A Scientometric Approach for Personalizing Research Paper Retrieval

Nedra Ibrahim, Anja Habacha Chaibi and Henda Ben Ghézala
RIADI Laboratory, ENSI, University of Manouba, Manouba, Tunisia

Keywords: Scientometric Indicators, Qualitative Search, Scientometric Annotation, Re-ranking, Similarity Score, User Profile, User Model.

Abstract: Scientific researchers are a special kind of users which know their objective. One of the challenges facing todays' researchers is how to find qualitative information that meets their needs. One potential method for assisting scientific researcher is to employ a personalized definition of quality to focus information search results. Scientific quality is measured by the mean of a set of scientometric indicators. This paper presents a personalized information retrieval approach based on scientometric indicators. The proposed approach includes a scientometric document annotator, a scientometric user model, a scientometric retrieval model and a scientometric ranking method. We discuss the feasibility of this approach by performing different experimentations on its different parts. The incorporation of scientometric indicators into the different parts of our approach has significantly improved retrieval performance which is rated for 41.66%. An important implication of this finding is the existence of correlation between research paper quality and paper relevance. The revelation of this correlation implies better retrieval performance.

1 INTRODUCTION

Current web search engines are built to serve all users, independent of the special needs of any individual user. When searching for scientific papers amongst the exponentially amount freely available, via bibliographic databases, it is becoming extremely difficult to find the best information that meets the researcher's requirements.

The researcher being the focus of the proposed approach, he aims to product a literature review or a scientific publication. From the online available information resources, when conducting an information search, he is facing a set of external factors. On the other hand, the information research must meet a set of requirements. The two main issues affecting researchers' search for information are the information overload and heterogeneity of information sources. In return, the researcher's scientific production should respond to his institution's qualitative requirements and have some quality indicator.

This paper discusses how a researcher creates his definition of quality that can be used to drive a specific information search. However, several practical questions arise when dealing with research paper retrieval: How to integrate the scientific quality into the personalized information retrieval (IR) process? Which quality elements should be integrated? At which level the quality should be integrated? What will be the contribution of quality integration? To answer all these questions, we propose a personalized retrieval system based on scientometric evaluation.

The remainder of the paper is organized as follows: Section 2 describes the existing approaches on personalized research papers' retrieval. Section 3 is devoted to present the proposed approach and the three modules of the system. In Section 4, the results of our experimentation will be discussed. Finally, Section 5 concludes with a summary.

2 PERSONALIZED RESEARCH PAPER RETRIEVAL

The web has greatly improved the access to scientific literature. The progress of science has often been hampered by the inefficiency of traditional methods of disseminating scientific

Ibrahim, N., Chaibi, A. and Ghézala, H.
A Scientometric Approach for Personalizing Research Paper Retrieval.
DOI: 10.5220/0006671204190428
In *Proceedings of the 20th International Conference on Enterprise Information Systems (ICEIS 2018)*, pages 419-428
ISBN: 978-989-758-298-1

information. We reviewed some personalized research paper's retrieving systems. We classified them into two categories: personalization of ranking and recommendation.

Singh et al. (2011) proposed ranking the research-papers based on citation network using a modified version of the PageRank algorithm (Plansangket and Gan, 2017). Tang et al. (2008) ranked authors on h-index and conferences' impact.

In research-paper recommendation, the Content-Based Filtering (CBF) was the predominant recommendation class. The majority utilized plain terms contained in the documents (Nascimento et al., 2011), others used n-grams, or topics based on Latent Dirichlet Allocation (LDA) (Beel et al., 2016). DLib9 (Machine Readable Digital Library) (Feyer et al., 2017) is a web-service that generates recommendations based on a single document. Moreover, it offers different recommendation approaches, such as stereotype-based and content-based algorithms with additional re-ranking using bibliometric data. Few approaches also utilized non-textual features, such as citations or authors. The CORE recommender (Knoth, 2015) uses collaborative filtering and content-based filtering. Another approach used co-citations to calculate document relatedness (Pohl et al., 2007). CiteSeer has a user profiling system which tracks the interests of users and recommends new citations and documents when they appear (Lawrence et al., 1999a). It used citations instead of words to find similar scientific articles. Some recommendation approaches built graphs to generate recommendations. Such graphs typically included papers that were connected via citations. Some graphs included authors, users/customers and publishing years of the papers (Huang et al., 2012).

