Investigating Domain Features for Scope Detection and Classification of Scientific Articles

Tirthankar Ghosal, Ravi Sonam, Sriparna Saha, Asif Ekbal, Pushpak Bhattacharyya
Indian Institute of Technology Patna
Bihta, Bihar, India, Pin : 801106
{tirthankar.pcs16,ravi.cs13,sriparna,asif,pb}@iitp.ac.in

Abstract

Editorial screening, better known as desk rejection is a common phenomenon in scholarly publishing. Many papers suffer desk rejection simply because they are not sent to the right journal. We propose a supervised machine learning system that could assist the editors in identifying out-of-scope manuscripts. Our approach is simple and learns feature representation from different sections of a research paper that contribute in adjudging the domain of that paper. On a certain journal our system outperforms the state-of-the-art by a wide margin ($\sim$37% in terms of accuracy). We believe that our approach is generic and with suitable adjustments could be applied to other journals having well-defined scope. Our feature set displays further potential for the development of a better journal recommender system for academic manuscripts.

Keywords: scope of an article, recommender system, desk rejection, bibliography analysis, classification

1. Introduction

Research articles are the manifestation of human scientific progress. The body of scientific knowledge moves forward with the advancements reported in the scientific articles published in different scientific journals or conferences. For years peer review has been the formal part of scientific communication that validates the quality of a scientific research article. To get published in a reputed journal or conference a certain research article goes through a series of discrete filtering steps. The first step in the peer review process is the initial screening usually performed by the editor(s). It is the job of the editor, who is also an expert in the particular field to take decisions whether an article should be rejected without further review or forwarded to expert reviewers for meticulous evaluation. The reviewers put their critical thinking, experience and knowledge for evaluation of the manuscript that finally decide the fate of the manuscript under review. The tasks of editors become too tedious to go through all the submissions in the initial screening phase as the number of submissions made are increasing day by day. Naturally that results in delay in processing an article in the initial screening phase itself. Careful observations and statistics reveal that there are five factors that play important role in the initial screening phase at the editors' desk :

1. Appropriateness of the article to the journal being sent (Aim and **Scope**).
2. Quality/Standard of the article under review.
3. Percentage overlap with existing articles (Plagiarism).
4. Spelling, grammar and language of the article under review.
5. Visually discriminative features of the article such as template mismatch (article not being prepared according to journal guidelines and formatting requirements), articles not having the standard components of a proper scientific communication.

The current work focuses on developing a **machine learning based automated system that could determine whether a submitted article falls into the scope of a given journal**. That is to say if accepted for publication, whether the submitted article is appropriate for readers of the particular journal. Finding the relevance of a prospective article to a particular venue (journal or conference) is a pre-requisite before further processing in peer review cycle. An AI assisted *scope check* system would benefit both the editors and the authors. Authors could try it and see how much his/her article is appropriate for the journal in consideration. Similarly editors could use the system to judge the appropriateness of a certain article under review. We view the problem as a two-class classification problem in machine learning with the classes in-scope (IS) or out-scope (OS) attributed to a research article for a given journal. We employ our methods on the articles of Computer Networks (COMNET), The International Journal of Computer and Telecommunications Networking and report the findings. We believe our proposed methodology is generic and with obvious exceptions can be adopted for many other journal(s) which has a similar view for scope. The superiority of our approach comes from the usage of bibliographic features that contribute to a greater extent in determining the scope of an article.

1.1. Background and Motivation

Finding the right journal for a certain research article is a mandatory pre-requisite that any researcher must have to ensure in order to increase its chances of acceptance. Suitability to the scope of a journal is an essential behavior that every research article has to exhibit for the readers of that particular journal. Submitting to an inappropriate journal can incur long delay until publication and a lot of efforts and time are wasted due to its reviewing cycle. So judging beforehand whether an article falls within the scope of a prospective journal is important from the point of view of an author. Also editors of a particular journal have to go through many irrelevant or *out-of-scope* submissions which

ultimately get rejected for not finding the right audience. Statistics reveal that a vast majority of rejections from the editor's desk are due to these *out-of-scope* submissions, even when the articles have enough significance and merit for publication. We carry our analysis on a subset of 5000 desk rejected articles from 10 different computer science journals provided by Elsevier. Figure 1 shows that a substantial amount of desk rejections accounts for the articles being *out-of-scope* of the respective journal(s). Most of the

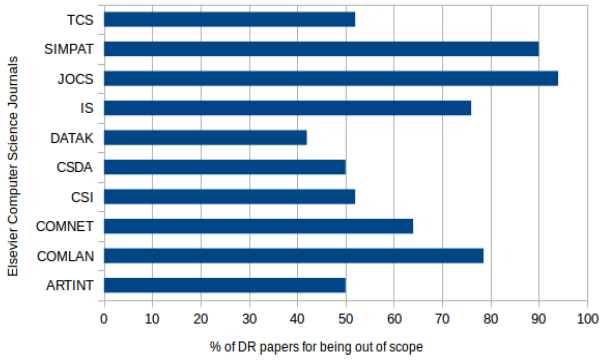

Figure 1: **Analysis of Desk Rejected (DR) papers due to *out-of-scope* for 10 different Elsevier Computer Science Journals where X : Percentage of desk rejections accounting for being *out-of-scope*, Y : Elsevier Journal Names (TCS → Theoretical Computer Science, SIMPAT → Simulation Modeling Practice and Theory, JOCS → Journal of Computational Sciences, IS → Information Sciences, DATAK → Data and Knowledge Engineering, CSDA → Computational Statistics and Data Analysis, CSI → Computer Standards and Interfaces, COMLAN → Computer Languages, ARTINT → Artificial Intelligence)**

