Evaluating the Usefulness of Citation Graph and Document Metadata in Scientific Document Recommendation for Neophytes

Bissan Audeh, Michel Beigbeder, Christine Largeron, Diana Ramirez-Cifuentes

Evaluating the Usefulness of Citation Graph and Document Metadata in Scientific Document Recommendation for Neophytes

Bissan Audeh
Bissan.Issa@chu-st-etienne.fr
Univ Lyon,
UJM-Saint-Etienne, CNRS,
Institut d Optique Graduate School,
Laboratoire Hubert Curien UMR 5516,
F-42023, SAINT-ETIENNE, France

Michel Beigbeder
Michel.Beigbeder@emse.fr
Univ Lyon, IMT Mines Saint-Etienne,
Institut Henri Fayol,
Univ Jean Monnet,
IOGS, CNRS, LHC, F - 42023
Saint-Etienne FRANCE

Christine Largeron
Christine.Largeron@univ-st-etienne.fr
Univ Lyon,
UJM-Saint-Etienne, CNRS,
Institut d Optique Graduate School,
Laboratoire Hubert Curien UMR 5516,
F-42023, SAINT-ETIENNE, France

Diana Ramírez-Cifuentes
diana.ramirez@upf.edu
Univ Lyon,
UJM-Saint-Etienne, CNRS,
Institut d Optique Graduate School,
Laboratoire Hubert Curien UMR 5516,
F-42023, SAINT-ETIENNE, France

ABSTRACT

Digital libraries have become an essential tool for researchers in all scientific domains. With almost unlimited storage capacities, current digital libraries hold a tremendous number of documents. Though some efforts have been made to facilitate access to documents relevant to a specific information need, for a new researcher who is discovering a research field, such a task remains a real challenge. Indeed neophytes do not necessarily use appropriate keywords to express their information need and they are not necessarily qualified to evaluate correctly the relevance of documents retrieved by the system. In this study, we suppose that the retrieval system in a digital library should take into consideration features other than content-based relevance. To test this hypothesis, we use machine learning methods and build new features from several metadata related to documents. More precisely, we propose to consider as features for machine learning: content-based scores, scores based on the citation graph and scores based on metadata extracted from external resources. As acquiring such features is not a trivial task, we analyze their usefulness and their capacity to detect relevant documents. Our analysis concludes that the use of these additional features improves the performance of the system for a neophyte user. In fact, by adding the new features we find more documents suitable for neophytes within the results returned by the system than when using content-based features alone.

1 INTRODUCTION

The digital revolution led to the emergence of digital libraries with almost unlimited data storage capacity. Finding information in such a huge mass of documents is not an easy task. While human support was more accessible in traditional libraries, in digital libraries new researchers discovering a new scientific field have to figure out alone the right query for their information need and judge the quality of the retrieved documents. These actions are a real challenge if the user is not assisted in the exploration of a scientific field new for him.

In this article, we focus on academic digital libraries whose collection is composed of scientific documents including books, journal articles and conference proceedings. These documents have the specificity of citing other documents by referring to the metadata that identifies them, notably the title, authors, publication date and venue (either a conference or a journal). These citations are essential for scientific activities and enable the evaluation of different aspects of science through scientometrics.

The goal of this paper is to evaluate the role of different types of features in suggesting documents that are not only relevant to neophytes' information need but also important in the research domain they are trying to explore. To achieve this goal, we extract documents' metadata from external ressources and use citation relationships between scientific documents. Indeed, we assume that review articles with high number of references and documents

frequently cited in the domain are suitable candidates for neophytes. Moreover, to begin a research, a neophyte needs to read introductory papers which explore the selected topic from a multidisciplinary point of view before considering more specialized publications. Finally, a neophyte cannot usually identify the experts in the domain, but needs to read their publications as well as papers published in prominent journals or conferences in this field. We believe that citation graph and metadata extracted from external resources provide helpful information to identify documents satisfying previous criteria. Based on this hypothesis, we define a set of features, that we expect to be good predictive features for suggesting scientific documents to neophytes. The usefulness of these features in identifying relevant documents is evaluated using machine learning methods.

A comparative study of digital libraries and related works is presented in Sec. 2. We describe the architecture of our system as well as the features studied in Sec. 3. Our evaluation protocol is detailed in Sec. 4 and the results in Sec. 5. Finally, Sec. 6 states our conclusion.

2 DIGITAL LIBRARIES: STATE OF THE ART AND CITATION USE

2.1 Digital Libraries

The purpose of digital libraries is to provide services to users about digital material which can be text, audio, or video. Most of these services imply processes of data storage, organization and retrieval in order to deliver relevant information to the final user [9]. We are particularly interested in the retrieval functionalities of digital libraries including bibliographic services. These services do not store nor provide the material itself but handle metadata to return only pointers to the actual material. We conducted a benchmark study in some well known digital libraries (cf. Table 1) to explore their retrieval functionalities and their ability to deliver relevant documents to neophytes, the users targeted by our research.

For our benchmark study we considered the following cases: Google Scholar[1], Microsoft Academic Search[2], Citeseer[3], ACM digital library[4], dblp[5], and Web of Science[6]. Three of them are limited to the computer science domain: Citeseer, ACM digital library and dblp while the other cover several scientific domains. We kept Microsoft Academic Search in our list because some of its functionalities were quite unique even though it is no longer available as a service.

