A Semantic Similarity Measure for Scholarly Document Based on the Study of n-gram

Yannick-Ulrich Tchantchou Samen

Department of Mathematic and Computer Science, Faculty of Science, University of Maroua, P.O Box: 814 Maroua, Cameroon
E-mail: yannick.samen@imsp-uac.org

Received 10 August 2022; Accepted 13 September 2022;
Publication 28 December 2022

Abstract

The performance of information retrieval systems is closely related to the ability of similarity measures to accurately determine the similarity value between documents or between a query and a document. In this paper, the issue of similarity measures in the context of scholarly documents is addressed. A semantic similarity measure is proposed. This similarity measure is able to exploit the metadata contained in the scientific articles, as well as the important n-grams identified in them. To evaluate the accuracy of our similarity measure, a dataset of articles is built as well as their similarity values manually estimated by human experts. Experiments performed on this dataset using Pearson correlation show that the similarity values obtained using the proposed measure are very close to those estimated by human experts.

Keywords: Semantic similarity, n-gram, natural language processing, scholarly document, similarity measure.

Journal of Web Engineering, Vol. 21_7, 2095–2114.
doi: 10.13052/jwe1540-9589.2175
© 2022 River Publishers

1 Introduction

Similarity measures are at the heart of the design and implementation process of information retrieval systems and search engines [1]. It is almost impossible to propose an adequate answer to a user's information need without using a similarity measure. Therefore, the precision of information systems is closely related to the ability of similarity measures to identify the right information, corresponding to the identified need [2].

This focus on similarity measures is also justified by the proportion of scientific contributions observed [3].

Over time, research has focused on the definition of similarity measures. Several works have classified similarity measures into several classes [4]. On the one hand, syntactic, semantic and fuzzy measures [5, 6]. Whereas, the hybrid similarity measures combining several approaches [7].

Studies conducted show that there is no single best similarity measure, as it depends on the application field and the use you want to make with [9]. However, the most commonly used similarity measure for information retrieval systems is cosine similarity [10]. Although this similarity measure gives good results, but it does not considers the context or the semantics between the words constituting each document. Because of the polysemy of words, to determine the similarity between two documents, it is necessary to consider the context in which these words are used [11].

The problem of measuring similarity is also important in the context of scholarly documents [12]. It is not always easy to determine precisely which scholarly documents address a given problem. However, in order to conduct thorough research on a problem, researchers need to investigate scholarly contributions related to these problems. However, most scholarly documents are not free: only their metadata are generally open access. This raises the following questions: How can we use natural language processing approaches to identify relevant concepts in the scholarly document? How can we define a good weighting measure able to determine the importance of a word in a scholarly document using the accessible metadata of that document? Finally, how to define a similarity measure, able to evaluate the similarity degree between two scholarly documents.

This paper makes several contributions. The proposal of an approach to identify important concepts in scholarly documents using n-grams and natural language processing algorithms. Next, the implementation of a semantic similarity measure using the identified concepts to define how similar two documents are. The experiments performed on the model show that it is able

to identify with accuracy the important concepts in a scholarly document. Also, the proposed similarity measure is able to determine the similarity degree between two scholarly documents, but also, to identify the cases of plagiarism in the documents.

The remain of this article is organized as follows: In the next section, the focus is on related work. The Section 3 presents a complete description of the model. While section 4 talks about the experimentation phase as well as the results obtained. This article ends with a conclusion and perspectives.

2 Related Works

Several works have been done to propose a similarity measure able to define the degree of similarity between documents and sentences. These contributions can be classify around large families according to the used approaches.

According to [5], we can distinguish: the string-based similarity composed by character-based similarity measures class and the term-based similarity measures class. Then, the corpus-based similarity which is a class of semantic similarity measure that determines the similarity between words according to information gained from large corpora. As well, the knowledge-based similarity which is one of semantic similarity measures that bases on identifying the degree of similarity between words using information derived from semantic networks [13]. Finally, hybrid similarity measures, obtained by combining several single similarity measures.

In recent years, several works have been done to exploit information on word meaning, context and semantics, to propose better textual similarity measures. In [14] Chen et al. proposed SIMCR a semantic similarity measure integrating multiple conceptual relationships for Web service discovery. SIMCR enables more accurate service-request comparison by treating different conceptual relationships in ontologies such as is-a, has-a and antonomy differently. Yousfi et al. in [15] proposed CSSM, a Context-based Semantic Similarity Measure able to perform semantic similarity comparisons properly, and obtains high accuracy. Little et al. [16] presented a new Semantic and Syntactic Similarity Measure (TSSSM) for political tweets. The approach uses word embeddings to determine semantic similarity and extracts syntactic features to overcome the limitations of current measures which may miss identical sequences of words. Meymandpour and Davis [17] addresses the problem of semantic similarity measure by developing a systematic measurement model of semantic similarity between resources in Linked Data.

