A Bibliometric Analysis of Distributed Incremental Clustering on Images

Dr. Preeti Mulay, Ayushi Agarwal, Krithika Iyer, Saloni Sarbhai

Symbiosis Institute of Technology, Symbiosis International University, Pune, India.

Abstract

Unstructured information is continuously irregular and streaming information from such a sequence is tedious because it lacks labels and accumulates with time. This is possible using Incremental Clustering algorithms that use previously learned information to accommodate new data and avoid retraining. This paper therefore seeks to understand the status of "Distributed Incremental Clustering" on images with text and numerical values, its limitations, scope, and other details to devise a better algorithm in future. To further enhance the analysis, we have also included methodology, which can be used to perform clustering on images or documents based on its content.

Keywords: Bibliometric Analysis; DICA; distributed; incremental; incremental clustering; text clustering; data clustering; Microsoft Azure

Introduction

Scopus and WOS (Web of Science) provide a vast number of published papers in this domain. To intensify the review process of research papers, we have chosen to perform a bibliometric analysis of all the work accomplished in “Distributed Incremental Clustering Algorithm on image datasets” using the database of Scopus and WOS. We in future would like to improve one of the most efficient methods of Incremental Clustering, i.e. “Closeness Factor-Based Algorithm”. CFBA is a data clustering technique that can form the raw data clusters instinctively and in parameter-free mode.

In future, we will be working on two problem statements :

1. Automatic Segregation of FDP Certificates based on Institute and Programs: In this case study, we will be clustering a large unlabelled set of images of the FDP certificates using CNN and Incremental Clustering algorithm. We chose this topic to analyze each staff with respect to the Institute they work for and the course they specialize in.
2. Electricity Smart Meter Data Extraction using Incremental Clustering: We are developing a system where a user can take a picture of the electricity meter and give this image as an input to the model, which in turn will calculate the bill based on the consumption and provide this information to the user.

This paper spans over four sections. The “Introduction” section gives introductory details about the research to be carried out. “Related work and bibliometric analysis” provides a literature review of work done on similar grounds from Scopus and WOS databases’ point of view. “Proposed methodology” section covers the research methodology in-depth related to distributed

incremental clustering algorithms on image datasets. “Conclusions” section consolidates concluding remarks followed by acknowledgement and references at last.

Related work and bibliometric analysis

This research aims to propose a novel approach to cluster images, documents or pdfs based on text content incrementally. In the field of unstructured data and incremental clustering, several researchers have contributed. This section presents a brief sketch of some work done in incremental clustering, including text and numerical data clustering. **Elie Aljalbout et al.** proposed a universal taxonomy of deep neural networking clustering methods.[1] The suggested taxonomy demonstrates a new clustering method by selectively rearranging previous aspects in order to transcend their respective task limits. This case study also shows that researchers and practitioners can systematically and analytically derive these clustering methods to suit their problem statement. This method has been shown to exceed the previous techniques on the MNIST dataset. The case study results showed that this taxonomy performs better, and the results are much balanced.

Nachiketa Sahoo et al. proposed a novel method for clustering unstructured documents.[2] In this paper, the authors have assessed a gradual hierarchical clustering algorithm often used with non-text datasets and introduced its variation, which can be more suitable for text documents. The variation of the Cobweb and Classit algorithm uses Katz’s distribution instead of using the Normal distribution as seen in the original Classit algorithm. The authors have used eleven existing text clustering datasets for the experiment and demonstrated that Katz’s distribution is more suitable and efficient for the word occurrence data. The novel incremental hierarchical clustering algorithm, which uses Katz’s distribution, has shown significant clusters and better results than using Normal distribution, especially among the larger datasets.

Sascha Hennig and Michael Wurst present a system that uses incremental clustering to manage and organize Newsgroup articles. This system is designed such that the news articles would be incrementally clustered as the streams of articles arrive. The proposed approach allows one to modify the clusters dynamically and rapidly with respect to the changing text streams. The model also enables users to change the cluster structure explicitly when required. Frequency term set clustering has shown terrific results for text document clustering, especially for a wide set of texts, and creates coherent clusters. Since news articles can often be to several topics, this algorithm also allows overlapping clusters, making the model much more scalable. In this news organizer system, each document is identified as a thread, and as a new article thread arrives, they are inserted into the appropriate hierarchical position automatically. It has been observed that there has been a trade-off between the permitted structural change. Nevertheless, high quality is achieved even with such strict restrictions, compared to the most non-incremental version of the algorithm.

M.Shashi and A.M.Sowjanya developed a more efficient approach for clustering incremental databases using cluster features. The proposed method has two modules one is initial clustering and other is incremental clustering. First, the initial clusters have been obtained by using the conventional k- means algorithm, and then cluster features have been used for clustering incremental databases based on a devised distance measure. They have used the mean value to cluster the closest pair after processing the set of data points. The experimentation has been performed using datasets from the UCI machine learning repository, including Iris dataset and Wine dataset.

Sajan Jaiswal proposed a method to extract the readings from the smart meter images. [5] In this paper, the authors have used Image Processing techniques and CNN to extract the meter image digits. The proposed algorithm is written in MATLAB. The following steps were carried out for the desired output (1) RGB to Grayscale conversion: the coloured image is converted into a white and black image. (2) Image noise removal: because of the conversion of the image from coloured to black and white, some noise can appear; therefore, a noise removal algorithm is applied. (3) Cropping the reading region: In this, the image is scanned for areas with a high white pixel density. When the area is found, a black pixel boundary is drawn around it, and then the image is cropped. (4) Image segmentation: This stage is where the actual digits are segmented. This is done using the VEDA algorithm. (5) A dataset for digit 0-9 is taken, and CNN is trained on that. (6) The cropped image is then run on the CNN, which is already trained on the digits 0-9, and each digit from the cropped image is extracted and saved in a text file. The digits extracted from the smart meter image can further be used to calculate the electricity bill.

