Incorporating Quality Measurement into Scientific Document Retrieval

Nedra Ibrahim, Anja Habacha Chaibi, Henda Ben Ghézala
RIADI Laboratory/ENSI
Tunisia
{nedra.ibrahim@ensi-uma.tn}
{anja.habacha@ensi.rnu.tn}
{henda.benghezala@ensi.rnu.tn}

Journal of Digital Information Management

ABSTRACT: *One of the challenges facing todays' researchers is how to find qualitative information that meets their needs. In scientific research, the quality of information is very important for institution quality improvement and research validation. The main purpose of the paper is the proposal of a scientometric annotation approach to improve retrieval system performance and meet researchers' needs. In this work, we discuss how to use scientometrics in document annotation to improve information quality. One possible solution to this problem is to automate and facilitate the selection of qualitative scientific documents by enriching the document annotation process with scientometric criterion. Our approach provided better performance for retrieval system compared to BM25 retrieval model. The best performance was supplied by the integration of document citation number and journal or conference ranking. The best improvement rate was 34.21% in F-measure, 52.22% in nDCG, 27.45% in MAP and 83.33% in P(k). An important implication of this finding is the existence of correlation between research paper quality and paper relevance.*

Subject Categories and Descriptors
[H.3.3 Information Search and Retrieval] I.7 Document and Text Processing

General Terms: Information Processing, Document Quality, Information Quality, Scientometric Evaluation

Keywords: Scientometric Retrieval, Scientometric Annotation, Scientific Quality, Qualitative Evaluation, Scientometric Indicator

DOI: 10.6025/jdim/2021/19/2/47-58

Received: 14 October 2020, Revised 2 January 2021, Accepted 13 January 2021

Review Metrics: Review Scale: 0-6, Review Score: 5.2, Inter-reviewer consistency: 89.5%

1. Introduction

We have observed a rapid and continuous growth in the number of scientific papers published each year (Harzing and Alakangas, 2016). Given a large number of publications, it is difficult to select papers that meet the expectations of researchers. The two main problems affecting scientific document retrieval are information overload and the heterogeneity of information sources (Haustein, 2016).

At this level, retrieval systems play an important role in enabling researchers to find and select the publications. However, current retrieval systems are designed to serve all users in the same way regardless of their particular needs (Ibrahim *et al.*, 2017). In the last few years, there has been a growing interest in scientific quality. That causes some problems since scientific research institutions give more importance to the scientific quality of their production. This quality is determined by a set of metrics that measure not only the quality of scientific papers but also the authors, laboratories and institutions quality (Hammarfelt and Rushforth, 2017). These measures are scientometric indicators used as tool of the scientific production evaluation. Several evaluation alternatives are being provided by various systems such as (Scopus, Google Scholar, SJR, Clarivate Analytics, Core, Microsoft Academic Search…) (Harzing and Alakangas, 2016).

So we are oriented to the proposal of qualitative annota-annotator to be used in scientific document retrieval. To do that, we are in direct relation with scientometrics which involves the application of quantitative methods that are devoted to scientific analysis and evaluation (Van Raan, 2013). Several indicators have been proposed as the basis of scientometric evaluation (Noyons *et al.*, 1999). Different workshops have taken place that intended to bring the Information Retrieval (IR) and bibliometrics/ scientometrics communities closer together and to enhance the link between domains (Mayr and Scharnhorst, 2015). The objective of this paper is to integrate scientometric indicators into scientific document retrieval process to improve retrieval performance by including information quality.

The paper is organized as follows: in Section 2, we describe the scientific quality measurement (methods and tools). In Section 3, we cover the essential related work in document annotation. In Section 4, we define our scientometric annotation approach. In Section 5, we present the proposed scientometric retrieval system. In Section 6, we present the evaluation of the scientometric annotation approach. We finish with a conclusion and future works in Section 7.

2. Scientific Quality Measurement

A scientific paper is considered to be an indicator of researchers scientific production. Thus, each group of researchers is interested in the evaluation of their scientific production. Recently, Schöpfel *et al.* (2019) and Azeroual *et al.* (2020) studied the influence of data quality on the success of the user acceptance of research information systems (RIS) by measuring satisfaction, perceived usefulness and ease of use. Other studies (Azeroual *et al.*, 2018 and Azeroual, 2019) evaluated the data quality used by research institutions. Their studies were based on error detection and data cleansing. In our context, we focus on scientific quality determined by quantitative and qualitative measures calculated after publication. So, we define scientometrics as all quantitative aspects of the science of science, communication science and science policy (Hood and Wilson, 2001). Publication and citation analysis have been used in the literature as very popular research evaluation tools (Zahedi *et al.*, 2014).

Scientometric indicators are identified as objective and useful research evaluation tools at different levels of analysis: macro level (countries), meso level (regions, areas, and centers) and micro level (research teams, individual researchers, research papers and journal/conference) (Noyons *et al.*, 1999). We classified the scientometric indicators according to their nature and use:

Production Indicators: are based on the quantification of publications number, citations number, self-citations number and download number.

• **Impact Indicators:** are based mainly on citations between articles. Indicators based on citations, measured at the micro, meso or macro level, are important in scientometrics (De Silva and Vance, 2017). Impact indicators include: journal Impact Factor (Bornmann and Williams, 2017), Citation success index (Milojevi *et al.*, 2017), H-index (Hirsch, 2005), SJR, Eigenfactor, SNIP (Walters, 2017) and others (Moed, 2017).

• **Composite Indicators:** are based on several measures such as H-index variants: g-index (Egghe, 2006), hi-index, hc-index, a-index, e-index (Zhang, 2009), m-index, c-index (Bras-Amorós *et al.*, 2011). Other composite indicators are AWCR, AWCRpA, and AW-index (Huggins-Hoyt, 2018), and the qualitative measure H_x (Ibrahim *et al.*, 2015).

The scientometric indicators have been used by bibliographic databases and classification systems. As bibliographic databases we cite Science Citation Index (SCI) (Moed, 2017), Google Scholar (Halevi *et al.*, 2017), CiteSeer (Harzing, 2011), Publish or Perish (Harzing, 2011), DBLP (Pal *et al.*, 2017), Crossref (Walker, 2002), Citebase (Brody, 2003), Scopus (Harzing, 2011) and Microsoft Academic Search (Thelwall, 2018). A key limitation of these bibliographic databases is that they use scientometrics to enrich search results when displaying results without considering it on their retrieval process. Moreover, we note the existing of several classification systems providing scientific journal ranking and conference ranking according to their impact. The leader of the ranking systems, and the oldest, is the Clarivate Analytics. It, annually, publishes the Journal Citation Reports (JCR[1]) which includes a number of indicators among which the journal impact factor (IF). The portal of the Association Core[2] provides access to the logs of journal classification and conference classification (A*, A, B and C). The SCImago Journal and Country Ranking portal (SJR[3]) provides a set of journal classification metrics and quality evaluation. SJR provides the journal classification (Q1, Q2, Q3 and Q4) based on the journal impact factor SJR.