Adhikari et al. [18] propose a generic intrinsic IC-based Semantic Similarity calculator. They also introduce a new structural aspect of an ontology called Disjoint Common Subsumers that plays a significant role in deciding the similarity between two concepts. Jiang, Wang and Zheng [19] propose a semantic similarity measure based on information distance for ontology alignment.

It appears that the similarity measures depend on their context of application. In the case of information retrieval systems, some similarity measures have been proposed: Zou and Valizadeh [20] proposed a query-sensitive similarity measure (QSSM) for information retrieval, to measure the similarity of two documents. Pushpalatha and Ananthanarayana [21] addresses the problem of similarity measure by proposing a information theoretic similarity measure for unified multimedia document retrieval. Gupta et al. [22] define a fuzzy-based approach to develop hybrid similarity measure. The proposed approach overcomes the limitations of extensively used similarity measures such as Cosine, Jaccard, Euclidean and Okapi-BM25. Ramya et al. [23] build a similarity measure which helps to retrieve more relevant documents from the repository, thus contributing considerably to the effectiveness of Web Information Retrieval system. Eminagaoglu [24] define a similarity measure for vector space models in text classification and information retrieval. This similarity measure can be effectively used for vector space models and related algorithms such as k-nearest neighbours (k-NN) and Rocchio as well as some clustering algorithms such as K-means.

However, few works have focused on semantic similarity measures exploiting the metadata contained in scholarly documents.

3 Overview of the Proposed Model

Scholarly documents are generally semi-structured documents. However, the metadata they contain is structured. Given the non-accessibility of the full text content of these documents, their metadata are used to infer the important information contained in each document.

In this work, the information exploited comes from the metadata such as the title, the abstract and the keywords of the article. To achieve this, the crossref[1] and Semantic scholar[2] APIs are used to access these metadata.

The complete architecture of the proposed model is shown in Figure 1. The proposed model consists of several steps. These steps range from

[1] https://api.crossref.org

[2] https://www.semanticscholar.org/product/api

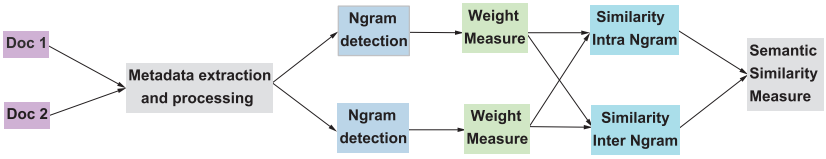

Figure 1 Main architecture of the proposed model.

metadata extraction, through the identification of n-grams and the calculation of the similarity degree between two documents.

3.1 An Indexing Approach Based on n-grams

Many indexing methods in the literature are limited to identifying important atomic or 1-gram concepts in a document. However, documents also contain n-grams ($n > 1$). By indexing a document without considering these terms, one loses contextual information, useful for understanding the content of the document.

For example, let us consider the words "natural", "language" and "processing". Each of these words has a polysemic meaning. By putting them together, the resulting expression conveys something different than if they were considered individually.

In this section, the focus is on identifying important n-grams in a scholarly document.

3.1.1 Metadata extraction and processing

In a scholarly document, there are several categories of metadata to describe the document [25]. We have among others the metadata on the authors, the bibliographic references and finally the metadata dealing with the content of the article. The Figure 2 presents the set of metadata dealing with the content of the article.

These information are extracted from the document and by using natural language processing algorithms, their textual content is cleaned.

A n-gram is a contiguous sequence of elements of length n. It can be a sequence of words, bytes, syllables or characters. The most commonly used n-gram models for text categorization are word and character based n-grams. Examples of commonly used n-grams are: unigram ($n = 1$), bigram ($n = 2$), trigram ($n = 3$) [26].

Our goal is to identify the important n-grams in the metadata. A n-gram is considered important in the document if its constituent words are also important.