It is often noticed that in machine learning and computer vision, image clustering is a critical yet daunting activity. Current techniques often overlook the significance of feature learning. **Jianlong Chang et al.** suggest Deep Adaptive Clustering (DAC) to resolve this issue, which transforms the clustering issue to a binary pair classification system which can decide whether pairs of image dataset belong to the same clusters.[6]. In this algorithm, the similarities between label characteristics of images produced by a deep convolutional network are calculated using the cosine distance (ConvNet). The learned mark characteristics appear to be one-hot vectors used for clustering images by inserting a constraint into DAC. The biggest problem is that in image clustering, the ground-truth comparisons are unknown. This problem is handled by presenting an alternating adaptive learning algorithm, which selects labelled samples alternatively and trains the ConvNet. Finally, images are grouped automatically based on the characteristics of features. The results show that the proposed technique achieves state-of-the-art performance on popular datasets. For example - it yields 97.75% accuracy on MNIST, 52.18% on CIFAR-10 and 46.99% on STL-10.They theoretically verified that their model could be used to represent images with one-hot vectors that can be utilized for clustering images. The paper concludes that the method proposed is not limited to simple image datasets but also performs efficiently with large-scale complex image datasets.

In 2019, Mulay and Joshi et al.[12] published a journal paper showing the use of Azure Private Cloud as a PaaS, which serves as an interface for healthcare professionals to evaluate health information via a CFBA-embedded web app. This CFBA implementation is based on Azure and has ninety percent precision. The implementation of CFBA on the public cloud of Microsoft Azure is intended to manage various datasets of chronic diseases, attain scalability with performance, and is independent of domain.

In 2018 **Tureczek, A., Nielsen, P. S., & Madsen, H.** published a paper which clustered the household consumption of electricity utilizing smart meter data from Esbjerg. This paper proposes four things:

- Presenting a cluster analysis of Esbjerg electricity consumption data.
- The data which is not predicted or clustered by K means is autocorrelated.
- Input data features are extracted and transformed, which will enable K means to account for autocorrelation in the clustering.
- The concept of cross-validation is extended to the unsupervised learning concept of cluster validation.

Before clustering, the data was pre-processed to remove the missing data. Then the data was segmented into smaller subsets to extract the features of the data. In this, three different methods were used to extract the features - normalization, wavelet transformation and autocorrelation feature extraction. Finally, the data were clustered using the K means algorithm.

Rongheng Lin, Zigui Jiang and Fangchun Yang proposed a machine learning model that combined unsupervised learning's clustering technique and supervised learning's classification. The clustering algorithm is used to perform consumer categorization, and then an algorithm is then used to further categorize consumers and their characteristics.

Designing the keyword search strings

Table 1. Document type and Top subject area publication count

Document Type	No of Documents	Subject Area	No of Documents
Article	25	Computer Science	36
Conference Paper	11	Business, Management and Accounting	5
Review	4	Engineering	5
Book Chapter	3	Decision Sciences	4

Table 2. Key terms searched

Search String No	Search String	Search result in scopus
1	"distributed incremental clustering" OR "incremental clustering"	3072
2	("distributed incremental clustering" OR "incremental clustering") AND ("text clustering" AND "data clustering")	43
3	("distributed incremental clustering" OR "incremental clustering") AND ("text clustering" AND "data clustering") AND (LIMIT-TO (FREETOREAD , "all"))	13
4	("distributed incremental clustering" OR "incremental clustering") AND ("text clustering" AND "data clustering") AND (LIMIT-TO (FREETOREAD , "all")) AND (LIMIT-TO (PUBYEAR , 2020) OR LIMIT-TO (PUBYEAR , 2019) OR LIMIT-TO (PUBYEAR , 2017) OR LIMIT-TO (PUBYEAR , 2016) OR LIMIT-TO (PUBYEAR , 2009) OR LIMIT-TO (PUBYEAR , 2006) OR LIMIT-TO (PUBYEAR , 2004))	13
5	("distributed incremental clustering" OR "incremental clustering") AND ("text clustering" AND "data clustering") AND (LIMIT-TO (FREETOREAD , "all")) AND (LIMIT-TO (PUBYEAR , 2020) OR LIMIT-TO (PUBYEAR , 2019) OR LIMIT-TO (PUBYEAR , 2017) OR LIMIT-TO (PUBYEAR , 2016) OR LIMIT-TO (PUBYEAR , 2009) OR LIMIT-TO (PUBYEAR , 2006) OR LIMIT-TO (PUBYEAR , 2004)) AND (LIMIT-TO (DOCTYPE , "ar"))	8

Table 3: Publication count and the top 5 sources

Year	No of Documents	Source Title	No of Documents
2020	3	IEEE Transactions On Knowledge And Data Engineering	2
2019	3	Knowledge And Information Systems	2
2018	3	ACM Computing Surveys	1
2017	3	ACM Transactions On Database Systems	1
2016	6	Advances In Intelligent Systems And Computing	1

Table 4. Affiliation and country publication count.