However, in the previous studies little attention has been given to the user. In (Singh et al., 2011), research-paper ranking approach didn't take into account the user preferences. In (Tang et al., 2008), the authors focused on ranking authors or conferences according to one of the impact criteria, which cannot match all users' preferences. The majority of research paper recommendation approaches was a content based (Nascimento et al., 2011), (Feyer et al., 2017) and (Knoth, 2015). In which, the authors focused on extracting text from the title, abstract, introduction, keywords, bibliography, body text and social tags. Some other approaches used different information such as citation or authors (Pohl et al., 2007), (Lawrence et al., 1999a) and (Huang et al., 2012). The problem with these approaches is in that they did not allow users to define their preferences. In fact, they did not take into account that researcher satisfaction might depend not only on accuracy or citations.

3 PROPOSED SCIENTOMETRIC APPROACH FOR PERSONALIZED RESEARCH PAPER RETRIEVAL

The researcher tries to produce a scientific qualitative production according to the strategy of his research institution. To validate its scientific production, the researcher must meet a set of qualitative criteria such as:

- Having publications in impacted journals and / or classified conferences.
- Having publications with a specific number of citations.
- Having a certain number of publications.
- Citing qualitative references.
- Citing trusted authors (belonging to well-known affiliations with a certain number of publications and citations).

Thus, the researcher needs to initiate a qualitative research according to his qualitative preferences after choosing his own definition of quality. When using the online bibliographic databases, the researcher finds some difficulties such as:

- Which conference ranking system to choose?
- Which impact indicator to consider?
- Which bibliographic database to choose?
- How to manage differences between the different bibliographic databases?
- How to validate his choice?

The quality of the information source is very important for institution quality improvement and literature review validation. The proposed system should be a solution to the researchers' problematic when searching for relevant information. We propose a personalized IR system dedicated to researchers to automate and facilitate the selection of qualitative research papers. We integrated scientific quality in the process of research and personalization of the system. The challenges of the proposed system are:

- Collecting researcher's preferences.
- Synchronizing between different online bibliographic databases to extract quality indicators.

- Selecting the most significant quality indicators.
- Extracting good quality indicators.
- Updating the various indicators.

Figure 1 presents a description of the proposed system. The proposed system is composed of three basic modules: a scientometric retrieval system, a user profile management module and a personalized access to information module. The first module is the scientometric retrieval system which is based on a scientometric annotator. The second module is the user profile management module. We enriched the user profile model by scientometric indicators to build the scientometric profile ontology. The third module is the user profile exploitation for which we propose a scientometric approach for re-ranking research papers. In the following, we detail each of the three modules.

3.1 Quality Measurement

A scientific paper is considered to be an indicator of researchers' scientific production. The assessment of research papers can be performed by a set of quantitative and qualitative measures. Scientometrics is defined as all quantitative aspects of the science of science, communication science and science policy (Hood and Wilson, 2004). Ibrahim et al. (2015) studied all the elements affecting the research paper quality. Amongst the large set of scientometric indicators existing in the literature, Ibrahim et al. selected the most ones reflecting the real paper impact. They showed that we can assess paper quality by combining a set of scientometric indicators which include: publications number, citations number, h-index, journal impact factor and conference ranking.

The scientometric indicators have been used by bibliographic databases, such as Science Citation Index (SCI) (Alireza, 2005), Google Scholar

