International Journal of Engineering Research And Management (IJERM)
ISSN: 2349- 2058, Volume-04, Issue-02, February 2017

Efficient Scientific Research Literature Ranking Model based on Text and Data Mining Technique

Ronald Brisebois, Alain Abran, Apollinaire Nadembega, Philippe N'techobo

Abstract— **With the rapidly increasing of the volume of scientific publications, find quickly the relevant papers for literature review (LR) about specific topic becomes a challenging task for researchers and students. In this vein, a new literature review assistant scheme (LRAS) (1) to evaluate scientific papers relevancy according to discipline and specific topic and (2) to find papers that match a specific research topic for LR is proposed in this work. More specifically, we propose an approach based on text and data mining (TDM) that computes paper index, called Dynamic Topic based Index (*DTb Index*), takes into account (i) venues impact, (ii) authors and their affiliated institutes impact, (iii) key findings and citations impact and (iv) papers references impact. We also implement efficient search prototype that find papers according to researcher selection parameters and his annotations. The required researcher selection parameters are (i) the main topic of his research, (ii) description of his research, (iii) the title and (iv) the keywords of the paper that he plans to provide in the context of his research and for which he needs to make a LR. Based on these parameters, the engine computes the literature corpus radius index (*LCR Index*) of each paper. The main contribution of LRAS search engine prototype is the fact that the *LCR Index* takes into account the area of research. We evaluated our proposed scheme and the simulation results show that the proposed scheme outperforms traditional schemes.**

Index Terms—**Research publications ranking, Bibliometrics, Scientometrics, Information Retrieval, Scientific literature evaluation, Reference analysis.**

I. INTRODUCTION

Literature review (LR) is one of the most important phases of research. Researchers must identify the limits and challenges about certain scientific domain. The problem is where to find the best and most relevant papers that guarantees to ascertain the state of the art on that specific domain. Certainly, the volume growth of scientific papers and the online availability of repositories allow researchers to discover, analyze and maintain an updated bibliography for specific research fields. However, in recent years, the crescent volume of scientific papers available is becoming a problem for researchers, who, unable to exploit the whole literature in a specific domain tend to follow ad-hoc approaches. In order to help researchers for the LR tasks, it becomes necessary to analyse a large volume of papers in a fairly short time. To do so, we need to evaluate paper relevance according to the scientific research domain and topic; this task refers to the ranking process of scientific papers. Ranking the relevance of scientific papers is an ongoing and a long-standing challenge.

Unfortunately, all the works about the scientific research impact are focused on the researchers ranking; however, a researcher impact is useful to rank scientific papers that he proposes. Some online academic search engines have already implemented several indices to evaluate the scientific impact of researchers, that is the case of the h-index and i10-index used in Google Scholar for evaluating researchers' impact. Most existing researchers' indexes computation algorithms are based on the number of citations received by each paper written by a researcher. For example, if a researcher has published more papers with more citations, the researcher's h-index increases. According to [1], there are four factors by which it is possible to measure the validity of scientific research: (1) number of papers, (2) impact factor of the journal, (3) the number and order of authors and (4) citations number. The number of papers speaks more about productivity than about quality while impact factor represents simple quantification of the data for scientific production. Citation analysis identifies the types of citations and measures the number of citations, self- citations. While peer-review and citation-based bibliometrics indicators have become global means of measuring research output and are playing a critical role in this process. However, citations have been criticized for limiting their scope within academic and neglecting the broader societal impact of research. Using these four factors, ranking the relevance of scientific papers cannot be done without text and data mining (TDM).

TDM can be defined as automated processing of large amounts of structured digital textual content, for purposes of information retrieval, extraction, interpretation, and analysis. Indeed, due to the large corpora of data accumulated, automated or semi-automated analysis of their contents reveals patterns that allows establishment of fact patterns invisible to the naked eye [2]. There are many reasons researchers might want to utilize TDM methods in their research. Clark [3] suggested, due to enormous growth of the volume of literature produced, that researchers should apply text mining technique to enrich the content and to perform the systematic review of literature. Indeed, mining can improve indexing, be deployed to create relevant links, to improve the reading experience. Specifically in the context of TDM, text mining is a subfield of data mining that seeks to extract

Ronald Brisebois, École de technologie supérieure, Université du Québec, Canada, (ronald.brisebois.1@ens.etsmtl.ca).
Alain Abran, École de technologie supérieure, Université du Québec, Canada, (alain.abran@etsmtl.ca).
Apollinaire Nadembega, Network Research Lab, University of Montreal, Canada, (apollinaire.nadembega@umontreal.ca).
Philippe N'techobo, École Polytechnique de Montréeal, Canada, (ntechobo.edoukou-philippe-armel@polymtl.ca).

valuable new information from unstructured (or semi-structured) sources. Text mining extracts information from within those documents and aggregates the extracted pieces over the entire collection of source documents to uncover or derive new information. This is the preferred view of the field that allows one to distinguish text mining from natural language processing (NLP).

TDM techniques are widely used for ranking algorithms. Ranking algorithms are defined as the procedure that search engines use to give priority to the returned results. Recent years have seen increased adoption of scientometrics techniques for assessing research impact of publications, researchers, institutions, and venues; scientometric can be defined as the science that deals with evaluation of a scientific article refers to the finding quantitative indicators (index) of the scientific research success; unfortunately, the field of scientometrics focuses on analyzing the quantitative aspects of the generation, propagation, and utilization of scientific information. Several approaches are proposed to rank scientific articles and measure the impact of research [1, 4-16]. Some approaches focused on journal ranking [15] while others focused on universities and research institution ranking [16]. Unfortunately, these approaches only consider publication-count or focused on citation analysis (citation-based approaches); the aggregate citation statistics are used to come up with evaluative metrics for measuring scientific impact. They ignored the quality of articles in term of new contribution and scientific impact and limited the evaluation to quantitative aspect. Despite several criticisms of citation-based measures for impact, it is still the subject of much scientometrics research; a highly cited paper for a given scientific research field has influenced many other researchers; new contribution includes methods for evaluating research institutions, journals and researchers. Indeed, the main approach for scientific articles ranking is the citation analysis that is mainly the number of time that a paper is cited; unfortunately, traditional approach does not consider the publisher, conference or workshop relevance. In addition, the social aspect is not taking into account; indeed, the peers' evaluations need to be considered to measure the quality of an article; the opinion of the scientific community of the research field may contribute to identify the relevant articles. Most of these approaches reduce a citation to a single edge between the citing and cited paper and treat all the edges equally. This is clearly an oversimplification since all citations are not equal and need to be considered distinctly.