Affiliations	No of Documents	Country	No of Documents
Yale University	2	China	12
Jilin University	2	India	7
Xi'an Jiaotong University	2	United States	7

Table 5. Top 5 authors and funding agencies (2016–2020)

Authors	No of Documents	Funding agencies	No of Documents
Kannan, R.	2	National Natural Science Foundation of China	4
Patil, H.	2	Consejo Nacional de Ciencia y Tecnología	1
Thakur, R.S.	2	Fundamental Research Funds for the Central Universities	1
Vempala, S.	2	Hong Kong Polytechnic University	1
Wang, G.	2	Institute for Critical Technology and Applied Science	1

Publication trends

Figure 1 and 2 shows the Document types for DICA, text clustering and data clustering from Scopus and Wos databases. Fig 3 and 4 different years in which the papers were published.

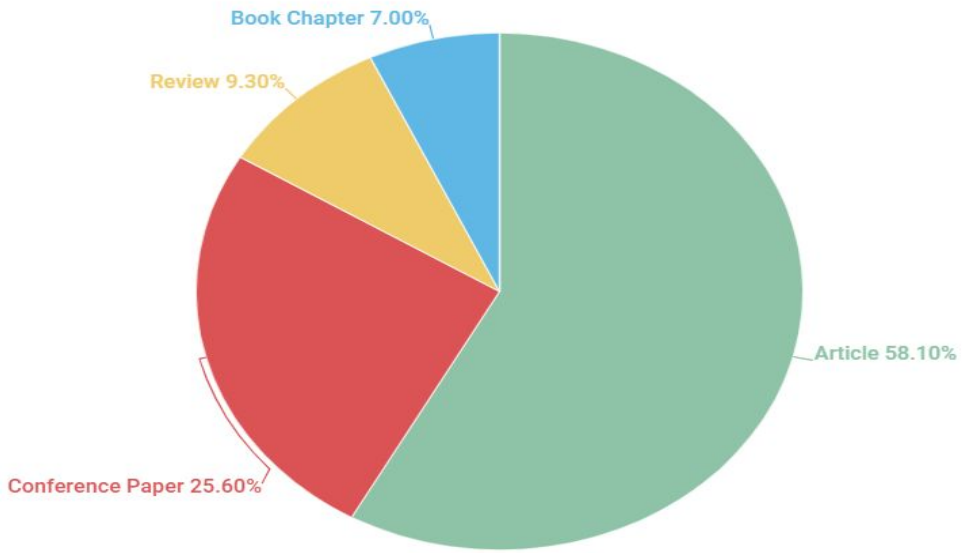

Figure 1. Publication types for DICA, text clustering and data clustering (Scopus)

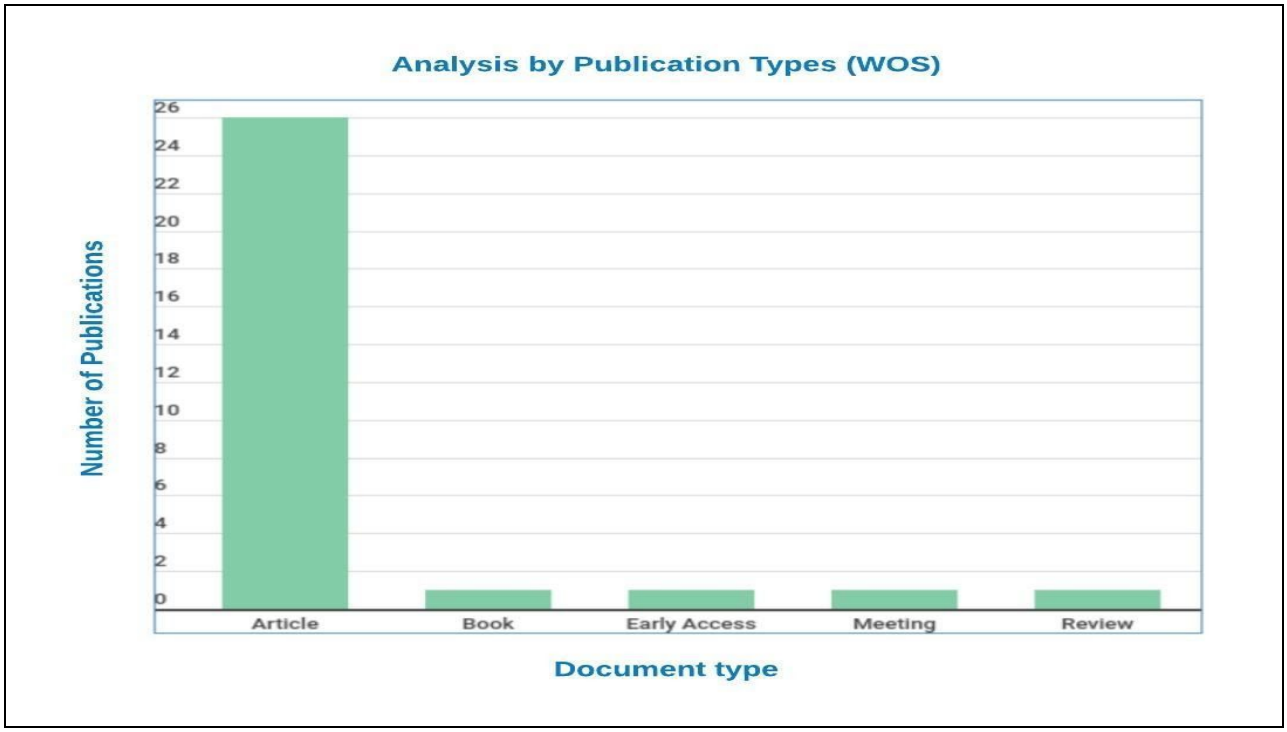

Figure 2: Document types in DICA, text clustering and data clustering (WOS)

