Data mining approach for identifying High-quality journals in Computer Science

J.K.D.B.G. Jayaneththi [1] and B.T.G.S Kumara[2]

[1] *Department of Computing and Information Systems, Faculty of Applied Sciences, Sabaragamuwa University of Sri Lanka, Belihuloya, Sri Lanka*

[2] *Department of Computing and Information Systems, Faculty of Applied Sciences, Sabaragamuwa University of Sri Lanka, Belihuloya, Sri Lanka*

[#] J.K.D.B.G. Jayaneththi; bgayashani@gmail.com

Abstract— *In the present scientific world, most of the authors of scientific literature are seeking effective ways to share their research findings with large peer groups. But finding a high-quality journal to publish is a huge challenge for them. Most of the journals available today are predatory and less-quality and most of them will publish almost everything that is sent to them without proper quality control. The main aim of this study is to help the researchers in identifying the quality level of Computer Science journals by introducing a data mining approach based on six journal quality metrics: Journal Impact Factor (JIF), SCImago Journal Rank (SJR), Eigenfactor, H-index, Source Normalized Impact per Paper (SNIP) and Article Influence (AI). Further, this aims to present the best metrics to measure the quality of those journals out of the six attributes and a more accurate data mining approach based on those metrics. A sample dataset of 200 journals was used for the study. Hence there were no former defined groups for the journals and they needed to be categorized into groups based on the distribution of values of the quality attributes the K-means clustering algorithm was applied for the dataset and it was clustered into five clusters as excellent, good, fair, poor and very poor using WEKA tool. When finding the best quality metrics, Pearson's and Spearman's correlation coefficients were calculated between each attribute against JIF using IBM SPSS Statistics 20 software and it was found that JIF, SJR, and SNIP are the best attributes to measure the quality of those journals based on the high coefficient values. Again a more effective clustering model with an accuracy of 0.9171, sensitivity of 1.0000, specificity of 0.9126, f-measure of 0.5556 and g-mean of 0.9553 was developed considering only those selected three attributes.*

Keywords— **Data mining, K-means clustering, predatory journals**

I. INTRODUCTION

When considering journals with the technological development and development of the World Wide Web, a large number of journals came into account and most of them were open access journals. With this emergence of open access journals, many predatory and low-quality journals came into operation. Most of the journals which are active now only focus on earning money, in truth they will publish almost everything that is sent to them (and paid for) with little to no quality control. Jeffrey Beall (a library scientist at the University of Colorado, Denver, USA) has coined the term 'predatory publishers' to describe publishers in the scholarly publishing business who collect article processing charges and provide rapid publishing without a proper peer-review process (J.Beall, 2014). So, authors faced a lot of problems when selecting the best and high-quality journals to publish their valuable inventions.

Because of the unavailability of a proper quality control process, poorly conducted and false research studies could be published in many journals. That causes greater consequences because these papers are cited and used as the basis for further studies. Another cause for concern is that honest researchers are sometimes duped into believing that journals are legitimate and may end up publishing their valuable efforts in them. Most of these journals have false editorial boards. People with a good reputation are named as editors without their knowledge. So, the need of finding ways to identify high-quality journals has become an important matter to concern and many researchers have conducted research studies to find a better solution for this matter.

A new campaign called Think, Check and Submit was introduced with the specific intent of providing researchers the information they need to become better informed about where to publish their work. With that the researchers were encouraged to be aware about some important facts such as is the publisher easily identifiable and contactable, are the article processing charges clearly described and does the editorial board consist of recognizable names (J.Roberts, 2016).

Bavdekar and Save have advised authors to first identify the main theme and the predominant message of the article they are intending to write and the target group for who the message is meant before selecting a journal to publish (S.B.Bavdekar and S.Save, 2015).

A novel approach for ranking journals in Science and Technology domain was introduced by Jangid and his team. They have used a new metric called Influence Score for ranking the journals. It was calculated using linear regression model by considering different journal quality metrics such as H-Index, Total docs (current year), Total refs, Total cites (3 years), Citable docs (3 years), Cites per doc (2 years), Total docs (3 years), References per doc and assigning weights to each factor based on the percentage of influence the factor might have in the calculation of the Influence Score (N.Jangid, S.Saha, S.Gupta and M.Rao, 2014).

