Abstract:

The evaluation of journals based on their influence is of interest for numerous reasons. Various methods of computing a score have been proposed for measuring the scientific influence of scholarly journals. Typically the computation of any of these scores involves compiling the citation information pertaining to the journal under consideration. This involves significant overhead since the article citation information of not only the journal under consideration but also that of other journals for the recent few years need to be stored. Our work is motivated by the idea of developing a computationally lightweight approach that does not require any data storage, yet yields a score which is useful for measuring the importance of journals. In this paper, a regression analysis based method is proposed to calculate Journal Influence Score. Proposed model is validated using historical data from the SCImago portal. The results show that the error is small between rankings obtained using the proposed method and the SCImago Journal Rank, thus proving that the proposed approach is a feasible and effective method of calculating scientific impact of journals.

1. Introduction and Background

Librarians and information scientists have been evaluating journals with regard to importance and popularity from last 75 years. The Journal Impact Factor proposed by Eugene Garfield \(^1\) was the founder and Chairman Emeritus, ISI (which later became Thomson Reuters) was a milestone in this regard.

The advent of the Thomson Reuters citation indexes made it possible to do computer-compiled statistical reports in terms of citation frequency. Thomson Reuters is credited with the invention of the journal “impact factor” in the 60s. Thomson Reuters began to publish Journal Citation Reports in 1975 as part of the SCI and the Social Sciences Citation Index \(^2\).

The impact factor is usually calculated over a period based on the number of citations in the current year to articles in the journal during the previous years considered. For instance, the two year impact factor is calculated for year \(n\) as below.

Let \(A\) = the number of times that all items published in that journal in years \((n-1)\) and \((n-2)\) that were cited by indexed publications during year \(n\).

Let \(B\) = the total number of “citable items” published by that journal in years \((n-1)\) and \((n-2)\). (“Citable items” for this calculation are usually articles, reviews, proceedings, or notes; not editorials or letters to the editor).

Then the impact factor for year \(n\) = \(A/B\).

Despite some of the criticisms, the Journal Impact Factor is widely used as a measure of the scientific influence of a journal. Most of the variants of the Journal Impact Factor require the citation data for the preceding few years from all the indexed journals. Also note that the impact factor for year \(n\) is only available in the next year \((n+1)\) because all the citations to previous years need to be available. Thus the computation of the impact factor of a journal involves significant overhead.

In this paper, a new approach for calculating Journal Influence Score (JIS) is proposed. This paper work is motivated by the idea of computing JIS directly without the overhead of citation data storage but at the same time yielding reasonable accuracy. We model it as a regression problem in which the individual weights corresponding to each of the input variables are computed from historical data, these weights reflect the relative influence the individual variables may have in the calculation of the influence score. These weights can then be used...
directly to compute the score of a journal without using the prestige of any other journal. The main advantage is a computationally lightweight scheme that does not require any data storage[3].

We conducted experiments on publicly available data to validate our approach. The data and experimental details are described in section 2. Results are described in detail in Section 3 and Conclusions are presented in section 4.

2. Experimental details

This section describes the details of our approach. We use a linear regression model where the response variable is the Journal Influence Score. The year under consideration was 2012. The input parameters (predictor variables) include the Quarter, H-Index, Total Docs 2012, Total Docs 3yrs, Total Cites 3yrs, Citable Docs 3yrs, Ref/Doc, Cites/Doc 3yrs and Total Ref.

Quarter is considered as one of the input variables. Intuitively, any journal to be evaluated in the first Quarter of the year has more probability of having greater influence, considering the number of publications is mostly limited. Hence the "quarter" of publication should be statistically significant. The probability of influence in our sample data validates the use of quarters in our model.

\[ \text{Quarter (Probability of Influence)} = \frac{Q_i}{\sum Q_i} \]

Where \( i = 1, \ldots, 4 \)

Starting with the initial set of input parameters, a two-phase approach was employed to obtain a more compact set of transformed variables. In the first phase, the number of variables were reduced using cross-correlation & MLR, and a down selected set of input variables was obtained. In the second phase, PCA was applied on this reduced set and the first few principal components that explained > 90% of the variability were retained. The final model was a MLR model on the principal components retained after the second phase.

