Assessing h- and g-Indices of Scientific Papers using k-Means Clustering

S. Govinda Rao
Associate Professor, Department of CSE,
Gokaraju Rangaraju Institute of Engineering & Technology

A. Govardhan, Ph.D
Director, School of Information Technology
JNTU Hyderabad

ABSTRACT

K-means clustering technique works as a greedy algorithm for partition the n-samples into k-clusters so as to reduce the sum of the squared distances to the centroids. A very familiar task in data analysis is that of grouping a set of objects into subsets such that all elements within a group are more related among them than they are to the others. K-means clustering is a method of grouping items into k groups. In this work, an attempt has been made to study the importance of clustering techniques on h- and g-indices, which are prominent markers of scientific excellence in the fields of publishing papers in various national and international journals. From the analysis, it is evidenced that k-means clustering algorithm has successfully partitioned the set of 18 observations into 3 clusters.

Keywords

K-means, clustering, h-index, g-index

1. INTRODUCTION

The existence of journals to publish scientific research or reviews on a specified topic has been in place since many years, which raised the alarm to build databases to disseminate literature information to everyone [1]. The number of papers published in journals has been on the rise for many years and they can be affiliated on account of their citations by the scientist's worldwide. The constant and increasing the volume of scientific literature and diversification of inter-disciplinary fields of science has created wealth of knowledge useful to many scientists [2-3] which intend to solve many problems. At the same time, the scientific field has also seen a gradual increase in the number of open access journals that publish specific streams of study [4-5].

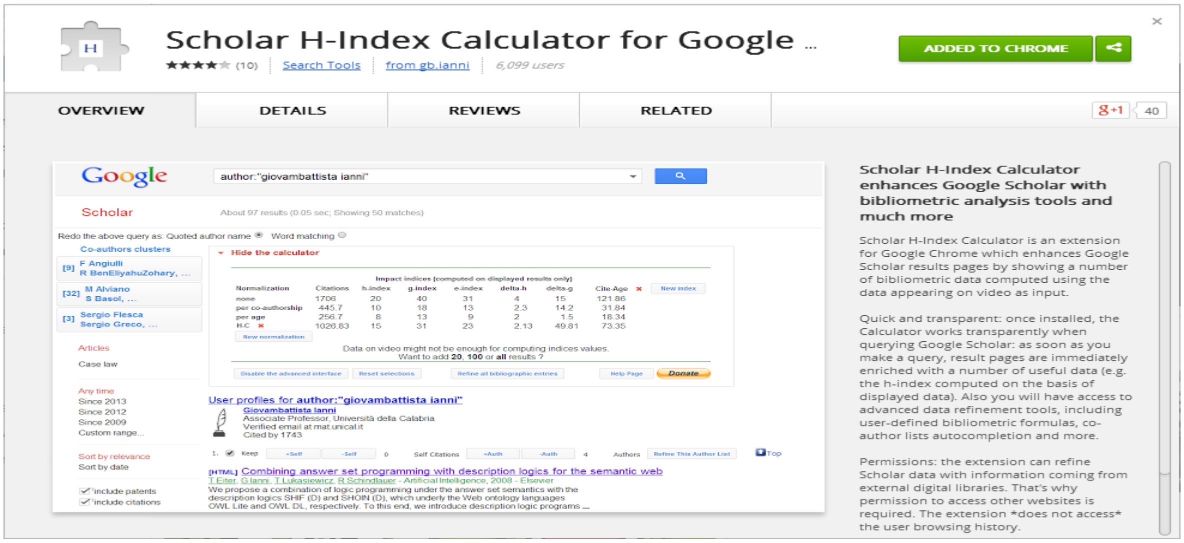
The best possible way to assess any journal is to follow the number of citations with respect to the number of papers published in a year, which is referred as Impact Factor [6]. Similarly, considering the importance of authorship of any work being cited by other works, h-index has been proposed by Hirsch [7]. This h-index evaluates the score generated from the papers published by the specific author as well as the number of papers published since the first publication [8]. However, h-index does not consider the specific field of work for instance, an author might publish papers on 'text mining', 'computer architecture', 'networking methods' etc. In such case, h-index is given for all papers published by the author, but not related to a specific field [9]. Therefore, the primary objective of this study is to calculate h-index of authors and cluster them using k-means clustering algorithm [10].