A system to evaluate academic electronic journals was introduced by Lopez-Ornelas and his team. They have used a survey which was validated by a judge-panel consisting of editors of online electronic journals and have proposed mainly seven criteria as quality of the content, continuity and periodicity, standardization, purpose and audience, timeliness and maintenance, external recognition of the publication's digital format and navigation and graphic design which can be used to evaluate and identify predatory journals (M.López-Ornelas, G.Cordero-Arroyo and E.Backhoff-Escudero, 2005).

So, the main purpose of this research study is to help the scholars and researchers to identify the high-quality journals in Computer Science subject field by developing a data mining model based on the main six quality metrics associated with journals namely JIF, SJR, Eigenfactor, H-index, SNIP, and AI score. Another important effort is to help the authors to get an idea about what are the most important quality metrics out of the main six attributes that should be concerned when measuring quality and selecting a Computer Science journal to publish.

II. Methodology and Experimental design

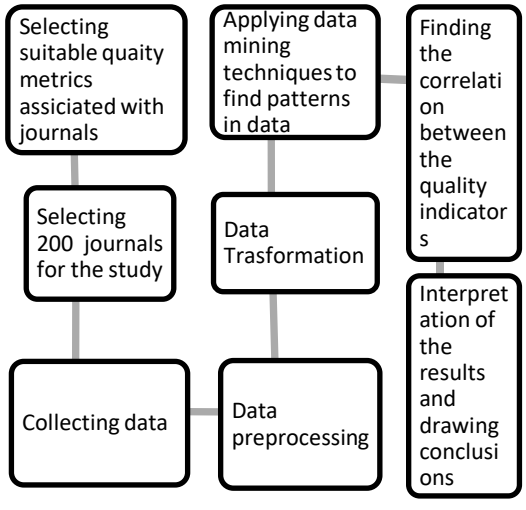

Figure 1. The design of the research

Fig 1 show the methodology which was used for the study. First suitable quality metrics were selected and they were JIF, H-index, SJR, SNIP, Eigenfactor Score and Article Influence Score. Then a sample of 200 journals was selected. After that data was collected, SJR and H-index values were collected using the SCimago Journal and Country Rank website, SNIP values were collected using the Scopus database, Eigenfactor Scores and Article Influence Scores were collected using the EIGENFACTOR.org website and the Journal Impact Factor values were collected from the SCIJOURNAL.ORG. All the values of the above attributes considered for the study were based on the year 2017.

After finishing data collecting step the data set was pre-processed. During pre-processing first the dataset was checked for missing values. As the data was manually collected by searching the databases there were no missing values in the dataset. Then the data set was tested for outliers and extreme values using the WEKA data mining tool. There were 5 outliers and 2 extreme values in the dataset and they were removed using WEKA.
Next step was data transformation. In data transformation, normalization technique was applied to transform the data into a form suitable for performing the data mining process. The min-max normalization technique was applied and data was transformed to fall between 0 and 1 for the ease of mining process.

After finishing the data pre-processing steps the suitable data mining technique was selected and applied to the dataset. The selected technique was K-means clustering. In K-means clustering defining the number of clusters is an important step. So, elbow method was used to define and validate the best fitting number of clusters. In elbow method K-means clustering algorithm was applied for a range of values of K and sum of squared errors (SSE) for each K value was calculated using WEKA. Then SSE was plotted against each K value. If the line chart looks like an arm, then the "elbow" on the arm is the best K value. (Knee point).

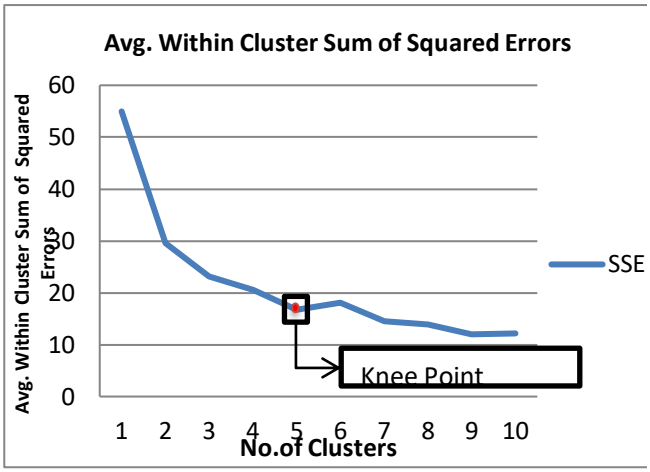

Figure 2. The elbow chart

According to the above fig. 2, it is clear that the Knee point lies at number 5. So, the optimum number of clusters is identified as five. Then the data set was clustered into 5 clusters based on the K-means clustering algorithm using the WEKA data mining tool.