All of these libraries provide the means to search and navigate through their documents' collection. Both dblp and Web of Science only work with metadata while the other libraries include the full text of documents. In the case of dblp, the only available metadata consists in articles' description. This description includes the article's title, authors, venue, the publisher and the year of publication. Although Web of Science built its notoriety on the creation and the curation of the citation data , Citeseer was the first to automatically extract the citation data from the papers themselves. This automatic extraction was firstly done using heuristics [7] and then with machine learning tools [6]. Later on, Google Scholar and Microsoft Academic Search also extracted citation from the textual content.

The search tools of these libraries return a list of pointers to documents that are related to the query. These lists can be sorted according to different criteria depending on the tool. For Google Scholar, Microsoft Academic Search, Citeseer and the ACM digital library, the sort criterion is strongly based on the content as in classical information retrieval. Nevertheless Google Scholar also uses the citations in a manner which is analogous to the PageRank that Google uses on the Web. Moreover, it is well known that Google uses many features to rank the pages in their search engines, but the exact features used are not public. Microsoft Academic Search uses an approach called PopRank[14] which is in some sense a generalization of PageRank that takes into account a graph with different types of nodes: documents, authors or venues (journals or conferences). The ranking of dblp and Web of Science is oriented more to the metadata rather than to the content. Both of them propose a ranking according to the publication date even if dblp can mix this criteria with relevance. Web of Science, which uses curated data, sorts documents according to the first author or to the venue but also according to the citation count which is based on the citation graph. Web of Science has a strong bias towards scientometrics and indicators about science. Google Scholar and Microsoft Academic Search also tend to compute such indicators. To do so, Google Scholar uses author profiles which are built by the authors themselves, and Microsoft Academic Search automatically builds author profiles but also connects the documents and authors to organizations and venues. Microsoft Academic Search is also able to compute indicators for these nodes. One major problem for both Google Scholar and Microsoft Academic Search is the quality of the extracted information, mainly because of the ambiguity of author names and venue titles.

The ranking proposed by the search engines of these libraries is either clearly not dedicated to relevance as in dblp and Web of Science, or not dedicated to neophytes even though it takes into account the relevance, which is the case of the other libraries in our benchmark. Moreover, although Google Scholar, Microsoft Academic Search and ACM digital library also use features based on citations, we do not know exactly what these features are and if their use is of interest for neophytes. Table 1 summarizes information on these digital libraries.

2.2 Citation Use

The interest of using relational information in addition to content has been pointed out by several studies since the beginning of information retrieval [18]. In the context of scientific paper retrieval, several types of graphs have been explored to enhance information retrieval or reference recommendation. To evaluate interesting scientific documents, some works focused on the influence of authors based on co-author graphs, where two authors are linked if they are authors of a same article [13]. Other studies were interested in the relatedness among scientific papers and thus they focused on the analysis of the co-citation graph, in which two documents are linked if they are both referenced by the same paper [5]. The

[1] https://scholar.google.fr/
[2] http://academic.research.microsoft.com/
[3] http://citeseerx.ist.psu.edu/
[4] http://dl.acm.org/
[5] http://dblp.uni-trier.de/
[6] http://webofknowledge.com/

	general	metadata	full text	citations	sort criteria
Google Scholar	✓	✓	✓	✓	relevance, publication date
Microsoft Academic Search	✓	✓	✓	✓	relevance, publication date, in-degree
Citeseer	✗	✓	✓	✓	relevance, publication date, in-degree, recency
ACM digital library	✗	✓	✓	✓	relevance, publication date, in-degree, downloads
dblp	✗	✓	✗	✗	relevance, publication date
Web of Science	✓	✓	✗	✓	publication date, in-degree, usage, author's name, venue name

Table 1: Comparison of some digital libraries. Column 2 indicates whether the library only contains documents related to a *specific* domain. Column 4 indicates if the full text of documents is available. Column 5 indicates if citations are taken into account and the last column lists the sort criteria.

citation graph, which links two documents if one cites the other, is another popular approach in scientific paper recommendation [2]. The dominant idea of this approach is to combine citation based scores extracted from this graph with content based scores. In information retrieval-like systems, where the recommendation is based on a query entered by a user, this combination can be achieved in two ways: by modeling the citation scores within the retrieval matching function, or by re-ranking an initial list of documents relevant for the query. As examples of the first solution, we can consider [23], [24] and [10]. The authors in [23] integrate relational information from a citation graph (degree, in-degree, out-degree) in a probabilistic matching function used for content-based retrieval. Similarly, [24] uses citation count, page-rank [15] and co-citation information to generate priors for a language model dedicated to information retrieval. Last but not least, [10] uses topic modeling to associate distribution-based priors to the nodes and edges of the citation graph, a page-rank with a priors algorithm is then used to select the most important papers within a topic.

Alternatively, proposals that use the citation graph to re-rank a list of documents focus mostly on generating a global score that combines a set of selected scores. For example, [20] uses a linear combination of the content based score with citation count, co-citation count, publication year, author feature and Katz distance. The authors conclude that content-based and Katz features are the most efficient in finding interesting documents. Another study [21] proposes an iterative algorithm that calculates a score of the popularity of venues in order to re-rank a list of relevant documents. The authors argue that this score overcomes the limitations of impact factors and exceeds the performance of the simple page-rank score. Recently, [19] proposed an approach to generate a reading list to help a new researcher in building a literature review. In this approach, the authors use author-keywords to generate sets of similar papers. A measure that calculates the citing and cited references for each set is then used to assign a value to each document in the collection. A list of documents retrieved for a query is then re-ranked according to this new measure in order to recommend top ranked papers to the user.