According to Wan and Liu [17], as a simple extension, taking into account the number of times a paper is cited in the citing paper often does a better job of measuring the impact of the cited paper; in order word, citations should be consider to evaluate papers impact. The text around a citation anchor can be used to assess the attitude of the citing paper towards the cited paper; for example, the citation category may define citing paper attitude. And aggregating the attitudes of all the citations to a paper can give us a quantitative measure of the attitude of the community towards that paper. However, in addition to citations, others aspects need to be consider such as: (1) analyzing of social aspects of scientific research, (2) analyzing history, (3) structure and progress of scientific fields and (4) measuring inter-disciplinary of scientific fields.

For example, the ranking of scientific journals is important because of the signal it sends to scientists about what is considered most vital for scientific progress. Journal rankings are also important because they provide a filter for researchers in the face of a rapidly growing scientific literature; they provide a way to quickly identify those articles that other researchers in a field are most likely to be familiar with.

In this paper, we propose scheme, called Literature Review Assistant Scheme (LRAS), that allows computing the ranking index of the relevance of scientific papers and subsequently, allows searching papers that best match with the researcher selection parameters. The main objective of LRAS is to assist the researchers in the LR redaction tasks that consist to, first, find papers which match with their research topic and secondly, evaluate the relevance of these papers. LRAS proposes two main processes:

1) The first process of LRAS allows evaluating the relevancy of a scientific paper for a given domain and research topic; to do that, LRAS computes the paper ranking index, called Dynamic Topic based Index (***DTb Index***) making used of TDM technique. Indeed, to compute the DTb Index, LRAS considers several criteria such as (i) venue age and impact, (ii) citation category and polarity, (iii) authors' impact, (iv) authors' institutes impact and (v) citing document of cited document. In contract to existing ranking algorithm, LRAS focuses on the paper age and author social activities in terms of researcher. Ranking algorithm also considers the number of time a paper is cited in the same documents.
2) The second process of LRAS allows finding the scientific papers that best match with the researchers' topics for their LR. Notice that the traditional search algorithms use only the titles of papers as selection parameter. In contract to them, LRAS search algorithm considers (i) the main topic of the research, (ii) description of the research, (iii) the title and (iv) the keywords of the paper that researcher plans to provide in the context of his research and for which he needs to make an LR. The LRAS search algorithm is based on TDM technique. The main contribution of LRAS search engine prototype is the fact that the algorithm takes into account the area of research.

The remainder of this paper is organized as follows. Section II presents some related work. Section III describes our proposed literature review assistant scheme (LRAS) using TDM approaches. Section IV evaluates the proposed literature review assistant scheme (LRAS) via simulations. Section V concludes this paper.

II. RELATED WORKS

The network-based analysis is a natural and common approach for evaluating the scientific credit of papers. Although the number of citations has been widely used as a metric to rank papers, recently some iterative processes such as the well-known PageRank algorithm have been applied to the citation networks to address this problem. In the context of this work, several existing approaches for scientific papers ranking [5, 6, 9-12, 14, 16-19] have been analysed.

Bornmann et al. [14, 16] proposed an web application to measure the performance of research institutions. They used two indicators to perform their measurement: best paper rate and best journal rate. Best paper rate is the proportion of the institutional publications which belong to the 10% most frequently cited publications in their subject area and publication year. The best journal rate is the proportion of publications which an institution publishes in the most influential journals worldwide. According to the authors, the most influential journals are those which are ranked in the first quartile (top 25%) of their subject areas as ordered by the indicator SCImago Journal Rank (SJR).

Ranking researchers, journals and institutions may not allow to evaluate the scientific papers relevancy; however, they may be use in this scientific papers relevancy index computation. Indeed, Marx and Bornmann [12] presented an overview of methods based on cited references, and examples of some empirical results from studies are presented. according to authors, the use of a selection for the analysis of references from the publications of specific research areas should enable the possibility of measuring citation impact target-oriented (i.e. limited to these areas). They mentioned that some empirical studies have shown that the identification of publications with a high creative content seems possible via the analysis of the cited references. For authors, cited reference analysis indicate the great potential of the data source. Authors also mentioned the new method, known as citing side normalization where each individual citation receives a field-specific weighting; to compute, each citation is divided by the particular number of references in the citing work.

Wan and Liu [17] proposed citation-based analysis to evaluate scientific impact of researchers in the context of Author-Level-Metric, called WL-index. Indeed, they raised the issue of the consideration of number of time that a cited paper is mentioned in a citing paper. According to authors, the counting based on the binary citation relationships is not appropriate; in a given article, some cited references appear only once, but others appear more than once. WL-index is a variant of h-index where the number of times cited paper is mentioned is considered. Indeed, take into account the number of times a cited paper is mentioned in citing paper is good idea; unfortunately, their proposed contribution cannot allow to measures impact of paper in order to identify relevant contribution; in addition, they do not consider the category of citation to evaluation scientific impact of researchers.

Hassan et al. [6] proposed a new ranking algorithm for scientific research papers, called Paper Time Ranking Algorithm (PTRA), that depends on three factors to rank its results: *paper age, citation index and publication venue*; they gave priority to each one of these parameters. Indeed, for a given paper, they computed its weight as the sum of the age of the conference or the journal impact factors, the number of citation of the paper and the age of paper. Unfortunately, they do not consider Author-Level-Metric and ignore the citation category in the computation of their citation index. Also, considering the number of citations is not good approach due to the age of paper; indeed, newspapers are penalized; they may use the average number of citations instead on the number of citations.

Rúbio and Gulo [11] proposed recommending papers based on known classification models including the paper's content and bibliometric features. Indeed, they combined text mining efforts and bibliometric measures to automatically classify the relevant papers. They made use paper's metadata such as *year of publication, citation number, references number and type of publication (journal, conference, workshop, etc.)* to measure the paper relevancy for specific science field. In they approach, they applied a ML algorithm ID3 for papers relevancy classification based on specialist annotation. Authors mentioned that their approach combines text mining and bibliometric; unfortunately, their approach only used bibliometric metrics. However, making use of machine learning (ML) technique is good things.

Madani and Weber [5] proposed an approach that applied bibliometrics analysis and keyword-based network analysis to recognize the main papers, authors, universities, and journals. Indeed, they made use bibliometrics (quantitative approach) analysis to find a general view about top authors, journals, universities, and countries; to find the most effective papers, they applied the 'eigenvector centrality' measure. For the patent evaluation, they extracted keywords from abstracts, created keyword-based network that is analyzed by cluster analysis to find groups of keywords making use of minimum spanning tree method. The list of limitations is: (1) authors do not explain how the keyword-based network is build; (2) they use only existing method and approach; and (3) paper manual annotated keywords (those given by authors of papers) are better than extracted keywords.

Wang *et al* [10] proposed a unified ranking model of scientific literature, called *MRFRank*, that employed the mutual reinforcement relationships across networks of papers, authors and text features. More specifically, *MRFRank* is proposed by incorporating the extracted text features and constructed weighted graphs. Indeed, for the same sentence, they extracted words and words co-occurrence form title and abstract. Then, they computed the TF-IDF of each word as the weight of this word. The main limitation of this approach is the fact that authors just consider the abstract to compute the weight of the word.