The h- and g- indices of few authors who have published scientific papers of excellence in the fields of computer science [11] are segregated. In order to collect and calculate manually, a reliable tool from Google Scholar [12] was used to perform the task. Google Chrome has developed an intuitive H-index calculator add-on to Chrome browser.

2. MATERIALS AND METHODS

h- and g- indices calculation

The h- and g- indices of few authors who have published scientific papers of excellence in the fields of computer science are segregated. In order to collect and calculate manually, which is a more tedious process than expected; a more reliable tool from Google Chrome was used to perform the task. Google Chrome has developed an intuitive H-index calculator add-on (Figure 1). Figure 2 represents an example of data obtained from Google Scholar.



author:"giovambattista ianni"

About 97 results (0.05 sec; Showing 50 matches)

Redo the above query as: Quoted author name ● Word matching ○

Impact indices (computed on displayed results only)

Normalization	Citations	h-index	g-index	e-index	delta-h	delta-g	Cite-Age
none	1706	20	40	31	4	15	121.86
per co-authorship	445.7	10	18	13	2.3	14.2	31.84
per age	256.7	8	13	9	2	1.5	18.34
H.C	1026.83	15	31	23	2.13	49.81	73.35

Data on video might not be enough for computing indices values.
Want to add **20**, **100** or **all** results ?

User profiles for author:"giovambattista ianni"
Giovambattista Ianni
Associate Professor, Università della Calabria
Verified email at mat.unical.it
Cited by 1743

1. ☑ Keep 0 Self Citations 4 Authors

[HTML] Combining answer set programming with description logics for the semantic web
T Eiter, G Ianni, T Lukasiewicz, R Schindlauer - Artificial Intelligence, 2008 - Elsevier
We propose a combination of logic programming under the answer set semantics with the description logics SHIF (D) and SHOIN (D), which underly the Web ontology languages OWL Lite and OWL DL, respectively. To this end, we introduce description logic programs ....

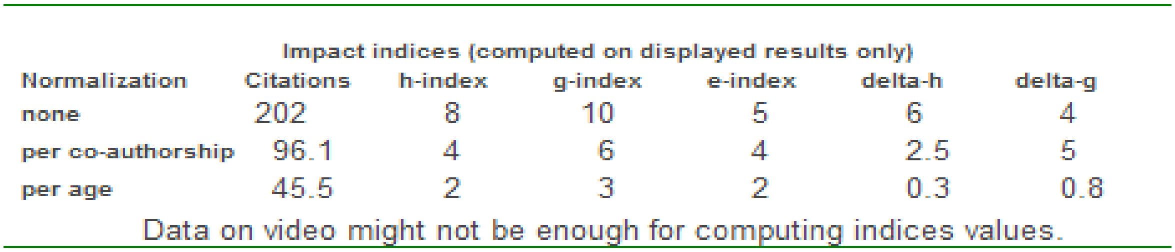
Scholar H-Index Calculator enhances Google Scholar with bibliometric analysis tools and much more

Scholar H-Index Calculator is an extension for Google Chrome which enhances Google Scholar results pages by showing a number of bibliometric data computed using the data appearing on video as input.

Quick and transparent: once installed, the Calculator works transparently when querying Google Scholar: as soon as you make a query, result pages are immediately enriched with a number of useful data (e.g. the h-index computed on the basis of displayed data). Also you will have access to advanced data refinement tools, including user-defined bibliometric formulas, co-author lists autocompletion and more.

Permissions: the extension can refine Scholar data with information coming from external digital libraries. That's why permission to access other websites is required. The extension *does not access* the user browsing history.

Figure 1. H-index calculator plug-in to Google Chrome browser



Impact indices (computed on displayed results only)

Normalization	Citations	h-index	g-index	e-index	delta-h	delta-g
none	202	8	10	5	6	4
per co-authorship	96.1	4	6	4	2.5	5
per age	45.5	2	3	2	0.3	0.8

Data on video might not be enough for computing indices values.

