Knowledge Discovery from Scientometrics Database

Muhammad Shaheen, *Member, IAENG,* Maliha Mehmood

***Abstract*—This paper proposed a unified ranking system for classification of scientific content. A list of parameters for unified ranking of scientific content (journals) is prepared first. A few quality indicators for evaluating the quality of journals are taken from existing parameters which are in use. These indicators include eigen factor, audience factor, impact factor, article influence, and citations. We also proposed one new metric, prestige of journal (PoJ) for the evaluation of journals. The values of different journals for the proposed indicators including the newer one is stored in an integrated database. A popular data mining technique for unsupervised classification named K Means clustering is applied to group the journals in different clusters. Clustering is an unsupervised classification technique in which the un-clustered classes do not bear any label. The clusters are labeled to find the exact rank of a science journal by using a state of the art technique of labeling clusters developed by the author of this paper. The experimental design of the paper is done and the parameters to evaluate the experiment are finalized. The results obtained from different experiments are in process yet and will be published in the extended version of this paper.**

***Index Terms*—Data mining, Scientometrics, Impact factor, Classification, Clustering.**

I. Introduction

Objective evaluation cannot be done without having measurable metrics. These metrics are good to be unbiased. Recent studies are made to address these biases to make the process of evaluation more reliable. Quality of a research journal is determined on the basis of one or many of the metrics including number of research publications in the journal, repute of the journal and database where the publications are indexed and citations etc. of the research papers. The values of these indicators are collected from the databases which are maintained by different organizations. These database include but are not limited to ISI web of Knowledge [3], Scopus [7] and SCImago Journal Rank (SJR) [13]. The measures which are mostly used for evaluation of the journals are impact factor (IF) [5], prestige of a journal [24], eigenfactor, article influence[**http://admin-apps.webofknowledge.com/JCR/help/h_impfact.htm**], 5-year impact factor and total citations [18].

Database technology in computer systems is used to store the larger contents related to these Scientometrics. Data mining and big data with its evolution are used for extracting useful knowledge from these databases. Data mining classifies clusters and extracts patterns from larger datasets through a computational process. [10]. Dis-sorted data is converted to meaningful information, patterns and classes that, after applying analytical techniques yield knowledge and implicit patterns [10]. These techniques can broadly be divided into supervised and unsupervised classification techniques. Supervised classification techniques are those in which labels of data classes are known and vice versa [10],[11]. In this study we used an unsupervised classification technique named clustering with a modification in the existing algorithm of the K Means clustering proposed in [10],[12].

In this paper, the quality evaluation parameters that are used to evaluate journals are stored in a database. These indicators rank a journal and different journal ranking systems are developed. One new factor that is "prestige of journal (PoJ)" is introduced in this paper. The journals are then ranked by using labeling clusters method which was proposed by the author of this paper in [10].

Rest of the paper is organized in a systematic way. The literature review is presented in Section 2. Data mining algorithm used in this study are given in Section 3. Proposed Method is given in Section 4. Section 5 contains Experiment and Results. The work is concluded in the final section).

II. Literature Review

A. Scientometrics

There has been significant debate and disagreement on evaluating the output of the research performed by groups and individuals. Deliberations are still needed for the multifaceted task of assessing the research. For example, Impact factor is widely used though it does not cover usage metric or prestige of a journal [19]. Impact Factor counts self-citation as a full citation in a research paper. It is calculated for one-year period by averaging number of citations per paper during the previous two years [13].

Carl Bergstrom introduced eigenfactor score in 2007, which uses data of the citations to measure the impact by a certain journal. Self-citation is not included in it and that's why it has lengthened the intended period to five

Prof. Dr. Muhammad Shaheenhas been serving as Dean of Faculty of Engineering and Information Technology at the Foundation University Islamabad (FUI) Pakistan, New Lalazar RawalpindiPakistan. He is a professional member of IAENG (phone: 92-331-4525045;e-mail: dr.shaheen@ fui.edu.pk).

Ms. Maliha Mehmood is working on her thesis towards completion of Masters in Science (Computer Science) under the supervision of Prof. Dr. M. Shaheen at Foundation University Islamabad. She is also working as TRA at the university. (phone: 92-51-5152268; email: maliha.mehmood@fui.edu.pk).