Our overview of the literature revealed very few papers addressing the specific problem of scientific paper recommendation for neophytes. The articles the most related to our subject are [19] and [1]. The paper of [19], confirms the specificity of neophyte scientific recommendation and provides an interesting framework for user case evaluation. Nevertheless, they use a heuristic method to combine relational and content data, which is the case of most of the studies that we reviewed in this domain. We consider that our machine learning approach can solve this issue efficiently since it allows the system to automatically learn how to combine the different scores, a task that is difficult even for an expert. The work in [1] compared the use of machine learning for document recommendation to list re-ranking and aggregation. The authors suggest that using citation scores as features for machine learning is promising for exploratory information retrieval. Based on these conclusions, we follow up the exploration of the utility of machine learning for neophyte document recommendation by evaluating supplementary features for machine learning in addition to content and citation based features. Furthermore, we pay special attention to the evaluation framework. As noted by [2], evaluation is a real problem that prevents an objective comparison of the different approaches. Two choices are essential in the evaluation strategy in this domain: the choice of ground truths and the choice of baselines. Although some works use an evaluation framework with predefined judgments provided by evaluation campaigns like TREC[7] [23], others prefer the use of more realistic and recent data like CiteSeer [19], ACM digital library [10] or Rexa [20]. In the works that we reviewed, the baseline is either a simple content-based information retrieval model [23, 24] or a more sophisticated baseline that already includes relational information [10].

Our goal in this paper is to outline what is important for recommending useful papers to neophytes. In particular we want to evaluate the added value of using citations based features and features based on other metadata in the recommendation process for neophytes. To achieve this, we built an evaluating framework in which we designed two realistic ground truths, and used the machine learning approach Random Forest. Our experiments took into consideration issues related to missing values, unbalanced classes and error costs. The proposed evaluation framework enabled us to measure the usefulness of three scenarios composed of different features' subsets.

3 FEATURE CONSTRUCTION MODEL TO IMPROVE SCIENTIFIC EXPLORATION IN DIGITAL LIBRARIES

3.1 System Architecture

In general, the goal of an information retrieval system is to select from a collection of documents a subset relevant to the information

[7] http://trec.nist.gov

need of a user. This information need is communicated to the system by means of a user-initiated query [11] and the selected documents are sorted according to their content based score which evaluates the similarity or some probability between the document and the query. In this work we evaluate the interest of taking into account new features in addition to content and citation based features notably by measuring their discriminative power. The proposed framework was built on top of a ISTEX[8] online library whose general architecture is depicted in the green frame in Figure 1.

In the first step, given a query q formulated by a user, the Lucene search engine computes a content based score $S_C(q,d)$ with the BM25 ranking function [17] for each document d belonging to the collection. It should be noticed that the content based search adopted by classical digital libraries, is reduced to this step (green frame in Figure 1) and considered as the baseline in our experimental evaluation. In the second step, features are extracted from the collection of documents, the citation graph or external resources. Finally, in the third step, all these features are combined with the content based score using machine learning models to predict the relevance of the documents for a query.

3.2 Feature Construction

In this work, our hypothesis is that a document suitable for neophytes must satisfy several criteria in addition to the content based relevance. Notably, as explained in the introduction, these users can be interested in documents mentioned by a large number of papers, having a multidisciplinary point of view or, at the opposite, more specialized, and finally in publications written by experts or published in reputable journals in the research field. For this reason, we propose to build from the citation graph or from meta-data the following set of features.

3.2.1 Citation based features. The citation based features are used to evaluate the document according to three criteria:

- importance: the document is mentioned by a large number of articles in the domain;
- coverage: the document mentions a large number of articles in the domain;
- popularity: popular documents mention this document in their bibliography.

Citation features are built using the citation graph in which the nodes are the documents of the whole collection, and the edges represent the citation relationships. In other words, there is a link from a document node A to a document node B if A cites B. We generated the citation graph on the basis of the titles found in the metadata of the library's documents, and the titles extracted from the PDF files in the Reference sections. It should be noted that the quality of the title extraction is poor, so we used a Locality sensitive hashing method to cluster the titles that had a Levenstein similarity greater than 0.85.

From our citation graph, three citation based scores are computed for document d: the in-degree $S_I(d)$, the out-degree $S_O(d)$ and the page-rank $S_{PR}(d)$ scores to respectively evaluate the importance, the coverage and the popularity of document d. These scores are independent of the query, thus they can be calculated off-line. Moreover, it should be stressed that other citation based scores could also be considered [3, 19].

3.2.2 Multidisciplinarity based features. When studying a new topic one might be interested in looking for very specialized publications or for articles considering the subject from the point of view of different disciplines. In our work, we consider that the number of disciplines to which a document belongs is a way to measure its multidisciplinarity and its specificity. For this purpose, we used the categorization which is based on Web of Science as it is available in the metadata of our documents. According to this categorization, a document can belong to three main categories that correspond to the different citation indexes of Thomson Reuters Web of Science: Science Citation Index, Social Science Citation Index, and Arts & Humanities Citation Index and 225 subcategories. Thus, the number of categories $Numcat$ and the number of subcategories $Numsubcat$ constitute two multidisciplinarity features. We consider that a document belonging to several categories is multidisciplinary, whereas a document that belongs to only one category but with subcategories is more specialized.