Figure 2. Index values computed by the calculator

Both the h,g-index are subjective to some scope by the number of papers that a journal publishes. Journal that publishes a larger number of papers has a higher likelihood of generating higher h and g-index since every article presents another opportunity for citations. The value for the indices depends on the range of papers being examined, and how comprehensively the citations for each have been indexed. The main power of the h-index is that it calculates quantity and impact by means of a single indicator. Egghe [13] defines g-index as “the highest rank such that the top g papers have, together, at least g2 citations. It also called that the top (g + 1) have less than (g + 1)2 papers”. The g-index is always greater than or equal to h-index.

Dataset

The dataset for k-means clustering analysis is given in the Table 1. The data was extracted from Google Scholar add-on H-index calculator.

Table 1. h- and g-indices calculated from the Google Scholar search

Name	h-index	g-index
A Govardhan	8	10
B Jalender	4	5
TB Reddy	17	32
RS Sisodia	21	48
SS Doshi	2	2
D Vasumathi	7	10
NMA Munassar	2	6
Y Sankarasubramaniam	9	70
MK Sundareshan	22	37
SI Sudharsanan	10	23
RC Rose	53	102
V Ganapathy	47	83
S Keshav	19	58
S Ur Rahman	8	16
MH Falaki	4	10
U Ismail	9	17
M Derakhshani	7	15
T LeNgoc	2	3
NH Ahmed	8	19
MK Pakhira	7	25
BB Jayasingh	2	3
BR Mohan	12	21
B Swathi	4	5
BV Swathi	2	2
MR Patra	9	15

Python program for k-means Clustering

The k-means clustering algorithm is well-liked because it can be applied to subjectively large sets of data. The user specifies the k-number of clusters to be found. The algorithm then separate the data into spherical clusters by finding a set of centroids, assigning each observation into a cluster and identifying new centroids, and repeat this process. The data file is formatted as a comma delimited file .csv and the file was called from python and any errors in the csv file such as white spaces/ new lines or quotes etc are rectified.

```
def lineToTuple(line):
    clearTrace = line.strip()
    clearTrace = clearTrace.replace('"', '')
    lineClear = clearTrace.split(",")
    convertString(lineClear)
    lineTuple = tuple(lineClear)
    return lineTuple
```

The number of centroids can be changed depending on the data being clustered. In this case, 3 centroids are considered.

```
showDataset2D(dataset)
clustering = kmeans(dataset, 3, True)
printTable(clustering["centroids"])
```

The k-means algorithm was implemented by finding distances between objects taken for the study. Once calculated, the sum of squares for the two objects from the centroid is calculated. Means are calculated for all objects which are nearer to each other making it to study all possible objects and the List of such objects will be created by the program to iterate the process. Centroids are assigned and the distance is calculated from each centroid to the objects and clusters are identified. To find out better clusters, a minimum distance is calculated from each object to the nearest assigned centroid.

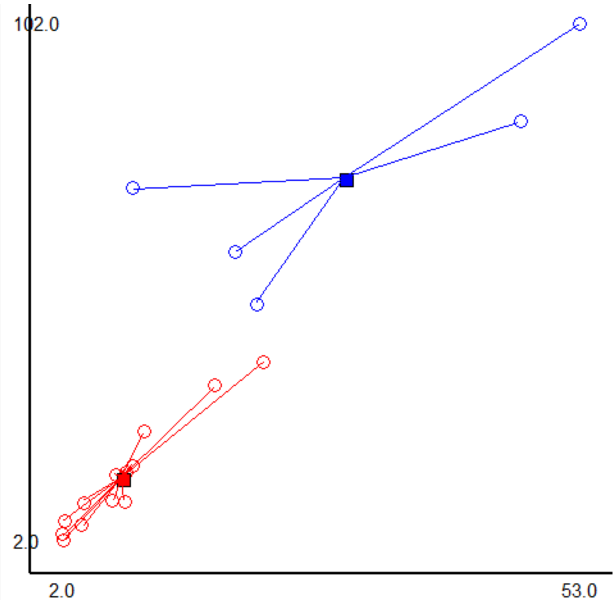
```
def distance(object1, object2):
    if object1 == None or object2 == None:
        return float("inf")
    SquareSum = 0
    for i in range(1, len(object1)):
        SquareSum += (object1[i] - object2[i])**2
    return SquareSum
```

Once each object is assigned to its nearest centroid, then the distances are recalculated by adjusting the centroids in such a way that the objects should have the shortest possible distance with the centroid. Likewise, a cluster index is created and appended to the objects. Finally, the k in k-means is calculated by randomly selecting k initial centroids. In order to visualize the procedure being followed, an animation was implemented to evaluate the cluster generation and objects are assign to each cluster based on the minimum distances from each centroid. The sum of squares (ss) of the distance of all points within a cluster to the centroid of the cluster is measured.