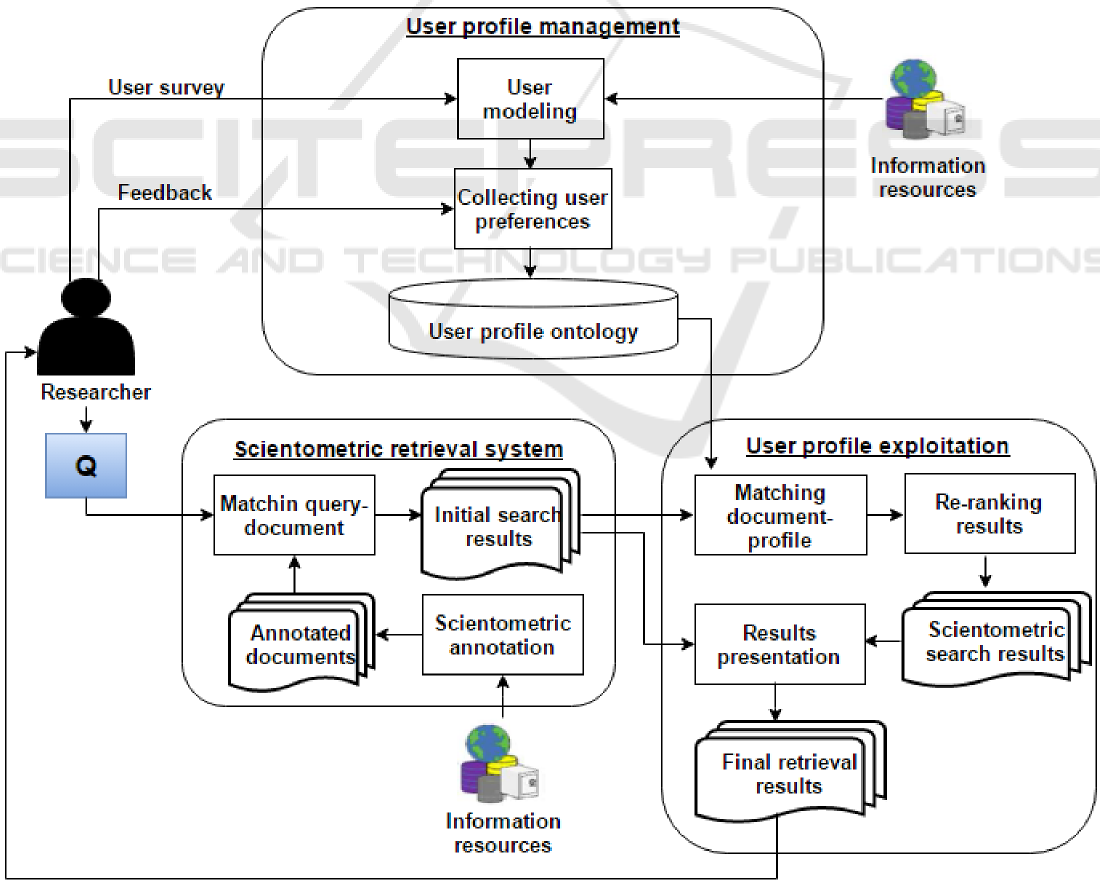

Figure 1: Proposed scientometric approach.

(Lawrence et al., 1999b), CiteSeer (Harzing, 2011) and Microsoft Academic Search[1]. Also, we note the existing of several ranking systems providing scientific journal ranking and conference ranking according to their impact. Thomson ISI annually publishes the Journal Citation Report (JCR[2]) which includes a number of indicators among which the Journal Impact Factor (JIF). The portal of the Association Core[3] provides access to the logs of journal and conference classification. The SCImago Journal & Country Ranking portal (SJR[4]) provides a set of journal classification metrics and quality evaluation.

3.2 Scientometric Retrieval System

To improve search results, we propose the application of scientometrics in the IR process. In this section, we specify how to integrate scientometrics at the indexing level.

We propose a scientometric annotator which is an automatic process. It allows the extraction of relevant indicators to each document from the online bibliographic databases.

A document can be a conference or a journal paper, thesis or master report. Amongst the large set of scientometric indicators existing in the literature, we selected the most ones reflecting the real paper impact.

We used the selected indicators to annotate research papers. Scientometric annotation is author-centered, document-centered, and venue-centered. It consists on representing and using a set of scientometric indicators:

- The impact of the author as an indicator of the researcher quality.
- The impact of the journal/conference as an indicator of the container quality.
- The impact of the research group as an indicator of the search environment quality.
- The impact of the paper as an indicator of the content quality.

The 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 affiliation. The back contains the references. We annotate research papers from online databases.

The annotation process consists of three data processing steps. The first step is the pre-treatment. It consisted on the construction of descriptive annotation from an online paper. The second step is the indicators' extraction. It consists on the extraction of the scientometric indicators corresponding to each document from the online database. The third step is the enrichment and the reconstruction of the Extensible Markup Language (XML) annotation file. It consists on the enrichment with the scientometric annotation and the reconstruction of the XML annotation file. The annotation file included the descriptive and scientometric annotations. Figure 2 gives an example of the produced XML annotation file.

The main limitations of the annotation process are:

- The diversity of information resources: we note the existence of several online bibliographic databases providing a large number of papers. In order to solve this problem, we have chosen the bibliographic database which provides the widest range of scientometric indicators.
- Updating scientometric indicators: after the annotation of the document, we must start a continuous updating process.
- The diversity of scientometric indicators: a single paper may have different values representing the same scientometric indicator in different bibliographic databases. To solve this problem, we propose a synchronization module. The synchronization consists on choosing the most recent value.

3.3 User Profile Management

Personalization aims to facilitate the expression of user needs and enables him/her to obtain relevant information. The user profile management module consists on the definition of a scientometric user model. Based on this model, we collect the user preferences to construct the user profile ontology.