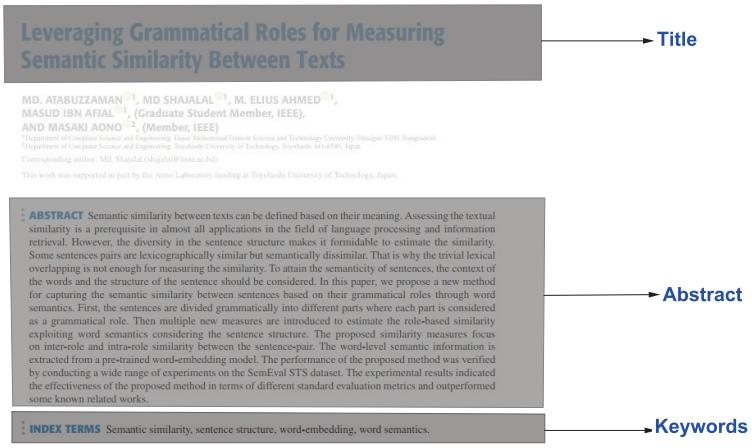

Figure 2 Some useful metadata in scholarly document.

In a document, the set of words constituting his textual content do not all have the same importance in this document. The punctuation and stop words are generally less important words. However, these words are not removed in our metadata as usually during the text mining process. The detection of n-grams is done in a gradual way: firstly, the identification of the important unigrams, then the bigrams, and finally the trigrams.

1. **unigram detection process**
   To determine the important unigram in the metadata, the POS tag (Part of speech tag) of each words is identified. POS tagging is a process which categorize words in a text in correspondence with a particular part of speech, depending on the definition of the word and its context.
   The POS tag of each words in each metadata is extracted (title, abstract or keywords). A unigram will be considered potentially important for a document, if it is either a subject, object complement, adjective or noun. By applying these conditions to each word, all potentially important unigrams are identified.
   Then, lemmatization is applied to each unigram to reduce each word to its lemma. In fact, for grammatical reasons, documents use different forms of a word, such as *language*, *languages*. The lemmatization is used to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form.
2. **bigram detection process**
   A bigram is an expression composed of two words that follow each other. However, a bigram will be considered potentially important for a

document if it consists of a succession of two nouns, or a succession of an adjective followed by a noun. In this process, stopwords and sentence delimiters are useful. Indeed, their presence allows the model to know which words follow each other in the document and their POS tagging. Two unigrams potentially important can be separated by a stopword. By removing all stopwords at the beginning of the process, the model can consider these words as following words and identify them as a potentially important bigram.

3. **trigrams detection process**
   A trigram is an expression consisting of three unigrams. As in the case of bigrams, it is imposed that a trigram will be considered as potentially important for a document, if it is made up of three potentially important unigrams. However, these unigrams must be either an adjective followed by two nouns or three consecutive nouns.
   By doing so, the set of potentially important n-grams for each document is obtained. The next step is to determine the degree of importance of each n-gram in the document.

3.1.2 Weight of a n-gram in a document

After the n-grams identification, the model need to know their weight in the document.

1. **Case of unigrams**
   The weight of a unigram depends on several parameters:
   (a) **The metadata in which the unigram is found.**
   If a unigram potentially important is found in any of the three metadata, this is an indicator of its importance in the document. Depending on the metadata in which a unigram is found, a weight is associated to it in proportion to the importance of this metadata in the structuring of a scholarly document [27].
   If the unigram is in the title, it is assigned a weight of 0.5. If on the other hand it appears in the Keywords, it is assigned a weight of 0.3. and if it is in the abstract, it is assigned a weight of 0.2.
   (b) **The frequency of this unigram in each metadata in which it appears.**
   The fact that a unigram is found in a metadata is not sufficient to conclude that it is important for the document. We must also consider its frequency of appearance in each metadata.

In that way, the weight of a unigram in a metadata by multiplying the weight associated to this metadata by the frequency of appearance of the unigram using the Equation (1).

$$P_{i,m_j} = \alpha_{i,m_j} * u_{m_j} \tag{1}$$

where
α_{i,m_j} is the frequency of appearance of i in the metadata m_j and u_{m_j} is the weight associated to the metadata m_j.

c) **Relationship between the unigram and the higher level n-grams ($n > 1$).**
It can happen that some potentially important unigrams are in potentially important n-grams ($N > 1$). This information is important and deserves to be consider in the process of determining the weight of this unigram in the document.
Considering these parameters, we determine the importance of each unigram by using the Equation (2).

$$\omega_{i,1-gram} = (\sum_{j=1}^{3} \alpha_{i,m_j} * u_{m_j}) e^{\frac{n_i}{n}} \tag{2}$$

where α_{i,m_j} is the frequency of appearance of i in the metadata m_j and u_{m_j} is the weight associated to the metadata m_j.
n_i is the number of n-gram ($n > 1$) potentially important in which i appear and n is the total number of n-gram ($n > 1$) potentially important.