reputed journal publishers have their own systems that suggest relevant journals to an author against her work. Examples could be given of Journal Finder by Elsevier[1], Springer Journal Suggester[2], EDANZ Journal Selector[3], etc. Also some web-services like JANE (Journal/Author Name Estimator)[4] (Schuemie and Kors, 2008), eTBLAST (Errami et al., 2007), GoPubMed (Doms and Schroeder, 2005), HubMed (Eaton, 2006), Pubfinder (Goetz and von der Lieth, 2005), etc. suggest relevant biomedical literatures from PubMed[5] or MEDLINE[6] databases upon user query (typically the title and abstract of the article for which the user wants to find a suitable journal). These systems mostly rely on domain specific vocabulary match between the prospective article and different journals to generate a suitable match. Users generally have to submit their article title, abstract and/or keywords to get a list of potential journals where they could submit their article. The present work significantly differs from these systems in the sense that it augments the keyword based approach with clustering and bibliography analysis for measuring the suitability of a certain research article to a specific journal. Also the present work relies on classification framework which was not possible with the existing systems because of the confidential nature of data(rejected articles) and inherent proprietary reasons. In fact the strength of our system stems from thorough analysis and usage of rejected articles. Our approach achieves significant improvements over the state-of-the-art system in terms of accuracy.

[1]http://journalfinder.elsevier.com/
[2]http://journalsuggester.springer.com/
[3]https://www.edanzediting.com/journal-selector
[4]http://jane.biosemantics.org/
[5]https://www.ncbi.nlm.nih.gov/pubmed/
[6]https://www.nlm.nih.gov/bsd/pmresources.html

1.2. Contribution Outline

The key-contributions of the current work are as follows :

- The problem of **scope-check** of a particular research article (Y) with respect to a journal (J) is modeled as a binary classification problem, which, to the best of our knowledge is a first attempt of its kind.
- Usage of a large set of features covering different aspects of similarity of a new research article with respect to the accepted papers in that journal. We have considered key-word based similarity, likeness in semantic space captured using a clustering process and bibliographic analysis based measurements as feature values for the new article.
- Emphasis on the idea that *bibliographic information plays a major role in determining the scope of a scientific article*.
- Experiment with a large collection of machine learning techniques to solve the binary classification problem with varying feature combinations. Almost all experiments established that existing journal recommender systems could be greatly enhanced with our defined feature set.
- Significant improvement in accuracy (**39.7 %** to be precise) over the state-of-the-art (**Elsevier Journal Finder**)(Kang et al., 2015).
- Proposal of a robust generic approach to determine the appropriateness of a manuscript to a given journal. The system could be beneficial to both the authors as well as the editors to avoid longer time delays in publication of a manuscript due to *out-of-scope* reasons.

2. DATA DESCRIPTION

We frame our investigation as a binary classification of research articles (IS : *in-scope* and OS : *out-of-scope* classes). Articles already published by a journal signify that they are within-the-scope of that journal. So accepted *in-scope* data were not difficult to acquire.But rejected articles pertaining to a given journal are confidential and difficult to obtain. **Elsevier** provided us a subset of desk rejected Computer Networks (COMNET) articles along with editor/reviewer comments signifying the reasons for rejection. We take only those desk-rejected articles which were rejected for not being within *scope* of COMNET. Topics covered in the accepted articles of a particular journal serve as the benchmark of reference which also defines the scope of the journal. Thus we reserve a substantial portion of accepted data

Table 1: **Article Statistics for COMNET Journal, Rejected OS articles → Articles rejected from desk for not being within the *scope* of COMNET**

Items	Statistics
Total Accepted COMNET articles	3878
Accepted articles used for extracting COMNET meta information	2878
Accepted articles used for training/testing	1000
Rejected **OS** articles	1000

Table 2: **Relevant information extracted from COMNET accepted articles**

COMNET Lists	Total entries
Keyword Dictionary	8887 keywords
Title List	32864 paper titles
Conference List	7276 conferences
Journal List	6124 journals
Author List	12134 authors

for generating the history information that defines the domain of operation of the journal. Data statistics are presented in Table 1.

2.1. Data Preprocessing

For the purpose of our experiments we process the accepted papers of COMNET and extract relevant information. The raw manuscripts (in .pdf) are parsed using GROBID[7] to generate the corresponding .xml versions from which the requisite information were extracted. GROBID (GeneRation Of BIbliographic Data) is a machine learning library for extracting, parsing and re-structuring raw documents such as PDF into structured TEI-encoded documents with a particular focus on technical and scientific publications. From the 2878 accepted articles of COMNET journal, we create certain exhaustive lists which are further utilized in the process of feature extraction. The statistics of relevant data extracted from the COMNET accepted articles 2016 are presented in Table 2.

2.2. Data Cleaning

The bibliographic data presented in research articles are not uniform. After extracting the conference and journal names from the .xml parsed output of GROBID (c.f. Section 2.1.), we perform the following tweaks while creating our reference lists.

- Removed editions from conference names and mapped different editions of the same conference into one.

