Implementation of K-Means Technique in Data Mining to Cluster Researchers Google Scholar Profile

Gigih Forda Nama, Lukmanul Hakim, Junaidi

Abstract: ***A university usually has many Lecturers that have an important role in improving the quality of Higher Education. The Lecturers should produce scientific publications at least 1 publication on each semester. The achievements of a Lecturer in research and publication become the main indicator that describes the professionalism of lecturers as scientists. Monitoring the improvement of publication trends is very important to do as an evaluation for organizational management in choosing the best strategy to strengthen the quality of publication, and one of the common tools used for analyzing the publication data is the Google Scholar system. This paper attempts to analyze the Google scholar data using Data Mining techniques (Text Mining) by R language, to collect Lecturer's profile and list of publications in a real-time, the aim of this research is to allow the Management for identifying the Cluster from total 1039 Lecturers on University of Lampung. The results of this research shown there were 5 Cluster of scholar profile data, with member details C0=102, C1=924, C2=1, C3=1, C4=11, total 88.93% of Lecturers are on cluster C1 with, centroid data was h_index=1.942, total_cites=20.89, i10_index=0.417.***

Keywords: Text Mining, r Language, Clustering, Data Mining, Google Scholar, Publication Analysis

I. INTRODUCTION

A university usually has many Lecturers that play an important role in improving the quality of Higher Education. In Indonesia, Lecturers should produce scientific publications at least 1 publication in 1 semester. The achievements of a Lecturer in research and publication become the main indicator that describes the professionalism of lecturers as scientists. Monitoring the improvement of publication trends is very important to do as an evaluation for organizational management in choosing the best strategy to strengthen the quality of publication, one of the common tools used for analyzing the publication data is The google Scholar system.

University of Lampung as a Public University in Bandar Lampung, Indonesia, has total 1039 Lecturers, each of Lecturers has a unique scientific reputation and produced various publications in several fields of science. Unfortunately the scholar profile identification process was done manually by checking author profile through google scholar web interface.

This research aims to recognize and analyze the knowledge pattern of the Lecturers on Google Scholar profile, for helping the University of Lampung Management in taking the right policy to improve the quality of scientific publications in the future.

II. LITERATURE REVIEW

The first work carried out in this research was to specify the suitable technology on the analysis of very large dimension of scientific publications data on the Google Scholar system. After going through the process of gathering and considering information from the Gartner's magic quadrant recommendation related to Machine Learning and Data Science comparison platforms, and also looking inside on work related to comparative technology analysis made by Gregor [1], it was decided to use Rapid miner technology. On his work, Gregor concluded that Rapid miner still plays a dominant role since the last five years. It provides lot of variety on smart technologies modeling for automated end to end development, also has very interactive visual workflow designer front-end, guided analytics, and it also supports for automatic retraining models, based on many platform data interchange.

Fig. 1 The Gartner Magic Quadrants on Machine Learning and Data Science and Platforms comparison, 2017 vs 2018 [1]

Revised Manuscript Received on September 22, 2019.

Gigih Forda Nama, Department of Informatics, University of Lampung
Lukmanul Hakim, Department of Electrical Engineering, University of Lampung
Junaidi, Department of Physics, University of Lampung

Fig.1 indicates the Gartner Magic Quadrant [MQ] for Machine Learning and Data Science and platforms comparison of years 2017 vs 2018. It can be concluded that there are five leading technology i.e. Rapid miner, Alteryx, SAS, KNIME, H2O.ai. There are also two companies trying to challenge the competition that are: Math Works, TIBCO Software (new), and five Visionaries: Databricks (new), IBM, Domino Data Lab, Dataiku, Microsoft, and there are four Niche Players: SAP, Anaconda (new), Angoss Teradata, and three new firms were added in 2017: TIBCO Software, Anaconda, and Databricks. Three companies shown on Magic Quadrant 2017 and disappeared in 2018 are: Alpine Data, FICO, and Quest.

Besides Rapid miner, R language was also popular tools as a text mining application. Some research conducted on R such; G. Wang et al [2] created the modeling problem formula of microbial fermentation evaluation and prediction using R language. R language was also implementing on Text Mining research and act as statistical analysis tool and running well on Ubuntu Linux LTS version 12.04 has been done by Agnihotri et al on works [3]. Data Analysis using R language for several purpose also found on research [4][5][6][7].