2. **Case of bigrams**
As in the case of unigrams, the importance of a bigram in a document depends on several parameters.

(a) **The metadata in which the unigram is found.**
A bigram is above all a concept of interest. It can be found in any of the metadata described above. As in the case of unigrams, its position gives information about its importance in the document.

(b) **The importance of the potentially important unigrams that constitute it.**
A bigram consists of two unigrams. If the unigrams that make it up are important for the document, this should have an influence on the weight of this bigram. This fact is considered during the process.
Similarly, if the bigram is made up of the unigrams of lower

importance in the document, then this should also have an impact on the weight of the bigram in the document.

(c) **Relationship between the unigram and the higher level n-grams ($n > 2$)**
As with unigrams, a bigram can be a part of a potentially important trigram. We also consider this eventuality during the process.

Using these parameters, the model define the importance of a bigram by using the Equation (3).

$$\gamma_j = \omega_{j,2-gram} + \frac{1}{2} * \eta_j \tag{3}$$

where $\omega_{j,2-gram}$ is obtained by using (2),

$$\eta_j = \sum_{i=1}^{2} \chi_i * \omega_{i,1-gram} \tag{4}$$

$\omega_{i,1-gram}$ is the weight of the unigram at position i in the bigram.
$\chi_i = 1$ if the unigram of the position i in the bigram is a potentially important unigram and $\chi_i = 0$ if not.
χ is used to select the unigrams whose weights will be considering for the calculation of the bigram weight. If the bigram is formed using a potentially important unigram, then to the weight of that bigram, half the weight of the unigram is added.

3. **Case of trigrams**
The weight of a trigram is obtained by generalizing the Equation (3). By doing so, we obtain the Equation (5):

$$\Gamma_{k,3-gram} = \omega_{k,3-gram} + \frac{1}{3} \sum_{l=1}^{2} l * \sum_{k=1}^{4-l} \chi_k * \omega_{k,lgram} \tag{5}$$

Where l is the size of the lower level n-grams. For the weight of a trigram, we add $\frac{1}{3}$ of the weights of the potentially important unigrams it contains, then $\frac{2}{3}$ of the weights of the bigrams it containss.

Let's illustrate this process to determine the weight of the trigrams "Natural Language Processing". The Figure 3 gives us an illustration of the procedure.

Assuming that "Natural Language" is a potentially important bigram, while "Language Processing" is not; and that "Natural", "Language" and "Processing" are potentially important unigrams.

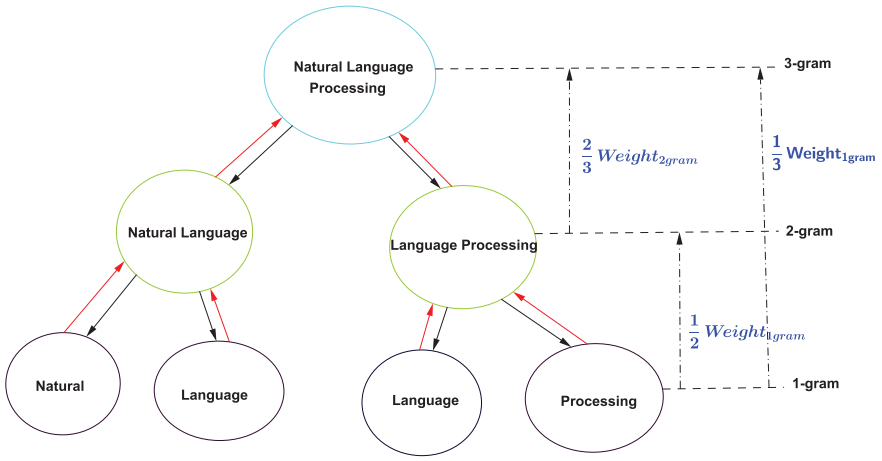

Figure 3 Process illustration of determining of trigram weight.

Because the bigram "Language Processing" is not potentially important for the document, it will not be considered in the trigram weight calculation. However, the bigram "Natural Language" will be considered. For this bigram, $\chi = 1$ and we add $\frac{2}{3}$ of its weight. For "Language Processing", $\chi = 0$. Proceeding in the same way with the unigrams, we obtain:

$$\Gamma_{NLP} = \omega_{NLP} + \frac{2}{3}\omega_{Natural\ Language} + \frac{1}{3}(\omega_{Natural} + \omega_{Language} + \omega_{Processing})$$

At the end of the process, we obtain each n-gram with its weight, representing its simportance in the document. These information are then used to define the similarity measure.