Gulo et al. [18] proposed a solution that automatically classifies and prioritizes the relevance of scientific papers; the solution combined text mining and ML techniques as support to identify the most relevant literature. According to authors, their approach allows to browse huge article collections and quickly find the appropriate publications of particular interest by using ML techniques. Indeed, based on previous samples manually classified by domain experts, they applied a Naive Bayes Classifier to get predicted articles; a human expert in a specific domain has analyzed each one of the training set of publications and classified the priority of the references regarding two main criteria: relevance of the reference and adequacy to the interested scientific domain. Then, based on the outputs of experts, the process of automatic classifying publications starts with a selected set of keywords that represent the context and the area of interest. As the entire supervised learning algorithm, manual contribution is highly required.

To conclude, we summarize the limitations of existing approaches for ranking the relevance of scientific papers as

follows:

1) they only use citations count; in addition, they do not consider the age of papers, penalizing the recent papers;
2) they do not consider the category and polarity of citations;
3) they do not consider the other types of venues, such as conferences and workshops. In addition, what about unpublished documents?
4) for those which are based on machine learning technique; they require a large manual contribution of specialists or experts for the training step of the learning model;
5) for those which are based on text analysis to identify relevant papers; they are limiting themselves to title and abstract.

In this paper, we propose a scheme that proposes solutions to overcome these limitations. The proposed LRAS considers several criteria such as venue age and impact, citation category and polarity, authors' impact, authors' institutes impact and citing document of cited document.

III. LRAS: Literature Review Assistant Scheme

Here, we present the details of the proposed scheme, called LRAS. More specifically, we present (A) the TDM process used by LRAS to compute the relevance ranking index that denotes the relevancy of a scientific paper for a research topic and (B) the TDM based process used by LRAS to find best papers for literature review (LR) of specific research topic.

A. Dynamic Topic based Index (DTb Index) computation process

As mentioned above, most of existing ranking approaches focus on measuring the influence of a scientific paper based on the citations analysis. In contract to these approaches, LRAS computes the DTb Index that denotes the paper relevancy according to a specific research domain and topic; that is why this index is called dynamic topic based.

More specifically, the DTb index is also computed as a weighted sum of the values that denote the importance of the different inputs considered. The DTb index is computed using a number of additional features:

1) Key findings and peer citations index (see equation 1),
2) Venue index (see equation 2 to 6),
3) Document references index (see equation 7 to 8),
4) Authors and their affiliated institutes (see equation 9 to 12).

In contrast to existing ranking approaches, the LRAS is not limited to journal-level metrics; it also considers conference proceedings and workshop metrics; making LRAS, a scheme based also on venue-level metric.

In the rest of this section, we show how the different concepts are used to compute the DTb Index (see equation 13).

1) Paper relevance according to researchers' key findings and peer citations

The Key Findings are the annotations in regards to important findings in the paper. Indeed, previous researchers who have already analyzed the paper have provided annotations called key findings. These key findings are identified and analyzed by the TDM approach. The TDM analysis consists in classifying the key findings into three categories:

1) *Very relevant:* indicates that the paper is very relevant and adequate for the LR,
2) *Adequate:* indicates that the paper is not relevant, but may be the focus of attention, if possible.
3) *Not relevant:* indicates that the paper is not relevant and not adequate for the search.

Let:

1) *Cat_annot* be the category of a key finding;
2) *Y* be the age of a paper *d;*
3) *X* be the publication date of *d*;

For example: for a paper published in 2000, Y =16 and X=2000.

The key findings index of paper *d* is computed as follows:

$$KeyFindingsIndex\left(d, Cat\_Annot, Y\right) = \frac{\sum_{i=0}^{Y-1}\left(\left(Y-i\right)\times Nb\left(d, Cat\_Annot, \left(X+Y-i\right)\right)\right)}{Y!} \quad (1)$$

where Nb(d,Cat_Annot,Z) denotes the number of times the key findings Cat_Annot= "very relevant" are detected in paper *d* at year Z.

The concept behind the computation of the key findings index is to give more importance to the more recent annotations instead of simply counting the number of considered key findings. This places more emphasis on recently published papers.

2) Paper relevance according to venue

The venue type is important in the ranking of scientific papers. The intent is to consider not only papers from academic journals, but also papers from other types of venues, such as conference proceedings and workshops, as well as unpublished papers such as research reports. In LRAS, four types of venue are considered:

1) Journal
2) Conference proceedings
3) Workshop
4) Unpublished.

Here, the venue types are ordered according to their importance in the researcher's opinion. For example: a researcher may consider that a journal paper is more important than a conference proceedings paper; thus, journal is first and conference is second. To compute the venue impact, LRAS evaluates the similarity between (1) the venue topic and the papers main topic and (2) the venue name and the papers title. The similarity matching of the paper's main topic (we assumed that the research topic of the paper is known in advance) with the venue's topics (where paper *d* is published or presented) is computed as follows:

$$sim\_topic\left(Td, Tv\right) = \max_{j\in[1,m]}\left(j-gram(Td, Tv)\right) \quad (2)$$

where *Td* and *Tv* denote the main topic of paper *d* and the main topic of venue *v*, respectively.

The similarity matching between paper title and venue name (where paper *d* is published or presented) is computed as follows:

$$sim\_name(Nd, Nv) = \max_{j \in [1,m]}(j - gram(Nd, Nv)) \quad (3)$$

where *Nd* and *Nv* denote the title of document *d* and the name of venue *v*,respectively.

Thus, the venue *v* impact for a specific paper *d* is given by:

$$VenueImpact(d, v) = age\_venue(v) + avg\_num\_pub(v) + rev\_num(v) + \frac{avg_{sub(v)}}{avg_{acc(v)}} + freq(v) + sim\_topic(Td, Tv) + sim\_name(Nd, Nv) \quad (4)$$

where

- *age_venue(v)* denotes the age of venue *v*,
- *avg_num_pu(v)* denotes the number of publications per year,
- *rev_num(v)* denotes the number of reviewers per submitted paper,
- *avg_sub(v)* denotes the average number of submitted papers per year,
- *avg_acc(v)* denotes the average number of accepted papers per year,
- *freq(v)* denotes the frequency of publication per year.

To take into account the type of venue, a weight is assigned to each of them according to its order and the couple (Vinit, Vunit), where:

- Vinit is an initial value and
- Vunit is the difference in weight between two consecutive types of venue.

For example: a venue type with order *i* will have the weight:

$$VtypeWeight(v) = Vinit + ((Q + 1 - i) \times Vunit) \quad (5)$$

where Q is the number of types of venue. Here, Q is equal to 4.