In the field area of Google Scholar research, Pratiba et al, conducted a research and trying to build an application that use web scraping and crawling techniques on Python language programming, to identifying the list of researcher's publications from the Google Scholar system and stored the data to a MySQL system and also Excel data [8]. While Yang et al on work [9] using Google Scholar and APIs technology to analyze the metadata of scientific publication, such as conferences and journals, the authors, title of publications and organizations affiliation. Other works on Google Scholar exploration found on work [10][11][12] using web scraping technology, Google's API, to analyze researcher profile, citation count. Some proven works implementing K-Means methodology for clustering founded on works [13][14][15] [16][17].

Clustering concept is the task on how to divide populations or data set into a specific number of groups that has similar pattern, final aim is to segregate the groups with similar traits and assign them into clusters. The K-means clustering was the well-known unsupervised machine learning algorithm, and widely used for partitioning the data into a set of specified k groups numbers. K-means algorithm process described as follows [18]:

1. Determine the amount number of clusters (K) to be created.
2. Choose randomly the k objects from data set as the centers of initial cluster or means.
3. The next is to assign each observation to their closest centroid object, based on Euclidean Distance between object and centroid, the formulation shown on formula (1)

$$d(x,y) = \|x - y\| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad ; i = 1,2,3 \dots n \qquad (1)$$

Where; xi = the object coordinat x on i
yi = the object coordinat y on i
n = the dimention of data

4. Perform iteration for each of k clusters, and update the cluster centroid by calculating the new means values of all data points in cluster.
5. Iterate steps three and four until the cluster assignments process stops, or the maximum number of iterations is reached.

III. METHODOLOGY

This research methodology divide into 2 Phase;

A. Web-based scholar application Development

In this phase R language was used for scraping the scholar data from google, PHP programming language was used for development scholar data visualization through web interface, MySQL system used for data store.

B. Data mining analysis from web-based scholar application

In this research, The Cross Industry Standard Process for Data Mining (CRISP DM) was employed, related researches using this method can be found on several works [19][20][21][22].

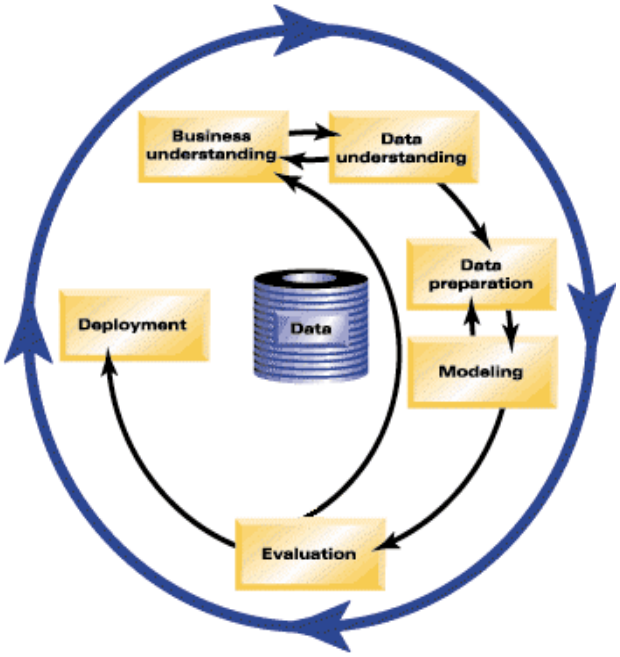


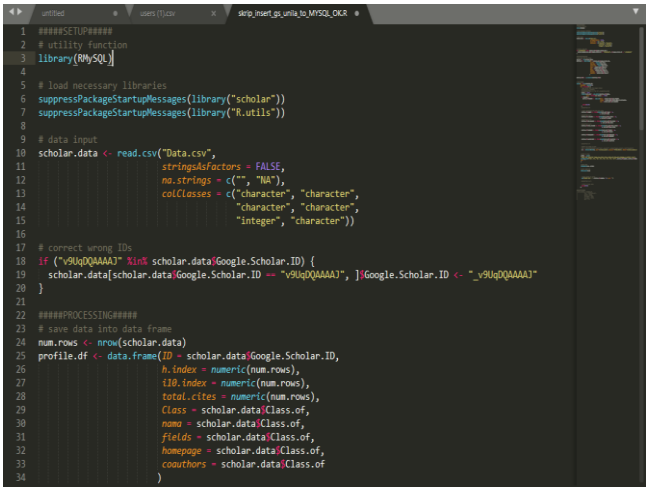
Fig. 2 The Cycle of CRISP DM [23]

There are six phases of CRISP DM life cycle which include understanding the business process, understanding the data flow, preparing data, modeling process, evaluating, and deploying), as in Figure 2. The arrow symbols indicate the importance and relation between each phase, while the sequence process on phases is not strict. This model is more flexible and customizable according to the needs. Instead of modeling, the operation shall focus on data visualization, exploration to identifying the knowledge pattern. It also allows to create model a data mining that fits with particular needs.