3.2 Proposal of a Semantic Similarity Measure

In this section, the focus is on the problem of similarity measure. Similarity measures are important in several ways: in information retrieval systems, to identify information or contents that meet an information need expressed or not by a user. They also allow to detect plagiarism in scientific works.

The approach that we propose aims at exploiting the information obtained during the previous phase on n-grams to establish a similarity measure able to give with precision the similarity degree between two scholarly documents. This similarity measure is also able to detect plagiarism in scholarly documents.

The the similarity degree identification between two documents is done in several steps. First, we define a similarity measure intra n-grams, then a similarity measure inter n-grams. Finally, the overall similarity value is a combination of these two similarity measures.

3.2.1 Intra n-grams similarity measure

The intra n-grams similarity measure aims at comparing n-grams of the same size between them, to evaluate how they are similar in both documents.

For the unigrams, we identify the concepts that appear at the same time in both documents; their weight and their grammatical role in each document. Is it a subject, adjective, indirect object or direct object. Then, the cosine similarity measure [28] in Equation (6), is used to compute the similaraty value between the unigrams.

$$cosinSim(A,B) = \frac{< A,B >}{||A||||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{6}$$

The process is repeated for the bigrams and the trigrams. Then, we determine the similarity value of each n-grams of the same size, in both documents.

By combining these different similarity values, we obtain the Intra n-grams similarity value using Equation (7).

$$Sim_{Intra}(d_1,d_2) = \alpha * Sim_{Intra1gram} + \beta * Sim_{Intra2gram} + \gamma * Sim_{Intra3gram} \tag{7}$$

Where $\alpha + \beta + \gamma = 1$ and $0 < \alpha, \beta, \gamma < 1$.

However, the value of α, β and γ depends on the use we want to do with. Depending on whether the similarity measure will be used for plagiarism detection or for information retrieval on documents, either unigrams or n-grams ($n > 1$) will be preferred.

Indeed, plagiarism detection aims at finding similarities in text from other sources or at finding copied work from other authors. At this level, the trigrams Intra similarity value will have more weight than the bigrams intra similarity value. In the same time, bigrams intra similarity value will have more weight than unigram intra similarity value. For this reason $\alpha < \beta < \gamma$.

In the context of information retrieval systems, the power balance is inverted. The documents similarity value will depend primarily on the unigrams intra similarity value. The n-grams ($n > 1$) bring semantic

and contextual information on the meaning of the 1grams found in each document. So, $\alpha > \beta > \gamma$.

3.2.2 n-gram Inter similarity measure

The Inter n-gram similarity only applies to the n-gram found in one document and not in both documents simultaneously.

The process of computing this similarity measure involves the following steps:

- In each document, the n-grams which are not found in the two documents is determined.
- Then, the first document d_1 is fixed, and for each n-gram of this document, the semantically closest m-gram is extracted in the other document.
- This m-gram is replaced in the second document by the n-gram with which it is closest. Its weight is obtained by multiplying the weight of the m-gram by its similarity value with the n-gram.

At the end of the process, the cosine similarity value of the two documents is computed.

This similarity value is called the "Inter n-gram Similarity Value".

At the end of the process, the two similarity measures are combined to determine the final similarity value between the two documents by using Equation (8).

$$finalSim(d_1, d_2) = \lambda * Sim_{Intran-gram}(d_1, d_2) + (1 - \lambda) * Sim_{Intern-gram}(d_1, d_2) \quad (8)$$

Where $0 \leq \lambda \leq 1$. $\lambda = 0$ if the two documents have no n-grams in common and $\lambda = 1$ if all the n-gram are in the two documents.

4 Experimentation of the Proposed Model

4.1 Description of the Experimentation Process

To implement our model, some python libraries were used (Spacy, NLTK) for the text processing. Then, Wordnet was used to identify the similar concepts in the different documents.

After the implementation of our model, we evaluated its ability to accurately define the similarity degree between two scholarly documents.

However, evaluating a similarity measure is a difficult task. The notion of similarity degree between two documents is difficult to quantify accurately.

To efficiently evaluate a similarity measure, we needs a predefined dataset containing the data as well as the similarity values estimated as correct for each document pair. In the context of similarity measures on scholarly documents, there is almost no such dataset.

To experiment the performance of this similarity measure, we first build a dataset containing a half thousand pairs of scholarly documents (their metadata), as well as the similarity value estimated by human experts. To build the dataset, we extracted the metadata from scholarly documents using the crossref api.[3] Then, we constituted pairs of documents by using only the articles for which all the interest metadata were available. The similarity value between the documents was evaluated by considering the similarity between the topics treated in each article, the taxonomic relationship between these topics and the important terms found in each documents.