Finally, the venue-based index of paper *d* is computed as follows:

$$VenueIndex(d, v) = VtypeWeight(v) \times VenueImpact(d, v) \quad (6)$$

3) Paper relevance according to authors and their affiliated institutes

Until now, a number of different indicators have been proposed for evaluating the scientific impact of a scientist or a researcher, most of which are variants and revisions of h-index. However, h-index is limited to number of citations without considering the author's social personality in terms of peer award, for example. As was done for the venue index, LARS computes the paper relevance based on the authors and their affiliated institutes.

Let:

1) *Td* be the main topic of paper *d*; we assumed that the research topic of the paper is known in advance;
2) a_i be an author.

The author a_i influence on the relevance of paper *d* is computed as follows:

$$AuthorImpact(d, a_i) = \frac{nb\_cited(Td)}{nb\_pub(Td)} + \frac{nb\_jour(Td)}{nb\_pub(Td)} + nb\_award(Td, a_i) + nb\_jour(Td, I_i) + nb\_award(Td, I_i) \quad (7)$$

where:

- *nb_cited(Td)* denotes the number of publications of author a_i cited on the topic *Td*,
- *nb_pub(Td)* denotes the number of publications of a_i on the topic *Td*,
- *nb_jour(Td)* denotes the number of journal publications by a_i on the topic *Td*,
- *nb_awar(Td,a_i)* denotes the number of awards of a_i on the topic *Td*,
- *nb_jour(Td,I_i)* denotes the number of publications which a_i's affiliated institute publishes in the most influential journals worldwide on the topic *Td*,
- *nb_awar(Td,I_i)* denotes the number of awards of a_i's affiliated institute on the topic *Td*.

The author index for paper *d* is computed as follows:

$$AuthorIndex(d) = \frac{\sum_{i=1}^{A}(A + 1 - i) \times AuthorImpact(d, a_i)}{A!} \quad (8)$$

where *A* denotes the number of authors of paper *d*. The idea is to give more importance to top authors; the first author therefore has greater weight than the second author.

4) Paper relevance according to document references

The paper's interaction with other papers on the topic is measured. Two groups of papers are defined: Citing documents and Cited documents.

For a better understanding, let *d* be a considered paper; a citing document is a document that cited the document *d*, while a cited document is a document cited by the paper *d*. Note that the number of cited documents is static while the number of citing documents may increase with time. These two terms are important for the evaluation of document relevance. Fig. 1 illustrates the two terms according to the publication date.

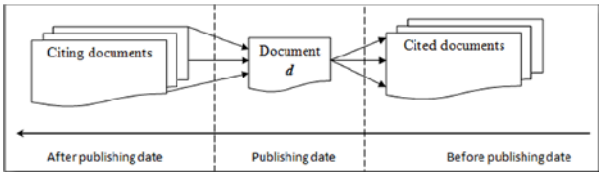

Fig. 1: Illustration of a paper reference documents

The paper's relevance based on citations includes three operands; the computation of paper's relevancy according to the references is based on the assumptions that (1) relevant papers very often cite relevant papers and (2) relevant papers are those that are frequently cited.

- Number of citing documents of paper *d* according to its age; it is computed as follows:

$$CitingImpact(d) = \frac{\sum_{i=0}^{Y-1}(Y-i)\times nb\_citing(i+1)}{Y!} \quad (9)$$

where *nb_citing(i)* denote the number of citing documents with age *i* and *Y* denotes the age of the document *d*. In addition, *CitingImpact (d)* gives more importance to recent citations.

- Average number of times a paper *d* is mentioned in citing documents; it is computed as follows:

$$CitingAvgImpact(d) = \frac{\sum_{j=1}^{P} nb\_time\_citing(d, D_j)}{P \times Y} \quad (10)$$

where $nb\_time\_citing(d,D_j)$, denotes the number of times the document *d* is cited in the citing document D_j, *P* is the total number of documents citing *d* and *Y* is the age of the document *d*.

- Number of citing documents of paper D_l (a cited document of paper *d*) according to the paper D_l age; it is computed as follows:

$$CitedCitingAvgImpact(d) = \left| \bigcup_{D_l \in L} \left\{ \frac{nb\_citing(D_l)}{age(D_l)} \geq 5 \right\} \right| \quad (11)$$

where *L* denotes the set of documents cited in *d*, $age(D_l)$ denotes the age of document D_l and $nb\_citing(D_l)$ denotes the number of times document D_l is cited.

Finally, the relevancy of paper *d* based on references is computed as follows:

$$ReferencesIndex(d) = CitedCitingAvgImpact(d) + CitingAvgImpact(d) + CitingImpact(d) \quad (12)$$

5) DTb index computation based on the previous computed index

As mentioned above, the DTb index is a weighted sum of the computed values for different features that impact the relevance of a paper.

Let the couple (Init, Unit) where:

- Init is an initial value, and
- Unit is the difference in weight between two consecutive aspects.

Init and Unit allow to assign different importance to each features. The DTb index of paper ***d*** is computed as follows:

$$DTb\ Index(d, RF, VN, AA, KF) = \frac{Val(RF,d) + Val(VN,d) + Val(AA,d) + Val(KF,d)}{\sum_{k=0}^{3}((Init + (Unit \times k)))} \quad (13)$$

where

$$Val(RF, d) = Init \times ReferenceIndex(d)$$

$$Val(VN, d) = (Init + (Unit \times 1)) \times VenueIndex(d)$$

$$Val(AA, d) = (Init + (Unit \times 2)) \times AuthorIndex(d)$$

$$Val(KF, d) = (Init + (Unit \times 3)) \times KeyFindingsIndex(d, Cat\_Annot, Y)$$

B. Papers corpus for literature review selection process

To identify an LR corpus, the selection parameters are classified into three categories (see Table 1):

1. Evaluation-based
2. Selection-based
3. Sort-based.

Table 1: STELLAR classification of selection parameters

Evaluation-based	Selection-based	Sort-based
Main Topic (MaT)	Discipline	MLTC (Yrs, %)
Keywords (KeW)	Languages	Number of References (<=)
Title (TiT)	LCR Index Threshold	Researcher Annotations (RA)
Description (DeC)		

Each class of the selection parameters is used for specific step on the selection process.

Selection based parameters are used to filter the papers repository in order to reduce the number of papers for the next steps; that allow to save computation cost. Sort based parameters are used to select the final list of papers for LR.

Evaluation-based parameters are used to compute the literature corpus radius (LCR) index. First, the value of each evaluation-based parameter is computed by determining the similarity of each evaluation-based selection with a predefined section of the document. The similarity matching value is in the range [0,1] where 1 means the most similar while 0 means the least similar. Next, based on the similarity matching value (e.g., the predefined weight of each of them), the LCR index is computed. Fig. 2 shows the process of LR corpus selection based on researcher's selection parameters and annotations.

