Enhancing Academic Literature Review through Relevance Recommendation

Using Bibliometric and Text-based Features for Classification

Thiago R.P.M. Rúbio
LIACC – Artificial Intelligence and Computer Science Laboratory
DEI - Faculty of Engineering, University of Porto
Porto, Portugal
reis.thiago@fe.up.pt

Carlos A. S. J. Gulo
PIXEL Research Group, UNEMAT Brazil
Faculty of Engineering, University of Porto
Porto, Portugal
sander@unemat.br

Abstract — The growing number of scientific publications and the availability of information in online repositories enable researchers to discover, analyze and maintain an updated state of the art bibliography. Indeed, few works explore this scenario in order to support researchers on the literature review step. Literature reviewing comprises a fundamental part of the scientific writing, in which publications are evaluated and selected by relevance. Different approaches for relevance are possible, whether a more qualitative (semantic) approach with text-based techniques either more quantitative (numerical) approaches that use article’s metadata, such as bibliometric measures. Bibliometrics provide direct evidences of relevance and could represent good attributes for automatic classification. Our insight is that if a bibliometric-based cannot outperform text-based approaches, a hybrid model using both could benefit from it enhancing the classification performance (in terms of accuracy, precision and recall). In this paper we presented a novel approach, using Machine Learning (ML), namely the ID3 algorithm for a classification model that learn from specialist annotated data and recommend relevant papers for a specific research. Experiments showed good results on learning performance when using a hybrid approach, increasing testing performance in 12%, achieving 89.05% in accuracy when classifying a paper as relevant.

Keywords – Systematic Literature Review (SLR); Machine Learning; Classification; Text Mining; Bibliometric

I. INTRODUCTION

In recent years, the crescent volume of scientific publications available is becoming a problem for researchers, who, unable to exploit the whole literature in a specific field tend to follow ad-hoc approaches [1]. Representing the strategy of not having a strategy, ad-hoc comprises the traditional way of literature reviewing, in which authors merge papers without clear criteria about the representativeness of the selected set. In fact, this scenario is worsen when considering the growth on web-based scientific repositories, that make searching a large number of publications even easier without assuring relevant content. In the opposite side, there is the Systematic Literature Review (SLR) [2], a methodological approach for selecting and evaluating papers in a research. In SLR, authors try to create metrics that will result on selecting the most relevant set of papers. Indeed, characterizing article’s relevance is a very important study field and could help all researchers to enhance their research.

Popular scientific repositories, such as ACM Portal, Engineering Village, IEEE Xplore, Science Direct, ArXiv, CiteSeer, and Web of Science recommend scientific articles based on the specific keyword occurrence, publication date or citations order. We believe that those features are somehow connected to paper’s relevance, although they are not sufficient to characterize an impact on the subject in order to be selected for the bibliography [3, 4]. In this sense, the discussion about relevance guide us in two different ways: firstly, a deeper, qualitative relevance meaning that would include the semantics for searching. When authors search for published works, they are generally looking for the same domain, application or problem, trying to reach methods and techniques that could apply in their specific case. On the other hand, a stricter, quantitative sense of relevance could be directly inferred from quality metrics, such as the number of citations or even author-related as h-index or conferences qualis.

References [5, 6] presented an approach which produces automatic article-based recommendations using a text mining classification technique, fitting in the more qualitative relevance approach discussed. Nevertheless, the work did not include important bibliometric measures, such as download totals or number of citations. We believe that a similar strategy, using such information for recommendation could present better results since these metrics are direct evidence of paper’s quantitative relevance. In the event that this strategy could not outperform the text mining classification approach aforementioned, at least combining the two models would enhance the performance of the system. In order to validate the hypothesis, we propose recommending papers based on known classification models including the paper’s content and bibliometric features. Combining text mining efforts and bibliometric measures to automatically classify the relevant papers could provide a new perspective for every researcher,
making it easier to analyze the state of the art and to improve research itself. To validate the performance of the classification, we propose a comparison with the qualitative model presented in [5, 6] using a dataset manually annotated by a domain specialist.

The rest of the paper is organized as follows: related works are reviewed in Section 2; an overview of the proposed solution is presented in Section 3 including a detailed description of the recommendation system; the performed experiments applying the selected classification algorithms on a data set of scientific articles and the obtained results are discussed in Section 4, which is followed by concluding remarks in Section 5.