ISBN: 978-988-14048-6-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

years. Article Influence (AI) [21] calculates the mean effect of journal articles in the first five years after being published. To calculate AI, a journal's eigenfactor score is divided by the total number of articles in that journal. Some authors are of the view that quantity should be taken into account as well, in [1] the author argued that the ranking of a journal should evaluate not only the quantity but its quality. Impact Factor (IF) measures the number of citations in a paper. Gonzalez et al. proposed ranking scheme of SJR that measures the significance of all the citations giving a realistic evaluation of the status of a journal in the community.

H-index is another indicator that is used to measure the scientific productivity of a journal. H-index calculates the deceptive scientific influence of a scientist or a journal as well as the actual scientific productivity [22].

From the above, it is concluded that there exist a number of metrics which can be used for evaluating the rank of a journal. But there still exist some factors which can add bias to the rank of a journal. The major biases are a bias of self-citation and bias of database indexing.

B. Data Mining

Data Mining (DM) is a technique of discovering knowledge from databases. The process of discovering knowledge through data mining is better explained in five stages given in [4],[8]. This process is applied to understandable data. In the context of data management, discovery is done on data to find meaningful patterns [9],[17]. While text mining, numerals mining, web mining, graph mining among others are increasing its acceptance in the field of data mining, knowledge discovery in databases that are now conventional has incorporated analytical capabilities into various data formats. Data-mining doesn't extend the existing databases; it applies analytical techniques to databases to extract hidden and non-trivial patterns from data. Its applications are broader and not limited to web click streams, human genome research, banking sector, telephone industry and space science [15],[16].

Data mining techniques when used for classification of datasets are separated into three broader categories. (1). Supervised classification. (2). Unsupervised classification. (3). Semi-Supervised Classification. If the class labels are provided to the user in input datasets, then classification technique used is supervised classification and in case of unavailability of class labels unsupervised classification techniques are used [20]. Semi-supervised classification is a mixture of both supervised and unsupervised techniques. One type of Unsupervised Classification is clustering. Clustering is a technique of grouping data items into pre-defined number of clusters on the basis of some implicit measure for which data is not needed. These implicit measures include distance, position and rank. Several clustering techniques have already been developed as well as reviewed in [6],[15]. K-Means clustering is a clustering technique in which the random cluster centers (pivot points) from the datasets are picked. The remaining data points are placed in different clusters on the basis of distance of the points from the pivotal points. The process of calculating Euclidean distance of the points from pivots is cycled until the convergence is achieved. The details of K-Means clustering is given in the next section.

III. Data Mining Algorithms

A. K-Means Algorithm

In K-Means clustering, data is clustered on the basis of similarities between the data elements. Euclidean distance is used as a metric to discover similarities. The equation of Euclidean distance is given as Eq 1.

$$D_i = a_0 + \sum_{n=1}^{k} \sum_{j=1}^{n} (C_k - X_j) \quad (1)$$

Every database record is represented by a point with multiple coordinates in Cartesian space. One database tuple is signified by only one point on the Cartesian system. Every coordinate in the system denotes one attribute. On the basis of the Euclidean distance from the cluster center of a certain point, specific points are allocated to K number of clusters. At the start of K Means process, cluster centers are chosen at random. Euclidean distance of each point from pivotal point is computed. The point is allocated to a cluster center on the basis of minimum distance. Once the points are allocated to the cluster center, mean of all allocated point is computed to come up with a new point which is considered as new cluster center. In this way, the cluster centers which were initially selected at random are re-adjusted to its expected position [23].

The process of adjustment of cluster centers to the point of convergence is the crux of the technique. All the data points are converted to a fixed number of clusters at the end of K Means clustering but all these clusters are unlabeled. The points are there in the form of groups but there isn't any tag of identification marked on these groups. The author of this paper proposed a technique for labeling unlabeled clusters on the basis of some ranking mechanism proposed in [10],[12]. The technique is given in the next section.