3. Document Annotation

The purpose of an IR system is to retrieve, from a database, the relevant document(s) corresponding to a user request. The fundamental process of an IR consists of three main phases: indexing, search and the results presentation.

Automatic document annotation is used for indexing documents to then retrieve the relevant ones. Annotating scientific documents, in particular, has recently received a lot of interest. Fisas *et al.* (Fisas *et al.*, 2016) developed a multi-layered annotated corpus of scientific documents in the domain of Computer Graphics. Sentences are an

[1]https://clarivate.com/webofsciencegroup/solutions/journal-citation-reports

[2]http://portal.core.edu.au/conf-ranks

[3]http://www.scimagojr.com/index.php

notated with respect to their role in the argumentative structure of the discourse. Singhal *et al.* (2013) proposed a novel approach to generate summary phrases for research documents. They incorporated new and popular scientific terminologies in document annotations using crowd-source knowledge bases like Wikipedia and WikiCFP. Gábor *et al.* (2016) proposed a process of creating a corpus annotated for concepts and semantic relations in the scientific domain. Concepts were identified and annotated fully automatically, based on a combination of terminology extraction and available ontological resources. De Ribaupierre and Falquet (2013) have developed a user-centric annotation model based on discourse elements. They defined OWL ontology and used to annotate a corpus of scientific articles in gender studies. Based on the use of Big Data technologies, Herrera *et al.* (2017) proposed a semantic annotation approach to facilitate the semantic annotation of large volumes of scientific documents with multiple domain ontologies. Galke *et al.* (2017) conducted a semantic annotation using just the metadata of the documents such as titles published as labels on the Linked Open Data cloud instead of using full-text. Boudin *et al.*, (2000) proposed a keyphrase generation method for scientific document retrieval. They show that predicted keyphrases are consistently helpful for document retrieval. Zhao *et al.* (2019) proposed a new annotation schema based on neural model to examine the revolution of scientific resources from trends in their function over time.

The existing studies aimed to improve the quality of search results in terms of relevance (thematic, contextual, cognitive or semantic relevance). None of the previous studies has addressed the problem of information quality, more precisely the quality of scientific information, in the IR process.

4. Scientometric Annotator

Studies on annotation approach that involves scientometrics by its set of quantitative indicators are still lacking. To overcome this lack and improve retrieval performance and information quality, we propose the application of scientometrics in document annotation. We propose a scientometric annotation which is an automatic process that allows the extraction of relevant indicators to each document from online bibliographic databases. A document can be a conference or a journal paper, thesis report, master or technical report. Scientometric annotation will be author-centred, document-centred, and venue-centred. It consists of representing, using a set of numerical indicators: the impact of the author (quality of the researcher), the impact of the journal/conference (quality of the container), the impact of the research group (quality of the search environment) and the quality of the content. The new method of scientometric annotation is carried out on different parts of the document structure: front, body and back. The body is the content of the document, the front contains the title, the authors, the conference/journal and the unit/research laboratory, and the back contains the references.

Figure 1 shows the architecture of the proposed scientometric annotation approach, which consists of three steps. The first step is the pre-treatment which consists on the extraction of scientific documents from online bibliographic databases such as Google Scholar, MS Academic Search, Scopus, etc. The second is the indicators extraction which consists on extracting scientometric information corresponding to each document from online resources. The third step consists on the enrichment and the reconstruction of the XML annotation document regrouping the different types of annotation. Below, we detail the three annotation steps.

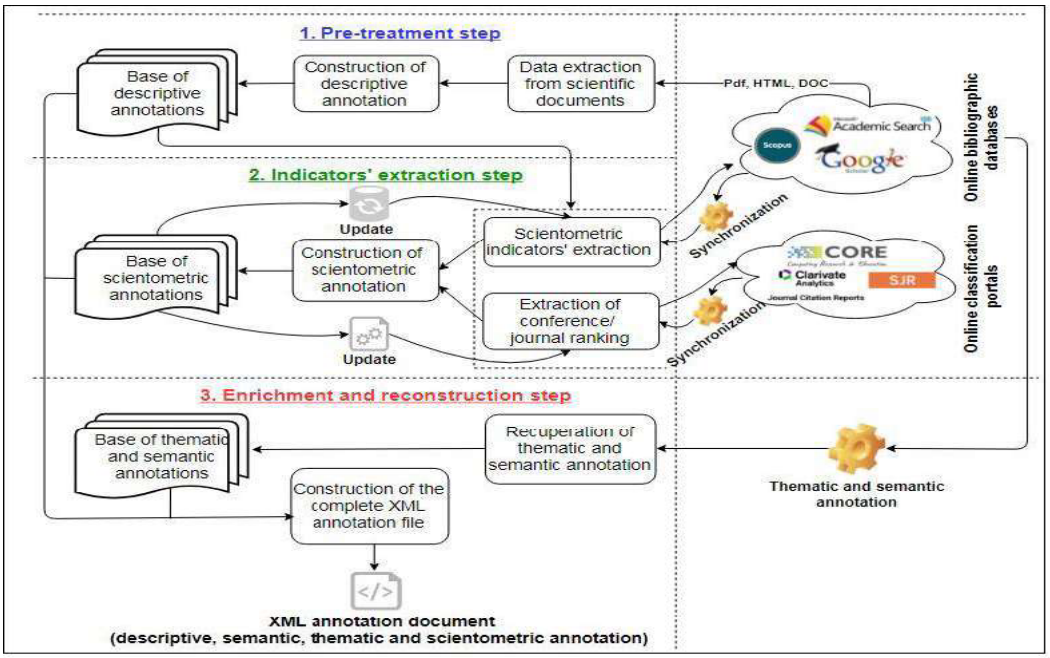


Figure 1. Architecture of the scientometric annotator

4.1. Pre-treatment Step

This module takes as input the title or the URL of scientific document available on the bibliographic database, and for each document the descriptive information is extracted. Subsequently consultations are made to different sources of bibliographic data and from the data consulted the necessary information of this publication is obtained. The descriptive information is transformed from plain text to XML format and is integrated into the XML annotation document. The descriptive annotation includes:

• **Bibliographic content:** For each document, we identify the authors, their affiliations, the conference or the journal name and information.

