Application of a Modified k-Means Algorithm to Evaluatesignificance of Bibliometric Indices of Journals

B. Annapurna
HoD, Computer Science
Ch.S.D.St. Theresa's College for WomenAutonomous
Eluru, India

M. M. Naidu
R.V.R &J.C.College of Engineering
Guntur
Andhra Pradesh
India

ABSTRACT
The indices such as h-index, e-index etc has received much attention from the scientific community owing to their prowess to impact journal quality. Many different indicatorshave been developed to overcome drawbacks of h-index. Nearly four indices which are of prime importance in the publishing industry were utilized.In this paper, we present a modified k-means algorithm to generate three clusters of datasets of scientific journals.

Keywords
h-index, m-index, A-index, e-index, k-means algorithm

1. INTRODUCTION
The recent expansion of knowledge and increasingly sophisticated scientific techniques has led to a proliferation of journals in various fields of disciplines. There has been a progressive increase in the scientific methods of journal citation quantification. To assess the quality of publications in journals of scientific discipline, several authors proposed and delineated the purpose of bibliometric indices for journals. Bibliometrics of journals has become an important and a promising tool for authors to submit their research papers. Hence, impact factor proposed to assess the quality of journals, however, several controversies existed [1]. The journal impact factor developed by Eugene Garfield and published by the Thomson Reuters is the first bibliometric evaluator. However, the potentialities and several limitations of the impact factor have been well discussed [2] [3] [4] [5]. Alternative journal rankings [6] has been proposed, however, they deal with a small subset of the literature in any discipline.

In recent years, several research and publications related indices were proposed to assess the quality of the academic research publications. Each one of those indices has its own strengths and weaknesses. The knowledge of research indices started when Hirsh proposed the h-index, designed to measure the impact of research publications to estimate the author influence [7]. h-index has been regarded as the most reliable, robust and easily computed [8] [9] [10]. h-index assesses both the quantity and importance or relevance of publications [11]. h-index has some limitations, and hence to overcome and provide enhancements to H-index, Egghe proposed the g-index [12]. Based on properties of h and g indices, Kosmulski [13] proposed the H(2)-index which concentrates on highly cited research publications. Other indices were proposed which concentrated on the publications that were located at (H-core) in its calculations [14] such as A-index proposed by Jin [15] where, the average number of citations for those publications in the H-core is evaluated. As a variation of A-index, Bornmann etal proposed m-index where, instead of arithmetic average, median is employed as the measure of central tendency [16]. m-index calculated by dividing the h index by the number of years of that journal's publication [17]. In this paper, we propose a method which implements modified k-means clustering algorithm on a set of computer science journals for which indices are evaluated.

2. MATERIALS AND METHODS
2.1 Dataset
A dataset of 150computer science journals were extracted from SCIMago [http://www.scimagojr.com/journalrank.php] website developed from the information contained in the Scopus database [18]. The parameters calculated by SCImago are Sci Journal Ranking (SJR), h-index, Total Docs, Total References, Total cites, citable docs, cites/doc and references/doc etc can be used to assess and analyze scientific domains.

2.2 Python
Using Python programming language, a program was written to perform k-means analysis. Python [19] has several advantages like open source, cross platform, object-oriented programming, dynamic typing features, simple and easy to learn and rich set of supporting libraries for mathematics, statistics, and visualization. Python modules like Numpy (Scientific Computing Tools for Python—Numpy), Scipy (Open Source Library of Scientific Tools), Python-Sklearn and matplotlib are used.

2.3 Modified k-means algorithm
A modification of k-means algorithm reported here, where, data scale to unit variance and principal component analysis used to reduce linear dimensionality using Singular Value Decomposition of the data or eigen value decomposition of a data covariance matrix and keeping only the most significant singular vectors to project the data to a lower dimensional space. A convergence criterion met until no new assignments appeared and until the centroids do not move significantly. If the dataset has more attributes, then each component represents one attribute and clustering will follow. Calculated inertia or within-cluster sum of squared criterion and clusters generated based on k. In the next step, cluster quality metrics applied to judge the goodness of fit of the cluster such as homogeneity score, completeness score, V measure, adjusted rand score and silhouette coefficient.

