*. Vienna, Austria, 591-603.

Bertin M., Atanassova I., Sugimoto CR., Larivière V., (2016). The linguistic patterns and rhetorical structure of citation context: an approach using n-grams. Scientometrics 109:1417-1434.

Blei, David (April 2012). "Probabilistic Topic Models". *Communications of the ACM*. 55 (4): 77–84.

Boyack, K.W., (2017). Investigating the effect of global data on topic detection. *Scientometrics*. 111 (2), 999-1015.

Chen, D., Mannig, C.D., (2014). A Fast and Accurate Dependency Parser using Neural Networks. In: *Proceedings of EMNLP 2014*. Doha, Qatar.

Gal, D., Thijs, B., Sipido, K., Glänzel, W., (2017) Topic modelling based network maps in cardiovascular research. In: *Proceedings of the 16th International Conference of the International Society for Scientometrics and Informetrics*. Wuhan, China, 591-603.

Gipp, B, Beel, J. (2007) Citation Proximity Analysis (CPA) - A New Approach for Identifying Related Work Based on Co-Citation Analysis. In: *Proceedings of the 12th International Conference of the International Society for Scientometrics and Informetrics*. Rio de Janeiro, Brazil, 571-575.

Glänzel, W., Thijs, B., (2017). Using hybrid methods and 'core documents' for the representation of clusters and topics: the astronomy dataset. Scientometrics, 111 (2), 1071-1087.

Harris, J.A., Arabzadeh, E., Fairhall, A.L., Benito, C., Diamond, M.E. (2006). Factors affecting frequency discrimination of vibrotactile stimuli: implications for cortical encoding. *PlosOne*, 1(1), e100.

Kiss, J.Z., Aanes, G., Schiefloe, M., Coelho, L.H.F., Millar, K.D.L., Edelmann, R.E., (2014). Changes in operational procedures to improve spaceflight experiments in plant biology in the European Modular Cultivation System. *Advances in Space Research*, 53 (5), 818-827.

Leydesdorff, L., Hellsten, I., (2006). Measuring the meaning of words in contexts: An automated analysis of controversies about 'Monarch butterflies,' 'Frankenfoods,' and 'stem cells'. *Scientometrics*, 67 (2), 231-258.

Leydesdorff, L., Rafols, I. (2011). Indicators of the interdisciplinarity of journals: Diversity, centrality, and citations. *Journal of Informetrics*, 5(1), 87–100.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. CoRR, abs/1301.3781.

Pennington, J., Socher, R., Manning, C.D., (2014). GloVe: Global Vectors for Word Representation. (available at: https://nlp.stanford.edu/pubs/glove.pdf)

Quoc, L. & Mikolov, T., (2014), Distributed Representations of Sentences and Documents. In: *Proceedings of the 31th International Conference on Machine Learning, ICML*. Beijing, China, 1188-1196.

Rehurek, R., Sojka, P. (2010). Software framework for topic modelling with large corpora. *Proc. LREC Workshop on New Challenges for NLP Frameworks*

Small, H., (1994). A SCI-map case-study – building a map of AIDS research. *Scientometrics*, 30 (1), 229-241.

Taşkin Z and Al U. (2018.) A content-based citation analysis study based on text categorization. *Scientometrics* 114(1):335-337

Thijs, B., Glänzel, W., Meyer, M.S. (2017) Improved lexical similarities for hybrid clustering through the use of noun phrases extraction. MSI Working Paper Series. University of Leuven, Leuven, Belgium

Thijs, B. & Glänzel, W. (2018), The contribution of the lexical component in hybrid clustering, the case of four decades of "Scientometrics". *Scientometrics*, 115(1), 21–33.

Wang, J., Thijs, B. & Glänzel, W. (2015). Interdisciplinarity and Impact: Distinct Effects of Variety, Balance and Disparity. *Plos One*, 10(5): e0127298

Wang, S., Koopman, R., (2017). Clustering articles based on semantic similarity. *Scientometrics*, 111 (2) 1017-1031.

MANAGEMENT, STRATEGY AND INNOVATION (MSI)
Naamsestraat 69 bus 3535
3000 LEUVEN, Belgium
tel. + 32 16 32 67 00
msi@econ.kuleuven.be
https://feb.kuleuven.be/research/MSI/