II. BACKGROUND AND RELATED WORKS

Literature review aims to synthesize the best-available empirical evidence to answer specific research questions, and also collecting information about studies, researchers, and hypothesis to identify the gaps and challenges in that field of study. Considered one of the most important tasks to locate, gather, appraise, and make use of state of the art, the literature review is a fundamental piece for the scientific writing. In fact, taking the common IMRaD (Introduction, Methods, Results and Discussion) structure for scientific writing, the literature review is generally placed on the introduction section, presenting a deep collection of related works that contribute for the fundamental bases of the research [7].

Relevance should depend on many other factors such as adequacy to the theme, specific tools used or even the test strategy, making the automatic recommendation of articles a challenging problem. There are some previous efforts to address this problem in the many stages of a recommendation system. Text processing has been considered in a fashion alternative using Natural Language Processing (NLP) approaches. Full-text processing enables the extraction of argumentative and rhetorical elements of scientific discourse, such as results and methodological discussions. Apache OpenNLP, Stanford CoreNLP, MALLET, GATE and CiteSpace [8, 9] are just some tools provided for text processing using NLP.

Information retrieval techniques enable to gain insights into a topic by contextualizing a search query, and can be applied for scholarly communication and related bibliographic databases. Mayr et al. [1] proposed a new approach to utilize bibliometric-enhanced information retrieval systems as recommendation service within a scholarly information retrieval environment. However, they considered a lot of research effort needs to be done to make progress in this theme in order to use the dynamic features in scholarly information retrieval.

Topic modelling are statistical methods, essentially used to analyze the words of the original documents to discover the topics that run through them and how those topics are connected to each other; moreover enables analyzing text of a large document collection and labelling correlations during the classification process. Blei and Lafferty [10], proposed an approach to automatically discover the topics from a collection of documents using latent Dirichlet Allocation (LDA) and probabilistic models.

On the other hand, bibliometric features are publications metadata that are related somehow to article’s relevance. De Bellis describes the importance of the citations on paper’s quality [3]. Other studies, as in [11] analyze academic writings and the impacts of the metadata on the indices calculated for papers or in other instances [12], authors’ relevance and country influence on paper visibility. Finally, [13] showed a strong correlation between the number of authors and the quality papers. As seen, publications metadata are often presented as numerical values and can be composed and associated to relevance through statistical methods or in more advanced scenarios, using machine learning techniques.

Many statistical algorithms can be used to create a model for classifying or labelling, such as [14, 15]: Latent Semantic Analysis (LSA) Language Model, Gaussian Model, and Bayesian Model, among others. Various techniques are used: Support Vector Machine (SVM), Naive Bayes classifier, K-Nearest Neighbor (K-NN), Rocchio Algorithm, Decision Trees, Ensemble Classifiers, and Inductive Logic Programming (ILP).

Reference [16] presented an approach to provide an automatic ranking system based on contents, combining semantic relevance, journal relevance and citation relevance. The references were retrieved using specialists’ sentences as content in order to ranking relevant references. Authors applied semantic, journal, and citation relevance as metrics for ranking the most important papers. In fact, this brief review highlights the current lack of efforts on using metadata to classify scientific papers. Next section presents our approach to select relevant works using bibliometric features.

III. A BIBLIOMETRIC RECOMMENDER SYSTEM FOR RELEVANT ARTICLES

Data-Mining (DM) techniques represent the next step on data analysis. In practice, Machine Learning (ML) algorithms extend the capabilities of humans when analyzing big volumes of data, both on number of records and number of attributes (complexity). It is practically impossible to perform some tasks, as in the literature review case, in order to evaluate the whole universe of data.

With that in mind, we followed a traditional DM methodology, also known as Knowledge Data Discovery (KDD). In Figure 1 KDD steps are explained aiming to our classification objective. First, data is retrieved from the literature repositories, then in a pre-processing step, outliers and dirty data is removed. The meaningful features are also selected as attributes for the ML algorithm. At this point, the data is presented to a Specialist that annotates the data, expressing a manual classification. Thus, the Data Mining step comprises dividing data in two subsets: training, passed to the algorithm to create the learning model and test set for evaluating the performance of the recommendation.
1) Data Set and Information Retrieval
Retrieving data from literature repositories is a difficult task. Although the publishers enable access to articles’ metadata and sometimes to full text, most of them restrict data collection or full access to the relevant metadata. In fact, even tools such as Google Scholar [16, 1] and other indexers do not make available many information. A complete literature retriever that crawls different repositories and creates a complete view of the literature given a query (relevant keywords for a research) is still missing. Thus, data collection comprises a manual process. In this work, we used the author’s domain knowledge to collect and process data from the IEEE Xplore API, the only platform that enabled to gather relevant paper’s metadata, namely: year of publication, citation number, references number and type of publication (journal, conference, etc.). Following we discuss how this information can be related to paper’s relevance.