IV. RESULTS AND DISCUSSION

A. Web-based scholar application Development

In this phase R language was used for scraping the scholar data from google, PHP programming language was used for development scholar data visualization through web interface, MySQL system used for data store.



```
#####SETUP#####
# utility function
library(RMySQL)

# load necessary libraries
suppressPackageStartupMessages(library("scholar"))
suppressPackageStartupMessages(library("R.utils"))

# data input
scholar.data <- read.csv("Data.csv",
                         stringsAsFactors = FALSE,
                         na.strings = c("", "NA"),
                         colClasses = c("character", "character",
                                        "character", "character",
                                        "integer", "character"))

# correct wrong IDs
if ("v9UqDQAAAAJ" %in% scholar.data$Google.Scholar.ID) {
  scholar.data[scholar.data$Google.Scholar.ID == "v9UqDQAAAAJ", ]$Google.Scholar.ID <- "_v9UqDQAAAAJ"
}

#####PROCESSING#####
# save data into data frame
num.rows <- nrow(scholar.data)
profile.df <- data.frame(ID = scholar.data$Google.Scholar.ID,
                         h.index = numeric(num.rows),
                         i10.index = numeric(num.rows),
                         total.cites = numeric(num.rows),
                         Class = scholar.data$Class.of,
                         nama = scholar.data$Class.of,
                         fields = scholar.data$Class.of,
                         homepage = scholar.data$Class.of,
                         coauthors = scholar.data$Class.of
                         )
```

Fig. 3 R-language used for scraping scholar data

Figure 3 shown snapshot of R-language for scrapping scholar information from google scholar system, several library was used on this program those are RMySQL, scholar, R.utils. All data was stored to MySQL system.

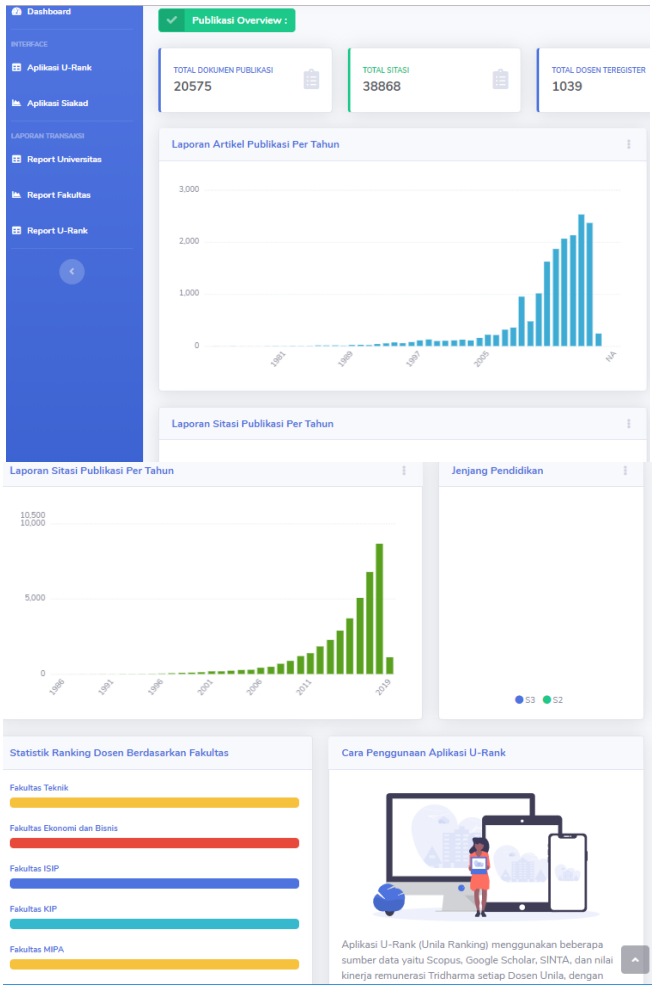


Fig. 4 Web-based scholar application report

Figure 4 shown Web-based scholar application report, this application developed by PHP programming language, this report contain publication information from all Lecturers.

B. Data mining analysis from web-based scholar application

1. Phase 1 - Business Understanding

The first stage to do in CRISP DM framework is to analyze a concrete primary business objective activity to specific data mining role. This research has business activity goal on how to identify the Lecturers profile and information about scientific publication from scholar application.