2.1 Data

The SCImago Journal & Country Rank (SJR) (http://www.scimagojr.com/) is a portal that includes the journals and country scientific indicators developed from the information contained in the Scopus database. These indicators can be used to assess and analyze scientific domains.

Our source data for this study were SCImago Journal and Country Rank’s portal which contained journals in Elsevier’s Scopus.

http://www.scimagojr.com/journalrank.php?area=0 &category=1706&country=all&year=2012&order=sjr&min=0&min_type=cd

2.2 Description of Model:

A multiple linear regression (MLR) model is used to predict a response variable \( y \), as a function of \( k \) predictor variables \( x_1, x_2, \ldots, x_k \) using a linear model of the form

\[ y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_kx_k + e \]

Where \( b_0, b_1, \ldots, b_k \) are fixed parameters that signify the weight of factors and \( e \) is the error.

Given a number of observations, the model consists of the following \( n \) equations:

\[ y_1 = b_0 + b_1x_{11} + b_2x_{21} + \ldots + b_kx_{k1} + e_1 \]
\[ y_2 = b_0 + b_1x_{12} + b_2x_{22} + \ldots + b_kx_{k2} + e_2 \]
\[ \vdots \]
\[ y_n = b_0 + b_1x_{1n} + b_2x_{2n} + \ldots + b_kx_{kn} + e_n \]

So, we have

\[ \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{nk} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} \]

So in matrix form: \( y = Xb + e \)

Where \( y \) = a column vector of \( n \) observed values of \( y = \{ y_1, \ldots, y_n \} \)

\( X \) = an \( n \) row by \( k + 1 \) column matrix whose \( (i, j + 1) \)th element \( X_{ij+1} \) = 1 if \( j = 0 \) else \( x_{ij} \)

\( b \) = a column vector with \( k + 1 \) elements= \( \{ b_0, b_1, \ldots, b_k \} \)

\( e \) = a column vector of \( n \) error terms= \( \{ e_1, \ldots, e_n \} \)

Next is Principal Component Analysis (PCA). In this analysis, the component which explains a small percentage of variation can be removed from further analysis.

2.3 Algorithm
Step 1. Import data from web (www.scimagojr.com)

Step 2. Find correlation of all factors with SJR

\[ R_{x,y} = \frac{\sum x_i y_i - \bar{x} \bar{y}}{\sqrt{\left( \sum x_i^2 - \bar{x}^2 \right) \left( \sum y_i^2 - \bar{y}^2 \right)}} \]

Step 3. Find Model equation of type

\[ y = b x + e \]

Step 4. Find parameter \( b \)

\[ b = (X^T X)^{-1} X^T y \]

Step 5. Derive Multiple Linear Regression equation to establish relationship between input factors and journal ranking output

Step 6. Extract p-values and correlation coefficient values from Multiple Linear Regression equation

Step 7. If input variables with P-value > 0.05 & Correlation Coefficient < 0.4, then remove parameter.

Step 8. Repeat Step 7 for all Input factors

Step 9. Repeat Step 2 – 8 till all parameters has P-value < 0.05 & Correlation Coefficient > 0.4

Step 10. Compute the mean and standard deviations of the variables.

\[ \bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{j,i} \]

\[ s^2_x = \frac{1}{n-1} \sum_{i=1}^{n} (x_{j,i} - \bar{x}_j)^2 \]

Step 11. Normalize the variables to zero mean and unit standard deviation.

\[ x'_j = \frac{x_j - \bar{x}_j}{s_x} \]

Step 12. Compute the correlation among the variables:

\[ r_{x_j x_k} = \frac{\sum_{i=1}^{n} (x'_{j,i} - \bar{x}_j) (x'_{k,i} - \bar{x}_k)}{s_{x_j} s_{x_k}} \]

Step 13. Prepare the correlation matrix:

Step 14. Compute the Eigen values & Eigen vectors of the correlation matrix.

Step 15. Obtain principal components by multiplying the eigenvectors by the normalized vectors

Step 16. Compute the values of the principal components.

Step 17. Compute the sum (the sum must be zero) and sum of squares of the principal components. The sum of squares gives the percentage of variation explained.

Step 18. Apply regression on the Principal factors to compute JIS.

Step 19. Calculate the quartile match

19.1: Check the samples in each quarter.

