An N-gram-based Information Retrieval Approach for Surveys on Scientific Articles

Yannick Ulrich Tchantchou Samen[1,2]
[1]Department of Mathematic and Computer Science, Faculty of Science, University of Maroua, P.O Box:814 Maroua, Cameroon
[2]Laboratoire de Recherche en Sciences Informatiques et Applications (LRSIA), UAC, Abomey-Calavi, Benin
E-mail: yannick.samen@imsp-uac.org

Keywords: NLP, information retrieval, scientific article, n-gram, similarity measure

Received: March 13, 2024

Humans are constantly searching for knowledge. This quest for knowledge has pushed back the boundaries of science. As a result, new scientific contributions are published daily in a variety of fields. However, it is not easy for a novice researcher to visualize all existing scientific contributions to a specific research problem in a short period of time. This study proposes an approach for extracting useful information from the metadata of scientific documents. Then, the design of an intelligent search system exploits the metadata contained in scholarly documents to provide an overview of scientific contributions to a research problem. The proposed model uses a new similarity measure based on the extraction of n-grams from the metadata of scientific articles. The model offers each user the possibility of visualizing the results of scientific contributions proposed by researchers in the form of a graph. Experiments carried out on a dataset of 126k data show that the model we propose achieves an overall precision of 0.89*, a recall of* 0.84 *and an F1-score of* 0.86*. This shows that the model can refine the search to provide scientific contributions that have a direct correlation with a user's need*

Povzetek: Predstavljen je nov pristop za iskanje informacij v znanstvenih člankih, ki uporablja n-gram tehniko in naravno jezikovno obdelavo za izboljšanje podobnosti med članki, omogočanje učinkovitejšega razvrščanja in prikaza raziskovalnih prispevkov ter zagotavljanje boljše vizualizacije in prepoznavanja relevantnih znanstvenih vsebin.

1 Introduction

Today, there is an exponential increase in the number of peer-reviewed scientific articles and journals [1]. It is increasingly difficult for a novice researcher to get up to speed on specific research questions. Sometimes, a brief overview of the field can be gained by using a published survey paper. However, the authors of the survey papers only address the issues from a specific angle of interest. Their concern is not always to provide a comprehensive review of the work in the requested area. Novice researchers face a number of difficulties. The first is information overload. The massive quantity of publications available on a given subject. Numerous articles, books, theses, reports and other sources may deal with similar aspects, but are not always directly relevant to the specific research question. Another major difficulty may be the lack of clear selection criteria. A novice researcher may not know how to refine his or her criteria to sort out the most relevant articles. For example, it may be difficult to determine what criteria to use to assess the quality and rigor of research: should articles be recent? What is the reputation of the journals in which they are published? Does the author have recognized expertise in the field? Should studies be quantitative or qualitative? All this does not allow the user to feel sufficiently immersed in the actual advances on the problem, and to form his or her own opinion about future work in the research area.

Moreover, extracting relevant data from scientific articles in Pdf format (Portable Document Format) remains a constant concern. Some authors have reflected on the question by suggesting approaches aimed at defining mechanisms for extracting important metadata in scholarly documents [2, 3, 4, 5, 6]. Traczyk et al. [7] proposed the CERMINE system to extract some metadata (title, author's names, keywords, reference) in certain scientific articles. However, several metadata such as the introduction and conclusion, which are important for understanding the textual content of an article, are not extracted.

In addition, the authors addressed the problem of clustering scientific papers based on citations. These works aim to bring out the graph of articles dealing with a problem by using citations [8, 9]. In [9], the authors proposed a graph model based on the clustering of scientific papers using an extended citation model. However, by clustering articles according to citations, there is a risk of including numerous articles which do not explicitly deal with the problem posed but just provide tools to tackle one aspect of the problem.

To design such systems, it is common to use similarity measures [10]. These measures allow evaluating the similarity between two or more documents in order to define which ones meet the user's needs. 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 [11].
The problem of measuring similarity is also important in the context of scientific papers [12]. It is not always easy to determine precisely which papers address a given problem.
In this paper, an information retrieval system for scientific articles is proposed. This system can automatically identify all the works published on a specific research problem. The system provides a graph to visualize the evolution of the scientific work done by the authors on the issue and then proposes a classification of the relevant articles on the topic. The classification takes into account the importance of the scholarly document concerning the problem posed and the link between all existing works on the question.
To do so, the model uses an approach that identifies the important concepts in a scientific article using n-grams and natural language processing algorithms. Then, it builds a semantic similarity measure using the identified concepts to define the similarity between two articles. This similarity measure is used to group articles into clusters. Each cluster represents a sub-field of computer science research. The rest of the paper is organized as follows: the next section discusses related work. Section 3 presents a semantic similarity measure based on the n-gram; Section 4 presents the overview of the proposed model as well as its different specificities. In the next section, the experimentation of the model components is performed. This work ends with a conclusion and perspectives.