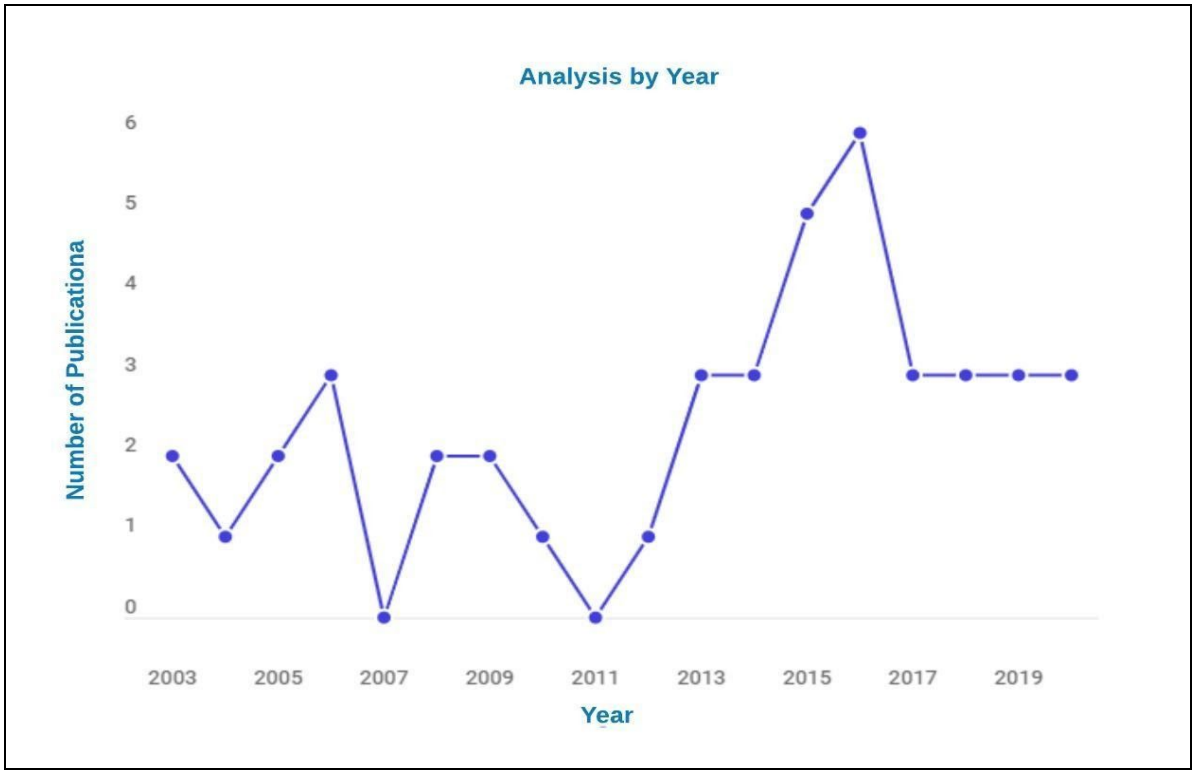

Figure 3. Yearly publishing trend in DICA, text clustering and data clustering(Scopus)

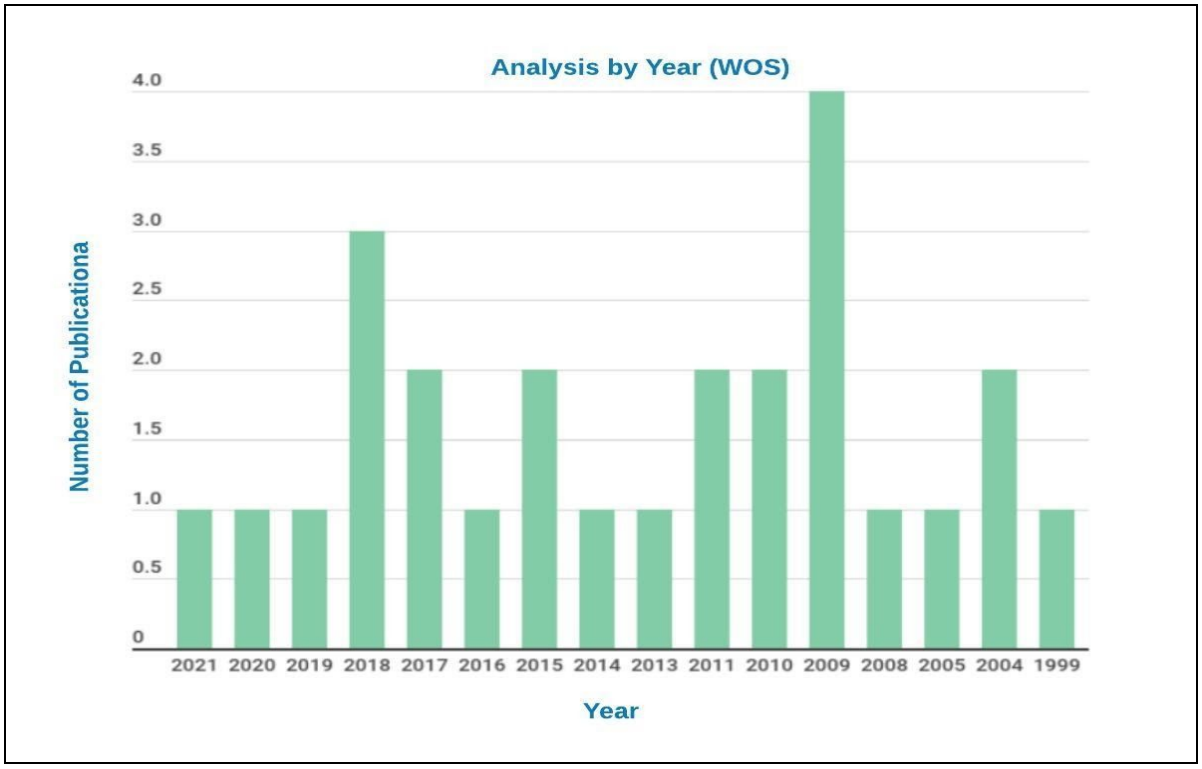

Figure 4: Yearly publishing trend in DICA, text clustering and data clustering (WOS)