2) Relevance attributes
As seen in [17, 18, 19], publication’s metadata are some of the main factors that will prescribe its relevance. In their work, [5, 6] text features were considered, such as title, abstract and full text. From the relevance characterization in Section 1, the quantitative relevance can be observed from the metadata, justified as the filtering criteria researchers often use to classify works as interesting literature for their own. Table 1 shows the main metadata features used to define relevant publications and a brief discussion about how they can imply in work’s relevance. In this work, only bibliometric features were considered, since they are a direct observation of the work influence on the scientific society.

3) Classification Algorithms
Using the previously explained attributes, the classification task can be performed by a large range of techniques, such as lazy modelling, bayesian algorithms, tree and rule induction or even neural networks. In order to assure the best performance, we have compared a range of algorithms, evaluating training and testing performances regarding accuracy, precision, and recall, which are known performance metrics for this task. Due to space restrictions, we abstain to show in Table 2 the results obtained with a set of selected algorithms, namely Decision Trees, Naïve Bayes, KNN, ID3 and Rule Induction. We used the Rapidminer tool for the process, using framework’s implemented algorithms. Results indicate that the ID3 algorithm outperformed the others with higher values in all metrics. With this process we have selected ID3 as the classifier to be used on our model (and experiments).

### Table 1. Impact of Bibliometric Attributes on Relevance

<table>
<thead>
<tr>
<th>Bibliometric attribute</th>
<th>Impact on Relevance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year of publication</td>
<td>Represents the energy of the publication. Recent publications have more updated state of the art considerations. Although some old references have deeper impact (books for example), the year is considered a measure of relevance.</td>
</tr>
<tr>
<td>Citation Number</td>
<td>Direct effect of the relevance: represents the fame of the paper. More relevant papers tend to be cited frequently.</td>
</tr>
<tr>
<td>References Number</td>
<td>A paper that has good references is, usually a more detailed work, with a more complex organization of the ideas. Thus, the references number could be observed a consistent measure of relevance, since the references are also linking edges to other relevant papers.</td>
</tr>
<tr>
<td>Media Type</td>
<td>Represents the target-group that will find this work relevant. For instance, in scientific research conference and journal papers tend to be more relevant than grey literature.</td>
</tr>
</tbody>
</table>

4) Performance Metrics
The quality of the results, i.e. the performance of the algorithm can be determined in terms of the metrics used to evaluate the classification. Thus, three main performance metrics can be analyzed for classification: Accuracy (1), Precision (2) and Recall (3). Summarzing, accuracy is related to the percentage of true classifications, precision is the percentage of the classified items that are relevant and recall the percentage of relevant items that were classified as it. Considering POS the set of relevant papers and NEG the non-relevant, true positives (TP) and false positives (FP) regards the records classified as POS when they really are and are not, respectively. Thus, with the same logic the papers classified as NEG can be true negatives (TN) and false negatives (FN). The metrics can be described in terms of this definitions, as seen in the following equations.

\[
\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)
\]
\[
\text{Precision} = \frac{TP}{TP+FP} \quad (2)
\]
\[
\text{Recall} = \frac{TP}{TP+FN} \quad (3)
\]
even neural networks. In order to assure the best performance, the framework’s implemented algorithms. Results indicate that the

We used the Rapidminer tool for the process, using

The quality of the results, i.e., the performance of the classification using the bibliometric attribute which are known performance metrics for this task. The selected ID3 as the classifier algorithm, but we wanted also to evaluate the performance of the ID3 using the bibliometric measures. In this sense, we have prepared three experiments in order to compare the performance of the classification using the strategy of [6] (Experiment 1), our strategy (Experiment 2), and a hybrid strategy combining both (Experiment 3).

1) Experiment I

Following the approach used by [6], the relevance is obtained by a text mining classification. The models learn the weights on representing a relevant work for the specific query. By using this methodology, the relevance is computed, although the results can introduce subjective recommendation. In this case, the performance analysis should occur in the training step, where the results are already known. Using our dataset under this strategy a benchmark can be established in order to serve as a comparison for our model. Indeed, in their work, the authors performed the literature review in another research domain, namely Medical Image Processing and Analysis powered by High Performance Computing.