2. Phase 2 - Data understanding

Very important to understand the scientific publication and store the data on to database system. Figure 5 shown the data attributed create on MySQL system.

Fields | Indexes | Foreign Keys | Triggers | Options | Comment | SQL Preview

Name	Type	Length	Decimals	Not null	Virtual	Key
id_user	int	255	0	☑	☐	1
nip	varchar	255	0	☐	☐	
nidn	varchar	20	0	☐	☐	
nama	varchar	255	0	☐	☐	
afiliasi	varchar	255	0	☐	☐	
fakultas	varchar	255	0	☐	☐	
jurusan	varchar	255	0	☐	☐	
program_studi	varchar	255	0	☐	☐	
foto	varchar	255	0	☐	☐	
id_gs	varchar	255	0	☑	☐	
total_cites	varchar	65	0	☐	☐	
h_index	varchar	255	0	☐	☐	
i10_index	varchar	255	0	☐	☐	
mail	varchar	255	0	☑	☐	
homepage	varchar	255	0	☐	☐	
co_authors	varchar	255	0	☐	☐	
flag_admin	varchar	255	0	☐	☐	
u_score	float	255	2	☐	☐	
sinta_id	int	30	0	☐	☐	
jabatan_akademik	varchar	255	0	☐	☐	
pendidikan_terakhir	varchar	255	0	☐	☐	
sinta_score_3	float	10	2	☐	☐	

Fig. 5 Dataset attribute of google scholar data

The dataset has 23 attributes those are; "id_user", "nip", "nidn", "nama", "afiliasi", "fakultas", "jurusan", "program_studi", "foto", "id_gs", "total_cites", "h_index", "i10_index", "mail", "homepage", "co_authors", "flag_admin", "u_score", "sinta_id", "jabatan_akademik", "pendidikan_terakhir", “sinta_score_3”, “sinta_score”.

3. Phase 3 - Data Preparation

This stage is one of the most important and usually takes more time than other phases in data mining stage. In reality, around 50-70% of research time table used for data preparation. The purpose of the preprocessing activity is to set the data into desired normal form for the next step of data mining process. In this this stage we should to perform the following as below:

- Probably merging the data sets or/and records.
- Select subset of data for analyze.
- Aggregating the records of data.
- Deriving the data into new attributes.

- Sorting the whole data for modeling activity.
- Replacing or remove blank data, or missing values, or outliers.
- Splitting the data and separate to training and test data sets

Choose the items dataset; The initial research will be limited to 1039 Lecturers data on google scholar.

Choose the attributes. The scholar database consist of various information about Lecturers, it is important to filter the attributes of data such removing some unnecessary attribute.

The dataset has total 23 attributes those are; "id_user", "nip", "nidn", "nama", "afiliasi", "fakultas", "jurusan", "program_studi", "foto", "id_gs", "total_cites", "h_index", "i10_index", "mail", "homepage", "co_authors", "flag_admin", "u_score", "sinta_id", "jabatan_akademik", "pendidikan_terakhir", "sinta_score_3", "sinta_score", we should eliminate (19) attribute except id_gs, total_cites, h_index, i10_index for dataset that will be used for Data Clustering analysis.

Row No.	id_gs	h_index	total_cites	i10_index
1	GeEBn64AAA...	7	312	4
2	Iqs5ptoAAAAJ	5	344	4
3	mwv95nUAA...	20	1618	32
4	nNGsWSUAA...	6	195	5
5	-sxXH5cAAAAJ	7	144	7
6	_6JQYc0AAAAJ	5	72	2
7	Mp5HK2UAA...	6	338	4
8	jFz1eZoAAAAJ	5	145	4
9	Tf8QNOkAAAAJ	3	31	0
10	AlyUxl8AAAAJ	6	170	4
11	jOHI7RYAAAAJ	6	163	5
12	fySkCBwAAAAJ	7	87	5
13	ov1Vc0EAAAAJ	5	158	4
14	U0h5EWsAA...	3	21	0

Fig. 6 Grouping the data and set the role of attribute

Figure 6 described attributes that used for scholar information analysis, each of attribute should have specified role, id_gs was set as a special attribute (label), h_index role set to regular attribute, total_cites set to regular attribute, and i10_index role also regular attribute.

4. Phase 4 - Modelling

After going through the process of preprocessing data and transform the data in to desired structure, the CRISP DM process continues to modeling phase. This stage are the main part of this research. Several model implemented to processing the scholar data.

- Modeling scholar cluster using K-Mean algorithm on Rapidminer.
- Identify scholar data cluster.