Keyword Statistics

Apt Keywords play an essential role in the lookup of targeted publications from the publication database. The top 10 important keywords list for Incremental Clustering, Distributed incremental clustering, text clustering and data clustering are presented in Table 1. It is clear from Table 1 that there is further scope for research in the areas viz., incremental clustering, incremental learning, text clustering, image processing etc.

Table 6.Important keywords for Distributed Incremental Clustering OR Incremental clustering on Images

Keywords	No of Publications	Keywords	No of Publications
Clustering Algorithm	225	Distributed	48
Incremental Clustering	38	Deep Learning	170
Incremental Learning	251	Image Segmentation	106
Text Clustering	20	Data Clustering	52
Image Processing	36	Feature Extraction	146

Source Statistics

Figure 5 and 6 covers the publication source types for DICA, text clustering and data clustering from Scopus and WOS from the year 2000 - 2021.

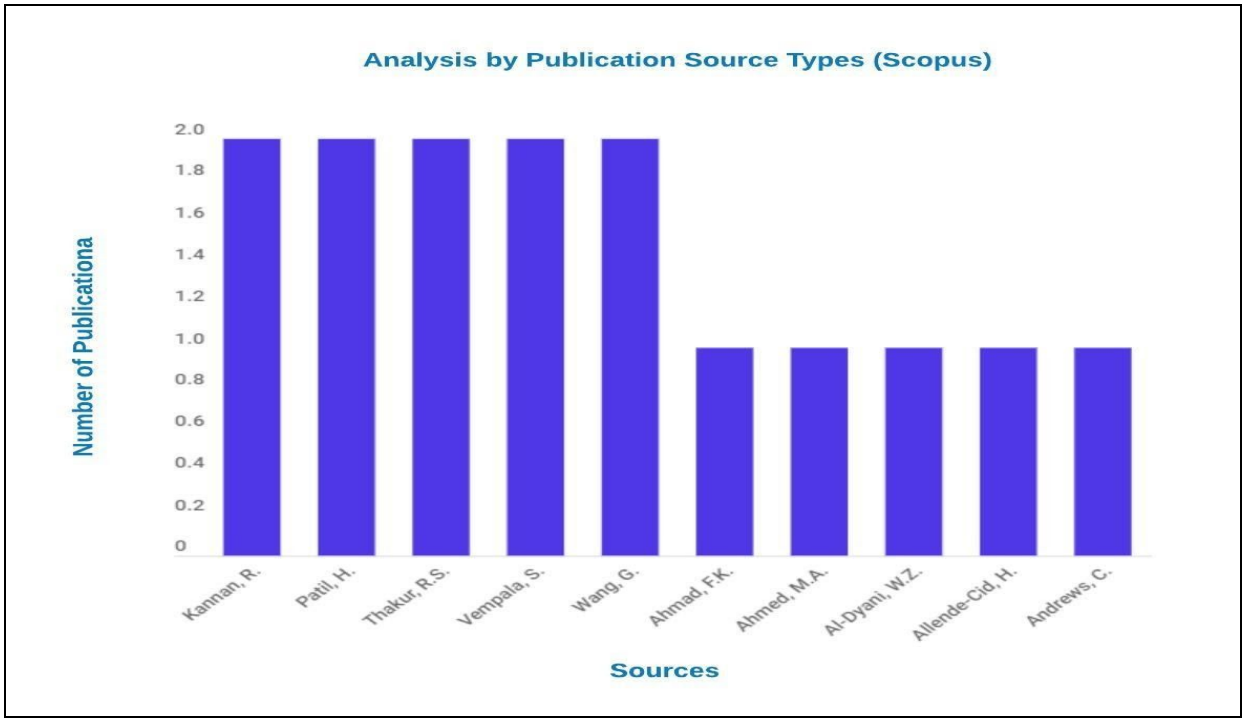

Figure 5: Top 10 affiliation statistics on DICA, text clustering and data clustering (Scopus)