For e.g. *Proceedings of the 8th ACM International Conference on Mobile Computing and Networking* → **ACM International Conference on Mobile Computing and Networking**

- Mapped the conference and journal names to their corresponding abbreviations[8]. .

For e.g. **JSAC → Journal on Selected Areas in Communications**

[7]https://github.com/kermitt2/grobid

[8]using Stanford Entity Resolution Framework (http://infolab.stanford.edu/serf/)

Table 3: **Top 10 author given keywords, $f(K_i)$ signifies the frequency of occurrence of keyword K_i across accepted articles of COMNET**

Keywords(K_i)	Frequency($f(K_i)$)
Remote authentication	19
Web science	15
Optimal traffic pattern	13
Packet filter	11
Survivability	10
Test criteria	10
Downlink–uplink asymmetric channel	9
High speed network	9
DoS attacks	8
Optical communications	8

CCS → ACM conference on Computer and Communications Security

- Mapped variants of certain words in conference or journal names via regular expressions.

For e.g.
Jour. → *Journal*, *Trans.* → *Transactions*, *Distrib.* → *Distributed*.
For most of the venue acronyms we refer to the JabRef[9] repository and ArnetMiner[10]. Many a times we perform cleaning manually and use regular expressions to generate the mappings.

2.3. Journal Specific Domain Information

As mentioned earlier we extract domain information from various sections of accepted articles and store them in certain lists. The history data of accepted articles would guide us to ascertain the relevance of a new article to the journal concerned.

2.3.1. Keyword Dictionary

We create a list of author given keywords (found in the Keywords section) from 2878 accepted COMNET articles and record their frequencies of occurrences across all the accepted articles of COMNET. Upon sorting by frequency we find that representative network terms appear at the top of the keyword list *viz.* Table3.

2.3.2. Title List

We create a list of all paper titles that have appeared in the reference section of all the accepted COMNET articles along with their frequencies of references from within the body of the individual article and occurrence across all the articles. Thus the value for an article title (T_i) in the exhaustive list is calculated as :

$$V(T_i) = \sum_{j=1}^{n} f_j(T_i) \quad (1)$$

where $f_j(T_i)$ corresponds to the number of times title T_i has been cited within an article j and n is the total number of accepted articles. So if a certain referred article X having title T_i has been cited 3 times within one article and 2 times within another article, then $V(T_i)$ for X would be 5.

[9]http://abbrv.jabref.org/

[10]https://www.aminer.cn/

2.3.3. Conference List

From the accepted articles of COMNET we create a list of all conferences in which articles referenced by COMNET accepted papers are published. We also record the frequency of appearance of such conferences in the reference section of the COMNET accepted articles. Thus the value for a conference (C_i) in the exhaustive list is calculated as :

$$V(C_i) = \sum_{j=1}^{n} f_j(C_i) \tag{2}$$

where $f_j(C_i)$ corresponds to the number of times conference C_i appears in the reference section of an article j and n is the total number of accepted articles. So if a certain conference C_i appears 3 times in the reference section of any particular article and 2 times in the reference section of another article, then $V(C_i)$ for C_i would be 5.

2.3.4. Journal List

From the accepted articles of COMNET we create a list of all journals in which articles referenced by COMNET accepted papers are published. We also record the frequency of appearance of such journals in the reference section of the COMNET accepted articles. Thus the value for a journal (J_i) in the exhaustive list is calculated as :

$$V(J_i) = \sum_{k=1}^{n} f_k(J_i) \tag{3}$$

where $f_k(J_i)$ corresponds to the number of times journal J_i appears in the reference section of an article k and n is the total number of accepted articles. So if a certain journal J_i appears 3 times in the reference section of one article and 2 times in the reference section of another article, then $V(J_i)$ for J_i would be 5.

2.3.5. Author List

We hypothesize that more an author publishes in a particular domain, greater is the chance that her prospective next would belong to that domain. Hence we record the publication frequency of authors in COMNET and create an author-frequency list. Author name disambiguation is a challenge here. For the current work we manually map the correct entries.

3. Methods

We propose methods that follow a culmination of unsupervised and supervised settings towards solving the problem. We view the problem as a two-class classification problem in supervised machine learning and hand craft features from the manuscripts. We experiment with several popular yet diverse classifiers from Support Vector Machines to Random Forest to classifier ensembles, investigate their performance on our extracted features and also compare with the *state-of-the-art* **Elsevier Journal Finder.**

3.1. Features

Here we discuss the features we employ for our machine learning experiments. Special focus was on bibliographic features extracted from the reference section of the candidate research article.

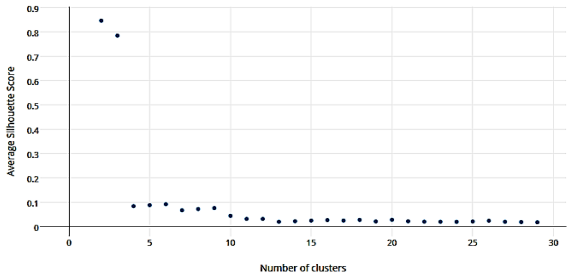

Figure 2: **Variation of Silhouette Scores for different** K

3.1.1. Keywords Score

We design this feature to emphasize the containment and relative importance of the keywords in the candidate article with respect to the Keyword Dictionary as in Section 2.3.1.. The value for this feature for a candidate article Y is thus calculated as :

$$KWScore_Y = \frac{|KW_Y \cap KW_D|}{|KW_Y|} \times \sum_{i=1}^{|KW_Y \cap KW_D|} f(K_i)$$