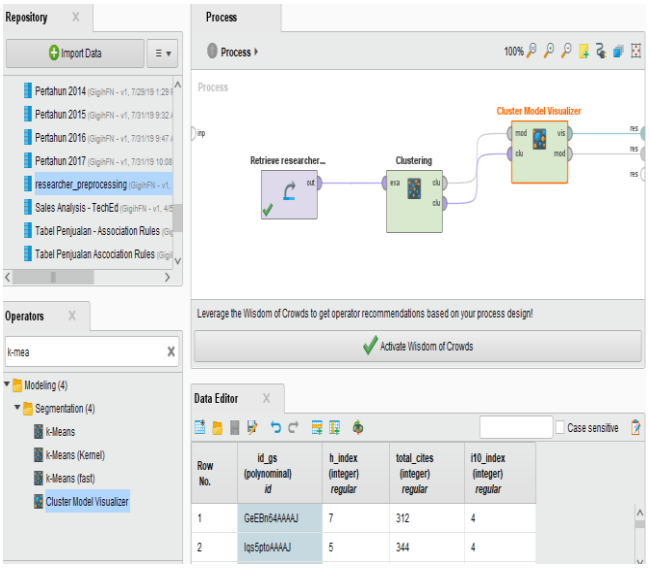


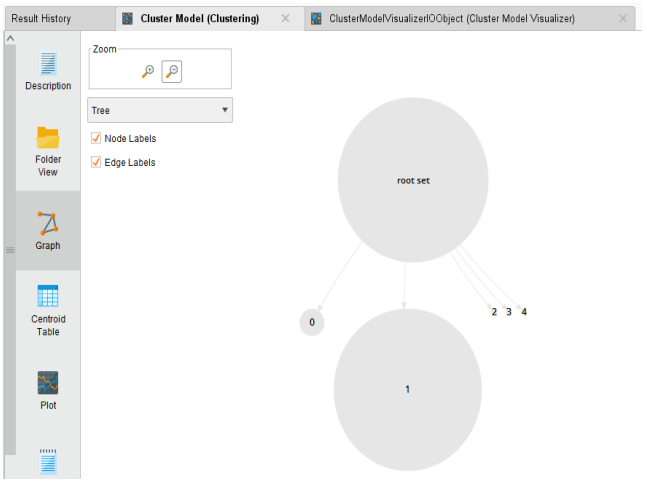
Fig. 7 The process creating operator model for dataset

Fig. 7 described the process for running the K-means algorithm on Rapid Miner, there were several operators involved (cluster model visualizer, k-means clustering) which aims to identified the centroids on each cluster. Here are the parameter implementing on K-Means cluster;

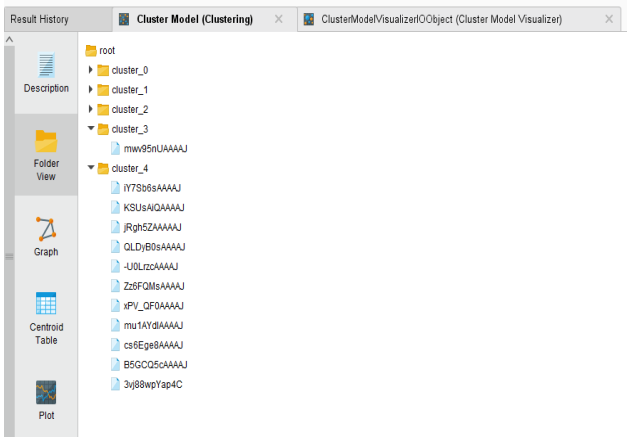
k = 5,
max runs = 10
measures type = Bregman Divergence
divergences = Square Eucledian Distance
max optimization steps = 100

DM Modeling Result

After running the model that build on previous stage of scholar data using K-Means algorithm, Rapidminer compiler produced the information of cluster model, the result was the cluster clasify to five cluster with members as follow ; Cluster 0 with 102 members, Cluster 1 with 924 members, Cluster 2 with 1 member, Cluster 3 with 1 member, Cluster 4 with 11 members, and total number of data was 1039, the graphical data visualization of each cluster member described on figure 8.