3. RESULTS AND DISCUSSION
The first step in k-means algorithm is to divide the given data set into user defined number of clusters. The initial choice of k in k-means is an interpretive decision and successive runs should be performed to obtain an optimized division of data for any chosen k value. A prior knowledge on the data structure would result in more appropriate clusters. However,

as the dimensionality of the data increases, it becomes increasingly difficult to determine a proper value for k. Hence, considerable attention has been given to the subject of cluster validation, a process which attempts to evaluate a particular division of data into clusters [20].

The outcome of modified k-means algorithm on a dataset of 150 journals resulted in 3 clusters (k=3). Each cluster data grouped based on the titles appeared in each group containing journal subjects such as Dataset 1 (Programming and Computing) Dataset 2(Data mining, knowledge discovery)Dataset 3(Information systems, Computing), respectively. Data was standardized and the output plot is given in Figure-1.

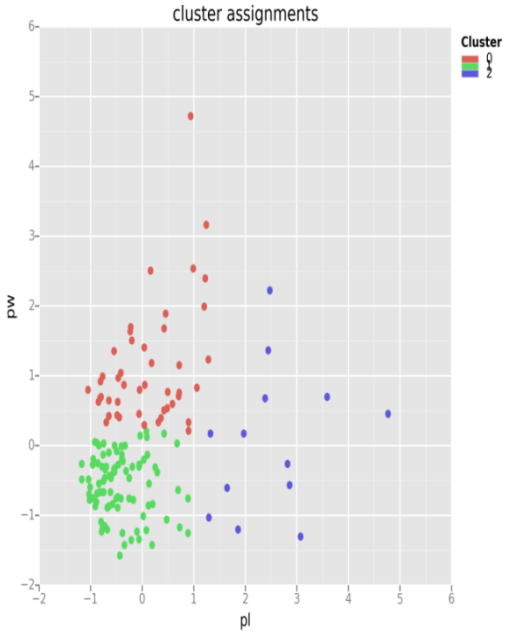

Figure 1: Cluster assignments of modified k-means clustering with n=3.

Quality of cluster validations is represented by homogeneity score, completeness score, V measure, adjusted rand score and silhouette coefficient. All these values are within the limits, given in Table-1.

Table 1: Validating the quality of clusters obtained from k-means

Quality Metric	**Obtained value from modified k-means, n=3**	**Obtained value from modified k-means, n=4**	**Obtained value from modified k-means, n=5**
Homogeneity score	0.008	0.013	0.013
completeness score	0.010	0.015	0.015
V measure	0.009	0.014	0.014
Silhouette Coefficient	0.452	0.266	0.353

From the results it is observed that the modified k-means algorithm is able to generate clusters based on user input and the validation metrics reported in Table-1 suggest that the scores of each metric with respect to k values are significant. A better Silhouette coefficient resulted in better clusters and hence k=3 represents the best cluster.

Several indicators which assess the scientific merits of researchers reported in literature quantify both the number of published papers and their citations in various other journals. To some extent, some indicators rely on the citation of articles published in journals. Few such important indices are-index, h-index, A-index and m-index. The three datasets (Tables 2-4) were obtained from modified k-means algorithm.

Table 2: Dataset-1 with parameters and index values.

Title	**H index**	**e-index**	**A index**	**m index**
Foundations and Trends in Computer Graphics and Vision	14	7	4.5	4.67
IEEE Transactions on Pattern Analysis and Machine Intelligence	221	77.79	28.38	73.67
Computer Methods in Applied Mechanics and Engineering	120	49.94	21.78	40
ACM Computing Surveys	90	28.04	9.73	30
IEEE Transactions on Evolutionary Computation	111	35.38	12.28	37