Indeed, the first step allows selecting a preliminary corpus of papers (C_0) based on researcher discipline and language. Then, based on the evaluation-based parameters, the LCR Index of each paper of the set of preliminary corpus of papers is computed. Then, based on the LCR Index threshold, the corpus of papers (C_1) is selected; C_1 represents the subset of C_0 where the LCR Index of papers is greater or equal to LCR Index threshold. Finally, based on the sort based parameters researcher and LCR Index, LRAS identifies the final corpus of papers (C_2) that will be used for the LR; C_2 is a subset of C_1.

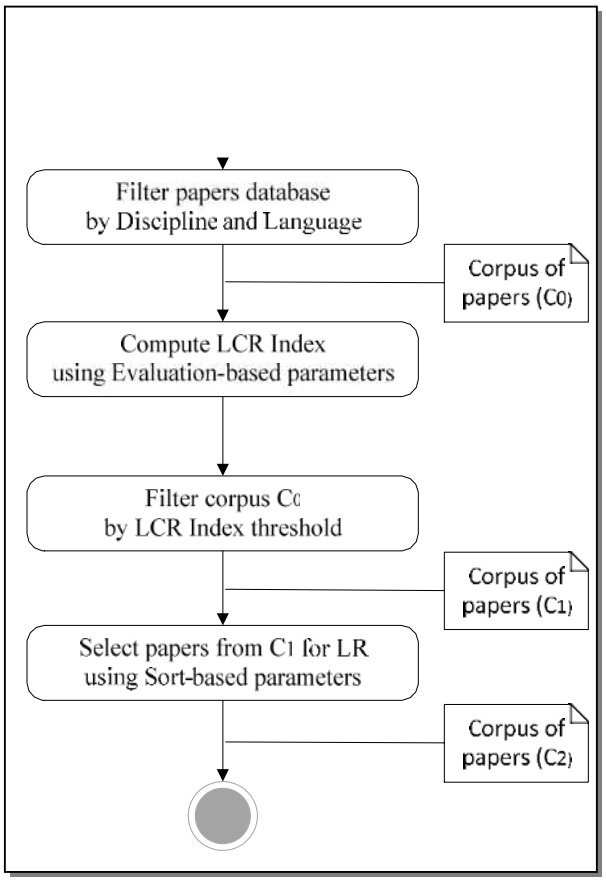

Fig. 2: Literature corpus radius (LCR) selection process

The step 1 and 3 can be performed by simply SQL request to the database using papers metadata discipline and language for step 1 and LCR Index for step 3; in the rest of this section, the details of step 2 and 4 are given.

1) Step 2 of LR corpus selection (LCR Index computation)

As the DTb Index, LCR Index computation is based on various features that match the researcher evaluation based selection parameters. For each feature, LRAS computes the similarity matching and performs weighted sum of these similarity values to obtain the LCR Index.

For each paper, equations (14) to (16) compute the similarity of paper with the researcher's main topic while equations (17) to (18) compute the similarity of each paper with the researcher selection parameters in terms of keywords. Equations (19) to (20) compute the similarity matching of each document with the RS parameters "Title" while equations (21) to (23) compute the similarity matching of each document with the RS parameters "Description". Finally, equation (24) allows computing the LCR Index.

- *Similarity matching of a researcher main topic with the topics extracted from paper abstract*

The similarity matching with the researcher main topic is computed from the abstracts. The abstract of each is recorded in the "ABSTRACT" metadata provided by the publisher. The similarity matching computation makes use of this metadata as input to determine the paper's similarity with the researcher-defined main topic.

Let *d* be the paper and *Ad* the abstract of *d*. Next, based on the topic detection algorithm, called BM-Scalable Annotation-based Topic Detection (BM-SATD), the topics of paper *d* are detected from *Ad*; we assume that BM-SATD exists. Thus, using paper's abstract as input, BM-SATD detects their topics.

Let:

1) *Ta* be the topic detected in the abstract of paper *d*;
2) *MT* be the main topic provided as the researcher selection parameters and *n* be the number of terms of MT = (w_1, w_2, …, w_i, …, w_n);
3) *SimMatch_MaT(MT,d)* be the function that evaluates the similarity of *MT* with the paper *d* abstract; note that the terms of *MT* are ordered.

First, the i-gram of MT is calculated:

$$f(i-gram, MT, Ad) = \sum_{k=1}^{n-(i+1)} nb\left(w_k, w_{k+1}, \ldots, w_{k+i-1}\right) \quad (14)$$

where nb(w_k, w_{k+1}, …, w_{k+i-1}) is the number of times that the i-gram (w_k, w_{k+1}, …, w_{k+i-1}) appear in *Ad* (the abstract of paper *d*).

Next, the weight of the researcher's main topic for paper *d* is computed using the following equation:

$$w\_Mat(MT, d) = \sum_{i=1}^{n} i \times f(i-gram, MT, Ad) \quad (15)$$

To obtain a similarity value between 0 and 1, normalization is applied. Let *Max_MaT* be the largest value of *w_MaT(MT,d)* among all the considered papers. SimMatch_MaT(MT,d) is computed by:

$$SimMatch\_Mat(MT, d) = \frac{w\_Mat(MT, d)}{Max\_Mat} \quad (16)$$

- *Similarity matching of researcher keywords with paper keywords*

The similarity matching based on the researcher keywords is computed using the paper keywords. The keywords of each paper are recorded in the "KEYWORDS" metadata provided by the publisher.

Let:

1) *Kd* be the *set* of keywords of paper *d*;
2) *KW* be the set of keywords provided in the researcher selection parameters;

3) *SimMatch_KeW(KW,Kd)* be the function that computes the similarity matching of *KW* with *Kd*.

First, the weight of *KW* according to paper *d* keywords *Kd* is computed as follows:

$$w\_KeW(KW,d)=|KW \cap Kd| \quad (17)$$

To obtain a similarity value between 0 and 1, normalization is applied; the SimMatch_KeW(KW,d) is computed as:

$$SimMatch\_KeW(KW,d)=\frac{w\_KeW(KW,d)}{KW} \quad (18)$$

- *Similarity matching of researcher's research title with paper title*

Before the similarity matching computation, the researcher title and paper titles are pre-processed. The objective of the pre-processing is to filter noise in order to obtain suitable text for performing the analysis. This consists in stemming, phrase extraction, part-of-speech filtering and removal of stop-words. More specifically, it includes the following operations:

1) Segmentation: the process of dividing a given document into sentences.
2) Stop-words removal: Stop-words are frequently occurring words (e.g., 'a' and 'the') that impart no meaning and generate noise. They are predefined and stored in an array. Note that the removal of stop-words follows specific rules. For example, in "prediction of mobility", removal of the stop-word "of" changes the expression to "mobility prediction".
3) Tokenization: the input text is separated into tokens.
4) Punctuation marks: the spaces and word terminators are identified and treated as word breaking characters.
5) Word stemming: each word is converted into its root form by removing its prefix and suffix for comparison with other words.