3.2.3 Author based features. We took into consideration papers' authors since they can be useful for measuring documents' quality. Indeed, metrics such as the number of publications an author has, the number of publications per year, and the number of co-authors, can be indicators of the productivity and level of collaboration of an author within the research field. Consequently, we used dblp for reasons of accessibility and good coverage in our field to extract the following metrics per author:

- publication number;
- co-author number;
- annual average publication number (the ratio of the publication number over the number of years when the author published).

It should be noted that this extraction is the most challenging task of the process, because of the author name ambiguity. To solve this issue, the GESTALT pattern matching method has been applied to match the records of the two databases: our digital library and dblp. This matching process is out of the scope of this article, interested reader is referred to [16] for more details. By combining these per author metrics, we generate four features at the document level:

- maximum author's publication number ($Maxpubnum$);
- maximum author's co-author number ($Maxcoauth$);
- maximum author's annual average publication number ($Maxpubperyear$);
- average authors' publication numbers ($Avgpubnum$).

Due to disambiguation difficulties in the matching process, the percentage of missing values for these features is equal to 50 %.

3.2.4 Journal based features. Journals' Impact Factor (JIF) aims to determine the reputation of a journal by measuring the average number of citations per article (published by the journal) over the previous two years. The JIF for a journal J in the year y is obtained with the following formula:

$$JIF(J,y) = \frac{Citations(J,y)}{Published(J,y_{-1}) + Published(J,y_{-2})} \quad (1)$$

[8] www.istex.fr

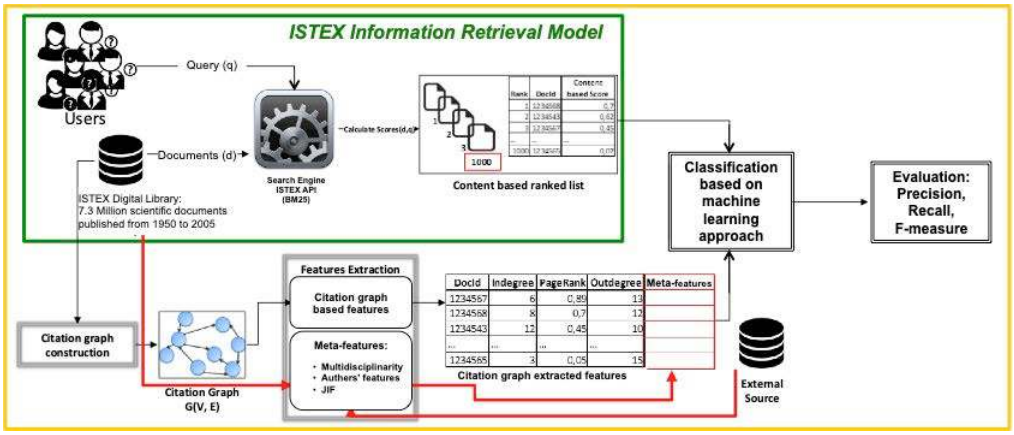

Figure 1: The structure of the system

where $Citations(J, y)$ is the number of citations made in the year y to the articles published in J in the two previous years ($y-1$ and $y-2$), $Published(J, y_{-1})$ and $Published(J, y_{-2})$ are the number of articles published in the journal J in the years (y-1) and (y-2) respectively. Although other metrics can be used such as Eigenfactor, it was found that the Eigenfactor is highly correlated with the JIF in our collection, with a value of 0.79 for Spearman's rank correlation coefficient.

3.3 Supervised Machine Learning Approach

In our system, the recommendation is formalized as a supervised machine learning problem in the following way: We consider a set of elements, where each element (q, d) is composed of a query q and a document d and is associated with the previously described scores: $S_C(q,d)$, $S_I(d)$, $S_O(d)$, $S_{PR}(d)$, $Numcat(d)$, $Numsubcat(d)$, $JIF(d)$, $Maxpubnum(d)$, $Maxcoauth(d)$, $Maxpubperyear(d)$, $Avgpubnum(d)$ that we consider as predictive features. Given these predictive features, the task consists in deciding whether the document d associated to the query q should be recommended to neophytes or not. If a learning set is available, in other words, if we have a subset of elements (q, d) for which both the predictive features and the binary label (relevant or non relevant) are available, the task can be defined as a classification problem and solved with supervised machine learning methods. Frequently and notably in our context, the difficulty lies in obtaining the labels. To do so, we designed two ground truth scenarios, called Popularity and Thesis, that will be described in section 4.2. Each of these scenarios is associated with a target variable that we want to predict.

Once the features and the labels are built for all the examples belonging to the learning set, this latter can be split into a training set and a test set. Then, a model can be learned, with machine learning algorithms, on the first set and it can be evaluated on the remaining elements belonging to the test set. In our experiments, we used Random Forest (*RF*) [4] as the machine learning model. This model was chosen because it allows to measure the discriminative power of each feature; which is our aim, and also because it achieved good results in preliminary comparative experiments. The main advantage of the machine learning approach compared to list re-ranking, lists aggregation or linear combination of scores is that it automatically selects the useful predictive features and combines them without having to set weights. Consequently, this approach allows the best predictive features to be identified. Theoretically, this should optimize the prediction of the right documents for the neophyte.