Once the dataset is constituted, our similarity measure is used on each pair of documents of the dataset to define their similarity value.

At the end of the process, a metric evaluation is used to assess the results obtained.

One of the most widely used metric for evaluating relatedness measures is Pearson correlation. It indicates how closely the results of a measurement resemble human judgments. A value of 0 means no correlation and 1 means perfect correlation [29].

The Pearson (r) product-moment correlation coefficient is calculated by using Equation (9)

$$r = \frac{n(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{\sqrt{(n(\sum x_i^2)(\sum x_i)^2}\sqrt{n(\sum y_i^2)(\sum y_i)^2}} \tag{9}$$

where x_i refers to the ith element in the list of human judgments, y_i refers to the corresponding ith element in the list of computed values, and n refers to the number of words paires.

We then use cosine similarity on our dataset to compare the results obtained using our similarity measure (OSSV) with those obtained using the traditional cosine similarity measure (CSV). Finally, we use an improved version of the cosine similarity (ICSV) measure using our n-gram-based indexing approach.

[3]https://www.crossref.org/documentation/retrieve-metadata/rest-api/

4.2 Experimental Results

At the end of these different phases of experimentation, the different similarity values obtained for each pair of documents were compared. The Figure 4 presents firstly, the correlation between the values obtained with the proposed similarity measure (OSSV) with those estimated manually (PSV). This correlation is represented in blue color. The second scatter plot in black color, represents the correlation between the values obtained using the traditional cosine similarity measure (CSV) and those defined by the experts. From this figure, it appears that the points distribution in the cloud tends to be further away from the diagonal in the case of the traditional cosine similarity measure than for the proposed similarity measure.

The Figure 5 presents the correlation between the values obtained with the improved cosine similarity measure using the proposed indexing approach (ICSV) and those estimated manually. This correlation is represented in green color. The second scatter plot, in black color, represents the correlation between the values obtained with the traditional cosine similarity measure (CSV) and those defined by humans. From this figure, it appears that the points distribution in the cloud is almost similar for the CSV and for the ICSV, when the evaluated document pairs are not very similar. However, this points distribution is closer to the diagonal in the case of ICSV than in the case of CSV, when the evaluated document pairs are similar. This shows the ability of the proposed indexing approach to improve the evaluation and the detection process of similar document pairs.

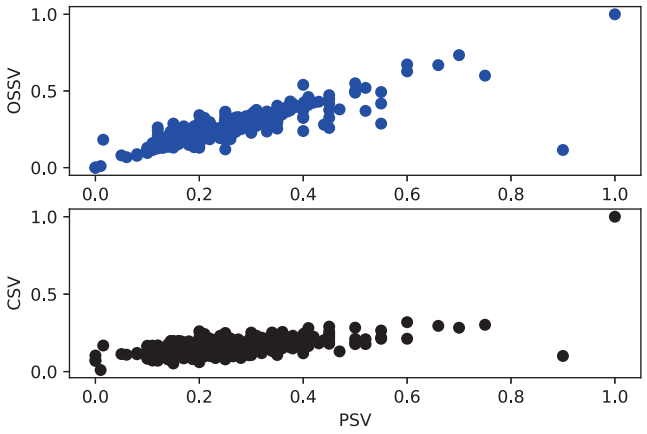

Figure 4 Correlation between (PSV, OSSV) and (PSV, CSV).

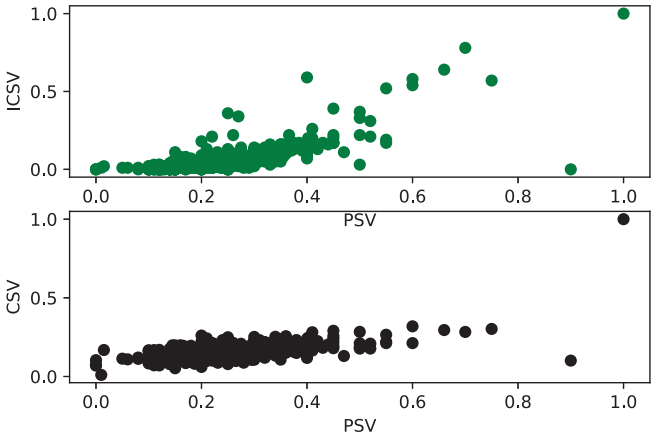

Figure 5 Correlation between (PSV, ICSV) and (PSV, CSV).