19.2: Compare these samples with same number of samples from the results of our model for each quartile.

19.3: Calculate the percentage of match.

| Basis of Comparison | Existing Method | Proposed Method |
|---------------------|-----------------|-----------------|
| Number of input variables | 13 | 5 |
| Procedure | Iterative | Weighted |
| Expected Time Complexity | More | Less |
| Database Size | Huge | Insignificant |
| Historical Data | 3-5 Years | 2 Years |
| Algorithm Used | Google Page Rank Algorithm | Weight based on regression |

Table: Comparison Between Existing & Proposed System

NOTE: It has been determined that regression model applied after the PCA, namely the Revised PCMLR (Principal Component based Multiple Linear Regression model), proposed here accounts for the fact that the principal components are linearly independent of each other in the domain space. K-means Clustering is implemented to achieve classification between “National” and “International” journals, a metric that libraries and academic institutions may use for measuring impact of scientific work.
On the input factors multiple linear regression is applied. After applying MLR on the computer science and application subject area, results are as follows:

**Analysis Phase-I:**

In this phase, all the input parameters are fed to analyze the Correlation and Regression statistics.

| Factor                  | P-value  | Correlation Coefficient | Optimization Decision |
|-------------------------|----------|-------------------------|------------------------|
| Quarter                 | 3.67E-09 | -0.76553                | Yes                    |
| H index                 | 0.197845 | 0.691082                | Yes                    |
| Total Docs. (2012)      | 8.84E-05 | 0.155572                | Yes                    |
| Total Docs. (3years)    | 0.902654 | 0.370305                | Yes                    |
| Total Refs.             | 0.008454 | 0.395752                | Yes                    |
| Total Cites (3years)    | 2.32E-07 | 0.554409                | Yes                    |
| Citable Docs. (3years)  | 0.607635 | 0.370423                | Yes                    |
| Cites / Doc. (2years)   | 2.48E-14 | 0.848587                | Yes                    |
| Ref. / Doc.             | 0.82342  | 0.170068                | No                     |

**Inference:**

- The value of coefficient of determination, \( R^2 \) is 0.7986 i.e. 79.86% variation in Journal Influence factor is explained by this regression.
- Significance F value is 8.18E-70 i.e. less than 0.05, which means it passes F-test.
- P-value for Ref/Docs is greater than 0.05 and has a weak correlation with SJR Score. So, Ref/Doc. is removed from further analysis.

**Analysis Phase-II:**

Result of Correlation and Regression after removing Ref/Doc is as shown in below table:

**Key Terms:**

- **Null Hypothesis considered for significance testing:** Input factors are not able to describe the Output.
- **F-Test:** This test is used to check the hypothesis that the proposed model fits the data well. The model is significant if MSR/MSE value is greater than \( F[k,n-k-1] \) which is taken from F-distribution table.
- **P-value:** P-value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true. Hypothesis is rejected if the value is less than 0.05 for 95% confidence interval.
- **Hypothesis Testing:** Hypothesis Testing is a test of significance and used to support a decision statistically. Hypothesis testing is divided into three steps.

**Step 1:** Define Null and alternative hypotheses

- **Null Hypothesis (Ho):** The null hypothesis is a claim of no difference or no impact
- **Alternative Hypothesis (Ha):** Opposite of null hypothesis, a claim of difference or impact

**Step 2:** Calculate test statistics and p-value

- **P(Probability) Value**
  - Statistical measure for the strength of Ho
  - Probability that it will be wrong to select the Alternate Hypothesis
  - Higher the p value, the more evidence we have to support Ho
  - P-value less than 0.05 means we reject the null hypothesis.

**Step 3:** Take statistical significance decision : By convention if \( p > .05 \), Accept H0

**F-value**: It is used for significance testing for ANOVA (Analysis of Variance)
Principal Component Analysis (PCA) deals with the process of reduction of features used in the computation of the Influence Score of the journals. This is ensured by removing one of the interdependent factors the model uses in the calculation process. If any factors are related to one another by a percentage greater than the user defined threshold value set for the feature removal, then the model selects one of these features and removes it from the selection procedure. After the feature is removed, data for the processing is again calculated in a recursive manner.