2 States of the art

2.1 Metadata extraction on scholarly documents

The day-to-day activities of researchers have helped push the boundaries of science. However, these advances open the door to many other problems such as the multiplication of the volume of data [13]. Because these articles are generally published in pdf format, the problem often arises of knowing how to automatically extract the metadata from these files. In addition, how to offer the possibility to a junior researcher to make a quick investigation on the work carried out on a research problem.
Scholars face various problems concerning metadata extraction in scientific papers. Many studies have proposed approaches for successfully extracting metadata [25, 15, 16, 17]. However, the main issue concerns the different varieties of metadata that can be extracted by utilizing these approaches. A. Souza, et al. [13] proposed the ARTIC model. ARTIC is a method for extracting metadata from scientific documents. It uses a two-layer probabilistic framework based on Conditional Random Fields to extract metadata. Tkaczyk et al. in [7] proposed an open-source system for extracting structured metadata from scholarly articles in a born-digital format. Their CERMINE system can extract some metadata types with an average F score of 77.5%. In [18], the authors used formatting templates and implicit formatting of semantic information for automatic metadata identification and segmentation. They built a pipeline program, namely PAXAT, to implement their approach for metadata extraction.
However, similar to the previously mentioned works, PAXAT cannot extract metadata such as introduction, conclusion, and important results. S. Qiu and T. Zhou in [19] studied the technologies and applications of metadata extraction in digital books, and proposed a metadata extraction method using information compensation from the web. Other authors have adopted rule-based techniques to address this issue. Hashimi et al. [20] presented a rule-based approach to extract metadata from the research articles while in [21], Zaman et al. proposed a novel ontological framework for information extraction (OFIE) using a fuzzy rule-base and word sense disambiguation to extract information from diverse scientific sources. Recently, S. Li and Q. Wang in [22] proposed a hybrid model for recognizing generic sections in scientific papers. This model considers both sections of headers and body text to automatically recognize generic sections in academic documents. The model achieved 91.67% F1-value in the generic section recognition of sections.
Table 1 presents a comparative study of the strengths and weaknesses of a number of metadata extraction models for scholarly documents.
Several authors have studied hybrid approaches combining metadata and semantic embeddings for processing information contained in scholarly documents. Bhagdev et al. [26] present Hybrid Search; a method that integrates ontology-based and keyword-based search techniques. They address the limitations of semantic search methods by combining the two approaches, thus improving precision and recall in document retrieval. Bodo and Csato [27] present a hybrid method for metadata extraction that combines classification and clustering techniques. It allows the desired information to be extracted without the need for conventional labeled datasets, making it applicable to a variety of document analysis tasks, including document layout analysis. In [28], Mitrov et al. Combine semantic matching, word embeddings, transformers, and LLMs for enhanced document ranking. They propose a new methodology, weighted semantic matching (WSM) combined with MiniLM, to improve document search performance. By integrating various semantic matching techniques, the approach achieves superior precision and recall measures, demonstrating the effectiveness of combining different methodologies. Raman et al. in [29] propose a robust representation of documents using latent topics and metadata. Their research addresses the challenges of document classification when

labeled examples are scarce. They propose generating document representations that capture both textual content and metadata artifacts, employing a self-supervised approach that learns a soft-partition of the input space, leading to improved classification performance with limited labeled data. In [30], Aman Ahluwalia et al. offer a hybrid semantic search approach capable of revealing user intent beyond keywords. They address the limitations of traditional keyword-based search in understanding user intent. It introduces a hybrid search approach that leverages non-semantic search engines, Large Language Models (LLMs), and embedding models to deliver highly relevant and contextually appropriate search results. In [31], the authors present a hybrid approach for metadata extraction from scholarly articles, merging structural and semantic data. The method enhances search accuracy by facilitating precise searches through machine-readable metadata, which supports semantic queries on document metadata and structural elements.

2.2 clustering scholarly documents

While the work on extracting metadata from scholarly documents continues to attract the attention of the scientific community, several researchers have instead focused on clustering scientific articles according to their similarity and/or their proximity using the citations [23, 24, 25, 32].
In [9], Zhang et al. (2019) proposed an extended citation model for scientific articles clustering. They considered various parameters like the frequency, integrated text, and wide distribution of a scientific document quoted in other documents. However, this work doesn't cover many kinds of scholarly documents (documents with letters, editorials, overviews). Yoon et al. [8] defined a probabilistic network graph for fine-grained document clustering and developed a probabilistic generative model and calculation method. They analyzed the relevance of the documents based on their content and rankings, and they proposed an innovative document-embedding approach that considers both the relevance and content of network-based document clustering. Rinartha and Kartika in [33] proposed a system able to classify articles using a computer system based on the contents of the article. They built their article clustering system by using the word frequency method adopted from the term frequency of TF*IDF and using cosine similarity to cluster scholarly articles according to the research topics.

3 A semantic similarity based n-gram

3.1 A scholarly document indexing approach

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.
An n-gram is a contiguous sequence of elements of length n. It can be a sequence of words, bytes, syllables, or characters.
The purpose of this step is to identify the important n-grams in the metadata. An n-gram is considered important in the document if its constituent words are also important.
In a document, the set of words constituting his textual content does 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 word 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 word in each metadata is extracted. 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*, and *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.

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 potentially important

Table 1: Comparative study of some metadata extraction approach on scholoarly documents

Contributions	Strength	Weakness	Precision	F1-value
A. Souza et al., [13], 2014	A probabilistic approach to extracting metadata such as article title, authors' names and affiliations	Very limited metadata. Impossible to extract the information needed to understand the content of each article		99%
D. Tkaczyk et al.[7], 2015	Extract a larger amount of structured metadata such as title, abstract and even bibliographic reference information.	Its accuracy in extracting metadata is average. In addition, it is difficult to extract textual information from different parts of each article.	81%	77.5%
C. Jiang et al., [18] 2018	A formatting templates and implicit formatting semantics informations to extract pure texts in article on PDF formatting	This template allows access to the article's overall textual content without being able to specify the individual metadata contained in the article.	94.07%	
S. Qiu and T. Zhou [19], 2019	Study the keys technologies and applications of metadata extraction in digital books and proposes a metadata extraction method using information compensation from Web	Focuses on digital books but is not interested in scientific articles. In addition, the model is obliged to use the Web to complete the information it has not been able to extract from the documents.		98% (in terms of title and authors extraction)
I. Safder et al. [16], 2020	Innovative approaches using a Bi-LSTM model to extract algorithmic pseudo-codes and sentences from algorithmic metadata.	This model is essentially concerned with the algorithms and pseudo-code found in the articles. It is unable to extract other metadata.	78.5%	93.32%
A.M. Hashmi et al., [20] 2020	Rule based approach to extract metadata from scientific Pdf documents	Don't consider full tex contains in scientific Pdf documents		93.33%
G. Zaman et al. [21] 2021	A novel ontological framework for information extraction using fuzzy rule base and word sense disambiguation. The approach is validated wide document domain publishing in IEEE, ACM, Elsevier and Springer.	However, it doesn't cover many other scientific article formats, nor does it allow real-time metadata extraction.	89.14%	89%
S. Li and Q. Wang [22], 2021	Hybrid model which considers both section headers and body texts to recognize generic sections in scholarly documents	The work is limited to a predefined scientific article format, but the structure of articles is not always the same for each journal and type of article.		91.67%

for a document if it is made up of three potentially important unigrams.
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.