B. Applying Labels on Clusters

As mentioned earlier, clusters do not contain class labels. The labels of clusters are not meaningful and do not give any clue about the class/ category of data points contained inside. A technique of labeling clusters is proposed in [10][12] which applies labels on clusters after applying K Means technique. In this technique, the prominence of an attribute is determined by finding its correlation with the independent variable by using equation 2. The step wise procedure for labeling clusters is given below;

1. Apply K means clustering to obtain number of required clusters from the provided dataset.
2. Identify independent variables from the dataset according to which class label could be allocated.
3. Apply correlation analysis on every attribute of the dataset with the variable selected in Step 2. The result of correlation analysis is called weight value of that attribute. The formula of correlation analysis is given in Eq 2.

ISBN: 978-988-14048-6-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

$$Corr\ (A_i, A_j) = \sum_{i=1}^{n} {}^{P(A_i, A_j)}/_{n} \qquad (2)$$

A_i is the dependent variable (journal indicators) and A_j is independent.

4. Multiply weight value obtained in Step3 to the corresponding dependent variables.

Allocate data point to that cluster for which frequent membership value is the highest. The details of frequent membership functions can be found in [10].

The resulting correlation value is multiplied by the actual value of the attribute to find its standing in descending order. The one with the highest new actual value is placed at the top of the list. Labels can be allocated to the cluster on the basis of top-down order. These labels are not designed according to the content of data within a cluster but actually sorted in order of preference with respect to the content of data.

IV. Proposed Method

As discussed in the literature review, the techniques which are used to rank journals of scientific contents or otherwise are mostly based upon the number of citations that a journal offers. Most of the existing journal ranking techniques only consider the positive examples of citations while the proposed technique includes negative examples as well. For instance, in current techniques, the rank of a journal X in 2010 is computed by taking into consideration those journal papers which have been cited in the year 2008, 2009 and 2010. These cited papers are known as positive journals by using K Means clustering for grouping and Labelling technique for naming clusters instances. Negative instances that have not been cited in 2008, 2009 and 2010 are not counted in finding the quality of a journal. The inclusion of negative examples enables one to find the exact rank of the journal as how many manuscripts of a particular journal are never cited. Papers which are not cited anywhere are evident on acceptance of manuscripts for publication with lesser worth hence affecting the overall score of journal rank. Secondly, this inclusion made the ranking of journals relative to each other and presented a real picture of interest of the community in the research articles of the journal. Other indicators related to the prestige of the journals are also available in the extant literature. A journal's prestige is calculated through the impact value of that journal. The impact of the journal unifies its importance in the community of circulation and its use by the researchers. In the proposed technique, a new indicator with the name of "Worth of the journal" is proposed which is used for calculating prestige of a journal. The worth of a journal is determined by the number and prestige of editorial board and review committee. It is computed by the number of recognized publications, the number of experience years and impact of publications of review committee members/editorial board members. The motive to include prestige is the frequency of journals. Some particular journals are published two or three times a year. Calculating the rank of such a journal using conventional methods would definitely yield lesser rank than the relatively more frequently published journal. Journal'sprestigereduces biasing. Factors including number of experience years of reviewers, number of research papers, impact of research and qualifications of the reviewers are used to calculate worth of a journal.

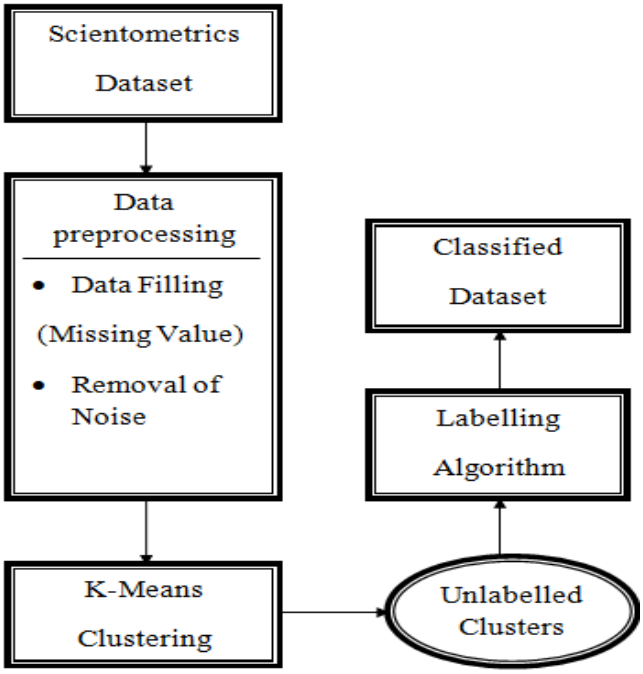

Fig. 1. Proposed Method for classification of science journals by using K Means clustering for grouping and Labelling technique for naming clusters