We proposed a scientometric user profile model in which we integrated the dimension: "scientometric preferences". This dimension represents the researchers' needs by incorporating different scientometric indicators to the user profile. The profile model is an instantiation of the generic model described in the work of Ibrahim et al. (2016).

[1] www.academic.research.microsoft.com/
[2] Thomson, R. (2017), Journal Citation Reports® Science Edition.
[3] www.portal.core.edu.au/conf-ranks/
[4] www.scimagojr.com/index.php

```xml
<Publication id="18">
  <DescriptiveAnnotation>
    <Title>An Information Retrieval Approach For Automatically Constructing Software Libraries </Title>
    <Author Name="Yoëlle S. Maarek" Affiliation="Yahoo Research Labs" />
    <ArticleContainer Type="Journal" Name="IEEE Transactions on Software Engineering - TSE" Info=", vol
    <Abstract>...a technology for automatically assembling large software libraries which promote softw
  </DescriptiveAnnotation>
  <scientometricAnnotation>
    <bodyLevel>
      <docCitationNumber> 308</docCitationNumber>
      <coAuthorNumber>3</coAuthorNumber>
    </bodyLevel>
    <authorLevel>
      <authorName>Yoëlle S. Maarek</authorName>
      <authorPosition>1</authorPosition>
      <authorPublications> 61 </authorPublications>
      <authorCitations> 1294 </authorCitations>
    </authorLevel>
    <journalLevel>
      <journalPublicationNumber> 3,916</journalPublicationNumber>
      <journalCitationNumber> 156,846</journalCitationNumber>
      <journalSelfCitationNumber> 5903</journalSelfCitationNumber>
      <journalImpactFactor>A*</journalImpactFactor>
    </journalLevel>
    <affiliationLevel>
      <affiliationHindex>Yahoo Research Labs</affiliationHindex>
      <groupCitations> 9,953</groupCitations>
      <groupPublications> 163,564
      </groupPublications>
    </affiliationLevel>
  </scientometricAnnotation>
</Publication>
```


Figure 2: Example of XML annotation file.

We performed a user study to select the indicators that interest the researchers. The selected indicators were incorporated into the user profile model. It stores the necessary information describing the quality of a research paper according to the researcher's needs. These preferences are organized into five SubDimensions which are the different entities affecting the paper's quality. The quality of each entity is measured by a set of scientometric indicators which represent the attributes of each SubDimension:

- Author quality: is measured by the mean of four attributes (h-index, citations number, publications number and author position).
- Content quality: is measured by the mean of the paper citations number and the co-authors number.
- Journal or conference quality: scientific journals or conferences are containers of research papers. A good quality of the journal promotes the selection of the document. The quality of the paper container is evaluated by its ranking, number of citations, number of publications and number of self-citations.
- Affiliation quality: we consider the quality of author's affiliation measured by the group h-index, the number of publications, the number of citations and the number of self-citations.

On the other hand, each SubDimension is extended on ExtSubDimension by moving to a higher level of abstraction. Each ExtSubDimension will be organized into attributes which represent the scientometric indicators measuring its quality:

- Career quality: We associate the quality of career to the author quality as an extension. The quality of author career is measured by the number of years spent by the author on research in a specific discipline, and his current title.
- Source quality: We designate by the source of scientific documents the bibliographic databases such as: Google Scholar, DBLP and MS Academic Search. The quality of information source is measured by the number of publications, the interval of time and the number of domains covered by the source.
- Publisher quality: the quality of the container can be extended to the evaluation of publisher quality which can affect the quality of papers.

This latter is measured by the number of specialties, the number of published journals or conferences.

- Organization quality: we extended the affiliation quality to the organization quality measured by the Shanghai ranking (in the case of academic organizations), the number of publications and the number of citations.
- Association quality: For each conference, we join his association (eg. IEEE). The quality of conference association is measured by the number of specialties covered by the association and the number of conferences organized by the association.

The proposed user profile is based on an implicit and an explicit interaction with the user. Collecting user preferences is based on the user navigation to measure his interest to a given entity. We collect user preferences from the number of pages the user reads, user's interaction with the papers (downloads, edits, views) and citations. Otherwise, the interactions are explicit because we ask the unknown user to define his quality preferences according to a set of scientometric preferences.

Based on the user preferences, we construct the user profile ontology. The profiles are containers of knowledge about the user. We opted for ontology to represent the scientometric preferences of the user. The ontology domain covers the scientometric domain (assessment tools, measures and indicators) conducted for a scientific research evaluation. In Figure 3, we present a portion of the proposed user profile ontology graph.