3.2 Weight of a n-gram in a document

After the n-grams identification, the model needs 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 [54].
   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. if the word is in the rest of the document, it is associated with a weight of 0.1.
   This weight distribution of the metadata is made by considering the importance of each metadata in the structure of a scholarly document. Others weighting are possible, but only if they reflect the predominance relation between the metadata in a scholarly document.

   b) **The frequency of this unigram in each metadata in which it appears**
   The fact that a unigram is found in metadata is not sufficient to conclude that it is important for the document. We must also consider its frequency of appearance in each metadata.

   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-g ($N > 1$). This information is important deserves to be considered in the proces determining the weight of this unigram i document.

2. **Case of bigrams**
   As in the case of unigrams, the importance of a bi 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. This eventuality is also consider during the process.

3. **Case of trigrams**
   The weight of a trigram is obtained by generalizing the same process for the bigram.

Let's illustrate this process to determine the weight of the trigrams "Natural Language Processing". The Figure 1 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.
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:

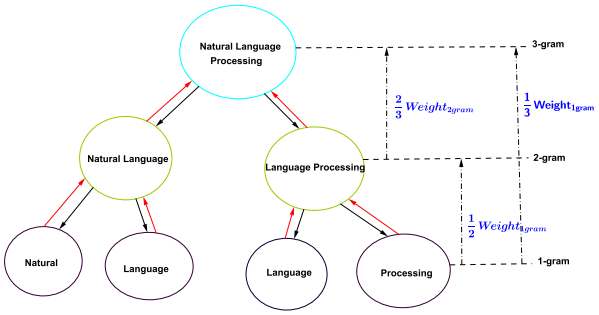

Figure 1: Process illustration of determining trigram weight

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

At the end of the process, each n-gram with its weight representing its importance in the document is defined. These informations are then used to define the similarity measure.

3.3 A semantic similarity measure based on n-gram

In this section, the focus is on the similarity measure problem. 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 the detection of 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 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.

3.3.1 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 sim-

ilarity measure [55] is used to compute the similarity value between the unigrams. 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 2.

$$Sim_{Intra}(d_1, d_2) = \alpha * Sim_{Intra1gram} + \beta * Sim_{Intra2gram} + \gamma * Sim_{Intra3gram} \quad (2)$$

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

3.3.2 Inter n-gram 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 are 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 3.

$$finalSim(d_1, d_2) = \lambda * Sim_{Intran-gram}(d_1, d_2) + (1 - \lambda) * Sim_{Intern-gram}(d_1, d_2) \quad (3)$$

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.
At the end of the process, an indexing vector is created for each document. This vector consists of the n-grams and their weight in the document. Then, using the proposed similarity measure, each document is classified in the cluster according to its degree of similarity with the scholarly document representing this cluster.

4 The main architecture of the proposed model

The proposed model provides a framework for performing an intelligent survey on a database of scientific articles. This model consists of several components. All of these components make it possible to determine the user needs, process them step by step, and offer a global overview of the scientific work on this question. The main architecture of the system is shown in Figure 2.

4.1 Pre-processing phase

The Pre-processing phase help to prepare information contained in the scientific articles. Initially, a series of scientific articles are collected and stored in a database. Each article dealt with a scientific problem concerning a specific research area. Since academic articles are usually in PDF format, it is important to extract metadata that will make it easier to understand the content of each article.
At the end of this phase, each article is stored in a unique cluster representing this research sub-field. The clustering makes it possible to conduct the research intelligently and quickly in the cluster that best represents the scientific problem posed by the user.

4.1.1 Metadata extraction

In this work, several metadata are useful: the title, the author’s names and affiliations, the keywords, the abstract, the introduction, the conclusion, the reference, and the textual content of each article. However, no work identified in the related works section makes it possible to extract all these pieces of information.
To extract the maximum amount of useful, available, and good-quality information, two approaches proposed in the literature are used. The objective is to exploit the strengths of each approach to retain only the functionalities that offer the most guarantees.
The CERMINE model [7] is used to extract certain metadata from scientific articles. CERMINE is a Java library and web service for extracting metadata and content from scientific papers in digital form. With the CERMINE system, some metadata such as the author’s names, the title, the abstract, the references of the article (Date, pages, DOI, name of the journal), and the bibliographic references are extracted.
Otherwise, the PAXAT tool (PDF Article extraction and Analyzer Tool)[18] is used to extract other important content in the article like the introduction, conclusion, and textual content of the article. This tool uses the PDFBox java library.
These two tools also allow comparing the common results obtained for the extracted metadata to retain the results which seem common to both approaches.