• **Content descriptors:** Keywords and abstract.

• **Technical description:** Format and size. At this step, the number of co-authors and the position of the author (subject of evaluation) are extracted.

4.2. Scientometric Indicator's Extraction Step

The second step is the extraction of the required scientometric indicators. One purpose of the scientometric annotation is to consider the quality of the different elements in scientific document (content, container, authors and research environment). For this purpose, different scientometric indicators are extracted from online bibliographic databases allowing the measurement of the different elements quality. The body of scientific document, which represents its content, is annotated which the number of co-authors and the number of citations. To include the quality of document s authors, we annotate the document by the author s publications number, author s citations number, author s self-citations number, author s H-index and authors H_x indicator. At the container level, we considered conferences and journals ranking as a measure of conference/journal impact. To integrate the container quality, we consider the following indicators in the document annotation: conference/journal ranking, number of conference/journal publications, number of conference/journal citations and number of conference/journal self-citations. Finally, to consider the quality of the work environment, we annotate scientific documents by the number of research group publications, number of research group citations, stability of the research group and the number of group self-citations.

4.3. Enrichment and Reconstruction Step

The third step is the enrichment with the scientometric annotation and the reconstruction of the final XML annotation document. At this step, we build the thematic and semantic annotation of the scientific documents using the annotator implemented by Kboubi *et al.* (2012). This step consists on storing and regrouping the different types of annotations into a single document. The annotation document included the descriptive, thematic, semantic and scientometric annotation. At the end of the annotation process, an XML annotation document is created. Figure 2 presents an example of an extract of the XML annotation document containing the descriptive and the scientometric annotation.

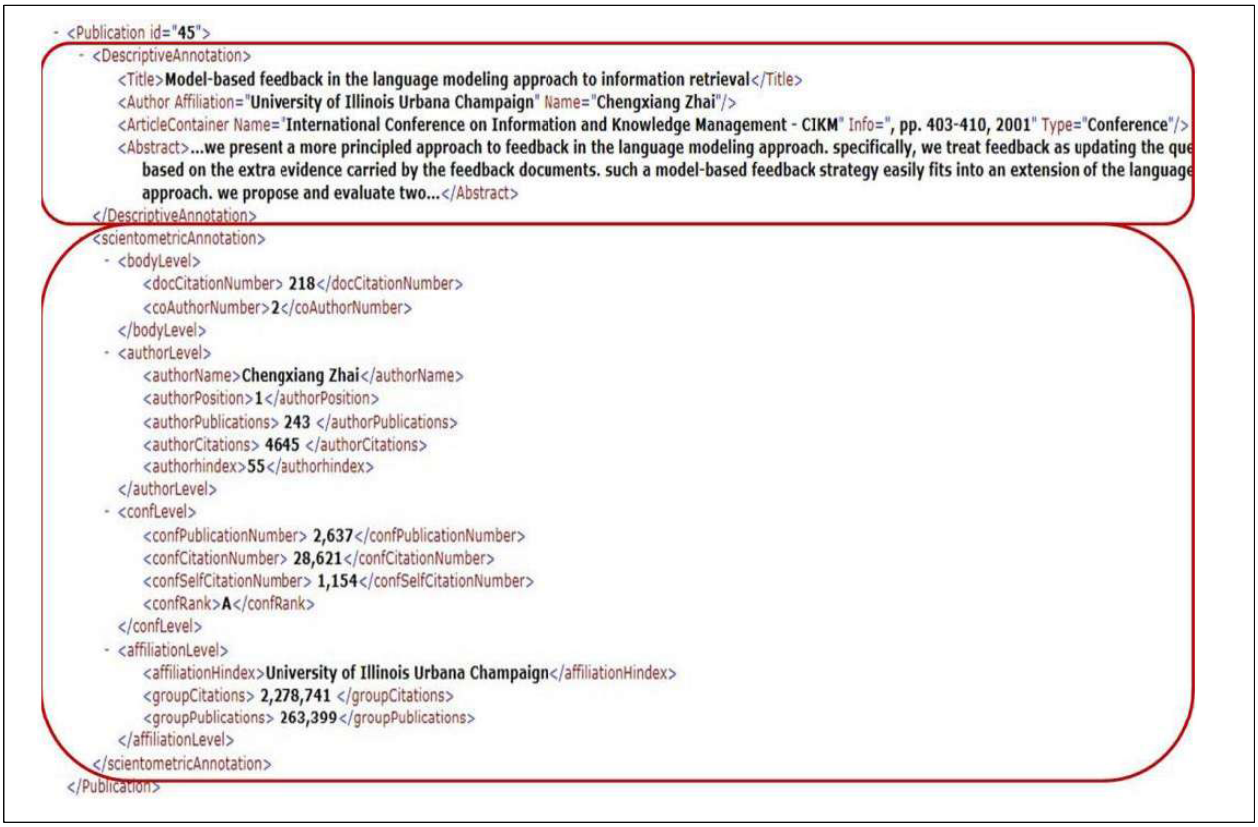


Figure 2. Example of an extract of the XML annotation document

5. Proposed Scientometric Retrieval System

The scientometric annotation is proposed to integrate scientific quality in the retrieval process. This will be carried out by representing the impact of the author, quality of document content, impact of the research environment and quality of the journal or conference. The scientometric annotator actually assists the retrieval of qualitative papers by adding additional information concerning the quality of each document. By considering this information, we can select the qualitative documents which are at the same time relevant to the user query.

In order to verify the contribution of scientometrics in document annotation, we designed and developed different retrieval models based on scientometrics. We proposed six scientometric retrieval models (ScientoRank, ScientoCite, ScientoRankCite, ScientoCiteRank, ScientoH and ScientoCiteH) based on an adaptation of the classic vector-space model and one classic vector-space retrieval model in which we did not integrate scientometrics (WoutSciento). We adapted the vector space model by integrating scientometric indicators. Table 2 shows the different retrieval models at two levels (search and ranking). The scientometric models differ by the criteria considered at search and ranking. These criteria depend on the strategy of the research institution. The selection of documents is according to its relevance and its quality. The quality of document is measured by the set of scientometric indicators. The selected documents are sorted according to the quality of each one (author quality, content quality, venue quality and container quality). For all proposed models, we measure the degree of document relevance which is the similarity between document and query, we considered the cosine similarity adapted with scientometric indicators. In classic keyword-based vector-space model, the query keywords are assigned a weight that represents the importance of the concept compared to the information need expressed by the query. We associate to each query a vector of its terms. To each term we associate a weight which measures the term frequency (tf.idf) in the document.