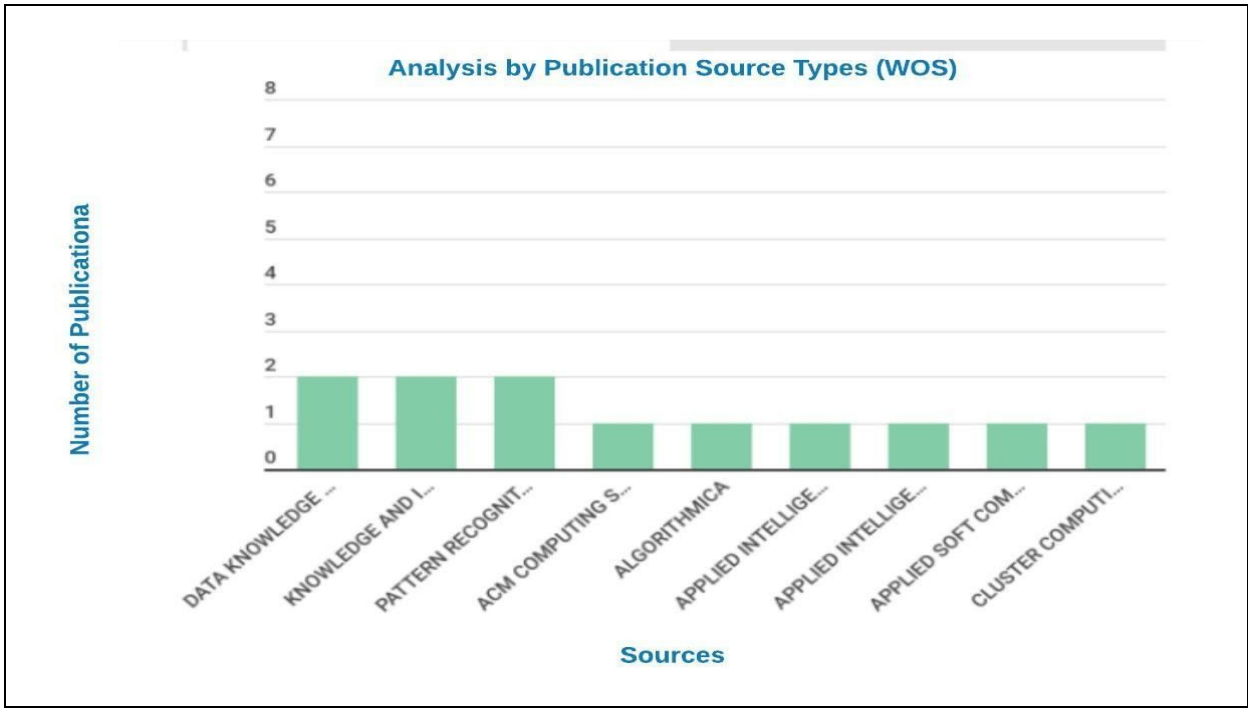

Figure 6: Top 10 sources publishing papers on DICA, text clustering and data clustering (WOS)

Authors and Affiliations Statistics

Figures 7 and 8 depict the top ten authors contributing and their affiliations to DICA, text clustering and data clustering from Scopus and Web of Science respectively. This statistics represents the influencing authors of DICA, Text and data clustering.

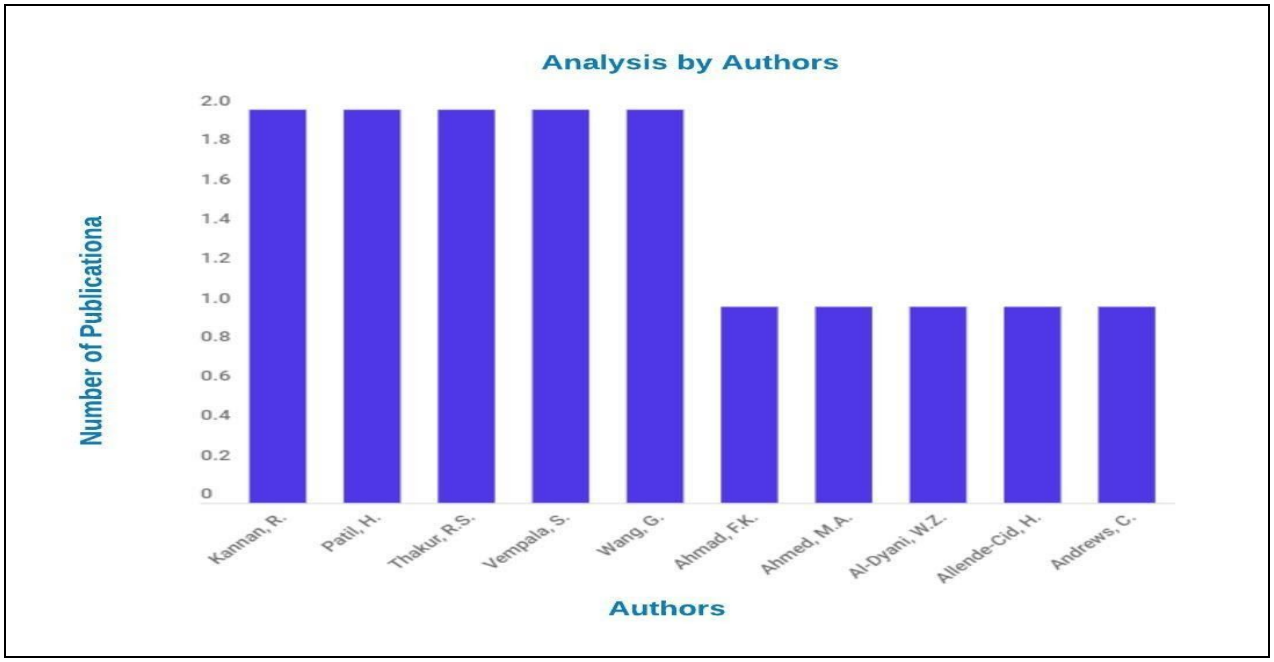

Figure 7. Leading authors contributing for DICA, text clustering and data clustering (Scopus)