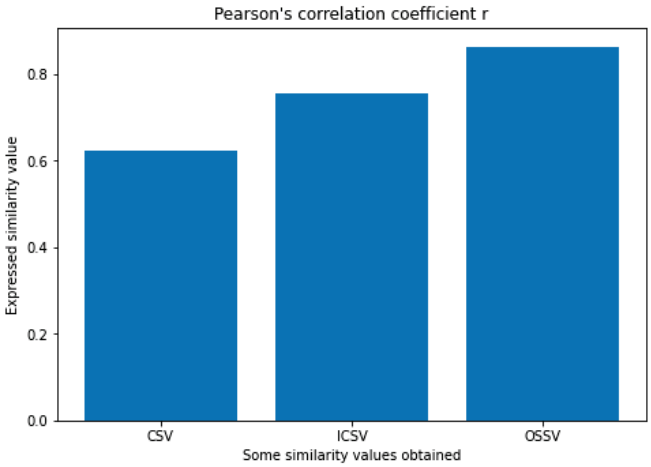

Figure 6 Pearson's correlation coefficient r of each similarity measure.

From these two figures, it can be seen that the joint use of the proposed indexing approach with the proposed similarity measure improves the detection process of similarity value between two scholarly documents.

The Pearson correlation coefficient r was allso calculted to assess the correlation between each similarity measure and the manually made estimates. The results are shown in Figure 6.

It emerges that the proposed similarity measure (OSSV) has a Pearson correlation coefficient of 0.8637, while the ICSV has a Pearson correlation

coefficient of 0.7544. Finally, it appears that CSV achieve a Pearson correlation coefficient of 0.6234.

5 Conclusion and Future Work

In this paper, the problem of similarity measures in the context of scholarly documents is addressed. To this end, a model for extracting metadata from scientific articles is proposed. These metadata are then processed using text mining algorithms, then the important n-grams are determined as well as their importance in each document. These n-grams are used to propose a semantic similarity measure. For the experimentation of our model, a dataset of metadata of scientific articles is built and their similarity values estimated by human experts. Experiments performed using this dataset show that the similarity values obtained with the proposed similarity measure are very close to those manually proposed by human experts.

In perspective, it will be interesting to see how to use this similarity measure to identify plagiarism in scientific articles. Furthermore, it will be useful to apply this similarity measure to build an information retrieval system able to automatically perform survey on given research problems.

References

[1] S. Giridhar, and K. Bhutani. Importance of Similarity Measures in Effective Web Information Retrieval. International Journal on Recent and Innovation Trends in Computing and Communication, 6, 29–33, 2018.

[2] D. Ifenthaler. Measures of Similarity. In: Seel N.M. (eds) Encyclopedia of the Sciences of Learning. Springer, Boston, MA. https://doi.org/10.1007/978-1-4419-1428-6_503, 2012.

[3] K.P. Reddy, T.R. Reddy, G.A. Naidu, and B. Vishnu, Impact of Similarity Measures in Information Retrieval. International Journal of Computational Engineering Research (IJCER), 8(6), 54–59, 2018.

[4] J. Wang and Y. Dong. Measurement of Text Similarity: A Survey. *Information*, 11, 421, 2020. https://doi.org/10.3390/info11090421.

[5] W.H. Gomaa and A.A. Fahmy. A survey of text similarity approaches. international journal of Computer Applications, 68(13), 13–18, 2013.

[6] Y. Song, X. Wang, W. Quan et al. A new approach to construct similarity measure for intuitionistic fuzzy sets. Soft Comput 23, 1985–1998, 2019. https://doi.org/10.1007/s00500-017-2912-0

[7] F. Lan. Research on Text Similarity Measurement Hybrid Algorithm with Term Semantic Information and TF-IDF Method. Advances in Multimedia, 2022. https://doi.org/10.1155/2022/7923262

[8] X. Wan. Beyond topical similarity: a structural similarity measure for retrieving highly similar documents. Knowl. Inf. Syst. 15, 1, 55–73, 2008.

[9] F.L. Liu, B.W. Zhang, D. Ciucci, W.Z. Wu and F. Min. A comparison study of similarity measures for covering-based neighborhood classifiers, Information Sciences, V. 448–449, pp. 1–17, 2018. https://doi.org/10.1016/j.ins.2018.03.030.

[10] R. Subhashini and V.J.S. Kumar. "Evaluating the Performance of Similarity Measures Used in Document Clustering and Information Retrieval", 2010 First International Conference on Integrated Intelligent Computing, pp. 27–31, 2010. doi:10.1109/ICIIC.2010.42.