Considering the features are highly co-related, we can easily say that any of these two features can be removed from the dataset without having much changes or uncertainties in the influence score calculation. Whichever feature is still kept in the dataset compensates for all the losses incurred due to the removal of the other feature. But, to ensure high efficiency standards and accuracy requirements, the model selects one of the highly co-related features that have lesser co-relation with all the other existing features. Hence, we tend to remove the one that is having high co-relation with one feature and comparatively lesser co-relation with the others existing in the dataset. This ensures lesser compensation requirements and even lesser effect in deflections made in the final influence score.

The percentage threshold value of correlation based on which the removal process takes place is set by the user keeping in mind the accuracy standards the model must pass. Hence the percentage match after which the model is accepted becomes the deciding element in making the decision for setting the final Threshold Value for the feature removal process.

### Analysis Phase-III:

Regression and Correlation is applied on the data after removing Total Docs (3years). Result is as shown in below table:

| Factor                  | P-value  | Correlation Coefficient | Optimization Decision |
|-------------------------|----------|-------------------------|-----------------------|
| Quarter                 | 3.29E-09 | -0.76553                | Yes                   |
| H index                 | 0.1994   | 0.691082                | Yes                   |
| Total Docs (2012)       | 8.73E-05 | 0.155572                | Yes                   |
| Total Docs (3years)     | 0.884709 | 0.370305                | No                    |
| Total Refs.             | 0.007956 | 0.395752                | Yes                   |
| Total Cites (3years)    | 2.16E-07 | 0.554409                | Yes                   |
| Citable Docs (3years)   | 0.61594  | 0.370423                | Yes                   |
| Cites / Doc. (2years)   | 3.11E-15 | 0.848587                | Yes                   |

### Inference:

- \( R^2 \) is 0.7986 i.e. 79.86% variation in JIF is explained by this regression, which is similar to the previous \( R^2 \) (before removing Ref/Doc).
- Significance F value is less than 0.05 i.e. it passes the F-test.
- P-value for Total Docs (3Years) is greater than 0.05 and has a weak correlation with SJR Score. So, Total Docs (3years) is removed from further analysis.

### Principal Component Analysis:

| Factor                  | P-value  | Correlation Coefficient | Optimization Decision |
|-------------------------|----------|-------------------------|-----------------------|
| Quarter                 | 3.29E-09 | -0.76553                | Yes                   |
| H index                 | 0.1994   | 0.691082                | Yes                   |
| Total Docs (2012)       | 8.73E-05 | 0.155572                | Yes                   |
| Total Docs (3years)     | 0.884709 | 0.370305                | No                    |
| Total Refs.             | 0.007956 | 0.395752                | Yes                   |
| Total Cites (3years)    | 2.16E-07 | 0.554409                | Yes                   |
| Citable Docs (3years)   | 0.61594  | 0.370423                | Yes                   |
| Cites / Doc. (2years)   | 3.11E-15 | 0.848587                | Yes                   |

Principal Component Analysis (PCA) was applied on the dataset to reduce the number of features used in the calculation of the Influence Score. The process involves the removal of one of the interdependent factors the model uses in the calculation process. If any factors are related to one another by a percentage greater than the user defined threshold value set for the feature removal, then the model selects one of these features and removes it from the selection procedure. After the feature is removed, data for the processing is again calculated in a recursive manner.

Considering the features are highly co-related, we can easily say that any of these two features can be removed from the dataset without having much changes or uncertainties in the influence score calculation. Whichever feature is still kept in the dataset compensates for all the losses incurred due to the removal of the other feature. But, to ensure high efficiency standards and accuracy requirements, the model selects one of the highly co-related features that have lesser co-relation with all the other existing features. Hence, we tend to remove the one that is having high co-relation with one feature and comparatively lesser co-relation with the others existing in the dataset. This ensures lesser compensation requirements and even lesser effect in deflections made in the final influence score.

The percentage threshold value of correlation based on which the removal process takes place is set by the user keeping in mind the accuracy standards the model must pass. Hence the percentage match after which the model is accepted becomes the deciding element in making the decision for setting the final Threshold Value for the feature removal process.

### PCA Model:

In this model, PCA generates a result that shows the percentage of variation explained by each factor. The Journal Influence Score is calculated after finding the weight of each principle factor by using Multiple Linear Regression.