Where:

- KW_Y: is the set of author defined keywords in the candidate article Y
- KW_D: is the set of keywords in the Keyword Dictionary D
- $f(K_i)$: is the frequency of keyword K_i as listed in D
- $K_i \in \{KW_Y \cap KW_D\}$

3.1.2. Distance from cluster of similar articles

Accepted articles of a certain journal characterize the scope of that journal and could be grouped into some clusters representing different sub-domains/sub-themes within the journal scope. Obviously each of the accepted articles is novel and may vary in their concepts but overall they relate to some sub-themes in the semantic space. Thus the distance of a given research article from the set of clusters formed on the accepted articles could contribute significantly for in-scope determination. Any outliers to such clustering may be considered as *out-of-scope*. With this intuition we perform the steps in Algorithm 1. We use the Rapid Automatic Keyword Extractor (RAKE) (Rose et al., 2010) algorithm to extract the keywords from the scientific articles. The extracted keywords are considered as the representation of the scientific article itself. We then use the semantic power of *word2vec*[11] to vectorize the extracted keywords and thereby computed the document vector of the scientific article by averaging the keyword vectors. *K*-means clustering technique is applied on these set of vectors varying the number of clusters (K). As it would be difficult to know the value of K *apriori*, we have executed *K-means* with different values of K. The quality of the obtained partitioning at each run is measured using a popular internal cluster validity index, *Silhouette score*

[11] 1000 dimensional English wikipedia vectors

Algorithm 1 Distance from cluster boundary

1: Use RAKE to automatically extract keywords from the title, abstract and introduction sections of an article X belonging to journal J.
2: Use *word2vec* to generate the vectors of the extracted keywords from X.
3: Calculate the document vector of X by averaging all the keyword vectors from *Step 2*.
4: Repeat *Steps 1-3* for all the accepted articles of the journal J.
5: Apply K-means on the document vectors obtained from *Step 4* to generate the clusters (C_i)
6: Find the radius(r_i) of a cluster C_i as:

$$r_i = mean(distance(c_i, p_j))$$

where c_i is the centroid of cluster C_i and p_j is any point within cluster C_i.
7: Find the document vector (p_Y) of a candidate article Y using *Steps 1-3*.
8: Distance of the candidate article Y from the boundary of cluster C_i is given as :

$$D_i = distance(c_i, p_Y) - r_i$$

9: Repeat *Step 8* for all the clusters (C_i) obtained from *Step 5* to get :

$$D_Y = minimum(D_i)$$

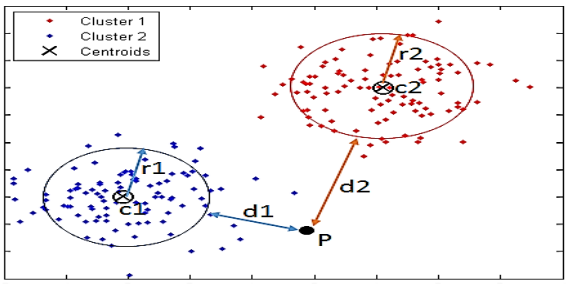

Figure 3: **Distance of article P from cluster boundary**

(Kaufman and Rousseeuw, 2009). The values of *Silhouette score* over different number of clusters are plotted (shown in Figure 2) and the K with maximum value of *Silhouette score* is selected as the optimal partitioning. Here to get the radius of a cluster we use the *mean* function instead of *maximum*, in order to neutralize the effect of extreme values in the cluster. We take the *Euclidean distance* as the distance measure in Algorithm 1. Finally we take the *minimum* of the distances of the candidate article Y from the cluster centers, in order to signify the closeness of Y to any of the clusters as in Figure 3.

3.1.3. Author Score

To calculate the author score, we take the average of the publication frequency of the authors in the concerned journal from the author list as in section 2.3.5.

Bibliographic Features

We found that the bibliography section in a research article consists certain information that contribute heavily in determining the scope of the article. We rely on the intuition that

If an article belongs to a certain domain then majority of its references would fall in that domain.

For this we create exhaustive lists of paper titles, journals and conferences from the reference section of all the accepted papers of the COMNET journal (c.f. Section 2.1.). Intuition is that information in accepted articles are the benchmark of reference. Manual inspection revealed that frequency of appearances of in-domain journals or conferences are always high in accepted articles. Bibliographic features measure the relative importance of an article or a journal or a conference with respect to the latent domain in concern.

Table 4: **Frequency of network related terms in conference and journal titles referenced by COMNET accepted articles**

Terms	$Freq_{Conf}$	$Freq_{Jour}$
Network/s/ing	7984	526
Communication/s	4256	1007
Mobile	1763	210
Wireless	2473	229
Internet	1365	314
Distributed	1055	80
Adhoc	658	52

3.1.4. Title Score

From the exhaustive list of paper titles as discussed in Section 2.3.2. we calculate the Title Score (T_Y) of a candidate article Y as :

$$T_Y = \sum_{i=1}^{m} V(T_i)$$

where m is the total number of references in Y. $V(T_i)$ is calculated using Equation 1.