3. RESULTS AND DISCUSSION

K-means suffers from drawback on the number of clusters k as an input argument. This is because of an inappropriate choice of k which might give up spurious outputs. Hence, it is always an important task to run diagnostic checks when using k-means algorithm to determine the number of clusters in the given dataset. Moreover, applying k-means value with values ranging from k=2, 3, 4 or 5 depends on the number of objects in the dataset and to avoid predictable clusters of similar size so that the observation of objects to the nearest cluster centroid will result in correct clusters. Therefore, an attempt has been made to cluster dataset using different k value ranges.

Option 1: k=2

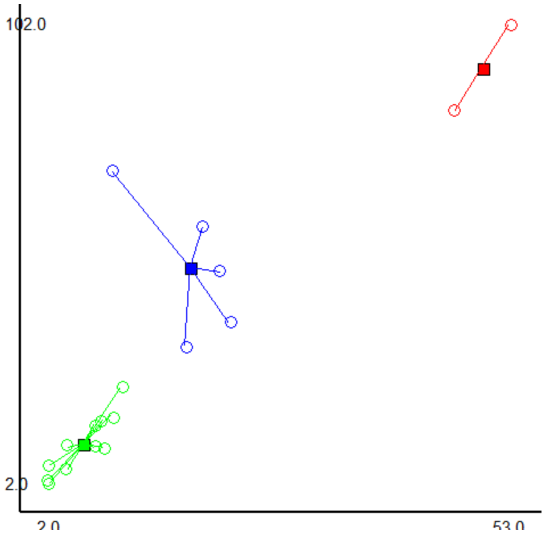


The two centroids data is

centroid0 29.80 72.20

centroid1 7.85 14.31

Option 2: k=3



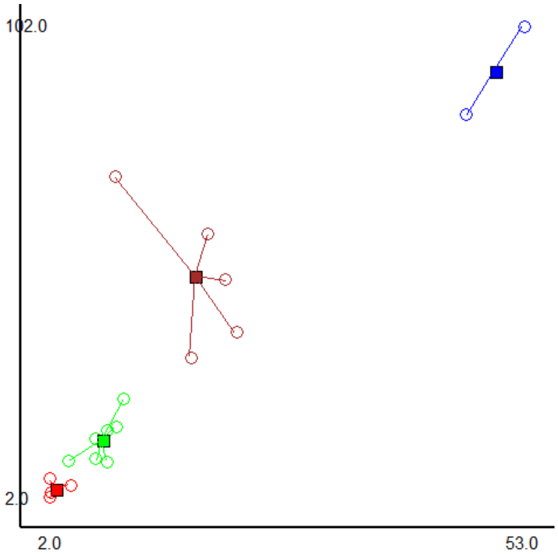
The three centroids are:

centroid0 50.00 92.50

centroid1 17.60 49.00

centroid2 5.73 10.64

Option 3: k=4



The four centroids are:

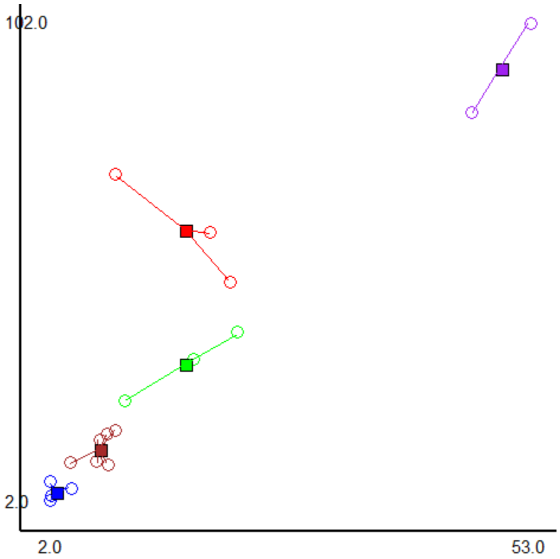
centroid0 50.00 92.50

centroid1 2.50 4.00

centroid2 7.57 14.43

centroid3 17.60 49.00

Option 4: k=5



The five centroids are:

centroid0 2.50 4.00

centroid1 16.33 58.67

centroid2 16.33 30.67

centroid3 7.17 13.00

centroid4 50.00 92.50

In Option-1, when k=2, the data was segregated into two clusters by calculating the median of points surrounding the centroids. From the image, the two clusters represented defined points being clustered as entities.