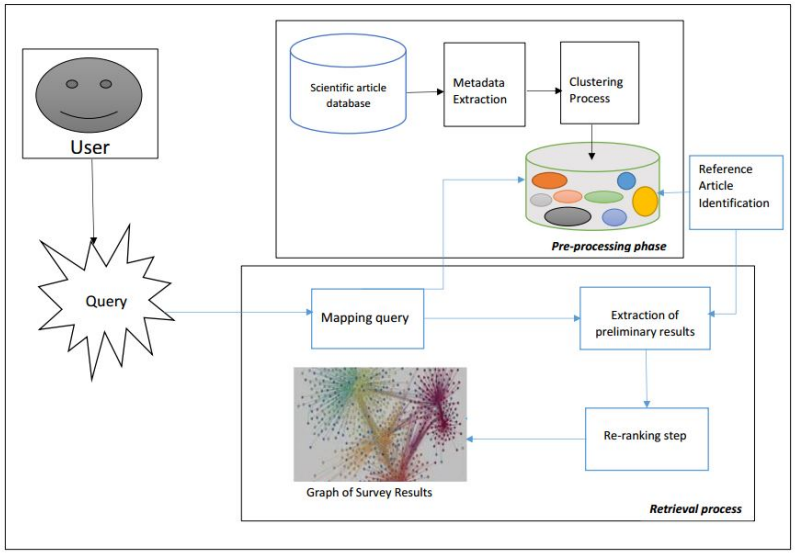

Figure 2: Main architecture of the proposed model

Once the metadata is extracted, it is processed by text mining algorithms to define an indexing vector. In this work, the indexing vector consists of the n-grams and their weight in each scholarly document.

4.1.2 Clustering process

Before clustering the articles, we need to determine their n-gram index vector using the indexing approach proposed in the previous section.
The clustering process of articles in clusters makes it possible to facilitate the investigation, thereby saving considerable time. The system no longer needs to search in the database completely to collect relevant articles corresponding to the scientific problem posed by user. It just needs to locate the right cluster, and makes research on.
To classify the documents in each cluster, it suffices to match each document with the vector of weighted terms representing each cluster and identify the one which is closest to the document (in term of the similarity value between its representative and the vector of terms representing the document).
By doing this, we classify each document in a unique cluster.

4.2 Retrieval process for scientific contributions

This phase begins with the issue of a query in natural language. The query is related to a scientific problem of which user wishes in a short time to make a complete view of the work carried out on the question, and the link between them.
As soon as the query is issued, the system applies the previous text mining algorithms to identify the terms vector representing the query in natural language.
Indeed, a query sent can be treated as the title of a scholarly document that the user is looking for.

4.2.1 Mapping query

After having identified the terms vector representing the query, it is necessary to match it with the terms vectors representing each cluster to identify in which cluster the query should be carried out. To do this, the model proceeds as follows: Since queries are research issues, they are considered potential scientific article titles. Thus, using the proposed similarity measure, the document whose title seems most similar to the scientific problem posed in the query is identified. This document is now considered ”the reference article” for this query.
Thereafter, the system identifies the cluster where the potential answers to the problems are found. It is in this cluster that the rest of the process will take place.
The ”reference article” is the unique document that best deals with the scientific problem posed by the user in his query. This reference article is the most similar to the query in the sense of their proposed similarity value.

4.2.2 Extraction of preliminary results

Articles are retrieved using the reference article identified previously.
The reference article is the article that best addresses the problem posed in the query. Thus, to complete the survey, it would be necessary to identify all the articles similar to the reference article in the sense of textual similarity. For this reason, the textual similarity value between each article and the reference article is computed using our proposed similarity measure. The most similar articles will be con-

sidered as the ”potentially relevant” article to the query.
Then, these articles are classified in descending order of their similarity value with the reference article.
Next, the potentially relevant articles are ranked using the following parameter:

$$\omega_i = (\alpha \times S_i(R_i) + (1 - \alpha) \times OriginalRank_i)^{-1} \quad (4)$$

Where α is the similarity value between the reference article title and the query.
$S_i(R_i)$ is the position of article i, obtained by ranking the articles in descending order of the similarity value between their title and the query.
$OriginalRank_i$ is the rank of the article i, obtained by ranking the articles in descending order of the similarity value between their vector of weighted terms and the query.
After having determined this weight value for each item, they are classified in ascending order.

4.2.3 Re-ranking step

To conclude that an article is relevant for a query, the model considers several parameters.

- the textual similarity value between this article and the reference article,
- The number of articles ”potentially relevant” to the query and citing this article. Indeed, among the ”potentially relevant” articles, if several refer to an article belonging to this family, then it means that this article is not only relevant but is also at the bottom of several works carried out on the scientific problem expressed in the query. This is an indicator to gauge the importance of the article for the problem expressed in the query.

From this, the model defines the ”relevance weight” using the equation 5. This parameter is used to identify the relevant documents for the survey.

$$\gamma_i = \omega_i \times e^{\frac{no_i}{n}} \quad (5)$$

Where:
ω_i is the similarity value between article i and the reference article.
no_i is the number of ”potentially relevant” articles which cite article i.
n is the number of all the ”potentially relevant” articles for the query.
By applying this parameter to each article, the system calculates its relevance weight. Then, these articles are ranked in descending order of their relevance weight.

4.2.4 Article survey graph

When a user, or a novice researcher, carries out a literature review on a specific research problem, he generally wants to obtain two things: firstly, that the system should propose a list of scientific articles dealing with the subject of the query; secondly, that the system should offer the possibility of visualising the evolution of scientific contributions over time, as well as the links existing between these articles. The article information retrieval system we are proposing offers the user these two possibilities.
Some heuristics are used to build the graph of the scientific articles survey.

- The graph only concerns the articles deemed relevant to the query. The system does not focus only on the citations between the different works to decide on the structure of the graph.
- An article can cite another appearing in the graph and not itself be part of the nodes of the graph. In other words, it is not the fact of citing or being cited by an article located in the graph that guarantees the presence in the graph.
- The presence of an article in the graph depends on the relevance of its content concerning the scientific problem identified in the query. Its position in the graph depends on the link existing between it and the other articles relevant for the query and in particular the reference article.

To build this graph, the system proceeds as follows:

1. The identifying all the articles relevant to the query and which must appear in the graph.
2. Next, for each article, it lists all the relevant articles cited by the latter. Each bibliographic reference is represented by the article’s title, the author’s names, the journal, and the date of publication.
3. Then, the graph building begins with the reference article. Starting from the reference article, the system identifies in its bibliographic references, all the relevant articles which are cited by it. It represents them in the graph by indicating the edge direction.
4. Next, it locates all the relevant articles citing the reference article. Once located, these articles are represented in the graph, specifying the relationship that links them with the reference article.
   Thereafter, the article represented on the vertice is deleted in the rest of the process.
5. When the system finishes with the reference article, it goes through the other vertices already represented in the graph to have the relevant articles which are cited or which cite them. It resumes the process for all the vertices and step by step, the graph is built.