3.1.5. Conference Score

Similarly from the exhaustive list of conferences as discussed in Section 2.3.3. we calculate the Conference Score (C_Y) of a candidate article Y as :

$$C_Y = \sum_{i=1}^{m} V(C_i)$$

where m is the total number of conference references in Y. $V(C_i)$ is calculated using Equation 2.

3.1.6. Journal Score

Likewise from the exhaustive list of journals as discussed in Section 2.3.4. we calculate the Journal Score (J_Y) of a candidate article Y as :

$$J_Y = \sum_{i=1}^{m} V(J_i)$$

where m is the total number of journal references in Y. $V(J_i)$ is calculated using Equation 3.

All of these features are normalized corresponding to their maximum values. Table 4 signifies the prevalence of domain information in the bibliography section of accepted articles for a particular journal.

4. Evaluation

We view the problem of determining *scope of a scientific article* as a binary classification problem. To evaluate the performance of our system we employ a range of classifiers on our feature set using WEKA(Hall et al., 2009).We use the following algorithms with their default parameters: Naive Bayes (NB)[12] (John and Langley, 1995), Decision Tree (DT)[13] (Quinlan, 1993), Support Vector Machine (SVM)[14] (Platt, 1998), Logistic Regression (LR)[15] (le Cessie and van Houwelingen, 1992), Multi Layer Feed Forward Neural Network(MLFN)[16] (Rumelhart et al., 1985) and Random Forest (RF) [17] (Breiman, 2001). Also we experiment with the stacked ensembles of NB, DT, MLFN as base with SVM as meta; NB, RF, MLFN base with LR as meta; Decision Tree and SVM as base with RF as meta; DT, MLFN, SVM as base with DT as meta with the later yielding the best performance. Majority voting ensemble with these classifiers yield comparable performance (see Table 5). We report the average of the *stratified* 10-fold cross validation results with 2000 instances (1000 accepted and 1000 rejected) in Table 5. We also compare the classification performance of our system with the *state-of-the-art* **Elsevier Journal Finder** (Kang et al., 2015) on the same dataset and report results.

4.1. Experimental Setup

Elsevier Journal Finder (Kang et al., 2015) is a *state-of-the-art* recommender system provided by Elsevier solutions to the academic fraternity that recommends highly relevant journals to the authors for their papers. **Elsevier Journal Finder** takes as input the *Title* and *Abstract* of a prospective scientific article and presents a list of 10 relevant Elsevier journals to the user as output which s/he may consider for submitting her/his article. Although the recommended journals are limited only to the Elsevier publishing house, but it is to be noted that Elsevier has more than 2900 peer-reviewed journals that cover almost all the major scientific domains. We follow the heuristics shown below in determining the predicted class label of a COMNET article Y subjected to **Elsevier Journal Finder**:

If COMNET appears in the list of suggested journals against the subjected article Y $\rightarrow$ Article is ***In-Scope*** of COMNET

otherwise, **Elsevier Journal Finder** deems Y to be ***Out-of-Scope*** for COMNET

We have the true class labels of the COMNET articles as:

If article Y is published in COMNET $\rightarrow$ Article is ***In-Scope*** of COMNET

[12] using Kernel Density Estimator and Supervised Discretization as all the attributes are numerical

[13] C4.5 Decision Tree with Confidence Threshold set to 0.25 and minimum number of instances per leaf to 2

[14] John Platt's implementation of Sequential Minimal Optimization with polynomial kernel

[15] multinomial LR with ridge estimator

[16] # of hidden units=2, loss function$\rightarrow$ Mean Squared Error, activation function$\rightarrow$ Sigmoid and Conjugate Gradient Descent as optimization function

[17] RF of 100 trees with minimum number of instances per leaf set to 1

		Predicted	
		COMNET in Top 10?	
Actual	COMNET	YES (In-Scope)	NO (Out-Scope)
	Accepted (In-Scope)	*True Positive*	*False Negative*
	Rejected (Out-Scope)	*False Positive*	*True Negative*

Figure 4: **Confusion Matrix for Elsevier Journal Finder**

If article Y is actually rejected by Editor of COMNET due to *out-of-scope* $\rightarrow$ Article is ***Out-Scope*** of COMNET

The scenario is best depicted in Figure 4. For our system we have the class labels predicted by the different classifiers against the true class labels.

4.2. Results and Discussion

Our results (c.f. Table 5) show the effectiveness of our feature set, especially the *Bibliographic features* upon the problem in hand. Our best performing algorithm corresponds to a stacked ensemble classifier (*Decision Tree, Multi Layer Perceptron and Support Vector Machine in the base with Decision Tree as meta learner*) which outperforms the **Elsevier Journal Finder** system by a margin of **39.7 %** in terms of accuracy, a significant improvement over the current *state-of-the-art*. Due to restriction in bulk access we could not experiment with other venue recommender systems as specified in Section 1.1. Also recommender systems are publisher specific, hence a specific journal would not be recommended by recommender systems of other publishers. The extensive set of results reported in Table 5 are to justify the superiority of our method using a wide and varied set of classification algorithms. Thorough and careful analysis of the results (c.f. Table 5) led us to the following observations:

(a) From our ablation study we see that for all the frameworks (individual classifiers or ensembles), augmenting *Bibliographic* features has induced significant improvement over the other features. This is due to the fact that *Bibliographic feature* values were deduced from within the body section of the scientific articles. *When a certain portion of a scientific article cites an in-domain reference, the scope of that portion is influenced by the domain of that reference. That is to say, the domain of the cited reference exerts local influence on that portion of the scientific article*. So if many in-domain references are cited in distributed portions of a scientific research article, quite possibly the entire scientific research article falls in the same domain. We measure *in-domain* or *in-scope* by simply counting occurrences of features (as discussed in Section 3.1.) across a certain journal (COMNET here).