SIAM Journal on Computing	68	19.54	6.62	22.67
Computers and Education	77	60.39	48.36	25.67
Mathematical Programming	75	24.39	8.93	25
Proceedings - IEEE Symposium on Security and Privacy	43	24.19	14.6	14.33
ACM Transactions on Mathematical Software	54	15.1	5.22	18
IEEE Transactions on Mobile Computing	80	44.2	25.43	26.67
Computers and Geotechnics	48	27.64	16.92	16
Journal of Machine Learning Research	94	47.8	25.31	31.33
Foundations of Computational Mathematics	29	13.23	7.03	9.67
Artificial Intelligence	101	26.42	7.91	33.67
Computational Intelligence and Neuroscience	23	19.9	18.22	7.67
Journal of Computer Assisted Learning	48	20.9	10.1	16
Proceedings of the Annual IEEE Conference on Computational Complexity	21	10.34	6.1	7
IEEE Computational Intelligence Magazine	26	16.94	12.04	8.67
INFORMS Journal on Computing	48	13.86	5	16
Statistics and Computing	41	16.85	7.93	13.67
Journal of Scientific Computing	42	23.9	14.6	14
International Journal of Machine Learning and Cybernetics	15	17.75	22	5
IEEE/ASME Transactions on Mechatronics	74	40.79	23.49	24.67
Automated Software Engineering	29	10.1	4.52	9.67
Computational Geometry: Theory and Applications	35	12.04	5.14	11.67
Journal of Graph Algorithms and Applications	24	7.21	3.17	8
Advanced Engineering Informatics	43	24.39	14.84	14.33
Artificial Intelligence and Law	22	6.56	2.95	7.33
Theory and Practice of Logic Programming	26	14	8.54	8.67
Fuzzy Optimization and Decision Making	29	11	5.17	9.67
Computers and Mathematics with Applications	69	74.37	81.16	23
Journal of Artificial Intelligence Research	76	20.47	6.51	25.33
Mechanism and Machine Theory	60	30.53	16.53	20
Empirical Software Engineering	39	16.25	7.77	13
Algorithmica	50	19.57	8.66	16.67
International Journal of Intelligent Systems	46	20.42	10.07	15.33
IEEE Transactions on Parallel and Distributed Systems	78	43.49	25.24	26
Control Engineering Practice	67	36.12	20.48	22.33
Artificial Intelligence Review	40	20.22	11.23	13.33
Mathematics of Control, Signals, and Systems	26	6.08	2.42	8.67
ACM Transactions on Software Engineering and Methodology	53	12.41	3.91	17.67
Engineering Applications of Artificial Intelligence	54	35.81	24.74	18
Mathematical and Computer Modelling	59	49.58	42.66	19.67
Integrated Computer-Aided Engineering	25	17.38	13.08	8.33
Artificial Intelligence in Medicine	53	19.31	8.04	17.67
Advances in Engineering Software	39	26.85	19.49	13

Table 3: Dataset-2 with parameters and index values.

Title	e index	H index	A index	m index
Information Systems Research	27.46	99	8.62	33
Journal of Operations Management	28.43	108	8.48	36
IEEE Transactions on Fuzzy Systems	48.66	119	20.9	39.67
International Journal of Robotics Research	38.69	89	17.82	29.67
IEEE Transactions on Automatic Control	72.63	175	31.14	58.33
Computers and Operations Research	49.34	84	29.98	28
IEEE Transactions on Signal Processing	88.86	162	49.74	54
IEEE Journal on Selected Topics in Signal Processing	41.55	45	39.36	15
IEEE Transactions on Robotics	42.05	77	23.96	25.67
IEEE Transactions on Image Processing	74.72	169	34.04	56.33
Medical Image Analysis	36.12	76	18.17	25.33
Computers and Structures	37.5	75	19.75	25
ACM Transactions on Database Systems	12.96	59	3.85	19.67
IEEE Signal Processing Magazine	37.43	106	14.22	35.33
Journal of Field Robotics	22.43	52	10.67	17.33
IEEE Transactions on Knowledge and Data Engineering	41.89	103	18.04	34.33
Mechanical Systems and Signal Processing	47.79	79	29.91	26.33
IEEE Transactions on Software Engineering	29.75	111	8.97	37
IEEE Transactions on Neural Networks	48.25	128	19.19	42.67
Pattern Recognition	63.95	121	34.79	40.33
Machine Learning	21.45	103	5.47	34.33
Journal of Computer and System Sciences	21.95	56	9.61	18.67
IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics	49.31	102	24.83	34
Fuzzy Sets and Systems	35.96	110	12.75	36.67
Data Mining and Knowledge Discovery	18.76	62	6.68	20.67
ACM Transactions on Information and System Security	13.86	41	5.68	13.67
Computational Statistics and Data Analysis	34.47	57	21.84	19
Networks	13.6	38	5.87	12.67
Mathematics of Operations Research	12.04	50	3.9	16.67
IEEE Robotics and Automation Magazine	22.54	51	10.96	17
Information and Computation	15	50	5.5	16.67
Topics in Cognitive Science	17.66	19	17.42	6.33
ACM Transactions on Knowledge Discovery from Data	12.92	21	8.95	7
Robotics and Computer-Integrated Manufacturing	27.15	51	15.45	17
Signal Processing	51.38	77	35.29	25.67
Computational Geosciences	17.94	34	10.47	11.33
IEEE Robotics and Automation Magazine	22.54	51	10.96	17
Computational Materials Science	55.98	59	54.12	19.67
ACM Transactions on Knowledge Discovery from Data	12.92	21	8.95	7
IEEE Signal Processing Letters	38.61	82	19.18	27.33
Pattern Recognition Letters	46.7	92	24.71	30.67