The influence score is mapped with all the feature values and the model generates the percentage by which one feature explains the scatter of the samples in the dataset in this plotting. This percentage is used further to select those elements that can be removed from the influence score calculation procedure so as to reduce time complexity while the model is still able to maintain the same accuracy, in compliance with the results generated by the standards.

The removal procedure is again dependent on the percentage explanation that these features give.
Making this an iterative process, the model removes the feature with the least percentage first and then moves up to the feature with the percentage just above in the percentage stack. Once a feature is removed, the model computes the percentage match that a feature has with the standard requirements. If the match percentage is higher than the desired threshold, then the removal is approved and the removal iteration proceeds, else the removal of the features is disapproved and the process stops.

| Factor                | Percentage of Variation Explained |
|-----------------------|----------------------------------|
| Quarter               | 62.7952%                         |
| H-index               | 21.3912%                         |
| Total Docs. (2012)    | 8.4489%                          |
| Total References      | 3.539%                           |
| Cites/Doc(2years)     | 1.3498%                          |
| Citable Docs. (3years)| 1.2379%                          |
| Total Cites(3years)   | 1.2379%                          |

Table: Principal Component Analysis

Inference:
- The above table shows that principle factors are Quarter, H-index, Total Docs. (2012) and Total References and Cites/Doc(2years).
- These factors explain approximately 97.5% of variation in Journal Influence Score.
- Total Cites(3years) and Citable Docs. (3years) explain only 2.5%. So these factors can be removed from further analysis.
- Although Cites/Doc (2years) also explains only 1.3498%, but removal of this leads to a drastic decrease in the match percentage, and it drops down to a value lesser than the threshold set due to the accuracy needs.

After removing these factors, MLR is applied on the remaining 5 factors. The results of regression are as follows:

**SUMMARY OUTPUT**

**Regression Statistics**

- Multiple R: 0.8774
- R Square: 0.76983
- Adjusted R Square: 0.764575
- Standard Error: 0.323211
- Observations: 225

**ANOVA**

|                | Df | SS     | MS     | F       | Significance F |
|----------------|----|--------|--------|---------|----------------|
| Regression     | 5  | 76.51797 | 15.30359 | 146.4943 | 8.28932E-68     |
| Residual       | 219| 22.87793 | 0.104465 |         |                |
| Total          | 224| 99.39591 |        |         |                |

**Coefficients**

| Factor                | Coefficients | Standard Error | t Stat | P-value | Lower 95% | Upper 95% |
|-----------------------|--------------|----------------|--------|---------|-----------|-----------|
| Intercept             | 0.513322     | 0.110512       | 4.644933 | 5.87E-06 | 0.295518325 | 0.731126  |
| Quarter               | -0.14076     | 0.030902       | -4.55526 | 8.69E-06 | -0.201667404 | -0.07986  |
| H index               | 0.004716     | 0.001247       | 3.781766 | 0.000201 | 0.0022588486 | 0.007174  |
| Total Docs. (2012)    | 0.000131     | 0.000121       | 0.084092 | 0.279556 | -0.000107049 | 0.000369  |
| Total Refs.           | -8.3E-06     | 7.75E-06       | -1.0703  | 0.285661 | -2.35727E-05 | 6.98E-06  |
| Cites/Doc. (2years)   | 0.301404     | 0.030123       | 10.00582 | 1.19E-19 | 0.242036313  | 0.360772  |
Snehanshu Saha, Neelam Jangid, Anand MN, Sidhant Gupta

Inference:

- $R^2$ is 0.7698 i.e. 76.98% variation in Journal Influence score is explained by this regression.
- Significance F value is less than 0.05, which means it passes F-test.

3. Results:

Regression equation:

The fitting model assumes the form:

$$\text{Journal Influence Score} = 0.513322 - (0.14076 \times \text{Quarter}) + (0.004716 \times H \text{ index}) + (0.000131 \times \text{Total Documents For Current Year}) - (8.3E-06 \times \text{Total References}) + (0.301404 \times \text{Cites/Doc. in previous 2 years})$$

The above equation shows that the journal base score is 0.513322, which explains the initial score of each journal, affected further by the factor values of that journal. The other values in the equation signify the weight to their corresponding factors.