3.4 User Profile Exploitation

The proposed personalization approach is based on the exploitation of the user profile to re-rank documents according to the user preferences. We proposed a scientometric re-ranking approach based on users' quality preferences. We define a scientometric score based on scientometric indicators deriving from user profile. This score is used to re-rank search results and to deliver qualitative information at the top ranks.

For each of the returned results (Ai), we calculate its similarity to the user profile. Then, we re-rank the search results according to the similarity score. We propose a scientometric score as a combination of the scientometric indicators of the user model. We calculate the scientometric score which we note as

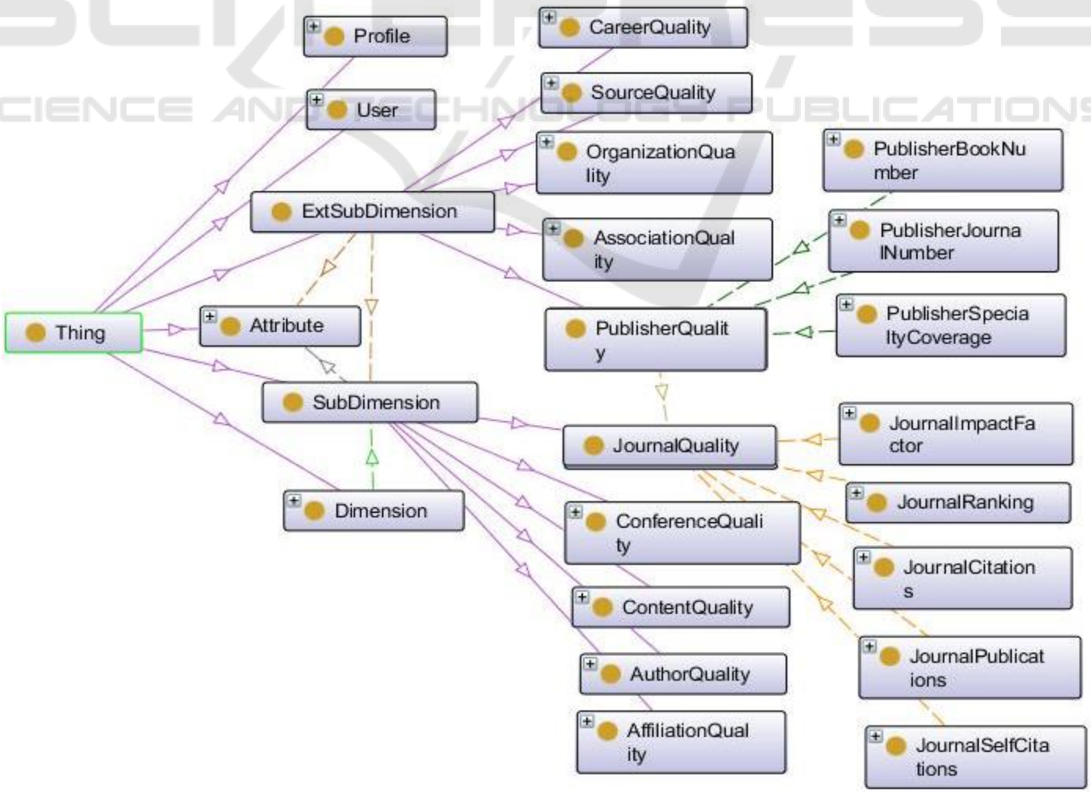

Figure 3: Portion of the user profile ontology graph.

Q. This scientometric score was the result of the application of an adapted mathematical model of weighted sums considering the scientometric preferences of the user. The equation that describes the proposed scientometric score is as follows:

$$Q(A_i) = \sum_{j=1}^{n} W_{SUB} * Q_{SUB}(A_i) + \sum_{j=1}^{n} (W_{SUB} + W_{EXT}) * Q_{EXT}(A_i), \quad \forall i \in [1, m] \tag{1}$$

Q_{SUB} and Q_{EXT} represent respectively the quality of each SubDimension and ExtSubDimension. W_{SUB} and W_{EXT} are the importance weights attributed by the user to each SubDimension and ExtSubDimension.

We calculate the scientometric rank based on the scientometric score. Then, we determine the final rank based on the initial rank and the scientometric rank. Equation (2) represents the formula of the final rank:

$$\text{FinalRank} = \alpha * \text{InitialRank} + (1 - \alpha) * \text{ScientometricRank}, \alpha \in [0,1] \tag{2}$$

The initial rank is the original rank returned by the retrieval system and the scientometric rank is calculated according to scientometric score.