In Table 1, we present the different proposed retrieval

Scientometric Retrival Model	**Search**		**Ranking Parameters**
	Search criteria	**Similarity score**	
ScientoRank	Cosine similarity + container ranking	$sim(d_j,q)=\frac{\sum_{i=1}^{N} w_{i,j}w_{i,q}}{\sqrt{\sum_{i=1}^{N} w_{i,j}^2}\sqrt{\sum_{i=1}^{N} w_{i,q}^2}}.rank_j$	Container ranking + document citation number
ScientoCite	Cosine similarity + document citation number	$sim(d_j,q)=\frac{\sum_{i=1}^{N} w_{i,j}w_{i,q}}{\sqrt{\sum_{i=1}^{N} w_{i,j}^2}\sqrt{\sum_{i=1}^{N} w_{i,q}^2}}.\frac{cite_j-cite_{thr}}{cite_{thr}}$	Document citation number + container ranking
ScientoRankCite	Cosine similarity + container ranking + document citation number	$sim(d_j,q)=\frac{\sum_{i=1}^{N} w_{i,j}w_{i,q}}{\sqrt{\sum_{i=1}^{N} w_{i,j}^2}\sqrt{\sum_{i=1}^{N} w_{i,q}^2}}.\frac{cite_j-cite_{thr}}{cite_{thr}}.rank_j$	Container ranking + document citation number
ScientoCiteRank	Cosine similarity + container ranking + document citation number	$sim(d_j,q)=\frac{\sum_{i=1}^{N} w_{i,j}w_{i,q}}{\sqrt{\sum_{i=1}^{N} w_{i,j}^2}\sqrt{\sum_{i=1}^{N} w_{i,q}^2}}.\frac{cite_j-cite_{thr}}{cite_{thr}}.rank_j$	Document citation number + container ranking
ScientoH	Cosine similarity + H-index	$sim(d_j,q)=\frac{\sum_{i=1}^{N} w_{i,j}w_{i,q}}{\sqrt{\sum_{i=1}^{N} w_{i,j}^2}\sqrt{\sum_{i=1}^{N} w_{i,q}^2}}.\frac{H_j-H_{thr}}{H_{thr}}$	H-index
ScientoCiteH	Cosine similarity + document citation number + H-index	$sim(d_j,q)=\begin{cases}-\frac{\sum_{i=1}^{N} w_{i,j}w_{i,q}}{\sqrt{\sum_{i=1}^{N} w_{i,j}^2}\sqrt{\sum_{i=1}^{N} w_{i,q}^2}}.\frac{cite_j-cite_{thr}}{cite_{thr}}.\frac{H_j-H_{thr}}{H_{thr}}, \\ \quad if\ cite_j<cite_{thr}\ and\ H_j<H_{thr} \\ \frac{\sum_{i=1}^{N} w_{i,j}w_{i,q}}{\sqrt{\sum_{i=1}^{N} w_{i,j}^2}\sqrt{\sum_{i=1}^{N} w_{i,q}^2}}.\frac{cite_j-cite_{thr}}{cite_{thr}}.\frac{H_j-H_{thr}}{H_{thr}}, \\ \quad otherwise\end{cases}$	Document citation number + H-index

Table 1. Scientometric retrieval models

models. In each model we integrated one or more scientometric indicator in search and results ranking. We specified in the search criteria and ranking parameters columns the criteria considered respectively in search and ranking results. We considered four search and ranking criteria (similarity, container ranking, document citation number and H-index) which intervene at the different retrieval models:

Similarity: The similarity between a document and a search query; we considered the cosine similarity referred to the cosine of the angle between document and query vectors. Equation (1) represents the classic similarity score of the vector space model:

$$sim(d_j, q) = \frac{d_j.q}{||d_j||.||q||} = \frac{\sum_{i=1}^{N} w_{i,j} w_{i,q}}{\sqrt{\sum_{i=1}^{N} w_{i,j}^2} \sqrt{\sum_{i=1}^{N} w_{i,q}^2}} \quad (1)$$

The attribute vectors d_j *and* q are the term frequency vectors of the documents and queries. $w_{i,j}$ is the frequency of term i in document j calculated using the weight *tf.idf*. $w_{i,q}$ is the frequency of term i in query q calculated using *tf.idf*. The cosine similarity of two documents will range from 0 to 1, since the term frequencies (*tf-idf* weights) cannot be negative. The angle between two term frequency vectors cannot be greater than 90°. For each proposed model, we adapted the cosine similarity to define different similarity scores enriched by scientometric indicators presented. Selected documents should have a positive similarity score. This means that selected documents should be at the same time relevant and qualitative.

• **Container Ranking (rank):** A document container is a journal or a conference. Selected documents must have been published by a ranked journal or conference. We associate a positive weight to each of container classes (*A*, A, B and C*) and a weight of 0 to non-ranked conferences or journals (by considering Core2020 ranking).

• **Document Citation Number (cite):** Selected documents must have a threshold number of citations. ($cite_{thr}$) . $cite_{thr}$ is determined by the user when executing his query.

• **H-index (H):** The first author in selected documents must have a threshold H-index. (H_{thr}). H_{thr} is determined by the user when executing his query.

In ScientoRank and ScientoRankCite, we rank search results according to the container ranking (A*, A, B, C) and then according to the document citation number. In ScientoCite and ScientoCiteRank, search results were ranked according to the document citation number and then according to the container ranking. The results returned by ScientoH are sorted according to the H-index of the first author. In ScientoCiteH, the results are sorted according to the document citations number then the author H-index. The order of ranking criteria makes the difference between the two retrieval models ScientoRankCite and ScientoCiteRank.

6. Evaluation of the Scientometric Annotation

Our evaluation environment consisted essentially of multi-model retrieval system, a data collection and a set of evaluation measures. On the basis of the implemented retrieval system with its different models, we conducted a series of experimentations to compare the performance

Element	Description
Data source	We have tested our system based on scientific documents available on an online bibliographic database. In our case, we opted for MS Academic search to extract published research papers. Our choice is justified by the broad set of scientometric indicators covered by MS Academic Search. We annotated 15000 scientific documents, a sample of annotations is available on this link5. We considered Core for conference and journal ranking.
Queries	The query set is a collection of 300 queries. The different queries were formulated by the members of our research laboratory to constitute valid search queries. Each query is represented by a vector of its terms. To each query, they associated a set of relevant documents from our corpus. All topics are in the information system domain. This choice is justified by the fact that the researchers in RIADI6 laboratory are specialized in this field.
Relevance judgment	The members of our research laboratory assessed typically 15000 documents from research domain. They used graduated relevance judgment from 0 to 5 to distinguish the documents entirely relevant to the query from the documents partially relevant. Our scientometric data collection will be used to evaluate the different retrieval models.
Baselines	For the baseline retrieval models, we considered the vector space and BM25 retrieval models.