Journal of Discrete Algorithms	12.17	19	8.79	6.33
Journal of Information Hiding and Multimedia Signal Processing	14.46	14	15.93	4.67
Robotics and Autonomous Systems	30.61	70	14.39	23.33
IEEE Software	20.81	72	7.01	24
International Journal of Sensor Networks	19.34	20	19.7	6.67

Table 4: Dataset-3 with parameters and index values.

Title	**e index**	**H index**	**A index**	**m index**
Archives of Computational Methods in Engineering	16.73	32	9.75	10.67
MIS Quarterly: Management Information Systems	40.48	132	13.42	44
Proceedings of the Annual ACM Symposium on Theory of Computing	26.4	40	18.43	13.33
Proceedings - Annual IEEE Symposium on Foundations of Computer Science, FOCS	24.56	62	10.73	20.67
IEEE Wireless Communications	39.94	98	17.28	32.67
IEEE Transactions on Information Theory	79.71	192	34.09	64
Journal of the ACM	17.55	88	4.5	29.33
IEEE Journal on Selected Areas in Communications	54.24	165	18.83	55
IEEE Communications Magazine	68.36	144	33.45	48
Journal of Strategic Information Systems	16.85	50	6.68	16.67
IEEE Transactions on Wireless Communications	76.86	118	51.06	39.33
Foundations and Trends in Information Retrieval	8.83	15	6.2	5
IEEE Transactions on Industrial Informatics	46.71	39	56.95	13
Information Sciences	77.68	91	67.31	30.33
Information and Organization	10.2	34	4.06	11.33
Annual Review of Information Science and Technology	6.86	38	2.24	12.67
Web Semantics	19.7	49	8.92	16.33
Proceedings - International Symposium on Computer Architecture	25.22	50	13.72	16.67
IEEE/ACM Transactions on Networking	39.59	124	13.64	41.33
Enterprise Information Systems	18.17	21	16.71	7
Journal of Management Information Systems	18.84	90	4.94	30
Journal of Computer-Mediated Communication	14.07	64	4.09	21.33
Communications of the ACM	53.14	131	22.56	43.67
Decision Support Systems	38.91	76	20.92	25.33
Information Systems Journal	13.3	52	4.4	17.33
IEEE Network	26.36	80	9.69	26.67
Journal of the American Society for Information Science and Technology	42.24	83	22.49	27.67
Journal of Information Technology	14.87	43	6.14	14.33
Journal of the Association of Information Systems	17.32	31	10.68	10.33
Knowledge and Information Systems	29.51	31	29.1	10.33
ACM Transactions on the Web	15.13	26	9.81	8.67
Information Systems	23.3	53	11.25	17.67
European Journal of Information Systems	19.24	58	7.38	19.33
ACM Transactions on Programming Languages and Systems	9.85	51	2.9	17
IEEE Communications Letters	54.17	96	31.56	32