4 EXPERIMENTATION AND EVALUATION

We performed different experimentations to evaluate the three system modules.

4.1 Evaluation of the Scientometric Retrieval

To evaluate the scientometric retrieval system, we propose a multi-model retrieval system. It consists of a scientometric annotator and several retrieval models that operate this annotator. These models differ by the criteria considered when matching the document to the query:

- Classic: is a classical retrieval model based on the similarity between a document and a search query; referred to as the term frequency (tf).
- Sciento1: the first scientometric model. It is based on the similarity between document and query in addition to the container ranking.
- Sciento2: the second scientometric model. It is based on the similarity between document and query in addition to the documents citation number.
- Sciento3: the third scientometric model. It is based on the similarity between document and query in addition to both container ranking and documents citation number.

In Classic, we have not integrated scientometrics. We integrated scientometrics into the three other models. We evaluated and compared the performance of the two retrieval categories based on a test collection and different evaluation measures. The test collection contains 1500 annotated research papers and 30 different queries. The annotation files are the result of the annotation of 1500 published papers extracted from MS Academic Search.

This evaluation is carried out to find out the effect of the integration of scientometrics on the performance of retrieval systems. Thus, we are interested to the comparison between classical retrieval models and scientometric retrieval ones. In order to verify the validity of scientometric retrieval models, we carried out several experiments. Fig. 4 and Fig. 5 show a recapitulation of the results of the performed experimentations. The results show that all the scientometric models performed an improvement in performance. This improvement is proved by the F-measure and Mean Average Precision (MAP) variations. Sciento3 realized the best improvement in F-measure which is rated for 41.66%. Sciento1 and Sciento2 realized an improvement in F-measure which is respectively rated for 33.33% and 30.55%. We note a best rate of MAP improvement is realized by Sciento3 which is rated for 14.03%. Sciento1 and Sciento2 realized an improvement in MAP rated for 5.26%.

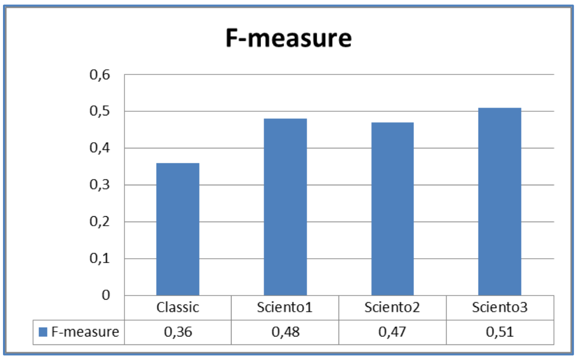

Figure 4: F-measure variation.

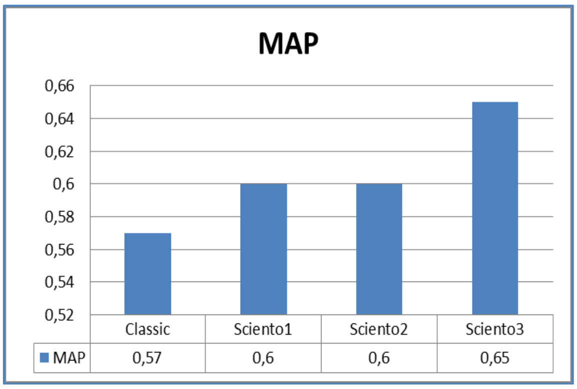

Figure 5: MAP variation.

It has been found that scientometrics has enhanced the relevance of results and has provided better performance to the retrieval system. The best performance is provided by Sciento3, in which both the number of document citations and container ranking were integrated.

4.2 User Profile Ontology Validation

To test the profile ontology, we used the Pellet reasoned available directly from PROTÉGÉ (Sirin et al., 2007). Pellet is a complete and capable OWL-DL reasoner with very good performance (Sirin et al., 2007). It has user defined data types, and debugging support for ontologies. We describe three tests provided by Pellet: consistency test, classification test and queries test.

- Consistency test: is made based on the class description, which ensures that ontology does not contain any contradictory facts. A class is considered inconsistent if it cannot have any instance. Inferred class hierarchy after invoking the reasoned showed that all classes are consistent.
- Classification test: can check whether a class is a subclass of another class or not. It computes the subclass relations between every named class to create the complete class hierarchy. The classification test shows that no suggestion has been produced by the reasoner Pellet and that "Asserted hierarchy" and "Inferred hierarchy" are identical. This indicates the validity of the ontology classification.
- Queries test: PROTÉGÉ allows querying the project and locating all instances that match the specified criteria. Queries are a way to identify the instances in the project, based on class and slot properties. To validate the queries test, we have created different queries using SPARQL (Pérez et al., 2009) tool.