(b) To emphasize detection of *out-of-scope* articles we look into the *Recall* for *out-of-scope* class, **R(OS)**, and observe that for all the algorithms, *Bibliographic features* are more sensitive in detecting *out-of-scope* articles. The contribution of each feature could be seen in Figure 5.

(c) For almost all the classifiers our feature combination outperforms the **Elsevier Journal Finder** in terms of precision, recall and accuracy values. This we could attribute to the fact that Elsevier Journal Finder only considers the *Title* and *Abstract* portions of a research article and uses

Table 5: **Classification results for different feature combinations, P → Precision, R → Recall, F → F-Score, Auth. → Author Score, IS → In-Scope, OS → Out-Scope. Figures in bold indicate the best performance achieved.**

Classifier & Features	P(IS)	P(OS)	P(avg)	R(IS)	R(OS)	R(avg)	F(IS)	F(OS)	F(avg)	Accuracy	Kappa
Elsevier Journal Finder	53.9	34.0	43.9	45.3	43.1	44.2	49.2	38.0	43.6	44.4	0.50
Naïve Bayes (NB)											
Keywords feature (KW)	69.1	79.0	73.1	93.0	38.7	71.0	79.3	52.0	68.2	71.0	0.34
Cluster Distance (CD)	61.2	50.7	57.0	88.3	17.6	59.7	72.3	26.2	53.6	59.7	0.06
Bibliographic features (Bib)	87.1	56.6	74.8	54.0	88.2	67.9	66.7	69.0	67.6	67.9	0.39
KW + Bib	88.3	59.0	76.4	58.0	88.7	70.4	70.0	70.8	70.4	70.4	0.43
CD + Bib	87.4	58.7	75.8	58.0	87.7	70.0	69.7	70.3	70.0	70.0	0.42
KW + CD + Bib.+ Auth.	87.2	61.5	76.8	63.3	86.3	72.6	73.4	71.8	72.7	72.6	0.47
Support Vector Machine (SVM)											
Keywords feature (KW)	67.8	85.9	75.1	96.3	32.8	70.6	79.6	47.5	66.6	70.6	0.32
Cluster Distance (CD)	59.6	100	76.0	100	0.5	59.7	74.7	1.00	44.9	59.7	0.01
Bibliographic features (Bib)	81.1	68.5	76.0	77.0	73.5	75.6	79.0	70.9	75.7	75.6	0.50
KW + Bib	78.6	78.6	78.6	88.0	64.7	78.6	83.0	71.0	78.1	78.6	0.54
CD + Bib	82.4	72.7	78.5	81.0	74.5	78.4	81.7	73.6	78.4	78.3	0.55
KW + CD + Bib. + Auth.	78.0	79.1	78.5	88.7	63.2	78.4	83.0	70.3	77.9	78.4	0.54
Multi Layer Perceptron (MLP)											
Keywords feature (KW)	68.8	70.6	69.5	88.3	41.2	69.2	77.4	52.0	67.1	69.3	0.32
Cluster Distance (CD)	63.1	55.7	60.1	85.7	26.5	61.7	72.7	35.9	57.8	61.7	0.13
Bibliographic features (Bib)	80.4	75.9	78.6	85.0	69.6	78.8	82.7	72.6	78.6	78.7	0.55
KW + Bib	83.1	74.0	79.4	82.0	75.5	79.4	82.6	79.8	79.4	79.4	0.57
CD + Bib	80.1	76.5	78.6	85.7	68.6	78.8	82.8	72.4	78.6	78.7	0.55
KW + CD + Bib + Auth.	81.4	76.0	79.2	84.7	71.6	79.4	83.0	73.7	79.3	79.4	0.57
Logistic Regression (LR)											
Keywords feature (KW)	71.9	62.8	68.2	77.7	55.4	68.7	74.7	58.9	68.3	68.7	0.34
Cluster Distance (CD)	62.2	52.7	58.4	85.7	23.5	60.5	72.1	32.5	56.1	60.5	0.10
Bibliographic features (Bib)	82.6	72.5	78.5	80.7	75.0	78.4	81.6	73.7	78.4	78.3	0.55
KW + Bib	82.3	78.7	80.8	86.7	72.5	81.0	84.4	75.5	80.8	80.9	0.59
CD + Bib	81.7	74.7	78.9	83.3	72.5	79.0	82.5	73.6	78.9	78.9	0.56
KW + CD + Bib + Auth.	80.6	76.8	79.0	85.7	69.6	79.2	83.0	73.0	79.0	79.2	0.56
Random Forest (RF)											
Keywords feature (KW)	67.8	85.9	75.1	96.3	32.8	70.6	79.6	47.5	66.6	70.6	0.32
Cluster Distance (CD)	65.1	49.2	58.7	66.7	47.5	58.9	65.9	48.4	58.8	58.9	0.14
Bibliographic features (Bib)	80.2	80.7	80.4	89.0	67.6	80.4	84.4	73.6	80.0	80.3	58.1
KW + Bib	83.6	83.9	**83.7**	90.3	74.0	**83.7**	86.9	78.6	**83.5**	**83.7**	0.65
CD + Bib	80.8	81.9	81.2	89.7	68.6	81.2	85.0	74.7	80.8	81.2	0.59
KW + CD + Bib + Auth.	83.8	82.6	83.3	89.3	74.5	83.3	86.5	78.4	83.2	83.3	0.64
Decision Tree (DT)											
Keywords feature (KW)	67.8	85.9	75.1	96.3	32.8	70.6	79.6	47.5	66.6	70.6	0.32
Cluster Distance (CD)	62.4	50.0	57.4	80.7	28.4	59.5	70.3	36.3	56.5	59.5	0.10
Bibliographic features (Bib)	80.3	85.5	82.4	92.3	66.7	81.9	85.9	74.9	81.5	81.9	0.61
KW + Bib	82.9	84.1	83.4	90.7	72.5	83.3	86.6	77.9	83.1	83.3	0.64
CD + Bib	79.8	85.4	82.1	92.3	65.7	81.5	85.6	74.2	81.0	81.5	0.60
KW + CD + Bib + Auth.	83.6	83.9	**83.7**	90.3	74.0	**83.7**	86.9	78.6	**83.5**	**83.7**	0.65
Voting Ensemble											
All features with NB+DT+MLP+RF+SVM+LR	83.6	81.7	82.9	88.7	74.5	82.9	86.1	77.9	82.8	82.9	0.64
Stacked Ensemble											
All features with **Base**: NB+DT+MLP **Meta**: SVM	83.0	80.6	82.1	88.0	73.5	82.1	85.4	76.9	82.0	82.1	0.62
Base: NB+RF+MLP **Meta**: LR	83.9	81.8	83.1	88.7	75.0	83.1	86.2	78.3	83.0	83.1	0.64
Base: DT+SVM **Meta**: RF	84.2	82.4	83.5	89.0	75.5	83.5	86.5	78.8	83.4	83.5	0.65
Base: DT+MLP+SVM Meta: DT	84.2	84.1	**84.1**	90.3	75.0	**84.1**	87.1	79.3	**84.0**	**84.1**	0.66