Table 2. Data set

of the different models based on a set of measures.

6.1. Data Collection

Retrieval test collections consist on a set of documents, a set of queries, and a subset of the document collection considered to be relevant to each query. Relevance judgments are generally provided by subject experts. This is to distinguish relevant documents associated with each query from all other documents in database.

We start by studing existing IR test collections (TREC, NTCIR, CLEF, FIRE and INEX) (Carterette and Voorhees, 2011). After this study, we determine our need for a different test collection in that it consists of scientific articles rather than newspaper text. Thus allows for IR experiments that include scientometric information. We created our new test collection, which involved a long and expensive process. That was necessary because no ready-made collection existed on which the types of experiments with scientometric information that we envisage could be run. We adapted the Cranfield method (Carterette and Voorhees, 2011) to build our test collection based on published scientific papers from the online bibliographic database "Microsoft Academic Search[4]". We describe the main elements of our scientometric test collection in table 2.

6.2. Experimental Results

Several approaches addressed the problem of efficiency, speed and performance of retrieval systems in the general case. In this paper, particular attention is paid to the study of the effect of the integration of scientometrics on the performance of retrieval systems. We conducted a series of experimentations on the different retrieval models at two levels:

• **Search level:** Evaluation of the performance of the different retrieval models at the search level returns to the relevance evaluation by means of different measures.

• **Ranking level:** Evaluation of the different retrieval models performance at the ranking level returns to the ranking evaluation, which is performed using different measures.

6.2.1. Relevance Evaluation

Figure 3 presents the P-R curves of the different retrieval models. The scientometric retrieval models generate approximately close results in terms of relevance performance, which are better than WoutSciento. The two models ScientoRankCite and ScientoCiteRank are the same at the search level and present the best performance showed by its P-R curve. ScientoRank, ScientoCite and ScientoCiteH show very similar results which are close to the optimal. WoutSciento model present the lower performance compared to the other models for Recall>0.3. The performance of ScientoH is lower than other scientometric models and close to WoutSciento retrieval model. The results show that the combination of document container ranking and document citation number made the best performance. The existing overlap between ScientoRank and ScientoCite curves highlights the correlation between the two indicators: citation number and container ranking. Thus, a good ranking involves a good citation number and vice versa. By analyzing the performance of ScientoH, we can conclude that H-index did not improve search relevance.

In Figure 4, the F-measure curves of ScientoRankCite and ScientoCiteRank show best results. Thus, the integration of the combination of document citation and container ranking contributes to the improvement of retrieval system performance. Moreover, the F-measure curves of ScientoRank, ScientoCite and ScientoCiteH are approximately the same and show a faster increase to the optimal. When comparing with WoutSciento, we realize an important improvement in retrieval performance.

	Retrieval models	F-measure
Baselines	WoutSciento (vector space)	0.36
	BM25	0.38
Proposed scientometric retrieval models	ScientoRank	0.48
	ScientoCite	0.47
	ScientoRankCite	0.51
	ScientoCiteRank	0.51
	ScientoH	0.52
	ScientoCiteH	0.49

Table 3. F-measure results

By refering to Table 3, the proposed scientometric annotation improved the performance of retrieval systems. ScientoRankCite and ScientoCiteRank present better improvement compared to ScientoRank and ScientoCite. All scientometric models show better performance than Vector space and BM25. Moreover, the F-measure curves of the ScientoRank, ScientoCite and ScientoCiteH models are approximately the same and the closest to the optimal. They achieve growth faster than the curves of Vector space and BM25.

6.2.2. Ranking Evaluation

We performed the experimentations over our scientometric test collection for nDCG and P(k), computed at rank 10, 20, 30, 40 and 50. We present the results in Figures 5 and 6.

One can observe that nDCG curves, in Figure 5, decrease when the rank increases for all models. That is to say that scientometric models return more relevant re

[4] http://academic.research.microsoft.com

[5] https://transferxl.com/0867WN0W4KKYw

[6] http://www.riadi.rnu.tn

sults at top ranks, which really matter for users. Figure 6 presents the variation of precision at rank k. We note a better performance of the scientometric models. Scientometric curves are increasing while WoutSciento s curves are decreasing. This indicates that scientometric models are more precise at greater ranks. As can be observed, ScientoRankCite and ScientoCiteRank show a stability of p(k) at different ranks. These fluctuations observed in Figures 5 and 6 show that the scientometric annotation improved retrieval performance. More precisely, ScientoRankCite and ScientoCiteRank have the best performance. Figure 7 illustrates the MAP rates determined for all different ranks. ScientoRankCite and ScientoCiteRank show the best performance in MAP.

By refering to Table 4, the obtained results are compatible with those of relevance evaluation. The results show that the integration of the combination (document citation number, container ranking) leads to a better performance of retrieval system at the top ranks and greater ones. More than that, the overlap between ScientoRank and ScientoCite curves confirms the correlation existing between the citation number and container ranking, in fact, each implies the other. The performance of ScientoH and ScientoCiteH are better than WoutSciento and closer to the performance of ScientoRank and ScientoCite. For all measures, the best performance is provided by ScientoRankCite and ScientoCiteRank, in which both the number of document citations and container ranking were integrated.