[11] S. Wan and R.A. Angryk, "Measuring semantic similarity using wordnet-based context vectors," 2007 IEEE International Conference on Systems, Man and Cybernetics, pp. 908–913, 2007. doi:10.1109/ICSMC.2007.4413585.

[12] R. Ibrahim, S. Zeebaree and K. Jacksi. Survey on semantic similarity based on document clustering. Adv. Sci. Technol. Eng. Syst. J, 4(5), 115–122, 2019.

[13] R. Mihalcea, C. Corley and C. Strapparava. Corpus based and knowledge-based measures of text semantic similarity. In Proceedings of the American Association for Artificial Intelligence. (Boston, MA), 2006.

[14] F. Chen, C. Lu, H. Wu, and M. Li. A semantic similarity measure integrating multiple conceptual relationships for web service discovery. Expert Systems with Applications, 67, 19–31, 2017.

[15] A. Yousfi, M.H. El Yazidi and A. Zellou. "CSSM: A Context-Based Semantic Similarity Measure." 2020 IEEE 2nd International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS). IEEE, 2020.

[16] C. Little, D. Mclean, K. Crockett and B. Edmonds. A semantic and syntactic similarity measure for political tweets. IEEE Access, 8, 154095–154113, 2020.

[17] R. Meymandpour and J.G. Davis. A semantic similarity measure for linked data: An information content-based approach. Knowledge-Based Systems, 109, 276–293, 2016.

[18] A. Adhikari, B. Dutta, A. Dutta, D. Mondal and S. Singh. An intrinsic information content-based semantic similarity measure considering

the disjoint common subsumers of concepts of an ontology. Journal of the Association for Information Science and Technology, 69(8), 1023–1034, 2018.

[19] Y. Jiang, X. Wang and H.T. Zheng. A semantic similarity measure based on information distance for ontology alignment. Information Sciences, 278, 76–87, 2014.

[20] A.J.M. Zou and M.R. Valizadeh. A proposed query-sensitive similarity measure for information retrieval, 2006.

[21] K. Pushpalatha and V.S. Ananthanarayana. "An information theoretic similarity measure for unified multimedia document retrieval." 7th International Conference on Information and Automation for Sustainability. IEEE, 2014.

[22] Y. Gupta, A. Saini and A.K. Saxena. Fuzzy logic-based approach to develop hybrid similarity measure for efficient information retrieval. Journal of Information science, 40(6), 846–857, 2014.

[23] C. Ramya, S.P. Paramesh and K. S. Shreedhara. "A New Similarity Measure for Web Information Retrieval using PSO Approach." 2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS). IEEE, 2018.

[24] M. Eminagaoglu. "A new similarity measure for vector space models in text classification and information retrieval." Journal of Information Science, 2020.

[25] D. Tkaczyk, P. Szostek, M. Fedoryszak et al. CERMINE: automatic extraction of structured metadata from scientific literature. IJDAR 18, 317–335, 2015.

[26] H. Ahmed. Detecting opinion spam and fake news using n-gram analysis and semantic similarity. PhD Thesis, University of Ahram Canadian, 2012.

[27] Y.U.T. Samen and E.C. Ezin. "An Improving Mapping Process Based on a Clustering Algorithm for Modeling Hybrid and Dynamic Ontological User Profile", 2017 13th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), pp. 1–8, 2017. doi:10 .1109/SITIS.2017.12.

[28] R. Baeza-Yates and B. Ribeiro-Neto. Modern information retrieval (Vol. 463). New York: ACM press, 1999.

[29] H.T. Mohamed Ali, T. Zesch and M.B. Aouicha. "A survey of semantic relatedness evaluation datasets and procedures." Artificial Intelligence Review 53.6, 4407–4448, 2020.

Biography

Yannick-Ulrich Tchantchou Samen received a BSc of pure Mathematics, a MSc of error correcting code from the Dept. of Mathematics, Faculty of Science, at the University of Yaounde 1, Cameroon, in 2011, and 2013 respectively. He received a PhD of Semantic Web at the Institute of Mathematics and Physical Sciences, University of Abomey-calavi, Benin in 2017. He has been with the Laboratory of Research in Computer science and Applications (LRSIA) since 2017 as a Researcher. Since 2021, he is a Lecturer of Computer Science from the Dept. of Mathematics and Computer Science at the University of Maroua. His current research areas include Semantic Web, Information Filtering, Natural Language Processing, and Web mining.