	Used features
Scenario 1	Content feature
Scenario 2	Content + Citation features
Scenario 3	Content + Citation + Meta features

Table 2: The three evaluation scenarios in our experiments.

4 EVALUATION

4.1 Evaluated Scenarios

The features proposed in this work aim to improve the performance of scientific document recommendation for a neophyte. They can be grouped in three categories according to their resources:

- Content feature: the relevance score of a document based on its content;
- Citation based features: in-degree, out-degree and PageRank values computed from the citation graph;
- Meta features: multidisciplinary features, author based features and Journal Impact Factor.

The performance of the machine learning approach is evaluated using the three scenarios described in Table 2. In the first scenario, only the content based score is used to predict the relevance of a document for a query. This scenario is our baseline and, as previously said, it corresponds also to the content based search adopted by classical digital libraries. In the second scenario, citation based features are also taken into account. Consequently, it corresponds to information retrieval models of digital libraries which exploit relational information, as those described in section 2. Finally in the third scenario, multidisciplinarity, journal based features and author based features are considered in addition to content and citation

Original Title	Extracted Query
Revisiter le couplage traitement automatique des langues et recherche d'information	Automatic language processing information retrieval
Data quality evaluation in data integration systems	Quality evaluation data integration system
Symbolic data mining methods with the Coron Platform	Symbolic Data Mining Coron Platform

Table 3: Examples of thesis titles and the extracted keywords used to generate queries of our experiments

features. This last scenario corresponds to our evaluated system, whose architecture is depicted in Figure 1. We wanted to check the added value of our features by measuring their discriminant power whatever the cost of their computation.

4.2 Data

The set of documents used for our evaluation is extracted from ISTEX digital library. ISTEX is a national project that acquires scientific literature to offer an online and unlimited access to educational institutions in France. As for April 2017, the ISTEX online archive contained over 18 million scientific documents. From this dataset, we extracted the documents published between 1950 and 2005. These dates were chosen arbitrarily, but the choice of a period in the past is important for the construction of our popularity ground truth. After eliminating documents that have an empty reference list, we obtained the collection used in our experiments, which contains 7 343 385 scientific documents of many types except thesis manuscripts. For our experiments, we built 25 queries from the titles of 25 French PhD theses in computer science published before 2005, examples of these queries are illustrated in Table 3. The main preprocessing on the thesis titles consisted of translating french titles to english and removing stop words. Although 25 queries is a limited number, each query will be associated to 1000 retrieved documents to generate our evaluation dataset of 25 000 examples as we will see in Sec.4.3.

The titles of a thesis could reflect expert knowledge in its formulation or with the terms used. Though we think this is not a problem in our framework, because in the true life a PhD student discovers the PhD subject as formulated by the thesis advisor, and it would be another research problem to deal with vocabulary discrepancies of people who are not aware of the expert vocabulary. The extracted queries are then executed against the search engine of the ISTEX digital library. For each query, the top 1000 retrieved documents published before 2005 by the content-based search engine are considered as the *Initial List L*. It is important to know that ISTEX collection does not contain Thesis manuscrits, which eliminates the bias related to using Thesis subjects as queries.

As mentioned in Sec. 3.3, based on the 25 extracted queries we established two ground truths to evaluate our work: Thesis and Popularity. Both ground truths fit the context in which a neophyte is exploring a new search field. For the Thesis ground truth, we suppose that a PhD student will cite in the thesis the references that helped in exploring the subject and the references that he or she has used after acquiring a higher experience level, and could have been useful when starting the thesis as neophyte. Thus, in this ground truth, a document is relevant for a query if it is cited by the thesis from which we extracted the query.

For the Popularity ground truth, we suppose that a document is relevant if it has a good citation level in the future. For our experiment, we evaluate the popularity relevance of a document published between 1950 and 2005 by the number of times this document was cited in 2017 in GoogleScholar ($Citations_{GS}$). In this ground truth, we consider as relevant the 30 documents that have the highest $Citations_{GS}$ score in the list L.

Feature	Feat. set	Percentage of examples with missing value
Relevance score	Content	0%
$S_I(d)$ In-degree	Citation	0%
$S_O(d)$ Out-degree	Citation	0%
$S_{PR}(d)$ PageRank	Citation	0%
Maxpubnum	Meta	49.74%
Maxpubperyear	Meta	49.74%
Maxcoauth	Meta	49.74%
Avgpubnum	Meta	49.74%
Avgpubperyear	Meta	49.74%
Avgcoauth	Meta	49.74%
JIF	Meta	2.55%
Numcat	Meta	54.77%
Numsubcat	Meta	54.77%

Table 4: Percentage of examples with missing values for each feature in the All Examples Dataset.