	Retrieval models	nDCG	MAP	P(k)
Baselines	WoutSciento	0.5	0.57	0.38
	BM25	0.45	0.51	0.3
Proposed scientometric retrieval models	ScientoRank	0.57	0.60	0.48
	ScientoCite	0.54	0.60	0.51
	ScientoRankCite	0.64	0.65	0.55
	ScientoCiteRank	0.63	0.64	0.55
	ScientoH	0.53	0.57	0.44
	ScientoCiteH	0.53	0.58	0.47

Table 4. nDCG, MAP and P(k) results

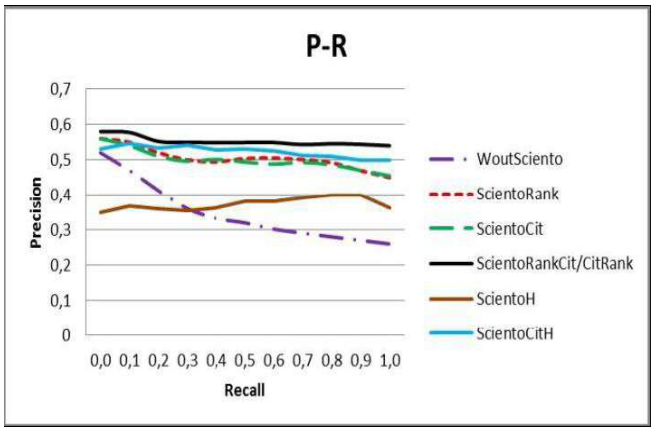


Figure 3. P-R curves

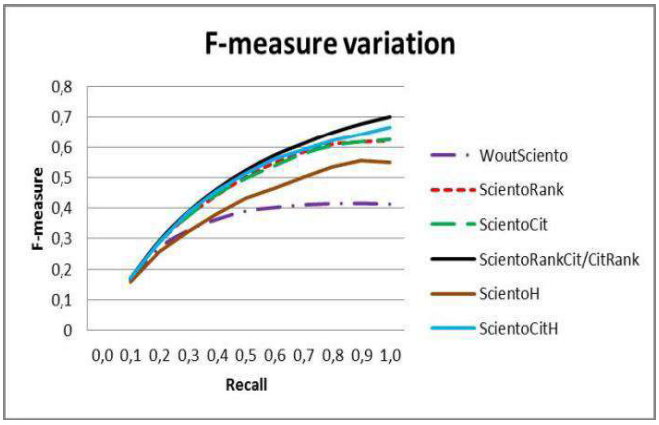


Figure 4. F-measure variation curves

6.3. Discussion and Comparison

As a baseline, we used Indri retrieval platform (Strohman *et al.*, 2005), which is developed under the Lemur project[7]. The basic Indri retrieval model is BM25 (Robertson, 1997) which is a probabilistic retrieval model. Figure 8 shows the improvement rates of scientometric

[7] http://www.lemurproject.org

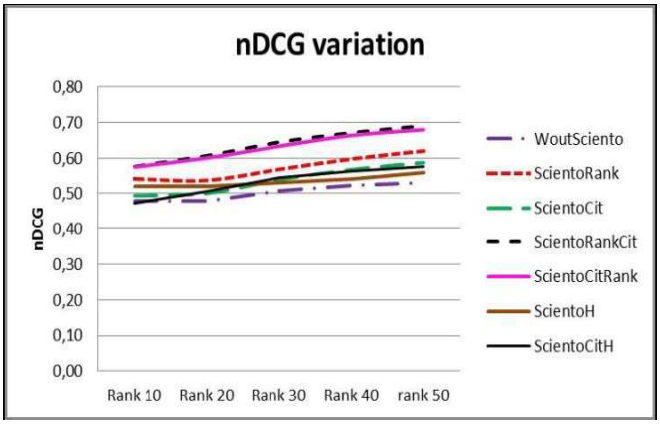


Figure 5. nDCG variation

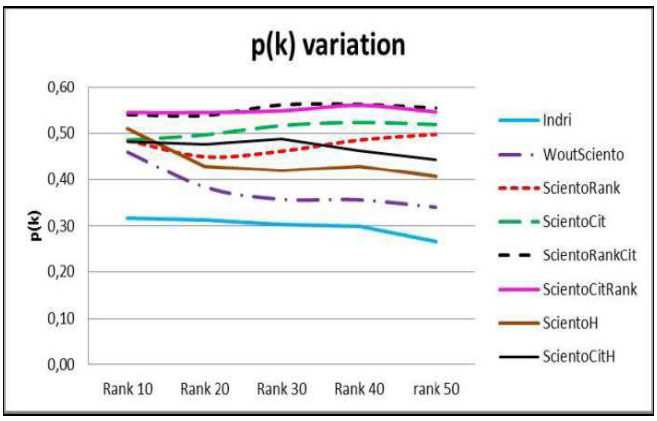


Figure 6. P(k) variation curves

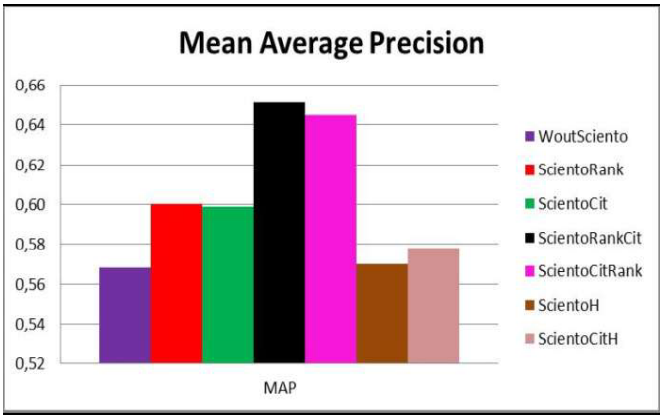
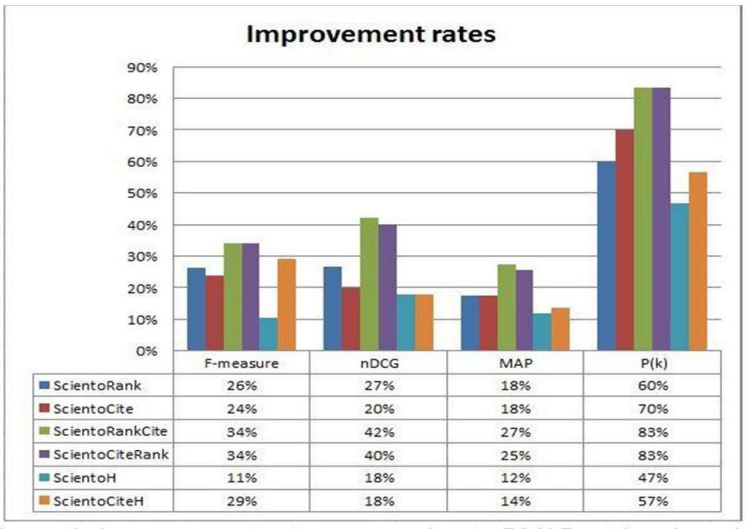