IEEE Transactions on Services Computing	21.54	27	18.19	9
Information and Software Technology	29.34	54	16.94	18
Software and Systems Modeling	14.35	28	8.36	9.33
IEEE Internet Computing	27.02	79	10.24	26.33
Information and Computation	15	50	5.5	16.67
Soft Computing	30.25	36	26.42	12
Interacting with Computers	19.34	47	8.96	15.67
International Journal of Human Computer Studies	20.71	76	6.64	25.33
Journal of Computer Security	11.96	40	4.58	13.33
Journal of Educational Computing Research	11.45	35	4.74	11.67
Computers and Security	21.31	51	9.9	17

4. CONCLUSION

Progressive increase in the scientific methods of journal citation witnessed the recent revolution of journal publishers intending to showcase their journal quality metrics. The dataset considered in clustering is based on few objectives that the bibliometric indices of computer science journals are of prime importance in evaluating the quality of a particular journal and the grouping of journal data is dependent on index values. The outcome of modified k-means algorithm on a dataset of 150 journals resulted in 3 clusters (k=3). Quality of cluster validations is within the limits. Further, work is in progress to study the influence of parameters on journal metrics.

5. REFERENCES

[1] M. Amin & M. Mabe. 2000. Impact Factors: Use & Abuse, Perspectives in Publishing, Vol.No. 1, 1-6.

[2] Garfield, E.1996. How can impact factors be improved? British Medical Journal, 313, 411–413.

[3] Glänzel, W., & Moed, H. F. 2002. Journal impact measures in bibliometric research. Scientometrics, 53(2), 171–194.

[4] Saha, S., Saint, S. & Christakis, D.A. 2003. Impact factor: a valid measure of journal quality? Journal of the Medical Library Association 91:42-46.

[5] Dong, P., Loh, M. & Mondry, A. 2005. The "impact factor" revisited. Biomedical Digital Libraries 2:7 doi:10.1186/1742-5581-2-7.

[6] Lim, A., Ma, H., Wen, Q., Xu, Z., Cheang, B., Tan, B. & Zhu, W. 2007. Journal-Ranking.com: An Online Interactive Journal Ranking System. Proceedings of the National Conference on Artificial Intelligence, Vol.22 (2):1723-1729.

[7] Hirsch, J.E. 2005. An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences, Vol.102:16569–16572.

[8] Olden, J.D. 2007. How do ecological journals stack-up? Ranking of scientific quality according to the h-index. Ecoscience, Vol. 14(3):370-376.

[9] Rousseau, R. 2007. The influence of missing publications on the Hirsch index. Journal of Informetrics,Vol. 1:2–7.

[10] Vanclay, J.K. 2007. On the robustness of the h-index. Journal of the American Society for Information Science and Technology,Vol. 58(10):1547-1550.

[11] Steven B. Bird.2008. Journal Impact Factors, h Indices, and Citation Analyses in Toxicology. JOURNAL OF MEDICAL TOXICOLOGY, Vol. 4(4): 261-274.

[12] Egghe, L. 2006. Theory and practice of the g-index. Scientometrics, Vol. 69(1), .131–152.

[13] Kosmulski, M. 2006. A new Hirsch-type index saves time and works equally well as the original h-index. ISSI Newsletter, Vol. 2(3), 4–6.

[14] Rousseau, R. 2007. The influence of missing publications on the Hirsch index. Journal of Informetrics, Vol.1(1), 2–7.

[15] Jin, B. 2007. The AR-index: complementing the h-index. ISSI Newsletter, Vol.3(1), 6.

[16] Bornmann, L., Mutz, R., & Daniel, H. 2008. Are there better indices for evaluation purposes than the h-index? A comparison of nine different variants of the h-index using data from biomedicine. Journal of the American Society for Information Science and Technology, Vol.59(5), 830–837.

[17] Steven B. Bird.2008. Journal Impact Factors, h Indices, and Citation Analyses in Toxicology. JOURNAL OF MEDICAL TOXICOLOGY, Vol.4(4): 261-274.

[18] http://www.elsevier.com/

[19] http://www.python.org

[20] Jonathan Baarsch and M. Emre Celebi.2012."Investigation of Internal Validity Measures for K-Means Clustering" in Proceedings of the International Multiconference of Engineers and Computer Scientists, Vol 1, IMCES, HongKong.