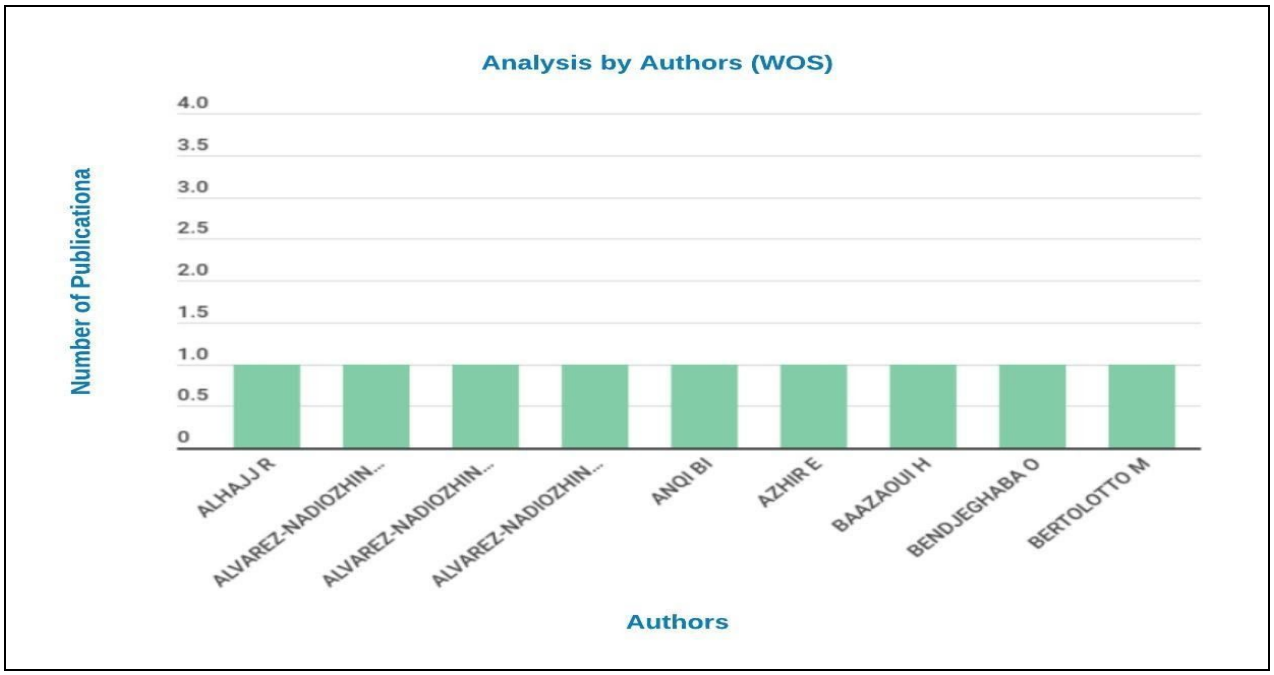

Figure 8. Leading authors contributing in DICA, text clustering and data clustering (WOS)

Top publishing countries

Figure 9 gives the top fifteen countries having publications in the area of DICA and text & data clustering from Scopus and WOS. It can be inferred that China leads followed by the United States and India according to Scopus , while China is leading followed by India in reference to the Web of Science (Figure 10).

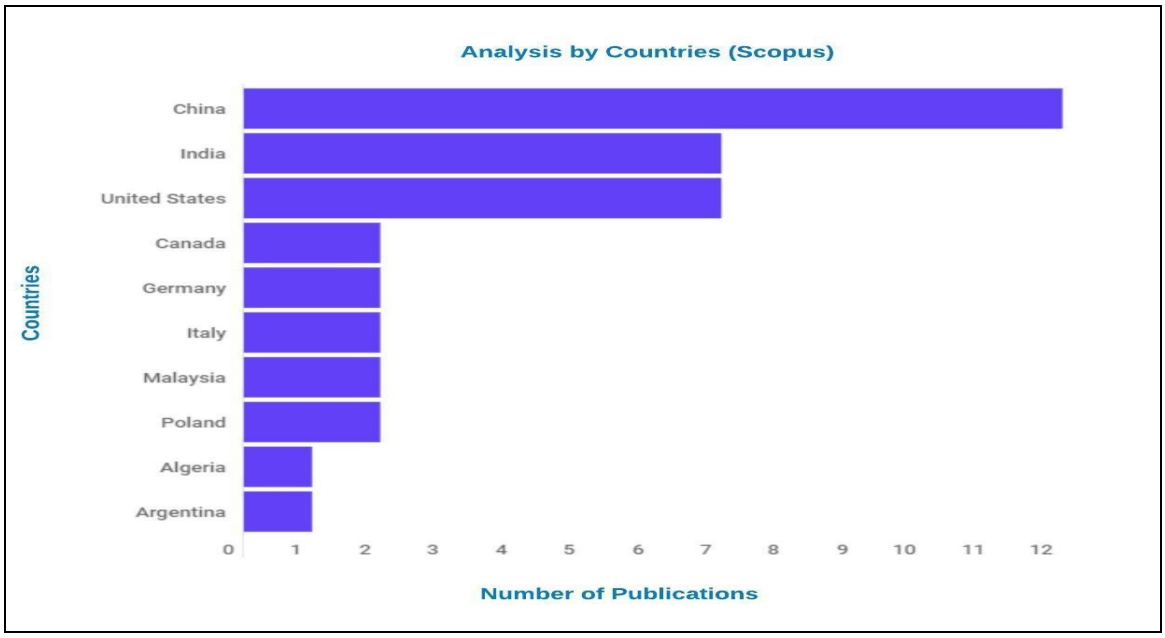

Figure 9. Top 10 countries publishing papers on DICA, text clustering and data clustering