the **Elsevier Finger Print Engine**[18] based on identification of *Noun Phrases* from those sections. Our method goes beyond this intuition and uses the *Bibliographic* information from within the body of the research article.

(d) We also carry our experiments with classifier ensembles (majority voting and stacked classifiers). Ensemble of (***Decision Tree, Multi Layer Perceptron and Support Vector Machine as base with Decision Tree as meta***) performs the best. Other ensembles deliver comparable performance to the best performing individual classifiers (Random Forest and Decision Tree).

(e) The **Keywords Feature** is more sensitive in identifying *in-scope* articles (*Recall* for IS is high in comparison to other features for almost all the classifiers). This feature looks for keyword match with the already published research articles (which we consider to be *in-scope*) and hence have a higher affinity towards detecting *in-scope* articles.

(f) Distance from the cluster boundary did not work well as well-defined compact clusters could not be formed due to the fact that each accepted article is *novel* and would have very less in common. Our observation also reveals that scope of a journal is time-variant. Clustering over more recent papers would have captured the compactness better. Also word embeddings generated on Computer Science domain specific vocabulary could improve performance. Our feature suffered from many out-of-vocabulary words.

(g) Author feature proved weak since an author belonging to the same domain may publish in related domain venues other than COMNET.

(h) Our features are not entirely independent to one another. Hence *Naive Bayes* do not perform as compared to the others but still manages to successfully outperform Elsevier Journal Finder on the same data.

[18]https://www.elsevier.com/solutions/elsevier-fingerprint-engine

Table 6: **Statistical t-test results of different classifiers against Elsevier Journal Finder**

Classifiers	$P(Avg)$	$R(Avg)$	$Acc.$
Random Forest (RF)	4.06E-102	5.43E-102	6.60E-102
Decision Tree (DT)	2.77E-102	3.69E-102	4.47E-102
Base: DT+MLP+SVM Meta: DT	1.89E-102	2.51E-102	3.04E-102

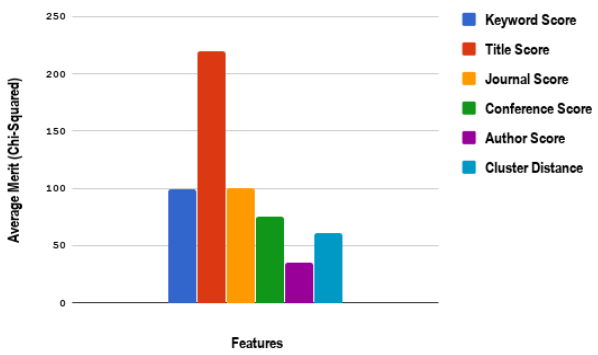

Figure 5: **Significance of features observed by ranking features based on Chi-Squared values**

4.3. Tests of Significance

From Table 5, we see that for all the classifiers, combination of our feature set surpasses the *state-of-the-art*. To prove the effectiveness of our feature set with different classifiers against the *state-of-the-art* **Elsevier Journal Finder**, we conduct statistical significance test (t-test) at 5 % significance level (Fisher, 1956). We subject the overall average accuracy, precision and recall on the same dataset produced by 20 consecutive runs of our best performing classifiers (1. Random Forest, 2. Decision Tree and 3. Classifier Ensemble (Base: DT+MLP+SVM, Meta: DT)) against the corresponding measures of Elsevier Journal Finder to verify for the statistical significance measures. Now between each two groups (a group corresponding to Elsevier Journal Finder and another group corresponding to any of the classification algorithm stated above) the p-values produced by **t- test** are reported in Table 6. As null hypothesis we assume that there is insignificant difference between mean values of two groups. According to alternative hypothesis there are significant differences in the mean values of two groups. It can be seen that all of the p-values in Table 6 are less than 0.05 (5 % significance level). It strongly indicates that the null hypothesis is wrong and the better mean values of the accuracy, precision and recall produced by the respective classification algorithms on our feature set are statistically significant and have not occurred by chance.