In this experiment, the papers are classified based on the abstract text, using a TF-IDF model. Experiment 1 comprises a reproduction of their models and the performance results were annotated. Since the whole dataset is annotated, we performed the experiment using the whole dataset for both training and test. Apart from test results, we assume that training performance is somehow the most important result to analyze, since it shows how well the algorithm can perform with previous known relevance. In fact, since the subject of research as well as the domain and problems can vary from research to research, we use this result to check whether the methodology used in our model, with bibliometric data, can achieve better training performances when compared to this experiment’s results.

Table 3 shows training and test performances using the text mining approach. Training accuracy, precision, and recall achieved 100%. Using a split validation technique (70% of the dataset for training and 30% for testing), performance on the test set was 76.35%. These results correspond to the predicted on the reference paper. In Figure 2 we can see the text mining model designed on Rapidminer tool, a software framework for machine learning, mining and predictive analytics, highlighting the process; Figure 3 shows a branch of the decision tree model resulting from the ID3 algorithm.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Accuracy (Train) %</th>
<th>Precision (Train) %</th>
<th>Recall (Train) %</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID3</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>76.83</td>
<td>53.72</td>
<td>50.86</td>
</tr>
<tr>
<td>K-NN</td>
<td>76.45</td>
<td>38.22</td>
<td>50.00</td>
</tr>
</tbody>
</table>

IV. EXPERIMENTAL SETUP AND VALIDATION

To better analyze the recommendation performance on literature review process we tested our model with a study-case of a realistic scenario. A specialist, doing the research survey, gathered a set of academic papers from the main scientific repositories and annotated the works, classifying whether the papers are relevant or not for its own research domain and problem. In this case, the researcher was interested on Multi-Agent Systems applied to the Smart Grids domain. The papers were obtained mainly from three repositories: IEEE Xplore, ACM Digital and Springer by querying: “Smart Grid Multi-Agent Systems”. For all the papers, the bibliometric features seen in Table 1 were recorded. In total, 790 papers were collected and the specialist manually annotated according to its own perception of the relevance. Finally, from the set of papers annotated, 186 (24%) were classified as relevant and on the other hand, 604 (76%) were considered as not relevant literature.

Machine Learning techniques can identify the patterns used by the researcher and try to recommend the relevant literature for the researcher’s work by using this manually verified information. As explained in Section 3, we have selected ID3 as the classifier algorithm, but we wanted also to evaluate the performance of the ID3 using the bibliometric measures. In this sense, we have prepared three experiments in order to compare the performance of the classification using the strategy of [6] (Experiment 1), our strategy (Experiment 2), and a hybrid strategy combining both (Experiment 3).

2) Experiment II

As seen, the results observed in Experiment 1 are similar to the reported on previous works. We expected to achieve higher performances and believed that bibliometric features could improve the classification, as stated in our hypothesis. For comparison, Experiment 2 was designed to test the performance of the classification using the bibliometric features strategy, applied to the same dataset. Accordingly to the results seen in Section 3, the ID3 algorithm was selected as classifier.

<table>
<thead>
<tr>
<th>Step</th>
<th>Accuracy %</th>
<th>Precision %</th>
<th>Recall %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training (split validation)</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Test (split validation)</td>
<td>76.13</td>
<td>38.06</td>
<td>50</td>
</tr>
</tbody>
</table>
Figure 4 shows a branch of the decision tree model resulting from the ID3 algorithm that the features are used for deciding paper’s relevance. In this sense, using a split validation of 70-30% for training and test, we obtained a very similar result as in Experiment 1: 90.15% of accuracy on training and 77.84% of accuracy on testing, as seen in Table 4. We expected higher values since, as discussed, the features used for the classification are evidences of the paper’s relevance.

Despite the fact that our classification model did not prove to provide better accuracy than with a text mining approach, we believe that a more refined model for relevance metrics could enhance the classification. The more complete set of features used, the better results on classification. We also understand that as a complex and text-relevant model, the relevance of the paper could be related also to the semantics of the search (a more qualitative approach) and more features in that direction could be used to enhance the model.

### TABLE IV. BIBLIOMETRIC-BASED CLASSIFICATION PERFORMANCE

<table>
<thead>
<tr>
<th>Step</th>
<th>Accuracy %</th>
<th>Precision %</th>
<th>Recall %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training (split validation)</td>
<td>90.15</td>
<td>89.07</td>
<td>80.30</td>
</tr>
<tr>
<td>Test (split validation)</td>
<td>61.71</td>
<td>49.26</td>
<td>49.40</td>
</tr>
<tr>
<td>Test (whole dataset)</td>
<td>77.84</td>
<td>76.19</td>
<td>79.14</td>
</tr>
</tbody>
</table>

Although, the bibliometric model did perform well, reaching the same accuracy as seen in previous models. The classification using the bibliometric model is suitable mainly when researchers do not have a complete set of metadata for the dataset and cannot provide text-based analysis. Our model could be seen as a substitution or an enhancement.