The output of the pre-processing is the set of terms.

Let:

1) *Td* be the set of terms of the title of paper *d*;
2) *TT* be the set of terms of the researcher selection title;
3) *SimMatch_TiT(TT,Td)* be the function that evaluates the similarity matching of TT with *Td*.

First, the weight of TT according to the paper *d* title *Td* is computed as follows:

$$w\_TiT(TT,d)=\max_{j\in[1,m]}(j-gram(TT,Td)) \quad (19)$$

where *m* denotes the number of terms of TT ($m=|TT|$). Indeed, w_TiT(TT,d) is the largest number of sequential terms of TT that appears in *Td*. To obtain a similarity value between 0 and 1, normalization is applied. The SimMatch_TiT(TT,d) is computed as follows:

$$SimMatch\_TiT(TT,d)=\frac{w\_TiT(TT,d)}{m} \quad (20)$$

- *Similarity matching of the researcher' research description with paper abstract*

The similarity matching of the researcher research description is performed using the paper abstract. To do this, the researcher description is semantically compared to the paper abstract in order to measure the similarity level. This similarity matching of a researcher description makes use of WordNet::Similarity, described in [20], which implements six measures of similarity and three measures of relatedness. Several terms may be semantically the same.

Let:

1) *DS* be the researcher description of the research topic as the selection;
2) *s* be the number of terms of $DS = (t_1, t_2, \ldots, t_i, \ldots, t_s)$;
3) *C* be the Literature Corpus where the papaers are of the same discipline;
4) *SimMatch_DeC(DS,d)* be the function that evaluates the similarity matching of *DS* with a paper abstract *Ad*.

First, the semantic similarity of each term in *DS* with those in *Ad* is determined on the basis of the semantic TF-ICF (term frequency – inverse corpus frequency) as follows:

$$SemSim\_T(t_i,d)=TF(t_i,d)\times\log\left(\frac{C}{ICF(t_i,C)}\right) \quad (21)$$

where C, $TF(t_i,d)$ and $ICF(t_i,C)$ denote the preliminary corpus of papers selected based on discipline and language, the number of occurrences of t_i in paper *d* and the number of papers in the corpus *C* where t_i appears.

Next, the semantic similarity of DS to the paper abstract is computed as follows:

$$SemSim\_DeC(DS,d)=\sum_{i=1}^{s}SemSim\_T(t_i,d) \quad (22)$$

To obtain a similarity value between 0 and 1, normalization is applied. The SimMatch_DeC(DS,d) is computed as:

$$SimMatch\_DeC(DS,d)=\frac{SemSim\_DeC(DS,d)}{Max\_DeC} \quad (23)$$

where *Max_DeC* denotes the largest value of SemSim_DeC(DS,d) among all the papers in *C* (i.e, preliminary corpus of papers selected based on discipline and language).

- *LCR Index computation*

Once the similarity matching of each evaluation-based selection is done, the LCR index can be computed. An LCR index value is within the range [0,1] where 0 means the least similar while 1 is the most similar. Note that the LCR index is a weighted sum of the computed value of each selection.

Let:

1) W_init be an initial value
2) W_unit be the difference in weight between two consecutive types of RS parameters.

The LCR index of a paper *d* of literature corpus *C* is computed as follows:

$$LCR\ \ Index\left(d, MT, KW, TT, DS\right) = 1 - \left(\frac{Val(DS,d) + Val(TT,d) + Val(KW,d) + Val(MT,d)}{\sum_{i=0}^{3}\left(W\_init + (W\_unit \times i)\right)} \right) \quad (24)$$

where:

$$Val\left(DS, d\right) = W\_init \times SimMatch\_DeC\left(DS, d\right)$$

$$Val(TT, d) = \left(W\_init + (W\_unit \times 1)\right) \times SimMatch\_TiT(TT, d)$$

$$Val\left(KW, d\right) = \left(W\_init + (W\_unit \times 2)\right) \times SimMatch\_KeW\left(KW, d\right)$$

$$Val\left(MT, d\right) = \left(W\_init + (W\_unit \times 3)\right) \times SimMatch\_MaT\left(MT, d\right)$$

2) Step 4 of LR corpus selection: MLTC, Number of references and "To be included in the LR"

This sub-section describes how LRAS takes into account the researcher's requirements in terms of MLTC (Mix of the Literature Temporal Coverage (Yrs, %), number of references and the specific annotation "To be included in the LR". The MLTC allows the researcher to include a certain percentage of papers whose age is greater than a given age (Yrs).

The idea here is to be able to include very relevant papers that are out of date. To take into account both the MLTC and the number of references without prioritizing either of them, a specific algorithm is needed, which is given by the following pseudo-code. In this pseudo-code, C_1 denotes the preliminary corpus of papers selected based on discipline, language and LCR Threshold while C_2 denotes the final corpus of papers for the LR.

```
New_C1 = ∅
Old_C1 = ∅

If (N ≤ Lengtj of All_C1)
   For the next document in All_C1
   If [(A ≠ 0) AND (B ≠ 0)]
      If [ (next document publication age ≤ y) ]
        Add next document to New_C1 ; A=A-1
      Else If [ (next document publication age > y) ]
            Add next document to Old_C1; B=B-1
          Else
            If [(A = 0) AND (B ≠ 0)]
              Add next document to Old_C1; B=B-1
            Else
              If [(A ≠ 0) AND (B = 0)]
                If [ (next document publication age ≤ y) ]
                  Add next document to New_C1; A=A-1
                Else
                  New_C1 = All_ C1
                  C2= New_C1 ∪ Old_C1
```

First, a list (in descending order) is created based on the LCR index applied to C_1 where the papers tagged "To be included in the LR" are at the top due to their priority; let All_C_1 be this list. Let MLTC (x, y) with its number of selections equal N: this means the researcher expects to have at most ***N*** documents, with a maximum of ***(100-x)%*** (i.e., $\frac{N}{100} \times (100 - x)$) that are at most ***y*** years old, and including all the papers tagged "To be included in the LR". Note that the latter papers have priority.

New_C_1 is defined as a sub-list of C_1 in which the paper age is less than or equal to *y*, and Old_C_1 contains papers older than *y*. Let $A = \frac{N}{100} \times x$ be the length of New_C_1 and $B = \frac{N}{100} \times (100 - x)$ be the length of Old_C_1.

Note that, when the number of papers in All_C_1 is less than N, all the documents are considered affinity matches for the LR; in that case, the MLTC selection is ignored.

However, when there are not enough papers whose age is less than or equal to *y* to satisfy the MLTC selection, a new MLTC is provided in order to reach the number *A*. But if the researcher requires the MLTC selection to be met, some papers are removed from New_C_1 in order to meet the selected MLTC(x, y).