Figure 7. MAP rates

models compared to Indri baseline. The results show that all the scientometric models provided an improvement in performance. ScientoRankCite and ScientoCiteRank realized the best improvement in F-measure which is rated for 34.21%. ScientoRank and ScientoCite realized an improvement in F-measure greater than 23% and ScientoCiteH realized an improvement of 28.95% and finally ScientoH realized the lowest improvement which is rated for 10.53%. The best improvement rate in nDCG was provided by ScientoRankCite (42.22%) and ScientoCiteRank (40%). ScientoRank and ScientoCite realized an improvement in nDCGp rated respectively for 26.66% and 20%. ScientoH and ScientoCiteH achieved the same improvement in nDCG evaluated for 17.78%. Same for p(k) improvement rates, the best results were realized by ScientoRankCite and ScientoCiteRank (83.33%). ScientoRank realized an improvement of p(k) rated for 60% and ScientoCite realized 70%. ScientoH and ScientoCiteH realized an improvement of 46.67% and 56.67%. By comparing the rates of MAP improvement, we note the best rate of improvement realized by ScientoRankCite which is rated for 27.45%. ScientoCiteRank realized the closest improvement to the optimal which is rated for 25.49%. ScientoRank and ScientoCite realized an improvement in MAP rated for 17.64%. The lowest improvement rates are 11.76% and 13.73% realized respectively by ScientoH and ScientoCiteH.

It has been found that integrating scientometrics at the system level has improved retrieval performance. ScientoRankCite and ScientoCiteRank show better improvement over to ScientoCite, ScientoRank and ScientoCiteH. ScientoH has achieved performance that is approximately close to that of Indri. This performance degradation can be justified by the fact the H-index only cannot improve the relevance of retrieval results. All other scientometric models performed better than Indri. In addition to quality improvement, scientometric annotation has enhanced the relevance of results. Scientometric annotation has provided better performance to the retrieval models at both search and ranking levels. The integration of scientometric indicators revealed an improvement of system performance at the top ranks and a good performance at greater ranks.

Summing up the results, the number of document citations and the container ranking were integrated into search and ranking levels. An important implication of these findings is that this combination better reflects the quality of the scientific article because these are good indicators of the impact of the article. These results are in good agreement with other studies which have shown that citation number is considered as a valid measure of



	F-measure	nDCG	MAP	P(k)
ScientoRank	26%	27%	18%	60%
ScientoCite	24%	20%	18%	70%
ScientoRankCite	34%	42%	27%	83%
ScientoCiteRank	34%	40%	25%	83%
ScientoH	11%	18%	12%	47%
ScientoCiteH	29%	18%	14%	57%

Figure 8. Improvement rates comparing to BM25 retrieval model

scientific quality (Manis, 1951). Our study has shown that there is a correlation between the scientific quality of an article and its relevance. If we consider that the relevance of the research is related to scientific quality, we can conclude document citations as well as the container ranking are good indicators of scientific document quality.

7. Conclusion and Future Work

The main purpose of the paper was the proposal of a scientometric annotation approach to improve IR performance. In this paper, our attention focused on the scientific document relevance as well as its scientific quality which is measured by scientometric indicators. The scientometric annotation included the quality of content, container, author(s) and the work environment. It was performed on the different documents parts.

To evaluate our scientometric annotator, we have designed and proposed six different scientometric retrieval models (ScientoRank, ScientoCite, ScientoRankCite, ScientoCiteRank, ScientoH and ScientoCiteH). These models differ by the criteria considered at search and ranking. The scientometric models were based on scientometric indicators among which we worked primarily on the quality of content (document citations) and container quality (journal or conference ranking). On the basis of these implemented models, we conducted a series of experiments (of relevance and ranking). These experiments were carried out to compare the performance of the different proposed models with Indri baseline based on a set of measures. It has been found that scientometrics provided an improvement in the performance of retrieval system in terms of relevance and scientific quality. The best improvement was provided by ScientoRankCite and ScientoCiteRank. The best improvement rate was 34.21% in F-measure, 52.22% in nDCG, 27.45% in MAP and 83.33% in P(k). This improvement was justified by the correlation between the scientific quality and the relevance of a document. The scientific quality was precisely reflected by a combination of the document citation number and container ranking. Our approach might be practical and convenient for researchers and institutions. These latter are interested in the improvement of their production scientific quality.

One of the big advantages of this annotation approach is its genericity; any bibliographic database can integrate it. Furthermore, new scientometric indicators can extend it. Our approach can be also considered as a state of art approach and a mean of literature review validation. Given the dynamic aspect of scientometrics domain, a synchronization module must be efficiently designed to synchronize the different scientometric indicators extracted from the different bibliographic databases.

References

[1] Azeroual, O., Saake, G., Abuosba, M. (2018). Data quality measures and data cleansing for research information systems, *Journal of digital information management*, 16 (1), *p* 12-21.

[2] Azeroual, O. (2019). Text and data quality mining in CRIS, *Information*, Vol. 10, *p* 374.

[3] Azeroual, O., Saake, G., Abuosba, M., Schöpfel, J. (2020). Data Quality as a Critical Success Factor for User Acceptance of Research Information Systems. *Data*, 5 (2), p 35.

[4] Bornmann, L., Williams, R. (2017). Can the journal impact factor be used as a criterion for the selection of junior researchers? A large-scale empirical study based on ResearcherID data. Journal of Informetrics, Vol. 11, p 788–799.

[5] Bornmann, L., Mutz, R., Hug, S., Daniel, H. D. (2011). A multilevel meta-analysis of studies reporting correlations between the h index and 37 different h index variants, *Journal of Informetrics*, Vol. 5, p 346–359.

[6] Boudin, F., Gallina, Y., Aizawa, A. (2020, July). Keyphrase Generation for Scientific Document Retrieval *In*: Proceedings of the 58th *Annual Meeting of the Association for Computational Linguistics (ACL)*.

[7] Bras-Amorós, M., Domingo-Ferrer, J., Torra, V. (2011). A bibliometric index based on the collaboration distance between cited and citing authors, *Journal of Informetrics*, Vol. 5, p 248–264.

[8] Brody, T. (2003). Citebase Search: Autonomous Citation Database, In Third international technical workshop and conference of the project SIN, Oldenburg, Germany.

[9] Carterette, B., Voorhees, E. (2011). Overview of Information Retrieval Evaluation, *Current Challenges in Patent Information Retrieval*, Vol. 29, p 69–85.

[10] De Ribaupierre, H., Falquet, G. (2013). A user-centric model to semantically annotate and retrieve scientific documents, *In*: Proceedings of the sixth international workshop on Exploiting semantic annotations in information retrieval, p 21–24. ACM.