6. If there are articles not citing any article appearing in the graph (or which are not cited by any article on the graph), then the system checks their similarity value with the reference article. If it is greater than a fixed threshold, it represents in the graph as an isolated vertice. Otherwise, they are not represented in the graph.

5 Implementation and experimentation of the proposed model

5.1 Description of the experimentation process of the proposed similarity measure

To implement the proposed similarity measure, some python libraries were used (Spacy, NLTK) for the text processing. Then, Wordnet was used to identify similar concepts in the different documents.
After the implementation of the proposed similarity measure, his ability to accurately define the similarity degree between two scientific paper is evaluated. 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, it is necessary to define 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 scientific documents, there is almost no such dataset.
The first step of this experimentation is the constitution of our dataset.
We have extracted articles from five research areas. These research areas correspond to the research fields of our experts. To be considered as an expert, you had to have published at least one scientific article in the field. These research areas are: ”Semantic Web and applications”, ”Algebraic coding theory”, ”Information retrieval system”, ”Fuzzy ontology modeling” and ”Question answering system”.
For each research area, two different experts in the field consulted articles to give an evaluation of the similarity value between two different articles.
This similarity value depended on the similarity between the keywords developed in each article and the research problems addressed by each article.
Two articles were said to be very similar if they dealt with the same problem and if their textual content was similar. The similarity value used was the average of the estimates made by each expert.
At the end of this process, we obtained a dataset containing the metadata on each article and a half thousand pairs of article with the similarity value.
Once the dataset is constituted, the proposed 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 metrics 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 [56].
The cosine similarity is used on the dataset to compare the results obtained using our semantic similarity measure (OSSM) with those obtained using the traditional cosine similarity measure (CSM). Finally, an improved version of the cosine similarity (ICSM) measure using the proposed n-gram-based indexing approach is applied on the dataset.

5.2 Description of the process for testing the automatic survey model

To test the performance of our survey retrieval system on scientific articles, we implemented three approaches: The first approach uses the indexing approach based on the weight of the words contained in the metadata by using the type of metadata in which the word appears (Title, abstract or keywords) and its proportion in each of this metadata. After this, we used the cosine similarity measure (CSM) as the similarity measure applied to our model.
In the second approach, we use our n-gram based indexing approach to index the different articles and then apply the cosine similarity measure (ICSM).
The last approach uses the indexing approach proposed in this article together with our semantic similarity measure as defined in Section 3 (OSSM).
For each approach, we evaluated the behaviour of the system in terms of the accuracy of the answers returned to the user’s queries.
For this experimental process, we used a dataset consisting of 126000 scientific articles extracted from the arXiv dataset [1]. Our dataset contained articles dealing with topics specific to five research subfields.These research subfields were used to represent the different clusters. For each subfields, queries were defined concerning a set of scientific problems that could be investigated scientifically. These research problems are divided into two groups: general problems on the one hand and specific problems on the other hand.
A problem is said to be general if its scope is broad and vague. This is the case, for example, if we want to study a problem such as: ”The semantic web and its applications”, ”The theory of algebraic coding”, ”The information retrieval system”, ”Question answering system” and ”fuzzy ontology”.
However, a problem is said to be specific if its field of action is precise, circumscribed and leaves no doubt about

[1] https://www.kaggle.com/datasets/Cornell-University/arxiv

the intentions of the research. This is the case of a query of the type: "Dynamic modeling of the user's profile for the personalization of information retrieval"; "Semantic similarity measurement for predicates in linked data".
Once the system was up and running, five users carried out two hundred queries on our system. These queries were made for each of the three approaches above using CSM, ICSM and OSSM.
Each user was a novice researcher (at least Phd students) with expertise in one of the five subflieds. Finally, the performances of the system have been compared using firstly the semantic similarity measure (OSSM) proposed in this paper, secondly using the cosine similarity measure (CSM) and finally, using the improving cosine similarity measure (ICSM).

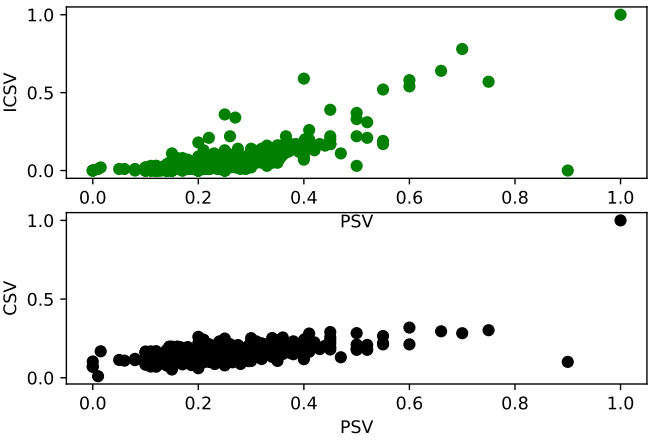

Figure 4: Correlation between (PSV, ICSV) and (PSV, CSV)

5.3 Discussion

5.3.1 Experimental results of similarity measure

At the end of these different phases of experimentation, the different similarity values obtained for each pair of documents were compared. Figure 3 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.