4.3 Machine Learning Framework

For our machine learning experience, for every query q, we consider as dataset all pairs of the form (q, d) where d is in the initial retrieved list L. Hence, from the 25 queries extracted from thesis' titles, we built our All Examples Dataset, which is composed of 25 000 examples. For each example represented by a couple (q, d), we assign the set of features described in Sec. 3 and a boolean label which is 1 if the document is relevant to the query, and 0 otherwise. The value of this label depends on the ground truth according to which the evaluation is done. In our dataset, the calculation of the feature values is not possible for all the data. Table 4 shows the percentage of missing values for each of our generated features.

It is important to remember that Scenario 1 only contains the feature that represents the relevance between a query and a document, and Scenario 2 contains the relational features that are calculated from our generated citation graph. Thus, any document in our collection will necessarily obtain values for all the features of these scenarios, which is not the case for the meta features of Scenario 3. For example, calculating categories counts for the multidisciplinary features was only possible for 74.29 % of ISTEX collection, which corresponds mainly to documents edited by Thomson Reuters, thus classified in at least one Web of Science category. In order to examine the effect of missing values, we constructed a dataset that only

	Fully Qualified Dataset		All Examples Dataset	
Ground truth	#label 1	#label 0	#label 1	#label 0
Popularity	122	3 954	750	24 250
Thesis	14	4 062	85	24 915

Table 5: Count of positive (label 1) and negative (label 0) examples in the Fully Qualified Dataset and All Examples Dataset for both ground truths.

contains the examples for which all meta features have assigned values. This new dataset contains 4 076 examples and is called Fully Qualified Dataset. Table 5 shows the number of positive and negative examples for each ground truth and each dataset. From this table, we note that our datasets are unbalanced. For the Popularity ground truth, 750 examples of the 25 000 are positive because the top 30 most cited for each of the 25 queries were considered as relevant. For the Thesis ground truth, the lack of positive examples is explained by the fact that the ISTEX database does not contain all the literature cited in the thesis, thus, only 85 examples are positive in the All Examples Dataset.

To analyze our proposed features, we choose the Random Forest method [4] for evaluation, which is widely used as a baseline in the literature [22]. The Random Forest algorithm builds several decision trees in the training step. For the classification tasks, it outputs the class that is the mode of the classes given by the individual decision trees. The implementation proposed in Weka[9] [8] for Random Forest was used to run our experiments. To verify how our features behave with other machine learning methods, we also used SVM to evaluate our three features' scenarios.

To better consider the generalization and independence of our results, we used two different cross validation methods: the K-fold cross validation and the repeated random sub-sampling validation that we refer to as random sampling validation. For the first validation method we use 10 folds, while for the random sampling method we use 10 subsamples with 70 % of the examples for training and 30 % for testing. Thus, our experiments aim to compare the three scenarios (*i.e.* the three sets of features) and the effect of missing values, class balancing and cost matrix application within the Thesis and Popularity ground truths. For this reason, default values set by Weka for the machine learning parameters were always applied, and no specific parameter optimisation was achieved.

4.4 Baseline and Evaluation Measures

As mentioned in Sec. 4.1, our baseline is Scenario 1, which corresponds to ISTEX retrieval system built on Lucene search engine. In other words, our baseline is a classical digital library that evaluates relevance based on matching the contents of documents and queries. Each of our experiments is conducted within the Thesis and Popularity ground truths, for both the Fully Qualified Dataset and the All Examples Dataset. After setting the training and testing sets, the results obtained by the machine learning classifier on the testing sets are evaluated with the usual measures: Precision, Recall and F-measure [11]. We tested two different ways of overcoming the issue of unbalanced data: using a class balancer filter for the training set and introducing a cost matrix in order to penalize false negative and false positive predictions.

[9]Weka is an open source software which provides a collection of machine learning algorithms and tools.

5 RESULTS

5.1 Results According to the Popularity Ground Truth

In this experiment we compare the precision, recall and F-measure when using the Random Forest algorithm with the Popularity ground truth. The first results presented in Table 6 concerns the Fully Qualified Dataset. Without the use of class balancing nor cost matrix, these results show the weak performance when using the content based feature alone (Scenario 1). The performance of the system is highly improved when adding the citation graph features (Scenario 2). The addition of the meta features (Scenario 3) has a positive impact on the results especially in precision, this impact is less important in recall and F-measure. This result means that the use of meta features with the graph and content features helps the system to predict less false positive values even though it does not retrieve more relevant documents when compared to the use of graph and content features alone.

With the same settings, we evaluated the effect of using class balancing. We also checked if normalizing the features into the range [0,1] has an effect on the results. Normalization led to insignificant but slightly better precision with Scenario 3. Thus, in the second set of results in Table 6 we present the effect of class balancing on the precision and the recall when all the features are normalized. We note from Table 6 that the use of class balancing improves the recall of the system, but has a negative influence on the precision whatever the feature group used is. In other words, training the model with a balanced set of positive and negative examples helps to recognize more examples as positive.

Furthermore, we examine the effect of using a cost matrix in the same settings without using class balancing. In order to find more relevant documents, we configured the cost matrix so that it assigns twice the cost for false negatives than for false positives. It is clear that, with our highly unbalanced data, penalizing the classification of a relevant document as a non-relevant could lead to a loss in precision. The third set of results in Tab. 6 shows, as expected, that using the cost matrix improved the recall but decreased the precision with the Fully Qualified Dataset. From these results we note that the use of a cost matrix improves the recall of the system when we have more than one feature (Scenario 2 and 3). In conclusion, with or without applying the cost matrix, adding the meta features (Scenario 3) improves the performance of the system when compared to the use of content based and graph based features only.