Weights are allocated to experience, research papers, the impact of research and qualifications on the theoretical basis. Since PoJ is determined for evaluation of research quality so two larger heuristic weights 35% and 30% are allocated to them. Educational qualification may not have a direct impact on research quality but still considerable, a weight of 10% is given to it.

Self-citations are included in the calculation of impact factor, which were removed in the eigenfactor value [2]. The proposed method removes all self-citations before calculating the rank. For instance, an author could cite their own 2008 published paper in their papers published in 2009 and 2010, but such citations get excluded in the working of the proposed technique because some of the authors would, without giving an actual citation, just mention their work textually.

ISBN: 978-988-14048-6-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

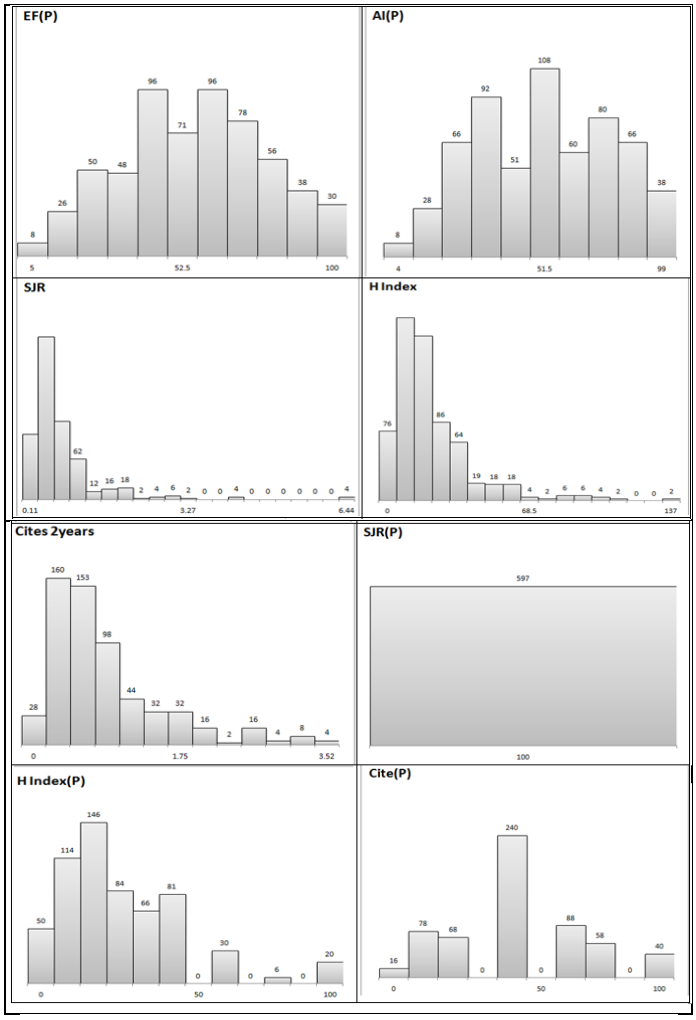

Fig. 2 Input data histograms (Dataset containing prestige, h-index, article influence, eigenfactor, citations of 600 journals)

After the worth of a journal has been determined, self citations are removed using self-citation removal algorithm. Values of prestige of journal, h-index, Article influence, eigenfactor and citations are stored in a shared integrated database. The database for this purpose is custom-built in this study and is designed in an open source software. Object oriented design model of database design is chosen. The interface/ front-end of the database is also added with analytical tools. Such databases can be analyzed by simple statistical tools which obviously needs human intervention. All the attributes stored in the database are evaluated to find similar journals to group them in one cluster by using K-Means clustering. Since these particular clusters happen to be unlabeled and the output does not have any meaning for an individual. So K means clustering is applied to the database to divide data in six clusters after whichthe clusters are labeled by using the method given in section 3.2. The labels on clusters mean the labels on each instance of cluster. Hence a new field "rank" is created in the database to store these newlyassigned labels. The following labels are applied on instances of the database. Although clustering and other data mining techniques are applied to classify the journals yet the method of labeled clustering used in this paper is demonstrating its first application. This method automatically ranks the journals without any human intervention whereas other unsupervised classification methods need human intervention for identifying the labels of each cluster.