In this study JIF, SJR, SNIP, H-index, Eigenfactor Score and AI were used for identifying the quality and categorizing the journals. The most acceptable parameter in the scientific world for evaluating journal quality is JIF which is calculated yearly by ISI and which is now a part of Thomson Reuters. So, comparing the other quality indicators with JIF is important for finding the most suitable metrics which can be used along with JIF to interpret the clusters.

For the comparison of the quality indicators, correlations between indices were evaluated using Pearson's correlation and Spearman's rank correlation. And all analyses were performed using IBM SPSS Statistics 20 software. Then the journals were again clustered by considering only the quality metrics which had a strong correlation with each other. And the clusters were re-interpreted for better results.

III. RESULTS

A. Results when considering all main six attributes

When considering all the six attributes the journals were clustered into 5 clusters and their cluster centroids were appeared as shown in the following table 1.

Table 1. The final cluster centroids when considering all attributes

Attribute	Cluster				
	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
SJR	0.0964	0.6306	0.2183	0.0786	0.3364
H-index	0.1684	0.6861	0.1427	0.0898	0.3073
SNIP	0.1642	0.4756	0.3265	0.1455	0.3401
EF	0.7075	0.8968	0.3602	0.2163	0.7312
AI	0.353	0.8893	0.5961	0.2366	0.7537
JIF	0.1068	0.4484	0.199	0.0889	0.2349

-The highest values
- 2nd place values
-3rd place values
- 4th place values
- The lowest values

After analysing the centroids of each cluster the clusters were interpreted as shown in the following table 2.

Table 2. Cluster interpretations

Cluster number	Cluster interpretation
Cluster 0	Poor quality journals
Cluster 1	Excellent quality journals
Cluster 2	Fair quality journals
Cluster 3	Very poor quality journals
Cluster 4	Good quality journals

The following table 3 shows sample of the journals which falls under different clusters.

Table 3. Journals appear in each cluster

Cluster	Examples
Excellent Quality Journals	• IEEE Transactions on Smart Grid • Communications of the ACM • Internet and Higher Education
Good Quality Journals	• Ocean Modelling • SIAM Journal on Computing • Computers in Industry
Fair Quality Journals	• Foundations and Trends in Information Retrieval • International Journal of Social Robotics • Computer Supported Cooperative Work
Poor Quality Journals	• Logical Methods in Computer Science

	• Control Engineering and Applied Informatics • Computer Journal
Very Poor Quality Journals	• Applied Categorical Structures • ICGA Journal • Computer Science and Information Systems

B. Results of the comparison between used quality metrics

Following table 4 shows the calculated Pearson's correlation coefficients for each of the quality attributes against JIF values and Spearman's correlation coefficients for each of the ranks against the JIF ranks.

Table 4. Correlation Statistics

Correlation statistics	**Coefficient values**
Pearson's r between JIF and SJR values	0.722
Pearson's r between JIF and H-index values	0.497
Pearson's r between JIF and SNIP values	0.731
Pearson's r between JIF and EF values	0.356
Pearson's r between JIF and AI values	0.474
Spearman's rho between JIF and SJR ranks	0.790
Spearman's rho between JIF and H-index ranks	0.472
Spearman's rho between JIF and SNIP ranks	0.771
Spearman's rho between JIF and EF ranks	0.369
Spearman's rho between JIF and AI ranks	0.355

According to the above table 4, it is clear that there is a high Pearson's statistical correlation between JIF and SNIP values (r = 0.731) as well as between JIF and SJR values (r = 0.722). And when considering Spearman's rho statistical correlation there is a high correlation between JIF and SJR ranks (r_s = 0.790), as well as between JIF and SNIP ranks (r_s = 0.771).

So, according to the correlation coefficient values it was clear that SJR and SNIP quality metrics are the best ones which can be used along with JIF, for evaluating the quality of Computer Science journals. Then the journals were again clustered by considering only the JIF, SJR and SNIP values.

C. Results when considering SJR, SNIP and JIF

When considering all the best three attributes the journals were clustered into 5 clusters and their cluster centroids were appeared as shown in the following fig 4.