If an "OR" has been used between the researcher sort-based selection' parameters, the LR corpus will be defined as the union of the subsets of papers provided by the MLTC process and the subsets of papers that are tagged "To be included in the LR".

Fig. 3 presents the LRAS prototype for LR corpus selection.

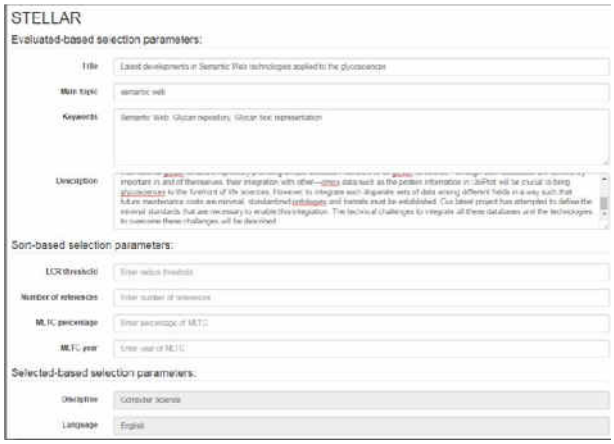

Fig. 3: LR corpus selection prototype

IV. PERFORMANCE EVALUATION

For the performance evaluation, we only measure the ranking relevance of papers. As comparison terms, we use the schemes described in [6] and [11], which are referred to as PTRA and ID3.

For the datasets harvest ring, LRAS prototype implements a crawler engine as [6]. This crawler consists of two main parts; automator and extractor. The main function of the automator is to retrieves search result from well-known scientific paper search engines: reseachGate, Academia, ScienceDirect, Scopus, Google scholar, Citeseerx and IEEE Xplore. The extractor extracts the useful information from the returned pages by the automator. This information's can be

summarized as: the title of the paper, the abstract of paper, the year of publication, the paper citation index, the venue of publication, the venue age and type, author award, author affiliation institute and venue impact. For each paper, the downloaded bibliographic files were parsed to extract the metadata.

Unfortunately, some information does not exist, such as, the venue age and type, author award, author affiliation institute and venue impact. To solve it; first, LRAS automator used the search engines mentioned above and Google with advance search.

For the simulations, 2,000 scientific papers were used. The papers dealt with various research topics in Computer Science. Two sub-domains were chosen, each with 1,000 papers: (1) artificial intelligence and (2) information systems. In the context of these simulations, the sub-domains are treated as domains. Here, a scenario was defined as a set of two simulator runs, one on each domain dataset. For the simulator run parameters, the metadata of one paper in the dataset (discipline, language, title, topic, keywords and abstract) were used as the researcher selection parameters.

Two performance criteria were used to assess the relevancy of the papers for the researchers:

1) Accuracy: the percentage of true classifications
2) Precision: the percentage of the classified items that are relevant.

Considering the sets of relevant papers (REL) and non-relevant papers, (NREL), true relevant (TR) denotes the papers classified as REL when they really are, while false relevant (FR) denote the papers classified as REL when they are not. Thus, with the same logic, the papers classified as NREL can be true non-relevant (TN) or false non-relevant (FN). Accuracy (denoted by a) and precision (denoted by p) were computed as follows for each scenario:

$$a = \frac{TR + FR}{TR + FR + TN + FN} \qquad p = \frac{TR}{TR + FR}$$

To identify TR, FR, TN and FN for each scenario, a target paper was chosen for the domain; next, the metadata of this target paper were used as the researcher selection parameters and the references papers in the output set of the prototypes were compared to the cited papers of the target paper. Through this comparison, TR, FR, TN and FN were defined. Let $a_{i,j}$ be the accuracy of the scenario i^{th} of the dataset j and $p_{i,j}$ be the precision of the scenario i^{th} of the dataset j; the average accuracy (denoted by Avg_a_i) and the average precision (denoted by Avg_p_i) are defined as follows:

$$Avg\_a_i = \frac{\sum_{j=0}^{D} a_{i,j}}{D} \qquad Avg\_p_i = \frac{\sum_{j=0}^{D} p_{i,j}}{D}$$

where D denotes the number of datasets.

Fig. 4 shows the average accuracy for the three different scenarios (LRAS, ID3 and PTRA): the horizontal axis represents the sequence number of the simulation scenario and the vertical axis represents the average accuracy of the associated scenario. It is observed that LRAS (in red) performs better than ID3 (in green) and PTRA (in blue): LRAS has an average accuracy of 0.91 per scenario while ID3, has an average of 0.60 per scenario. The average relative improvement in accuracy (defined as [Avg_a of LRAS – Avg_a of ID3]) of LRAS in comparison to ID3 is 0.32 (32%) per scenario.

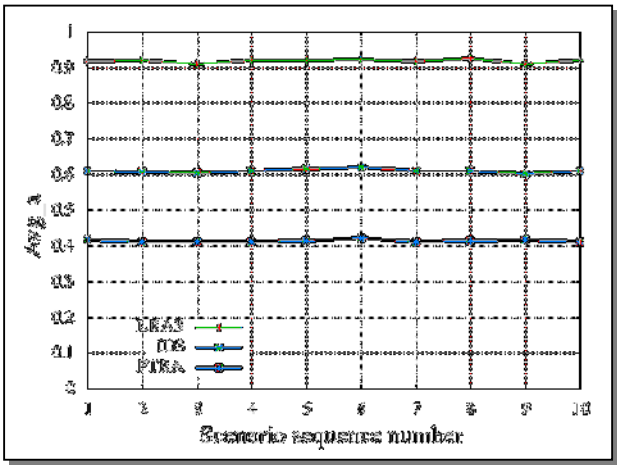

Fig. 4: Average accuracy Vs Scenario sequence number

Fig. 5 shows the average precision for the same scenarios of Fig. 4: the x Axis represents the simulations scenario sequence number while the Y axis represents the average precision of the associated scenario. LRAS performs better than ID3 and PTRA. LRAS produced an average precision of 0.96 per scenario while ID3, the best among the two works used for comparison, has an average of 0.65 per scenario. The average relative improvement in precision (defined as [Avg_p of LRAS – Avg_p of ID3]) of LRAS in comparison to ID3 is 0.31 (31%) per scenario.