A comparison has been done to analyze Match % among Journal Ranking using SJR Model, 5-parameter regression model and 9-parameter regression model. Results are as follows:

Another validation has been performed on a different subject category (Computer Networks and Communications) to analyze match % and the results are as follows:

![Accuracy Test between SJR model & proposed model](image1)

Computed errors in the proposed model reflect reasonable level of accuracy in all the quarters possible. This comparison is made with the SJR data. Hence, the model performs well in all possible cases and quarters.

![Scatter Plot between Quarter and JIS](image2)

![Scatter Plot between H-index and JIS](image3)
Scatter Plot between Total Docs(current year) and JIS

Scatter Plot between Total References and JIS

Scatter Plot between Cites/Documents and JIS

Snapshot of web based tool

Journal Influence score calculator is available at http://www.4shared.com/document/VYBmLQF Gce/Journal_Influence_Score_Calcul.html

Video tutorial is available at https://www.youtube.com/watch?v=rG_gsRLK N8I

High Level Design of Journal Influence Score Calculator

Classification Process:

After the influence score for all the journals in the sample set or a new upcoming journal is calculated, the process of classification starts where the intention is to classify the journal into one of the categories, which are National Journal and International Journal. This process is based on the value of the influence score. The higher the influence score, the more the journal is valued and accepted throughout. This is hence used as the means for classification.
The challenge here is to define a boundary that separates the two given classes based on which the class is to be assigned to any journal. This is done by making use of the K-Means Clustering algorithm. Considering the sample set this problem deals with is without the class instances and the training of the system or the machine is unsupervised, the clustering algorithm to be used also has to be unsupervised. Hence, Unsupervised K-Means Clustering Algorithm is put to use.

The samples are clustered and then rearranged iteratively until we have an instance where the change in the cluster means for both the classes is minimal after which the system attains stability. Another point to be kept in mind during this iteration step is that the system should not keep looking for such changes throughout and turn into an infinite loop. This would lead to system deadlock and the system would in-turn fail to perform in the desired manner. This highlights the requirement of an upper limit to the number of iterations performed for this process. Hence maintaining the system stability and also performing well too give the desired results.

**K-Means Clustering :**

Once the process of reducing the number of factors we take into consideration for the final influence score computation using the Principal Component Analysis model, we move forward to the part where we differentiate the entities in the structure into National and International categories. We achieve this by making use of the Unsupervised K-Means Clustering method.

**Detailing :**

The samples are clustered into two separate groups in this case by taking two distinct mean points for the clusters initially. These clusters continually keep changing after every iteration based on the evaluation of the change observed in the cluster mean after all the samples in the sample are collected into them. A sample goes into that cluster which lies closer to the position of the sample. Once the clustering is done, we recomputed the means for both the clusters. These steps are put into iteration until we either see negligible change in both the cluster means or the iteration step exceeds a certain limit. This is how we handle clustering of the complete sample set.

**Clustering Algorithm :**

Step 1: Calculate the influence score of all the journals in the sample set.

Step 2: Select two distinct cluster means arbitrarily.

Step 3: Initialize the variables (Iteration no =0, maxiterations =100, changed = 1)

Step 4: Loop until both the conditions are satisfied

While(changed ==1 & iteration no. < maxiterations)

Step 4.1 Increment iteration no, make changed=0

Step 4.2 For all samples in the dataset , classify all into the class with the nearest cluster mean.

Step 4.3 Initialize variables to 0 (ele0 = 0, ele1 = 0, sum0 = 0, sum1 = 0)

Step 4.4 Re-compute cluster means

For all samples in the dataset

If (class == 0)

Add influence score to sum0 , increment ele0

else

Add influence score to sum1 , increment ele1

new0 = sum0 / ele0;

ew1 = sum1 / ele1;

Step 4.5 Check for any significant change in the Cluster Means

(The threshold value to define a change as stable is anything less than 0.01 [square of 0.1])

\[ \text{If } ((\text{new}_0 - \text{old}_0)^2 > 0.01 \text{ or } (\text{new}_1 - \text{old}_1)^2 > 0.0) \]

changed = 1

Step 4.6 Store new values of the cluster means

\[ u_0 = \text{new}_0 \quad \text{and} \quad u_1 = \text{new}_1 \]

Step 5: Once the loop terminates, \( u_0 \) and \( u_1 \) are the final Cluster Means.