In order to evaluate the effect of missing data, we repeat our experiment with the All Examples Dataset that contains all the examples with several missing values for the meta features. In our dataset, content-based and graph based features always contain values for any document in the collection, that is why the results of our experiment with the All Examples Dataset only concerns Scenario 3 (cf. Table 7). In Weka, the Random Forest algorithm handles missing values as following: while instances with known feature

	no specific treatement			using class balancing and normalization			using the cost matrix		
Measure	Sc. 1	Sc. 2	Sc. 3	Sc. 1	Sc. 2	Sc. 3	Sc. 1	Sc. 2	Sc. 3
Precision	0.09	0.59	**0.65**	0.07	0.32	**0.50**	0.09	0.54	**0.57**
Recall	0.07	0.19	**0.21**	0.13	0.29	**0.36**	0.07	0.26	**0.29**
F-measure	0.08	0.29	**0.33**	0.09	0.30	**0.42**	0.08	0.35	**0.38**

Table 6: Popularity ground truth evaluation results with the Fully Qualified Dataset.In bold, best results in the corresponding configuration, underlined values are the best results over all configurations.

Measure	Scenario 3 in Popularity Ground Truth
Precision	0.82
Recall	0.23
F-measure	0.36

Table 7: Popularity ground truth evaluation results with the All Examples Dataset and Scenario 3 (no class balancing, no cost matrix).

Measure	Scenario 1	Scenario 2	Scenario 3
Precision	0	0.50	0.50
Recall	0	0.06	0.06
F-measure	-	0.11	0.11

Table 8: Thesis ground truth evaluation results with the All Examples Dataset.

values are split within branches according to their actual values, instances with unknown feature values are split in proportion to the split of known values. At the time of testing, test instances with a missing feature value are distributed into branches according to the proportions of training examples falling into those branches. As all the presented experiments achieved good precision without using class balancing and normalization, we reuse the same settings with the All Examples Dataset, because we assume that precision is a priority in exploratory search [12]. Comparing the results of Scenario 3 in the Popularity ground truth using the All Examples Dataset (Table 7) against the use of Fully Qualified Dataset (Table 6) shows that considering all the examples achieves better precision even when these examples contain missing values. The result in recall and F-measure is relatively close to the ones obtained using the Fully Qualified Dataset. This result is consistent with the way Weka handles missing values with random forest as explained earlier.

5.2 Results According to the Thesis Ground Truth

With the encouraging results in terms of precision in the popularity ground truth, we repeated the same experiment with the All Examples Dataset with the Thesis ground truth. As class balancing had negative effect on precision in our experiments with the Popularity ground truth, and as we consider the precision as a priority in the neophyte information retrieval problem, we did not use class balancing with Thesis ground truth. It should be noted that the Thesis ground truth is very restrictive, the percentage of positive examples does not exceed 0.34 %, which makes the possibility of predicting true positive examples very small. Tab. 8 shows clearly the difficulty of the Thesis ground truth, as the system cannot easily find true positive examples and obtains null results for precision and recall which leaves the F-measure undefined. Despite this difficulty, adding citation features achieves an impressive enhancement in precision. This result means that a document retrieved by the system with high values of graph based features is more likely to be considered as a citation by a PhD student than a document which is relevant based on the content.

Popularity ground truth		
Feature	Impurity decrease	# Nodes in RF
content	0.21	155409
JIF	0.19	106489
out-degree	0.16	126421
in-degree	0.13	69487
avg. pub.per year	0.08	3823

Table 9: Feature discriminative power in Popularity ground truth (10-fold cross validation).

Thesis ground truth		
Feature	Impurity decrease	# Nodes in RF
out-degree	0.19	9587
PageRank	0.17	7735
JIF	0.16	7918
in-degree	0.14	4748

Table 10: Feature discriminative power in Thesis ground truth (10-fold cross validation).

5.3 Discussion about the Discriminative Power of the Features

With the settings of the best results we achieved in terms of precision for both Popularity and Thesis ground truths, we analyze the importance of each of the features in our All Examples Dataset. We measure the feature importance by calculating both the average impurity decrease and the number of nodes using the features with the random forest algorithm.

From Tab. 9 we note that the most discriminant feature is the content feature. For both ground truths, graph based features are

Popularity ground truth			
	Sc. 1	Sc. 2	Sc. 3
Popularity ground truth	0	0.98	0.98
Thesis ground truth	0	0	0

Table 11: The precision of the three scenarios with SVM

in the top 5 discriminant features. From the meta features, the journal impact factor appears to have an important role in detecting the relevance of a document for a neophyte. From these results we conclude that although the content feature is discriminant in exploratory search, the relevance for a neophyte is considerably related to the citation graph and the quality of the journal in which the document is published. In addition, the Popularity ground truth reveals that the activity level of the authors is also an important factor in the popularity of scientific documents.

5.4 Performance with other Machine Learning Methods

In this work, we also explored the effect of our engineered features on SVM. Tab. 11 show that in the Popularity ground truth, SVM was improved with the use of graph and meta features. These results are encouraging as they confirm that, whatever the machine learning approach used, using the features that we proposed in this paper leads to better results than using content-based relevance alone. Nevertheless, the very low number of positive examples in the Thesis ground truth prevented these algorithms from recognizing relevant documents for a neophyte whatever the set of features used.