Table 5. Final cluster centroids when considering JIF, SJR, and SNIP

Attribute	**Cluster**				
	Cluster 0	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**
SJR	0.1767	0.8173	0.4716	0.0678	0.2834
SNIP	0.2565	0.5747	0.3212	0.1303	0.4885
JIF	0.1709	0.5649	0.2521	0.0799	0.2928

-The highest values

- 2nd place values

- 3rd place values

- 4th place values

- The lowest values

After analysing the centroids of each cluster the clusters were interpreted as shown in the following table 6.

Table 6. New cluster interpretations

Cluster number	**Cluster interpretation**
Cluster 0	Poor quality journals
Cluster 1	Excellent quality journals
Cluster 2	Fair quality journals
Cluster 3	Very poor quality journals
Cluster 4	Good quality journals

The following table 7 shows sample of the journals which falls under the newly interpreted clusters.

Table 7. Journals appear in new clusters

Cluster	**Examples**
Excellent Quality Journals	• IEEE Transactions on Smart Grid • IEEE Communications Magazine • IEEE Wireless Communications
Good Quality Journals	• Communications of the ACM • International Journal of Social Robotics • Future Generation Computer Systems

Fair Quality Journals	• Information Systems Journal • BIT Numerical Mathematics • Telematics and Informatics
Poor Quality Journals	• Memetic Computing • Biological Cybernetics • Journal of Computational Science I
Very Poor Quality Journals	• Applied Categorical Structures • Logical Methods in Computer Science • Theoretical Computer Science

IV. DISCUSSION AND CONCLUSION

Hence clustering is an unsupervised evaluation process the validation of the clustering algorithm is very important. It is very difficult to find whether the cluster configuration is acceptable or not (M. Halkidi, Y. Batistakis and M. Vazirgiannis, 2002). So, several clustering validity indexes have been developed and in this study, the obtained clusters were validated by calculating accuracy, specificity, f-measure and g-mean.

Following table 8 shows the obtained cluster validation results when considering all the six quality attributes.

Table 8. Cluster validation results

Accuracy	0.8394
Sensitivity	1.0000
Specificity	0.8218
Recall	1.0000
F-measure	0.5507
G-mean	0.9066

According to the above table 8, it is clear that the used clustering model is good and best suits for clustering the selected journals. And at most of the time, it gives the results as we predicted.

Following table 9 shows the obtained cluster validation results when considering the selected best three attributes.

Table 9: Validation results of the new clusters

Accuracy	0.9171
Sensitivity	1.0000
Specificity	0.9126
Recall	1.0000
F-measure	0.5556
G-mean	0.9553

According to the above table 9, it is clear that that the clustering model which uses JIF, SJR and SNIP to cluster the Computer Science journals is the most suitable and best fitting clustering model for clustering the selected Computer Science journals. That clustering model provides better results than the model which uses all the six attributes to cluster the Computer Science journals.

According to the obtained key results, we can conclude that the researchers and authors of Computer Science subject field can use the data mining models proposed in this study to find the best and most suitable journals to publish their valuable and innovative research findings. As well as they can gain an idea of what are the most suitable quality attributes they should consider when measuring the quality of Computer Science journals. So, they can publish their research papers in a well-recognized and high-quality journal and contribute to the global knowledge in an effective way. As well as they can achieve their academic career life targets and develop a reputed profile for them.

REFERENCES

S.B.Bavdekar and S.Save, 2015. Choosing the Right Journal for a Scientific Paper. *Journal of The Association of Physicians of India,* June, Volume 63, pp. 56-59.

J.Beall, 2014. Scholarly open-access publishing and the problem of predatory publishers. *Journal of Biological Physics and Chemistry ,* March, Volume 14, p. 22–24.

M. Halkidi, Y. Batistakis and M. Vazirgiannis, 2002 . Cluster Validity Methods : Part I. *SIGMOD Rec,* July, 31(2), pp. 40-45.

N.Jangid, S.Saha, S.Gupta and M.Rao, 2014. *Ranking of journals in Science and Technology Domain: A novel and*

computationally lightweight approach. Bangalore, s.n., p. 57 – 62.

M.López-Ornelas, G.Cordero-Arroyo and E.Backhoff-Escudero, 2005. Measuring the Quality of Electronic Journals. *Electronic Journal of Information Systems Evaluation,* 8(2), pp. 133-142.

J.Roberts, 2016. Predatory Journals: Illegitimate Publishing and Its Threat to All Readers. *The journal of Sexual Medicine,* October, Volume 13, pp. 1830-1833.