After the execution of the clustering process, the cluster means can be used further for any upcoming new journal entry to make the classification process more and more simple and reducing the complexity. Any new journal just undergoes the influence score calculation process proceeded by a check on this score, if the score lies closer to the cluster mean to the National category, then the journal is termed to be of National Standards , else the journal attains International Standard as the Influence Score lies closer to the cluster that corresponds to the cluster formed of samples of International standard. Hence, classification via the
Clustering mechanism is easily achieved and it still maintains the accuracy measures.

**Outputs and Graphical Plots:**

Here are the plots generated for 1084 samples given for the computer science category, taken from the SJR (SCImago Journal Ranking) database globally available. The plots show the scatter of the samples and also show the classes in different colours so as to understand how these factors influence the selection process for any sample and to what extent this effect is seen.

Looking at the graphs it is clear that H-Index explains more percentage of the scatter as shown as a result by PCA as the scatter is follows a linear pattern and also because we can say that the number of scatter points that lie closer to the decision boundary or the class boundary are very less. This in-turn reduces the risk involved in the class selection process while taking this factor into consideration. Hence it is also clear that H-Index defines the scatter to the Maximum extent along with the quarter in which the journal is published.

This is not the case with the other two factors here. The number of scatter points closer to the decision boundary are more and hence more risk is involved while making the class decision based on this factors.

4. **Conclusions**

In this paper, a regression analysis based method is proposed to calculate the Journal Influence Score. Comparison has done between the rankings using SCImago Journal Rank(SJR) and proposed method. The results show that error is minimal.

The model depicts significantly high accuracy levels as each quartile match varies from 78% - 92%. This is achieved without any iterative approach and/or requirements of data storage.

In this model, Principal Component Analysis(PCA) generates a result that shows what percentage of the variability is explained by the given dataset. PCA helps in reduction of input factors while maintaining the reasonable accuracy. The final regression equation has been derived with the principle factors.
Finally a classification scheme is used for categorisation of journals into “National” and “International” in order to help Libraries and repositories across the scientific and technical communities.

Proposed approach may be extended to a universal weight vector for all subject categories rather than evaluating separate weights for each. This might increase the percentage of errors in the model as the weights vary from categorically, but the approach would produce generic results.

NOTE: All supplementary files containing figures and tables are available here in a public folder:
https://drive.google.com/a/pes.edu/?tab=mo#folder/s/0B9K3zpr0Pox8bEZia1FDNjc4VW8

References

1. Citation indexes for science, Garfield, Science (AAAS),122 (3159):108–111. doi:10.1126/science.122.3159.108

2. [Garfield94] The Thomson Reuters Impact Factor, Web of Science, http://wokinfo.com/essays/impact-factor/ (originally published in print edition June 20, 1994, when Thomson Reuters was known as The Institute for Scientific Information®)

3. Neelam Jangid, Snehanshu Saha, Siddhant Gupta, Mukunda Rao J*, Ranking of journals in science and technology domain: a novel and computationally lightweight approach, IERI Procedia 10(2014)57–62 doi: 10.1016/j.ieri.2014.09.091

4. Raj Jain, The art of Computer System Performance Analysis, Edition 2009 (Wiley-India)

5. Chiang Kao, The authorship and internationality of industrial engineering journals, Scientometrics, 80 (3) (2009) 123-136.

6. Gualberto Buela-Casal, Pandelis Perkakis, Michael Taylor and Purificacion Checha, Measuring internationality: reflections and perspectives on academic journals, Scientometrics, 67 (1) (2006) 45-65.

7. Goran Svensson, Scholarly journal ranking(s) in marketing: single- or multi-item measures?, Marketing intelligence and planning, 26 (4) (2008) 340-352.

8. Gaby Haddow; Paul Genoni, Citation analysis and peer ranking of Australian social science journals, Scientometrics, 85 (2) (2010) 471–487.

9. Langville, A. N., & Meyer, C. D. (2006). Google's PageRank and Beyond: The Science of Search

10. Engine Rankings, New Jersey: Princeton University Press

11. https://www.youtube.com/watch?v=nEkwxfI3CxA

12. http://www.4shared.com/folder/sj31vMfu/online.html