6 CONCLUSION

Our aim was to explore the capacity of feature engineering combined with machine learning to improve information retrieval and recommendations for neophytes in digital libraries. In this view, the main contributions of our work are: the definition of predictive features to capture what makes a publication relevant for a neophyte, the proposal of a system for recommending documents to neophytes, the design of ground truths for testing the proposed model and the evaluation of this model on the French national digital library *ISTEX*.

The results we obtained confirmed the findings of previous works about the necessity of using relational and contextual information in addition to content-based matching to evaluate the relevance of a document. We demonstrated within realistic ground truths that meta information about a document such as an author's popularity or journal impact factor could reveal additional relevant documents suitable for a new researcher. The main limitation of this study is the difficulty in disambiguating authors, journals or venue names. Using an external resource to overcome such difficulty demands taking into consideration the particularity of the different scientific domains.

We believe that facilitating accessibility to relevant resources for neophytes does not only concern new researchers, it is also a major issue for designing efficient knowledge management systems in various fields. Notably, this subject concerns information access optimizing for neophytes in companies to enable full exploitation of information assets by their members including their newcomers.

ACKNOWLEDGMENT

This work has been partially funded by the ISTEX Project: http://www.istex.fr

REFERENCES

[1] Bissan Audeh, Michel Beigbeder, and Christine Largeron. 2017. A Machine Learning System for Assisting Neophyte Researchers in Digital Libraries. In *IAPR International Conference on Document Analysis and Recognition, ICDAR 2017.*
[2] Joeran Beel, Bela Gipp, Stefan Langer, and Corinna Breitinger. 2016. Research-paper recommender systems: a literature survey. *International Journal on Digital Libraries* 17, 4 (2016).
[3] Francesco Bonchi, Gianmarco De Francisci Morales, and Matteo Riondato. 2016. Centrality measures on big graphs: Exact, approximated, and distributed algorithms. In *Proceedings of the 25th International Conference Companion on World Wide Web.* International World Wide Web Conferences Steering Committee.
[4] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001).
[5] Zeljko Carevic and Philipp Schaer. 2014. On the connection between citation-based and topical relevance ranking: Results of a Pretest using isearch. *CEUR Workshop Proceedings* 1143 (2014).
[6] Isaac G. Councill, C. Lee Giles, and Min-Yen Kan. 2008. ParsCit: an Open-source CRF Reference String Parsing Package. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco.*
[7] C. Lee Giles, Kurt D. Bollacker, and Steve Lawrence. 1998. CiteSeer: An Automatic Citation Indexing System. In *Proceedings of the 3rd ACM International Conference on Digital Libraries, June 23-26, 1998, Pittsburgh, PA, USA.*
[8] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter* 11, 1 (2009).
[9] Michael Lesk. 1997. *Practical Digital Libraries: Books, Bytes, and Bucks.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.
[10] Xiaozhong Liu, Jinsong Zhang, and Chun Guo. 2013. Full-text citation analysis: A new method to enhance scholarly networks. *Journal of the American Society for Information Science and Technology* 64, 9 (2013).
[11] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval.*
[12] Gary Marchionini. 2006. Exploratory search: from finding to understanding. *Commun. ACM* 49, 4 (2006).
[13] David Mimno and Andrew McCallum. 2007. Mining a digital library for influential authors. *Proceedings of the 2007 conference on Digital libraries - JCDL '07* 2 (2007).
[14] Zaiqing Nie, Yuanzhi Zhang, Ji-Rong Wen, and Wei-Ying Ma. 2005. Object-level ranking: bringing order to Web objects. In *Proceedings of the 14th international conference on World Wide Web, WWW 2005, Chiba, Japan, May 10-14, 2005.*
[15] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. *The PageRank citation ranking: Bringing order to the web.* Technical Report. Stanford InfoLab.
[16] J. W. Ratcliff and D. E. Metzener. 1988. Pattern Matching: the Gestalt Approach.
[17] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In *Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994.*
[18] Gerard Salton. 1963. Associative document retrieval techniques using bibliographic information. *Journal of the ACM (JACM)* 10, 4 (1963).
[19] A. Sesagiri Raamkumar, S. Foo, and N. Pang. 2017. Using author-specified keywords in building an initial reading list of research papers in scientific paper retrieval and recommender systems. *Information Processing and Management* 53, 3 (2017).
[20] Trevor Strohman, W Bruce Croft, and David Jensen. 2007. Recommending Citations for Academic Papers. *Evaluation* (2007).
[21] Yang Sun and CL Lee Giles. 2007. Popularity weighted ranking for academic digital libraries. *Advances in Information Retrieval* (2007).
[22] Antanas Verikas, Adas Gelzinis, and Marija Bacauskiene. 2011. Mining data with random forests: A survey and results of new tests. *Pattern Recognition* 44, 2 (2011).
[23] Xiaoshi Yin, Jimmy Xiangji, and Zhoujun Li. 2011. Mining and modeling linkage information from citation context for improving biomedical literature retrieval. *Information Processing and Management* 47, 1 (2011).
[24] Haozhen Zhao and Xiaohua Hu. 2014. Language Model Document Priors based on Citation and Co-citation Analysis. In *ECIR.*