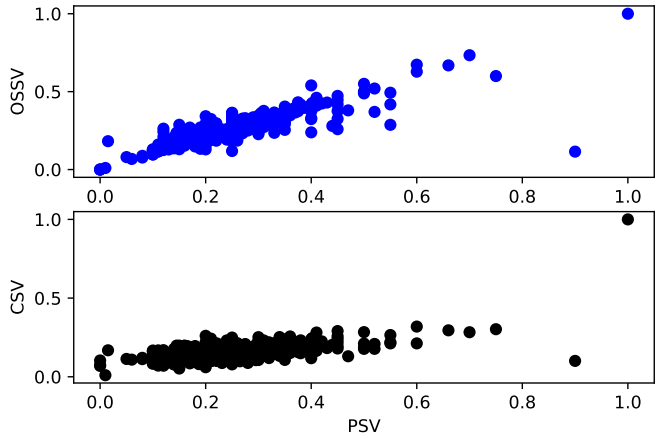

Figure 3: Correlation between (PSV, OSSV) and (PSV, CSV)

The Figure 4 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 improving the evaluation and the detection process of similar document pairs.
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 also calculated to assess the correlation between each similarity measure and the manually made estimates. The results are shown in Figure 5.

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.
Having a similarity measure that accurately assesses the similarity value between two documents is very important for the information retrieval system. The more accurate the similarity measure, the more certain we can be that the system will be able to return the results that best match the need expressed for any query.
At the end of this phase, we can conclude that the semantic similarity measure we have proposed is more accurate than the traditional cosine similarity measure.
In the following, we will discuss the performance of our information retrieval system when using our similarity mea-

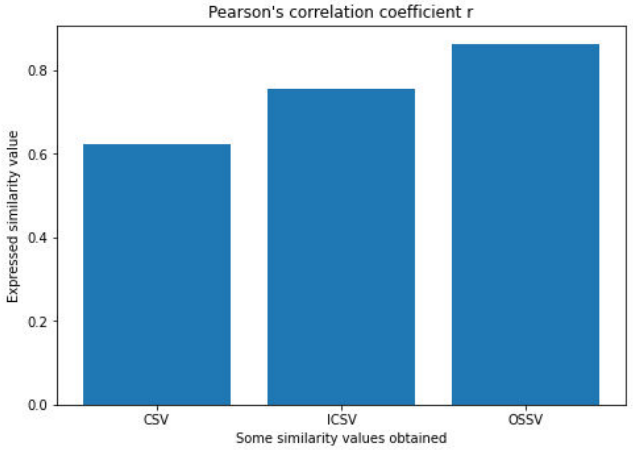

Figure 5: Pearson's correlation coefficient r of each similarity measure

sure and when using the cosine similarity measure. This will confirm the correlation between the accuracy of a similarity measure and the precision of the information retrieval system.

5.3.2 Experiment results of the proposed information retrieval system

The precision of a system is a parameter used to evaluate the proportion of good results returned in relation to the total number of results for a query.
A precision value close to 1 guarantees that all the results returned by the system meet the user's need expressed by the query.
For queries on general search problems, the comparison of the results obtained for each of the three indexing approaches and similarity measures can be found in Table 2.
For queries on specific (or precise) search problems, the

Table 2: Comparison of the precision of the system for queries concerning general search problems

	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
CSM	0.79	0.84	0.7	0.8	0.63
ICSM	0.88	0.9	0.809	0.86	0.77
OSSM	0.86	0.909	0.92	0.88	0.87

comparison of the results obtained for each of the three indexing approaches and similarity measures can be found in the Table 3.
From these two tables, it can be seen that the first approach

Table 3: Comparison of the precision of the system for queries concerning specific search problems

	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
CSM	0.64	0.87	0.92	0.76	0.648
ICSM	0.937	0.9	0.84	0.91	0.887
OSSM	0.9	0.89	0.93	0.97	0.88

is the least efficient overall for both "specific" and "general" queries. This is because the approach does not exploit the semantic relatedness that exists between the words in the query or in the various article metadata.
Using our indexing approach based on n-grams coupled with the cosine similarity measure, we obtain better overall performance than with the first approach. However, we note that this performance is rather unstable depending on whether we are dealing with questions related to one cluster or another. In Table 1, the accuracy for this approach is between 0.77 and 0.9. We would have expected to see a lower range. This is because with the second approach, the system returns very few responses to queries. Of all the possible results expected for a query, it returns only a small proportion. So the proportion of bad responses returned has a major influence on the accuracy value. This is not advantageous when you want to do a literature review on a specific research problem. Indeed, returning few results may suggest that there are not enough scientific contributions on the subject of the query. This is generally not the case.
Using the last approach (OSSM), the system's overall accuracy is better than with the two other approaches. This is due to the fact that with n-grams, the system is able to exploit the semantic relatedness between the concepts found in the articles. Not only can it exploit these semantic relationships, but it can also exploit the similarity between concepts, particularly when they appear to mean the same thing or opposite things.
Figure 6 summarizes the overall accuracy of the system when each of the two previous similarity measures is used. This figure shows the accuracy of the system when the

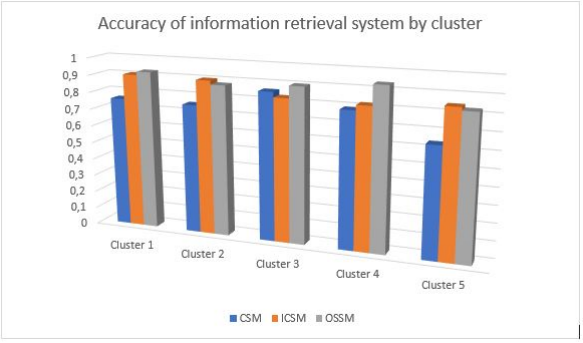

Figure 6: Comparison of results obtained with Cosine similarity measure (CSM), Improving Cosine similarity measure (ICSM) and with our semantic similarity measure (OSSM)

queries concern each research subfield represented by each cluster. This accuracy reflects the system's performance regardless of the type of problem submitted in the query (general or specific). Overall, the first approach performs less well than the two others. And the second less well than the third.
The figure 7 shows the behavior of the system by type of problem. It appears that the approach proposed in this paper offers better performance than the other two. On average, the system has an accuracy of 0.77 for the first approach, 0.85 for the second, and 0.89 for the approach developed in

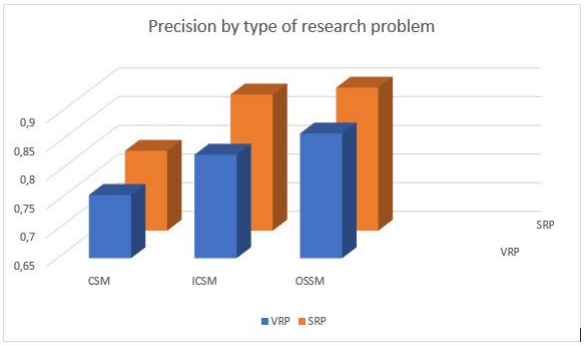

Figure 7: Comparison of accuracy per type of problem with CSM, ICSM and with OSSM

this paper.