4.3 Evaluation of the Scientometric Re-ranking

Our objective is to evaluate the proposed scientometric re-ranking algorithm among an initial ranking. We produce the personalized results and compare it to initial ones. We used the nDCGp (Jurafsky and Martin, 2008) as a measure of ranking performance. We performed the evaluation based on users' database containing 171 researchers working in our research laboratory (20 known users and 151 unknown users). We collected the user's scientometric preferences by launching a survey. We opted for the bibliographic database "MS Academic Search" to extract the initial ranking and the corresponding scientometric data. Our choice is justified by the broad set of scientometric indicators covered by MS Academic Search. We used keywords based queries to perform the experimentations. All the known users executed 30 queries on the MS Academic Search.

We consider the initial rank corresponding to the top hundred results returned by MS Academic Search. Then, were-rank top hundred initial results according to the scientometric score. Finally, we calculate nDCGp for the initial ranked list and the scientometric ranked list to compare between them. By considering the mean nDCGp of the obtained results, we observe that scientometric rank realized an improvement in performance. The improvement was rated for 14.75% compared to the MS Academic Search ranking.

4.4 Significance Test

A significance test allows the researcher to detect significant improvements even when the improvements are small. We want to promote retrieval models that truly are better rather than methods that by chance performed better. We opted for performing significance test to validate our experimentation on IR models. It turned out that several significance tests exist in the literature. An important question then is: what statistical significance test should IR researchers use?

Smucker et al. (2007) experimented the different significance tests on IR. They discovered that Student t-test have a good ability to detect significance in IR. The t-test is only applicable for measuring the significance of the difference between means. Student t-test consists of the following essential ingredients:

- A test statistic or criterion: IR researchers commonly use the difference in MAP or the difference in another IR metric.
- A null hypothesis: is that there is no difference in the two compared systems.
- A significance level: is computed by taking the value of the test statistic for the experimental systems. Then, determining how likely a value that larger could have occurred under the null hypothesis. This probability is known as the p-value. According to the p-value we distinguish three levels of significance. Low significance when $p \leq 0.1$. High significance when $p \leq 0.05$. Very high significance when $p \leq 0.01$.

As is measured by mean average precision, scientometric retrieval models (Sciento1, Sciento2, and Sciento3) performed an improvement rated for (5.26%, 5.26% and 14.03%) compared to the classical model. However, is this statistically significant improvement? The executed experimentations produced MAPs of 0.57 for classical retrieval model, 0.6 for both Sciento1 and Sciento2 and 0.65 for Sciento3. The differences in MAP are between 0.05 and 0.08. In order to test the significance of the difference in MAP performance, we used student t-test. We report the results in Table 1.

Table 1: Student T-test on MAP.

	Classic vs. Sciento1	Classic vs. Sciento2	Classic vs. Sciento3
p-value	0,003338	0,000269	0,000731

We consider the high significance level ($p \leq 0.05$) to interpret our results. Table 1 summarizes the results corresponding to the student t-test performed on our different retrieval models. The p-values correspond to the difference between classical retrieval model and respectively Sciento1, Sciento2 and Sciento3. The difference in MAP performance between the three pairs is significant at $p \leq 0.05$.

Given the obtained results, we can validate our experimentations. We approved the difference in performance between the scientometric retrieval models and the classical retrieval model.

5 CONCLUSION AND FUTURE WORK

In this paper, we focused on the research paper retrieval. This field essentially interests researchers which aim to produce qualitative papers. Researchers are interested to the information quality. The research paper's impact is measured by the means of scientometric indicators. We demonstrated that quality of research paper can be measured by a combination of scientometric indicators.

The researchers are using the online bibliographic databases to perform their IR. They are facing several difficulties when searching for relevant papers. To resolve these difficulties, we proposed a personalized retrieval system dedicated to researchers. To respond to the researchers' needs, we integrated the quality into the three modules of the system. We proposed a scientometric annotator which was the base of the retrieval system. For the retrieval personalization, we proposed a profile management module and a module to personalize access to information. The user profile management module consisted on user modeling and profile ontology construction. The personalized access to information consists on re-ranking search results according to the user preferences.

To validate the proposed approach, we performed an evaluation of the different system's modules. From the research that has been performed, it is possible to conclude that the integration of scientometrics enhanced the performance of the different modules. We approved the significance of our results by performing a student t-test. Summing up the results, it can be concluded that the application of scientometrics in the IR process was an effective way to improve search results.