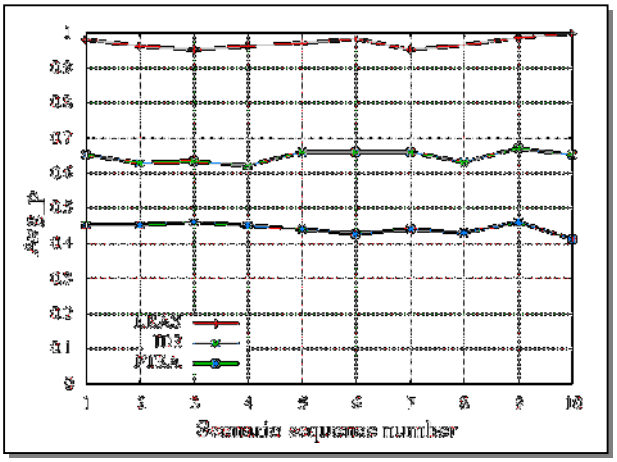

Fig. 5: Average precision Vs Scenario sequence number

V. CONCLUSION

In this paper, we have introduced a new scheme, which is called literature review assistant scheme (LRAS) for (1) ranking the relevance of scientific papers and (2) find the relevant papers that best match with the research topic, description and keywords of the researchers or students. More specifically, based on TDM technique, LRAS computed paper relevance index, called Dynamic Topic based Index

(*DTb Index*), taking into account (i) venues impact, (ii) authors and their affiliated institutes impact, (iii) key findings and citations impact and (iv) papers references impact. To select the papers for the literature review, LRAS used the LCR Index; LRAS computed the LCR Index based on TDM technique and using (i) the main topic of his research, (ii) description of his research, (iii) the title and (iv) the keywords of the paper that he plans to provide in the context of his research and for which he needs to make a literature review. The main contribution of LRAS search engine prototype is the fact that the algorithm takes into account the area of research.

We evaluated, via simulations, LRAS and compared it against two recent related schemes proposed in [6] and [11]. The simulation results demonstrated that LRAS achieved better accuracy and precision regardless of the sequence number of the simulation scenario. For example, in comparison to ID3 proposed in [11], LRAS yielded an average relative improvement in accuracy of 32% per scenario and an average relative improvement in precision of 31%. This superior performance might be attributable to the use of additional bibliometric metadata to evaluate the relevancy of papers.

REFERENCES

[1] I. MASIC, and E. BEGIC, "Evaluation of Scientific Journal Validity, It's Articles and Their Authors," *Stud Health Technol Inform.,* vol. 226, pp. 9-14, 2016.

[2] A. Okerson, "Text & Data Mining - A Librarian Overview," in 79th IFLA World Library and Information Congress, Singapore, Malaysia, 2013, pp. 1-6.

[3] J. Clark, "Text Mining and Scholarly Publishing," *Publishing Research Consortium*, 2013.

[4] M. Zhang, X. Zhang, and Y. Hu, "Ranking of Collaborative Research Teams Based on Social Network Analysis and Bibliometrics," in 12th International Conference on Cooperative Design, Visualization, and Engineering (CDVE), Mallorca, Spain, 2015, pp. 236-242.

[5] F. Madani, and C. Weber, "The evolution of patent mining: Applying bibliometrics analysis and keyword network analysis," *World Patent Information,* vol. 46, pp. 32-48, 9//, 2016.

[6] M. A. Hasson, S. F. Lu, and B. A. Hassoon, "Scientific Research Paper Ranking Algorithm PTRA: A Tradeoff between Time and Citation Network," *Applied Mechanics and Materials,* vol. 551, pp. 603-611, 2014.

[7] J. Beel, S. Langer, M. Genzmehr, B. Gipp, C. Breitinger, and A. Nurnberger, "Research paper recommender system evaluation: a quantitative literature survey," in International Workshop on Reproducibility and Replication in Recommender Systems Evaluation, Hong Kong, China, 2013, pp. 15-22.

[8] M. Cataldi, L. Di Caro, and C. Schifanella, "Ranking Researchers Through Collaboration Pattern Analysis," in European Conference on Machine Learning and Knowledge Discovery in Databases, Riva del Garda, Italy, 2016, pp. 50-54.

[9] F. Franceschini, D. Maisano, and L. Mastrogiacomo, "Influence of omitted citations on the bibliometric statistics of the major Manufacturing journals," *Scientometrics,* vol. 103, no. 3, pp. 1083-1122, 2015.

[10] S. Wang, S. Xie, X. Zhang, Z. Li, P. S. Yu, and X. Shu, "Future Influence Ranking of Scientific Literature," in Society for Industrial and Applied Mathematics (SIAM) International Conference on Data Mining, Philadelphia, Pennsylvania, USA, 2014, pp. 749-757.

[11] T. R. P. M. Rúbio, and C. A. S. J. Gulo, "Enhancing Academic Literature Review through Relevance Recommendation," in 11th Iberian Conference on Information Systems and Technologies, Gran Canaria, Canary Islands, Spain, 2016, pp. 70-75.

[12] W. Marx, and L. Bornmann, "Change of perspective: bibliometrics from the point of view of cited references—a literature overview on approaches to the evaluation of cited references in bibliometrics," *Scientometrics,* vol. 109, no. 2, pp. 1397-1415, 2016.

[13] Y. Dong, R. A. Johnson, and N. V. Chawla, "Can Scientific Impact Be Predicted?," *IEEE Transactions on Big Data,* vol. 2, no. 1, pp. 18-30, 2016.

[14] L. Bornmann, M. Stefaner, F. d. M. Anegón, and R. Mutz, "Ranking and mapping of universities and research-focused institutions worldwide based on highly-cited papers: A visualisation of results from multi-level models," *Online Information Review,* vol. 38, no. 1, pp. 43-58, 2014.

[15] M. Packalen, and J. Bhattacharya, "Neophilia Ranking of Scientific Journals," *National Bureau of Economic Research Working Paper Series,* vol. 21579, 2015.

[16] L. Bornmann, M. Stefaner, F. d. M. Anegón, and R. Mutz, "Ranking and mappping of universities and research-focused institutions worldwide: The third release of excellencemapping.net," *COLLNET Journal of Scientometrics and Information Management,* vol. 9, no. 1, pp. 65-72, 2015/01/02, 2015.

[17] X. Wan, and F. Liu, "WL-index: Leveraging citation mention number to quantify an individual's scientific impact," *Journal of the Association for Information Science and Technology,* vol. 65, no. 12, pp. 2330-1643, 2014.

[18] C. A. S. J. Gulo, T. R. P. M. Rubio, S. Tabassum, and S. G. D. Prado, "Mining Scientific Articles Powered by Machine Learning Techniques," *OASIcs-OpenAccess Series in Informatics,* vol. 49, pp. 21-28, 2015.

[19] P. Mayr, A. Scharnhorst, B. Larsen, P. Schaer, and P. Mutschke, "Bibliometric-Enhanced Information Retrieval," in 36th European Conference on IR Research (ECIR), Amsterdam, The Netherlands, 2014, pp. 798-801.

[20] T. Pedersen, S. Patwardhan, and J. Michelizzi, "WordNet::Similarity: measuring the relatedness of concepts," in Demonstration Papers at Human Language Technology conference/North American chapter of the Association for Computational Linguistics (HLT-NAACL), Boston, Massachusetts, USA, 2004, pp. 38-41.