The table 4 gives an overview of our system's performance in terms of precision, recall and F1-score using our proposed semantic similarity measure. From this table, it appears that our system has an overall recall of 0.85, while its F1-score is around 0.87.

Table 4: Global precision, recall and F1-score of the system

	Precision	Recall	F1-score
Cluster1	0.88	0.92	0.8995
Cluster2	0.9	0.81	0.8526
Cluster3	0.926	0.85	0.8863
Cluster4	0.925	0.76	0.8344
Cluster5	0.828	0.89	0.8578
Overall performance	0.89	0.8421	0.86

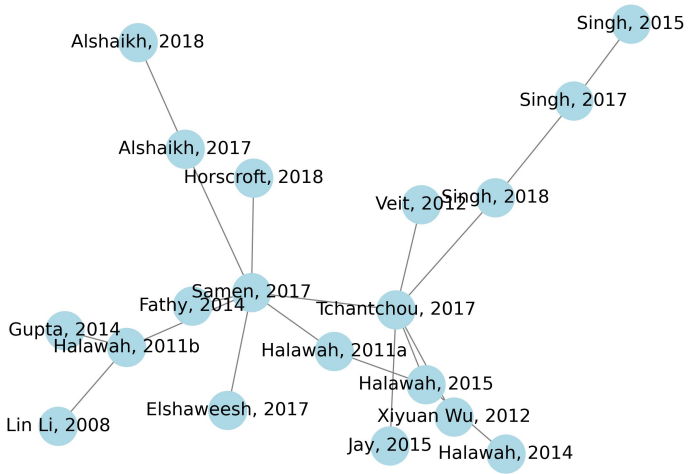

Figure 8: An example of survey graph generated by the model

Figure 8 shows an example of the results generated by the model for the following query: Dynamic user profile modeling approach to personalize information retrieval. This query aims to understand existing work on dynamic user profile modeling, given that users' interests while browsing may change over time. The novice researcher would like to have a clear idea of existing approaches and how they use dynamically collected information about the user to personalize the information to be sent back to the user. From this graph, the user can appreciate which works best address his concern, just by observing the different connections that exist between works over time. his concern.

6 Conclusion

In this paper, the problem of the automatic survey of scientific articles is addressed. The objective was to propose an information retrieval approach allowing a novice/confirmed researcher to visualize in a short time the evolution of contributions to a particular research problem. To this end, an approach to extract useful information from scholarly documents in Pdf format (metadata) is proposed. By using the information containing in useful metadata, a semantic similarity measure based on n-gram is defined. Finally, this similarity measure is used to implement an information retrieval system, allowing a user to investigate a problem by submitting a query to the system. The proposed system offers the option of presenting the survey results in graphical form.

From the implementation of the model, it is found that the proposed similarity measure can estimate with a good accuracy the similarity value between two scientific articles. For the experimentation of the similarity measure, a metadata dataset of scientific articles is constructed and their similarity values are estimated by human experts. Experiments with this dataset show that the similarity values obtained with the proposed similarity measure are very close to those manually proposed by human experts. Using this similarity measure, it finds that the system better refines the survey results to propose scientific contributions that are exclusively about the user's needs. Thus, the system's overall precision is about 0.9 while its recall is 0.8421, for an overall F1-score of 0.8652.

However, our system has a few limitations. The threshold of 0.5 set to be considered as potentially relevant for a query does not make it possible to identify articles that are relevant but whose textual content seems different from that of the reference article for a given query. This threshold was set to reduce the proportion of bad results returned by the system. It would be important to carry out a comparative study to identify the best threshold in terms of the system's performance.

Furthermore, in this work, we have only used metadata that is freely available. This metadata is not always sufficient to assess the similarity between two articles. It would be important in the future to develop an approach allowing to identifiy and extract the entire contents of a scientific article to be in PDF format (Introduction, conclusion, related work, and the other parts of the article). This will

enable the model to be readjusted to incorporate all this useful information into the information retrieval process. In the future, it would be interesting to address these limitations.

References

[1] P. Larsen and M. Von Ins. The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. Scientometrics, 84(3), 575-603, 2010.

[2] H. Han, C.L. Giles, E. Manavoglu, H. Zha, Z. Zhang and E.A. Fox. Automatic document metadata extraction using support vector machines. In 2003 Joint Conference on Digital Libraries, 2003. Proceedings. (pp. 37-48). IEEE.

[3] S. Marinai. Metadata extraction from PDF papers for digital library ingest. In 2009 10th International conference on document analysis and recognition (pp. 251-255). IEEE.

[4] P. Lopez. GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In International conference on theory and practice of digital libraries (pp. 473-474), 2009. Springer, Berlin, Heidelberg.

[5] A. Constantin, S. Pettifer and A. Voronkov. PDFX: fully-automated PDF-to-XML conversion of scientific literature. In Proceedings of the 2013 ACM symposium on Document engineering (pp. 177-180).

[6] J. Liu, T. Tang, W. Wang, B. Xu, X. Kong and F. Xia. A survey of scholarly data visualization. IEEE Access, 6, 19205-19221, 2018.