[11] De Silva, P. U., Vance, C. K. (2017). Measuring the impact of scientific research, *In Scientific Scholarly Communication*, p 101–115. Springer, Cham.

[12] Egghe, L. (2006). Theory and practise of the g-index, *Scientometrics*, Vol. 69, p 121–129.

[13] Fisas, B., Ronzano, F., Saggion, H. (2016). A multilayered annotated corpus of scientific papers, *In*: Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC2016, Paris, France.

[14] Gábor, K., Zargayouna, H., Buscaldi, D., Tellier, I. and Charnois, T. (2016). Semantic Annotation of the ACL Anthology Corpus for the Automatic Analysis of Scientific Literature, *In*: Proceedings of the LREC Conference, Portoroz, Slovenia.

[15] Galke, L., Mai, F., Schelten, A., Brunsch, D., Scherp, A. (2017). Using Titles vs. Full-text as Source for Automated Semantic Document Annotation, *In*: Proceedings of the Knowledge Capture Conference, Austin, USA, *p* 20–28. ACM.

[16] Halevi, G., Moed, H., Bar-Ilan, J. (2017). Suitability of Google Scholar as a source of scientific information and as a source of data for scientific evaluation-Review of the literature, *Journal of Informetrics*, Volume 11, *p* 823–834.

[17] Hammarfelt, B., Rushforth, A. D. (2017). Indicators as judgment devices: An empirical study of citizen bibliometrics in research evaluation, *Research Evaluation*, Volume 26, *p* 169–180.

[18] Harzing, A. W. (2011). The publish or perish book: your guide to effective and responsible citation analysis. Tarma software research, Australia.

[19] Harzing, A. W., Alakangas, S. (2016). Google Scholar, Scopus and the Web of Science: a longitudinal and cross-disciplinary comparison, Scientometrics, Volume 106, *p* 787–804.

[20] Haustein, S. (2016). Grand challenges in altmetrics: heterogeneity, data quality and dependencies, *Scientometrics*, Volume 108, p 413–423.

[21] Herrera, N. P., Gomez, F. L., Bucheli, V. A., Pabón, O. S. (2017). Semantic annotation and retrieval of scientific documents in a big data environment, In the 7th Latin American Conference on Networked and Electronic Media (LACNEM2017). Valparaiso, Chile.

[22] Hirsch, J. (2005). An index to quantify an individual?s scientific research output, *In*: Proceedings of the National Academy of Science, Volume 102, *p* 16569–16572.

[23] Hood, W. W., Wilson, C. (2001). *The literature of bibliometrics, scientometrics, and informetrics, Scientometrics*, 52, p 291–314.

[24] Huggins-Hoyt, K. Y. (2018). African American Faculty in Social Work Schools: A Citation Analysis of Scholarship, Research on Social Work Practice, Volume 28, p 300-308.

[25] Ibrahim, N., Habacha Chaibi, A., Ben Ghézela, H. (2017). Scientometric re-ranking approach to improve search results, In the 21st International Conference on Knowledge-Based and Intelligent Information Engineering Systems, Marseille, France, *p* 447–456. IEEE.

[26] Ibrahim, N., Habacha Chaibi, A., Ben Ahmed, M. (2015). New scientometric indicator for the qualitative evaluation of scientific production, *New Library World Journal*, Volume 116, p 661–676.

[27] Kboubi, F., Habacha, A C., Ben Ahmed, M. (2012). Semantic Visualization and Navigation in Textual Corpus, *International Journal of Information Sciences and Techniques (IJIST)*. Vol. 2, p 53–63.

[28] Manis, J. G. (1951). Some academic influences upon publication productivity Social Forces, Vol. 29, p 267–272.

[29] Mayr, P., Scharnhorst, A. (2015). Scientometrics and information retrieval: weak-links revitalized, Scientometrics, Vol. 102, p 2193–2199.

[30] Milojevic, S., Radicchi, F., Bar-Ilan, J. (2017). Citation success index- An intuitive pair-wise journal comparison metric, *Journal of Informetrics*, Vol. 11, p 223–231.

[31] Moed, H. F. (2017). From *Journal Impact Factor to SJR, Eigenfactor, SNIP, CiteScore and Usage Factor, in Applied Evaluative Informetrics*, p 229–244. Springer, Cham.

[32] Noyons, E. C., Moed, H. F., Van Raan, A. F. (1999). Integrating research performance analysis and science mapping, Scientometrics, Vol. 46, p 591–604.

[33] Pal, S., Moore, T. J., Ramanathan, R., Swami, A. (2017). Comparative Topological Signatures of Growing Collaboration Networks, in Workshop on Complex Networks CompleNet , p 201–209. Springer, Cham.

[34] Robertson, S. E. (1997). The probability ranking principle in IR, *Journal of documentation*, Volume 33, p 294–304.

[35] Schöpfel, J., Azeroual, O., Saake, G. (2019). Implementation and user acceptance of research information systems: An empirical survey of German universities and research organisations, *Data Technologies and Applications*, 54 (1), p 1-15.

[36] Singhal, A., Kasturi, R., Srivastava, J. (2013). Automating document annotation using open source knowledge, *In*: Proceedings of the IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT). p 199–204. IEEE Computer Society.

[37] Strohman, T., Metzler, D., Turtle, H., Croft, W.B. (2005). Indri: A language model-based search engine for complex queries, *In*: Proceedings of the International Conference on Intelligent Analysis, *p* 2–6.

[38] Thelwall, M. (2018). Microsoft Academic automatic document searches: accuracy for journal articles and suitability for citation analysis, *Journal of Informetrics*, Volume 12, *p* 1–9.

[39] Van Raan, A. F. J. (2013). Handbook of quantitative studies of science and technology. Elsevier.

[40] Walker, J. (2002). CrossRef and SFX: complementary linking services for libraries, New Library World, Vol. 3, p 83–89.

[41] Walters, W. H. (2017). Do subjective journal ratings represent whole journals or typical articles?, *Journal of Informetrics*, Vol. 11, p 730–744.

[42] Zahedi, Z., Costas, R., Wouters, P. (2014). How well developed are altmetrics? A crossdisciplinary analysis of the presence of alternative metrics? *In*: Scientific publications, *Scientometrics*, Volume 101, *p* 1491–1513.

[43] Zhang, C. T. (2009). The e-index, complementing the h-index for excess citations, *PLoS One*, Volume 4, p 29–54.

[44] Zhao, H., Luo, Z., Feng, C., Ye, Y. (2019, July). A context-based framework for resource citation classification in scientific literatures, *In*: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, p 1041-1044.