In our future research we intend to concentrate on the time factor by considering the publication year of the papers. The next stage of our research will be the experimentation on other samples and the consideration of other research disciplines such as medicine and bio-medications. Then, we will study the effect of varying disciplines on the results.

REFERENCES

Alireza, N., 2005. Google Scholar: The New Generation of Citation Indexes. *Libri, International Journal of Libraries and Information Studies*, 55(4), pp. 170–180.

Beel, J., Gipp, B., Langer, S., Breitinger, C., 2016. Research-paper recommender systems: a literature survey. *International Journal on Digital Libraries*, 17(4), pp. 305-338.

Feyer, S., Siebert, S., Gipp, B., Aizawa, A., Beel, J., 2017. Integration of the Scientific Recommender System Mr. DLib into the Reference Manager JabRef. In: *European Conference on Information Retrieval*, Springer, Cham, pp. 770-774.

Harzing, A., 2011. *The publish or perish book: your guide to effective and responsible citation analysis*, Tarma software research. Australia.

Hood, W., Wilson, C., 2004. The literature of bibliometrics, scientometrics, and informetrics. *Scientometrics*. 52(2), pp. 291-314.

Huang, W., Kataria, S., Caragea, C., Mitra, P., Giles, C. L., Rokach, L., 2012. Recommending citations: translating papers into references. In *CIKM'12, the 21st ACM international conference on Information and knowledge managemen*, Maui, Hawaii, USA, pp. 1910–1914.

Ibrahim, N., Habacha Chaibi, A., Ben Ahmed M., 2015. New scientometric indicator for the qualitative evaluation of scientific production. *New Library World Journal*. 116(11/12), pp. 661-676.

Ibrahim, N., Habacha Chaibi, A., Ben Ghézala, H., 2016. A new Scientometric Dimension for User Profile. In *ACHI'16, the 9th International Conference on Advances in Computer-Human Interactions*, Venice, Italy, pp. 261-267.

Jurafsky, D., Martin, J. H., 2008. *Speech and language processing: an introduction to natural language processing*, Prentice Hall.

Knoth, P., 2015. *Linking Textual Resources to Support Information Discovery*. PhD thesis, The Open University.

Lawrence, S., Bollacker, K., Giles, C. L., 1999a. Indexing and retrieval of scientific literature. In *CIKM'99, the eighth ACM international conference on Information and knowledge management*, Kansas City, Missouri, USA, pp. 139-146.

Lawrence, S., Lee, C. G., and Bollacker, K., 1999b. Digital libraries and autonomous citation indexing. *Computer*, 32(6), pp. 67-71.

Nascimento, C., Laender, A. H., da Silva, A.S., Gonçalves, M.A., 2011. A source independent framework for research paper recommendation. In *JCDL'11, the 11th annual international ACM/IEEE joint conference on Digital libraries*, Ottawa, Ontario, Canada, pp. 297–306.

Pérez, J., Arenas, M., and Gutierrez, C., 2009. Semantics and complexity of SPARQL. *ACM Transactions on Database Systems (TODS)*, 34(3), pp. 16-43.

Pohl, S., Radlinski, F., Joachims, T., 2007. Recommending related papers based on digital library access records. In *JCDL'07, the 7th ACM/IEEE-CS joint conference on Digital libraries*, Vancouver, BC, Canada, pp. 417–418.

Plansangket, S., Gan, J., Q., 2017. Re-ranking Google search returned web documents using document classification scores. *Artificial Intelligence Research*, 6(1), pp. 59-68.

Singh, A., P., Shubhankar, K., Pudi, V., 2011. An efficient algorithm for ranking research papers based on citation network. In *DMO'11, the 3rd IEEE Conference on Data Mining and Optimization*, Putrajaya, Malaysia, pp. 88-95.

Sirin, E., Parsia, B., Grau, B., Kalyanpur, A., and Katz, Y., 2007. Pellet: A practical OWL-DL reasoner. Web Semantics: Science, Services and Agents on the World Wide Web. *Software Engineering and the Semantic Web*, 5(2), pp. 51-53.

Smucker, M. D., Allan, J., and Carterette, B., 2007. A comparison of statistical significance tests for information retrieval evaluation. In *CIKM'07, the sixteenth ACM conference on information and knowledge management*, Lisbon, Portugal, pp. 623-632.

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z., 2008. Arnetminer: Extraction and Mining of Academic Social Networks. In *KDD'08, 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Las Vegas, Nevada, USA, pp. 990-998.