[7] D. Tkaczyk, P. Szostek, M. Fedoryszak, P.J. Dendek and Bolikowski. CERMINE: automatic extraction of structured metadata from scientific literature. International Journal on Document Analysis and Recognition (IJDAR), 18(4), 317-335, 2015.

[8] Y.C. Yoon, H.K. Gee and H. Lim. Network Based Document Clustering Using External Ranking Loss for Network Embedding. IEEE Access, 7, 155412-155423, 2019.

[9] S. Zhang, Y. Xu and W. Zhang. Clustering scientific document based on an extended citation model. IEEE Access,7, 57037-57046, 2019.

[10] 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.

[11] 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.

[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] A. Souza, v. Moreira and C. Heuser. ARCTIC: metadata extraction from scientific papers in pdf using two layer CRF. In Proceedings of the 2014 ACM symposium on document engineering (pp. 121-130), 2014.

[14] S. Liu and C. Chen. "The proximity of co-citation", Scientometrics, vol. 91, no. 2, pp. 495-511, 2012.

[15] S.A. Salloum, C. Mhamdi, M. Al-Emran and K. Shaalan. Analysis and classification of Arabic newspapers Facebook pages using text mining techniques. International Journal of Information Technology and Language Studies, 1(2), 8-17, 2017.

[16] I. Safder, S.U. Hassan, A. Visvizi, T. Noraset, R. Nawaz and S. Tuarob. Deep learning-based extraction of algorithmic metadata in full-text scholarly documents. Information processing & management, 57(6), 102269, 2020.

[17] Z. Nasar, S.W. Jaffry and M.K. Malik. Information extraction from scientific articles: a survey. Scientometrics, 117(3), 1931-1990, 2018.

[18] C. Jiang, J. Liu, D. Ou, Y. Wang and L. Yu. Implicit semantics based metadata extraction and matching of scholarly documents. Journal of Database Management (JDM), 29(2), 1-22, 2018.

[19] S. Qiu and T. Zhou. A method of extracting metadata information in digital books. In 2019 10th International Conference on Information Technology in Medicine and Education (ITME) (pp. 583-586), 2019. IEEE.

[20] A.M. Hashmi, M.T. Afzal and S. ur Rehman. Rule Based Approach to Extract Metadata from Scientific PDF Documents. In 2020 5th International Conference on Innovative Technologies in Intelligent Systems and Industrial Applications (CITISIA) (pp. 1 - 4), 2020. IEEE.

[21] G. Zaman, H. Mahdin, K. Hussain, J. Abawajy and S.A. Mostafa. An Ontological Framework for Information Extraction From Diverse Scientific Sources. IEEE access, 9, 42111-42124, 2021.

[22] S. Li and Q. Wang. A hybrid approach to recognize generic sections in scholarly documents. International Journal on Document Analysis and Recognition (IJDAR), 1-10, 2021

[23] B. Gipp and J. Beel. Citation proximity analysis (CPA): A new approach for identifying related work based on co-citation analysis. In ISSI'09: 12th international conference on scientometrics and informetrics (pp. 571-575), 2009.

[24] K.W. Boyack, D. Newman, R.J. Duhon, R. Klavans, M. Patek, J.R. Biberstine,... & K. Brner. Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches. PloS one, 6(3), 2011.

[25] S. Li and Q. Wang. A hybrid approach to recognize generic sections in scholarly documents. International Journal on Document Analysis and Recognition (IJDAR), 1-10, 2021.

[26] R. Bhagdev, S. Chapman, F. Ciravegna, V. Lanfranchi and D. Petrelli. Hybrid Search: Effectively Combining Keywords and Semantic Searches. In: Bechhofer, S., Hauswirth, M., Hoffmann, J., Koubarakis, M. (eds) The Semantic Web: Research and Applications. ESWC 2008. Lecture Notes in Computer Science, vol 5021. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-68234-9_-41, 2008.

[27] Z. BODO, L. CSATO. A Hybrid Approach for Scholarly Information Extraction. Studia Universitatis Babe?-Bolyai Informatica, [S.l.], v. 62, n. 2, p. 5-16, dec. 2017. ISSN 2065-9601. https://doi.org/10.24193/subbi.2017.2.01.

[28] G. Mitrov, B. Stanoev, S. Gievska, G. Mirceva, E. Zdravevski. Combining Semantic Matching, Word Embeddings, Transformers, and LLMs for Enhanced Document Ranking: Application in Systematic Reviews. Big Data and Cognitive Computing. 2024; 8(9):110. https://doi.org/10.3390/bdcc8090110

[29] N. RAMAN, A. NOURBAKHSH, S. SHAH, et al. Robust Document Representations using Latent Topics and Metadata. arXiv preprint arXiv:2010.12681, 2020.

[30] A. AHLUWALIA, B. SUTRADHAR, K. GHOSH, et al. Hybrid Semantic Search: Unveiling User Intent Beyond Keywords. arXiv preprint arXiv:2408.09236, 2024.

[31] M. Waqas, N. Anjum, and M.T. Afzal. A hybrid strategy to extract metadata from scholarly articles by utilizing support vector machine and heuristics. Scientometrics 128, 4349□4382 (2023). https://doi.org/10.1007/s11192-023-04774-7.

[32] K.W. Boyack, H. Small and R. Klavans. Improving the accuracy of cocitation clustering using full text. Journal of the American Society for Information Science and Technology, 64(9), 1759-1767, 2013.

[33] K. Rinartha and L.G.S. Kartika. Scientific article clustering using string similarity concept. In 2019 1st International Conference on Cybernetics and Intelligent System (ICORIS), Vol. 1, pp. 13-17, 2019. IEEE.

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

[35] 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

[36] 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

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

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

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

[45] 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.

[46] 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.

[47] 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.

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

[49] 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.

[50] 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.

[51] 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.

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

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

[54] 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.

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

[56] 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.