TABLE I
LABELS/ CLASSES OF JOURNALS

SNo	Criteria	Category
1	Clusters with highest correlation	Outstanding
2	Clusters with 2nd highest correlation	Excellent
3	Clusters with 3rd highest correlation	Good
4	Clusters with average correlation	Satisfactory
5	Clusters with 2nd lowest correlation	Unsatisfactory
6	Clusters with lowest correlation	Poor

V. EXPERIMENTS AND RESULTS

The experiment for the proposed method will be performed on a dataset of 600 science journals. Following five indicator values for each science journal will be calculated and stored in the database.

1. Prestige of the Journal (PoJ)
2. H-Index
3. Article Influence (AI)
4. Eigenfactor (EF)
5. Citations

The structure of the tables designed for performing the experiment is given below in table 2.

TABLE II
STRUCTURE OF THE DATABASE

Attribute ID	Attribute_Name	Data Type	Length
First table			
1.	Journal ID	Varchar	10
2.	Journal name	Varchar	50
3.	Indexing	Varchar	30
Second table			
1.	Journal ID	Varchar	10
2.	Prestige	Number	10
3.	H-Index	Number	10
4.	AI	Number	10
5.	EF	Number	10
6.	No_of_cit	Number	10

The values of these indicators are calculated by using the formuales already discussed. Data of the journals for this study is collected from multiple databases to avoid biasness. A histogram view of the dataset is given in Fig 1.

The indicators used for each journal will be calculated by using the equations discussed before, however prestige of a journal will be computed according to the formula proposed in this study. The obtained output will be applied with K Means Clustering the output of which will then be used for finding labels of clusters. The complete experiment of the study is yet to complete which will be published in the next version of the manuscript.

ISBN: 978-988-14048-6-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

VI. Conclusion

In this study, the indicators which are used to measure the rank and performance of a science journal are studied. These indicators are classified into different categories. The indicators are reviewed with respect to their categorization. A method for classification of these research journals on statistical basis is proposed then. The study claims an improvementin neutral assessment of research journal classification through clustering techniques and statistical correlation. It is attempted to remove biases from existing journal classification methods by adding negative value of citations which has not been considered in the previous studies. The reposition of research is being affected by the biasing in existing journal ranking algorithms affects on a big level. User's perspective of research integration can also be affected by this. A list of indicators for journal evaluation is finalized after comparing them with rest of the indicators. A new indicator with the name of prestige of the journal is introduced. The performance of the proposed method is yet to be computed on the following basis;

1. Classification Accuracy
2. Number of journals selected.
3. User Acceptance.

The extended version of the manuscript will be published in the near future.

The work can be extended to include scientometrics of different journal ranking agencies. On the basis of the implicit features these scientometrics can also be classified in future. There are numereous clustering techniques which can be compared on the results of journal ranking and the technique which suits the most can be proposed.

References

[1] B. González-Pereira, V. Guerrero-Bote and F. Moya-Anegon, "The SJR indicator: A new indicator of journals." scientific prestige.CoRR abs/0912.4141, 2009 .

[2] D. Butler, "Free journal-ranking tool enters citation market.", Nature, 451(6), doi:10.1038/451006a, 2008.

[3] E. Garfield, "The history and meaning of the journal impact factor.", The journal of the American Medical Association, 295(1), pp. 90–93, 2006.

[4] J. Poelmans, P. Elzinga, S. Viaene and G. Dedene, "Formal concept analysis in knowledge discovery: a survey." Conceptual Structures: From Information to Intelligence, pp. 139–153, 2010.

[5] J. M. Campanario and A. Molina, "Surviving bad times: The role of citations, self-citations and numbers of citable items in recovery of the journal impact factor after at least four years of continuous decreases.", Scientometrics, 81(3), pp. 859–864, 2009.