In Option-2, k=3, the three clusters represented a more feasible form of groups as the first cluster has two objects and middle cluster has five objects and the rest in the third cluster.

Option-3, k=4 does not represent a better classification group when compared with the remaining options because the third cluster has been split into further two classes whereas the first and second clusters remained as such.

Finally, option-4, k=5 showed a new cluster group formed by splitting the second cluster into two groups

From the above options and plots, it is evidenced that the k=3 represents the most appropriate option to cluster the given dataset. Because, considering the data representing the 3rd cluster in all the cases is justified, however, the first and middle clusters showed variance with respect to k value and the divergence of sets are little more non-compliance to the data except when the k value is 3.

4. CONCLUSION

Considering the importance of h- and g-indices for each author as a parameter to assess the quality of published papers in various journals, a k-means algorithm was implemented to study the objects used. A option was provided to choose the number of centroids and objects are assigned to each cluster based on the minimum distances from each centroid. It was observed that k=3 represents the best option to cluster the given dataset. k-means clustering has successfully partitioned the set of 18 observations into k clusters (3 clusters) in which each observation or an object belongs to one of the three clusters with the nearest mean. The work shall be extended to include cluster plots of varying significance.

5. REFERENCES

[1] http://www.sagepub.com/upm-data/29986_Chapter3.pdf

[2] G Charles Babu and Dr. A.GOVARDHAN, “Mining Scientific Data from Pub-Med Database” International Journal of Advanced Computer Science and Applications(IJACSA), 3(4), 2012.

[3] Richard Van Noorden (2013). Open access: The true cost of science publishing. Nature 495, 426–429

[4] Solomon, D. J. & Björk, B.-C. J. Am. Soc. Inf. Sci. Technol. 63, 1485–1495 (2012)

[5] Jerry A. Jacobs and Scott Frickel. Interdisciplinarity: A Critical Assessment. Annual Review of Sociology, 35: 43 -65 (2009)

[6] http://en.wikipedia.org/wiki/Impact_factor

[7] Hirsch, J. E. (2005). "An index to quantify an individual's scientific research output". PNAS 102 (46): 16569–16572

[8] Jacso, P. (2008b). The pros and cons of computing the h-index using Google Scholar. Online Information Review, 32(3), 437–452

[9] Jin, B. (2006). h-Index: An evaluation indicator proposed by scientist. Science Focus, 1(1), 8–9

[10] J. B. MacQueen (1967): "Some Methods for classification and Analysis of Multivariate Observations, *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability"*, Berkeley, University of California Press, 1:281-297

[11] S. Alonso, F.J. Cabrerizo, E. Herrera-Viedma, F. Herrera (2009). h-Index: A review focused in its variants, computation and standardization for different scientific fields. Journal of Informetrics 3: 273–289

[12] Google Scholar. (online resource). http://scholar.google.com/

[13] EGGHE, L. (2006), Theory and practise of the g-index. Scientometrics, 69 (1) : 131–152

6. AUTHOR’S PROFILE

S Govinda Rao working as Associate Professor in Department of CSE, GRIET Hyderabad.

He was completed M.Tech [IT] from Andhra University ,Visakhapatnam,india in 2006.Now he doing Ph.D(CSE) from JNTU Kakinada.He is around 10 years of Teaching experience.

Dr.AGovardhan working as Director, School of Information Technology, JNTU Hyderabad. He did B.E in computer science and engineering from Osmania University College of Engineering, Hyderabad, India in 1992, M.Tech from Jawaharlal Nehru University (JNU). Delhi in 1994 and he earned his Ph.D from Jawaharlal Nehru Technological University, Hyderabad (JNTUH) in 2003.He is around 20 years of teaching experience.