(a) Cluster visualization



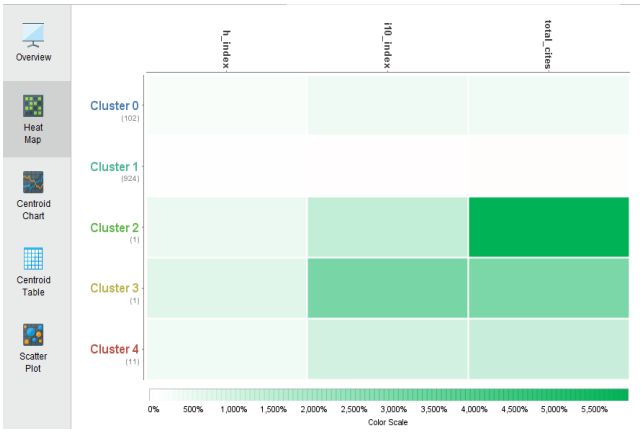
(b) Cluster member

Fig. 8 (a) (b). Graph of scholar data cluster

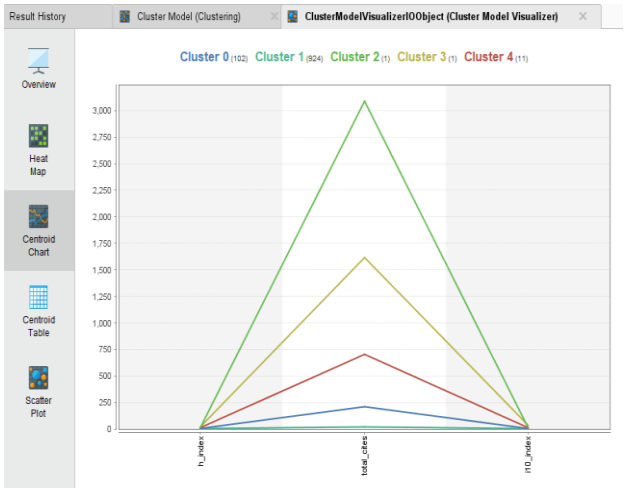
Cluster Model (Clustering) | ClusterModelVisualizerIOObject (Cluster Model Visualizer)

Cluster	h_index	total_cites	i10_index
Cluster 0	6.402	212.471	4.618
Cluster 1	1.942	20.898	0.417
Cluster 2	14	3091	15
Cluster 3	20	1618	32
Cluster 4	10.545	705.364	11.364

(a) Centroid table of clusters



(b) The heat-map of cluster visualization



(c) Graph of centroid chart on each cluster

Fig. 9 (a) (b) (c) Cluster data visualization

Figure 9 describe the cluster of scholar data visualization, especially on figure 8 part (a) explains the centroid table of scholar data cluster, in these results, rapidminer clustering from total 1039 record, clasify the data into 5 clusters, based on initial partition that defined on previous modeling stage. Cluster 0 until Cluster 4 has 3 observations attributes, those are h_index attribute, total_cites attribute, h10_index attribute. It can be concluded that the result of cluster produced already adequate and represented the actual data.

Fig. 10 Result of google scholar cluster overview

Figure 10 explain the whole data cluster overview, with total amount of cluster is five, the clustered was created by using algoritm distance measure squared euclidean distance, with average data cluster distance result is 1538.20, and Davies Bouldin index is 0.298, following explanation detail is;

1. The Cluster 1 is the largest cluster with 924 total member, and it has average distance 757.184.
2. The Cluster 2 has total 102 member, and it has average distance is 6139.235.
3. The Cluster 4 has total 11 member, and it has average distance is 24759.62
4. The Cluster 2 has total only 1 member, and it has average distance is 0
5. The Cluster 3 has also only 1 member, and it has average distance is 0

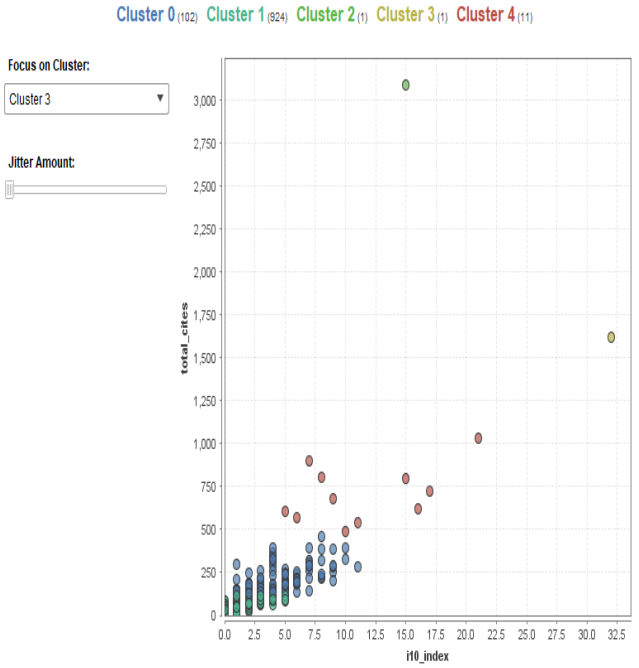


Fig. 11 Scholar data cluster scatter-plot visualization

Figure 11 explain the scholar cluster on scatter plot data visualization, from the chart it can be concluded that scholar data dominan on cluster 1 with centroid value is h_index=1.942, total_cites=20.89, i10_index=0.417. While the data statistic of scholar data show on figure 12.