[6] J. Gao and D. B. Hitchcock, "James-stein shrinkage to improve K-means cluster analysis.", Computational Statistics and Data Analysis, pp. 2113-2127, 2010.

[7] J. Bar-Ilan, "Which h-index?—a comparison of wos, scopus and google scholar.", Scientometrics, 74(2), pp. 257–271, 2008.

[8] K. J. Cios, W. Pedrycz, W.S Roman and L. A. Kurgan, "Data mining: a knowledge discovery approach." Springer Publishing Company, pp. 1-35, 2010.

[9] M. Shaheen and M. Z. Khan, "A method of data mining for selection of site for wind turbines.", Renewable and Sustainable Energy Reviews. In Press, 2015.

[10] M. Shaheen and M. Shahbaz, A. Guergachi and Z. Rehman, "Mining sustainability indicators to classify hydrocarbon development.", Knowledge-Based Systems, 24(8), pp. 1159–1168, 2011.

[11] M. Shaheen, M. Shahbaz and A. Guergachi, "Context based positive and negative spatio-temporal association rule mining.", Knowledge-Based Systems, 37, pp. 261-273, 2013.

[12] M. Shaheen, F. Basit and S. Iqbal, "Labeled Clustering: A Unique Method for Labeling Unsupervised Classes.", In: Proc. 8th International Conference on Internet and Secured Transaction., London, UK, 2013.

[13] M. E. Falagas, V. D. Kouranos, R. Arencibia-Jorge and D. E. Karageorgopoulos, "Comparison of scimago journal rank indicator with journal impact factor.", The FASEB Journal, 22(8), pp.2623–2628, 2008.

[14] M. Shaheen, M. Shahbaz, A. Guergachi and Z. Rehman, "Mining sustainability indicators to predict optimal hydrocarbon exploration rate.", In: IASTED proceedings of artificial intelligence and applications, Austria, pp. 394–400, 2010 b.

[15] M. Shaheen, M. Shahbaz and Z. Rehman, "Data mining applications in hydrocarbon exploration.", Artificial Intelligence Review, 35(1), pp.1–18, 2010 a.

[16] M. Shaheen and M. Shahbaz, "An Algorithm of Association Rule Mining for Microbial Energy Prospection." Sci. Rep. 7, 46108; doi:10.1038/srep46108, 2017.

[17] M. Shaheen, "CBPNARM: An algorithm for spatial association rule mining.", 2015 International conference on intelligent informatic and biomedical Sciences (ICIIBMS), Okinawa, Japan, 2015, pp.227-232. doi:10.1109/ICIIBMS.2015.7439484, 2016.

[18] M. McAleer, "Journal impact versus eigenfactor and article influence.", Kier discussion paper series, paper no. 737, 2010.

[19] N. Sombatsompop, T. Markpin and N. Premkamolnetr, "A modified method for calculating the impact factors of journals in ISI journal citation reports: Polymer science category in 1997–2001." Scientometrics, 60(2), pp. 217–235, 2004.

[20] S. Amreshi andC. Conati "Combining unsupervised and supervised classification to build user models for exploratory learning environments" Journal of Educational Data Mining, pp.1-54, 2009.

[21] S. J. Liebowitz and J.P. Palmer, "Assessing the relative impacts of economics journals." Journal of Economic Literature,pp. 77–88, 1984.

[22] S. Iqbal, M. Shaheen andF. Basit, "A machine learning based method for optimal journal classification.", In:Proc. 8th International Conference on internet technology and secured transactions, pp. 259-264, 2013.

[23] T. Kanungo, D. M. Mount, N. Netanyahu, C. D.Piatko, R. Silverman andA. Y. Wu, "An efficient K-means clustering algorithm: analysis and implementation.", IEEE Transactions on pattern analysis and Machine Intelligence, pp.881-892, 2002.

[24] P. Vicente, V. Guerrero-Bote and F. Moya-Anegon, "A further step forward in measuring journals' scientific prestige: The SJR2 Indicator.", Journal of Informetrics, 6, pp. 674-688, 2012.

ISBN: 978-988-14048-6-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)