#### 3) Experiment III

Our insights are that the two models combined could provide better results, since they both have good and similar results. To do so, a final Experiment 3 was designed to evaluate the combined performance using both models from Experiment 1 and 2. We have used the previous learned models combined in the Rapidminer tool using the “group models” operator. This operator is commonly used to join two or more models into a new model, equivalent to applying the original models in their respective order. Since the models are not trained again, the performance is evaluated on the testing results by applying the model to the whole dataset. Surprisingly, this hybrid approach seems to really enhance the performance. Table 5 shows an increase in 12% on accuracy, 11% on precision, and 3% on recall.

### TABLE V. HYBRID APPROACH PERFORMANCE

<table>
<thead>
<tr>
<th>Step</th>
<th>Accuracy %</th>
<th>Precision %</th>
<th>Recall %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test (whole dataset)</td>
<td>89.05</td>
<td>87.30</td>
<td>82.86</td>
</tr>
</tbody>
</table>

The hybrid approach, which provides more information for the classification, outperformed the standalone models, reducing the gap between qualitative and quantitative approaches. As preliminary study, we expected to use the bibliometric features to enhance the relevance classification, which could not be done isolated, but together with the text mining approach. Our sense is that other techniques, namely on the semantics and argumentation fields could bring yet more interesting information for the academic work’s relevance analysis.

### V. CONCLUSIONS AND FUTURE WORK

Literature Review is an important step on every academic research, but selecting the most representative set of articles that are relevant for a specific study is becoming a huge problem since the growth of scientific publication number and the wide spectrum of domains, problems and subjects of work. Recommendation systems are expected to help researchers on finding the State of The Art.

In this paper, we have presented a novel approach for recommending papers for literature review, based on mixed recommender models, namely text mining and bibliometric features-based. The preliminary results show improvement on recommendation. The contributions of our work are as follows.

We have discussed the relevance impact of bibliometric features in scientific works, which, together can imply on work's classification in two classes (relevant vs. non-relevant) given a specific context. Moreover, we have annotated a dataset of approximately 790 publications with scientific articles and its bibliometric measures (metadata) on the Smart Grids domain, namely papers that use Multi-Agent strategies for Smart Grids. The dataset is publicly available in the hope that it will foster further research on this topic. We propose a supervised classification approach that separates relevant papers for a specific domain using both qualitative (text-based) and quantitative (bibliometric-based) relevance models. After a battery of tests, the ID3 classification algorithm was selected, presenting best results over other classifiers (NB, KNN, DT, and NN).

Results showed that separately, both qualitative and quantitative approaches have similar classification
performances (accuracy of 77%), but combined, the performance increases to almost 90%, due to more complex information consideration. Combining qualitative and quantitative features, our approach improved the learning models by providing a more complex approach for classifying relevant scientific papers.

The contributions of our work are not restricted to this paper results. Machine Learning techniques are powerful tools to help researchers with the literature review process. We expect to see more ML-based approaches in this topic, since its relevance and soundness.

Finally, this paper ground floor for a deeper approach towards a semi-automatic literature review process. Using ML techniques to support the literature review could lead to interesting directions, such as summarization and automatic content generation. Moreover, online learning mechanisms can provide more detailed classifications, meaning more focused state-of-the-art. This kind of strategy enables, from a Systematic Literature Review perspective, to easily update literature databases and keep research aligned with the most recent works, which is very important in long research works such as thesis and research projects.

ACKNOWLEDGMENT

The first author is founded through an IBRASIL grant, a full doctorate Erasmus Mundus Action 2 STRAND 1, Lot 16 and coordinated by University of Lille. The second author thanks Universidade do Estado do Mato Grosso (UNEMAT – Brazil) and the National Scientific and Technological Development Council (CNPq) – process 234360/2014-9, grants 2010/15691-0. Authors thanks also Faculty of Engineering of University of Porto for the support given with partially founding by the Operational Programme for Competitiveness and Internationalization – COMPETE 2020 and by FCT – Portuguese Foundation for Science and Technology.

REFERENCES