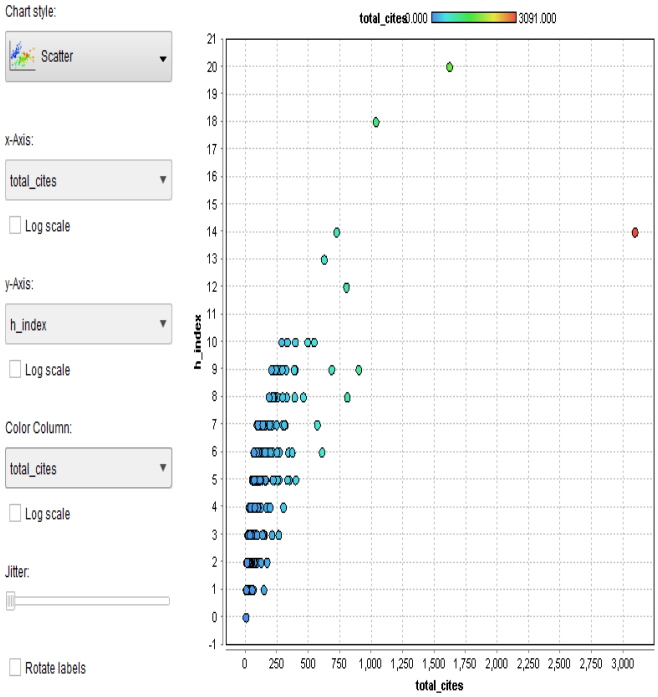


Fig. 12 Scholar Data Scatter Plot correlation between h_index and total_cites

Fig. 12 explained Data Scatter Plot correlation between h_index and total_cites, from the scatter plot graphic indicated that h_index directly linear proporsional to total citation.

Table. 1 Descriptive Statistic of scholar data

	h_index	total_cites	i10_index
Mean	2,499518768	51,44369586	0,989412897
Standard Error	0,073358069	4,464658651	0,071171238
Median	2	12	0
Mode	1	0	0
Standard Deviation	2,364588987	143,911677	2,294099733
Sample Variance	5,591281078	20710,57077	5,262893586
Kurtosis	5,916477521	210,2068834	45,55223798
Skewness	1,782152014	11,74821491	5,24872581
Range	20	3091	32
Minimum	0	0	0
Maximum	20	3091	32
Sum	2597	53450	1028
Count	1039	1039	1039

Table 1 described the descriptive statistic of total data, some important information founded that the Mean of h_index=2.49, total_cites=51.44, i10_index=0.98

5. Evaluation

This research using Rapidminer licensed for education, with this allowed us to process unlimited rows of data, even though only single processor can be used, this version also including the premium features such as Rapidminer Turbo Prep, and auto model feature, the rapidminer system running on experiment environment with Processor type is Intel (R) Core (TM) i7-3632 QM with CPU clock is 2.20 Ghz, also installed memory is 16 Giga Byte, with Window 10 version 64-bit professional edition, Solid State Drive (SSD) storage with capacity 1 Tera Byte. Data mining can run well on this environment.

6. Deployment

The final stage of CRISP DM framework is deployment phase, in this stage conducted information dissemination and took new insights of scholar data pattern founded during the research, be taken into consideration in decision making for improve the quality of research publications. Based on cluster pattern result analysis, it shown that the scholar data was so varied between one researcher another, we made some reported to the stake-holders and recommended them to conduct proper policy to improve the quality of scientific publication of each Lectures.

V. CONCLUSION

The main aim of this research is to recognize and analyze the knowledge pattern of the Lecturers on Google Scholar profile using the well-known frame work on Data Mining called Cross-Industry Standard Process for Data Mining (CRISP DM), it is necessary for helping the stake holders to conduct proper policy to improve the quality of scientific publication, the initial research limited to total 1039 Lecturers scholar data.