5. Error Analysis

Thorough analysis revealed that errors committed by our system mostly arise from the following cases:
(a) Presence of substantial amount of uncited references within the body of the scientific research article, although the references are very much in-domain. Our system predicts such cases as *in-scope* but true label is *out-of-scope*.
(b) Accepted articles having content and references very distantly related to Computer Networks. Our survey and understanding reveal that *scope* of a journal changes with time and gets defined after some period. So there are some initial cases where certain seemingly *out-of-scope* articles, remotely related to computer networks got accepted. Evidently our system fails to detect those cases.
(c) Research articles having very less number of references. In such cases the very significant *bibliographic* features did not contribute.
(d) References not in proper format. These references are not parsed appropriately by GROBID. Hence our system could not capture requisite information.
(e) Some research articles related to *Graph Theory* or *Cryptography* or *Cloud Computing*, having network related terms in their body as well as referencing *networks* conferences or journals, but deemed *out-of-scope* by the editors of COMNET for their content.
Our approach is not suitable for journals which have a very broad scope or journals which accept review papers from different subjects of a discipline such as *Computer Science Review* or *ACM Computing Surveys*. Also the definition of *Scope* is very much journal dependent. Not always *Scope* of an article implies plain *textual relevance* of the article to the accepted articles of the intended journal. Some journals (for e.g., *Simulation Modelling Practice and Theory*) are multi-disciplinary and cater to a wide-diversity of topics but look into certain specific characteristics (experiments, simulation, applicability) within a research article.

6. Conclusions and Future Works

The current work unravels an important insight into determination of *scope* of a scientific article by looking into its several aspects like extracted key-words, references cited, similarity with respect to the set of accepted articles. Our claim has been backed by sufficient empirical evidences, and to the best of our knowledge there has been no such work reported so far which has investigated this phenomenon. The proposed approach is generic and could be applied across other journals with the exception to those which accept survey articles on different aspects of a particular discipline. The proposed system could aid in initial screening of a large number *out-of-scope* articles that reaches the editor's desk and hence speed-up the overall process of peer review. Publishing houses could internally employ these proposed methods to design a system for both the authors and editors to curb *Desk Rejections* and proceed towards the more ambitious intent of employing artificial intelligence in peer review. In future, the authors would like to concentrate on deriving more features from the reference section, look into the avenues for improvement in the clustering approach and apply these knowledge across some journals of other disciplines as well.

7. Acknowledgements

We are thankful to Elsevier for providing us the required data to conduct our experiments. The first author is supported by Visvesvaraya PhD Scheme, an initiative of Ministry of Electronics and Information Technology (MeitY), Government of India.

8. Bibliographical References

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Doms, A. and Schroeder, M. (2005). Gopubmed: exploring pubmed with the gene ontology. *Nucleic acids research*, 33(suppl 2):W783–W786.

Eaton, A. D. (2006). Hubmed: a web-based biomedical literature search interface. *Nucleic acids research*, 34(suppl 2):W745–W747.

Errami, M., Wren, J. D., Hicks, J. M., and Garner, H. R. (2007). etblast: a web server to identify expert reviewers, appropriate journals and similar publications. *Nucleic acids research*, 35(suppl 2):W12–W15.

Fisher, R. A. (1956). Mathematics of a lady tasting tea. *The world of mathematics*, 3:1512–1521.

Goetz, T. and von der Lieth, C.-W. (2005). Pubfinder: a tool for improving retrieval rate of relevant pubmed abstracts. *Nucleic acids research*, 33(suppl 2):W774–W778.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

John, G. H. and Langley, P. (1995). Estimating continuous distributions in bayesian classifiers. In *Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345, San Mateo. Morgan Kaufmann.

Kang, N., Doornenbal, M. A., and Schijvenaars, R. J. (2015). Elsevier journal finder: recommending journals for your paper. In *Proceedings of the 9th ACM Conference on Recommender Systems*, pages 261–264. ACM.

Kaufman, L. and Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons.

le Cessie, S. and van Houwelingen, J. (1992). Ridge estimators in logistic regression. *Applied Statistics*, 41(1):191–201.

Platt, J. (1998). Fast training of support vector machines using sequential minimal optimization. In B. Schoelkopf, et al., editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press.

Quinlan, R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA.

Rose, S., Engel, D., Cramer, N., and Cowley, W. (2010). Automatic keyword extraction from individual documents. *Text Mining*, pages 1–20.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1985). Learning internal representations by error propagation. Technical report, DTIC Document.

Schuemie, M. J. and Kors, J. A. (2008). Jane: suggesting journals, finding experts. *Bioinformatics*, 24(5):727–728.