Based on the data mining analysis, it is shown that the Lecturers clasification divided into 5 Clusters, mostly the Lecturers are on Cluster 1 with total 924 members, and centroid h_index=1.942, total_cites=20.898, i10_index=0.417

ACKNOWLEDGMENT

We are also grateful to Kemenristekdikti and Lembaga Penelitian dan Pengabdian Universitas Lampung for supporting funding this research by Hibah Institusi scheme.

REFERENCES

1. P.Gregor. Gartner 2018 Magic Quadrant for Advanced Analytics Platforms: who gained and who lost.
2. G. Wang, Y. Xu, Q. Duan, M. Zhang, B. Xu, Prediction model of glutamic acid production of data mining based on R language, 2017 29th Chinese Control And Decision Conference (CCDC).
3. D. Agnihotri, K. Verma, P. Tripathi, Pattern and Cluster Mining on Text Data, 2014 Fourth International Conference on Communication Systems and Network Technologies.
4. J. Rajanikanth, T. V. R. Kanth, An Explorative Data Analysis on Bangalore City Weather with Hybrid Data Mining Techniques using R, 2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC).
5. A. Nasridinov, Y. Park, Visual Analytics for Big Data Using R, 2013 International Conference on Cloud and Green Computing.
6. D. Pant, V. Kumar, J. Kishore, R. Pal, Healthcare data modeling in R, 2017 1st International Conference on Intelligent Systems and Information Management (ICISIM).
7. P. R. Visali Lakshmi, G. Shwetha, N. Sri Madhava Raja, Preliminary big data analytics of hepatitis disease by random forest and SVM using r-tool, 2017 Third International Conference on Biosignals, Images and Instrumentation (ICBSII).
8. D. Pratiba, A. M S, A.L Dua, G. K. Shanbhag, N. Bhandari, U. Singh, 2018, Web Scraping And Data Acquisition Using Google Scholar, 2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS).
9. D. Yang, A. N. Zhang, W. Yan, 2017, performing literature review using text mining, Part I: Retrieving technology infrastructure using Google Scholar and APIs, 2017 IEEE International Conference on Big Data (Big Data).
10. P. T. Breuer, J. P. Bowen, Empirical Patterns in Google Scholar Citation Counts, 2014 IEEE 8th International Symposium on Service Oriented System Engineering.
11. Q. T. Le, D. Pishva, Application of Web Scraping and Google API service to optimize convenience stores' distribution, 2015 17th International Conference on Advanced Communication Technology (ICACT).
12. A. F. Rochim, A. Muis, R. F. Sari, Comparison of the H-Index of Researchers of Google Scholar and Scopus, World Academy of Science, Engineering and Technology International Journal of Educational and Pedagogical Sciences, Vol.11, No.10, 2017.
13. M. R. Mahmud, M. A. Mamun, M. A. Hossain, M. P. Uddin, Comparative Analysis of K-Means and Bisecting K-Means Algorithms for Brain Tumor Detection, International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), 2018.
14. M. S. Rahim, T. Ahmed, An initial centroid selection method based on radial and angular coordinates for K-means algorithm, 20th International Conference of Computer and Information Technology (ICCIT), 2017.
15. M. A. Altuncu, B. Türkoğlu, M. A. Çavuşlu, S. SahIn, Implementation of K-means algorithm on FGGA, 26th Signal Processing and Communications Applications Conference (SIU), 2018.
16. P. Manivannan, P. Isakki Devi, Dengue fever prediction using K-means clustering algorithm, IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), 2017.
17. D. Despa, G. F. Nama, An Enhanced K-Means Clustering Algorithm for Pattern Discovery in Big Data Analysis of 3-Phase Electrical Quantities, International Journal of Engineering & Technology, Vol 7, No 4.44, 2018.
18. UC Business Analytics R Programming Guide - K-mean Algorithm. https://uc-r.github.io/kmeans_clustering, 2018.
19. P. Kalgotra, R. Sharda, Progression analysis of signals: Extending CRISP-DM to stream analytics, IEEE International Conference on Big Data (Big Data), 2016.
20. F. Chiheb, F. Boumahdi, H. Bouarfa, D. Boukraa, Predicting students performance using decision trees: Case of an Algerian University, International Conference on Mathematics and Information Technology (ICMIT), 2017
21. L. C. Chinchilla, K. A. R. Ferreira, Analysis of the behavior of customers in the social networks using data mining techniques, IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2016.
22. Z. Hou, Data Mining Method and Empirical Research for Extension Architecture Design, International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), 2018.
23. C.Shearer. The CRISP-DM Model: The New Blueprint for Data Mining, Journal of Data Warehousing, Volume 5, Number 4, pag. 13-22, 2000.