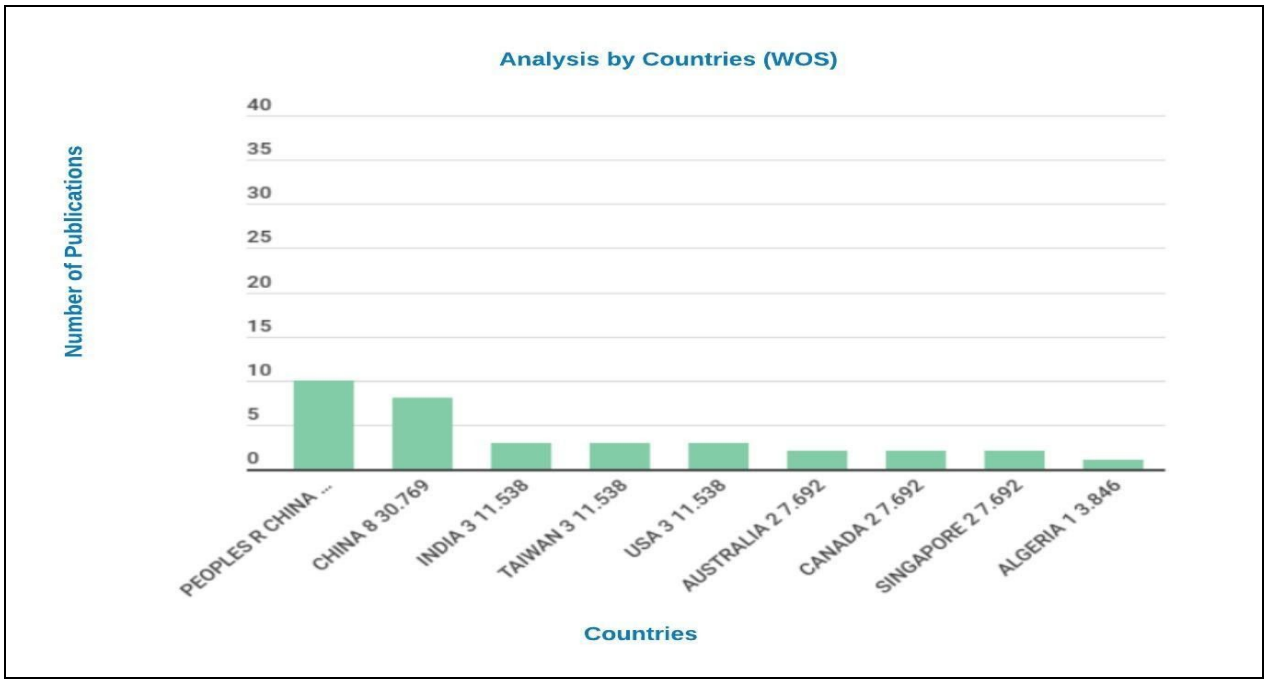

Figure 10: Top 10 countries publishing papers on DICA, text clustering and data clustering

Prominent subject areas

Figure 11 and 12 shows the maximum number of publications for DICA, text clustering and data clustering in different subject areas.

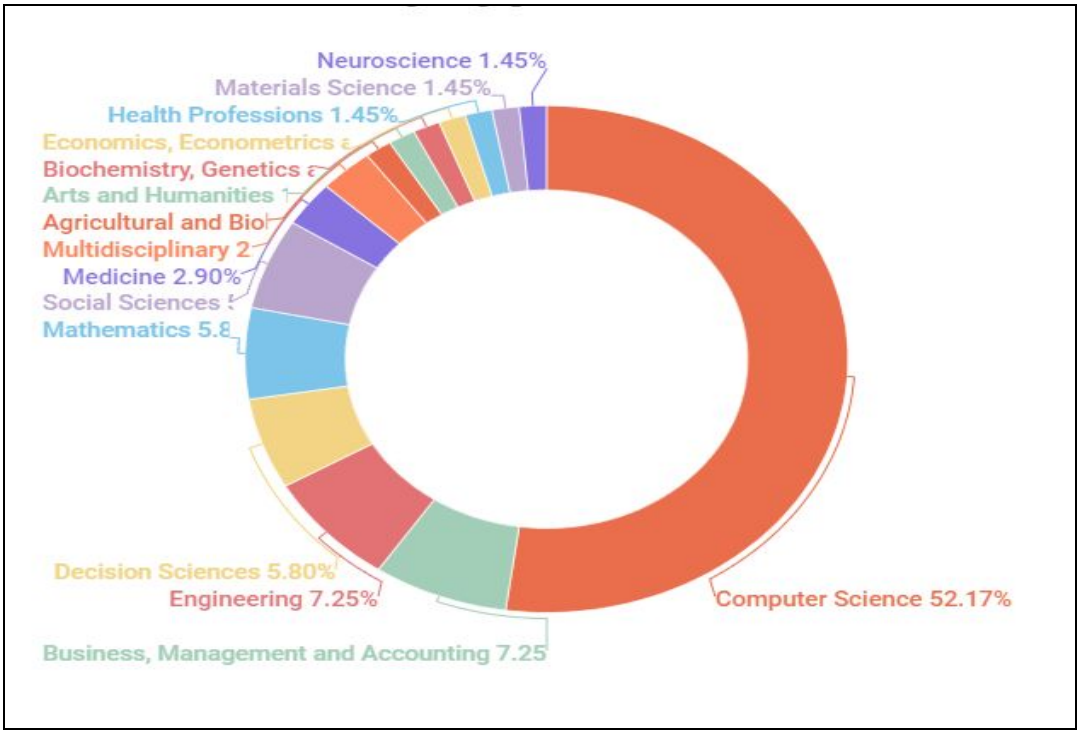

Figure 11: Prominent subject areas for DICA,text clustering and data clustering (Scopus)

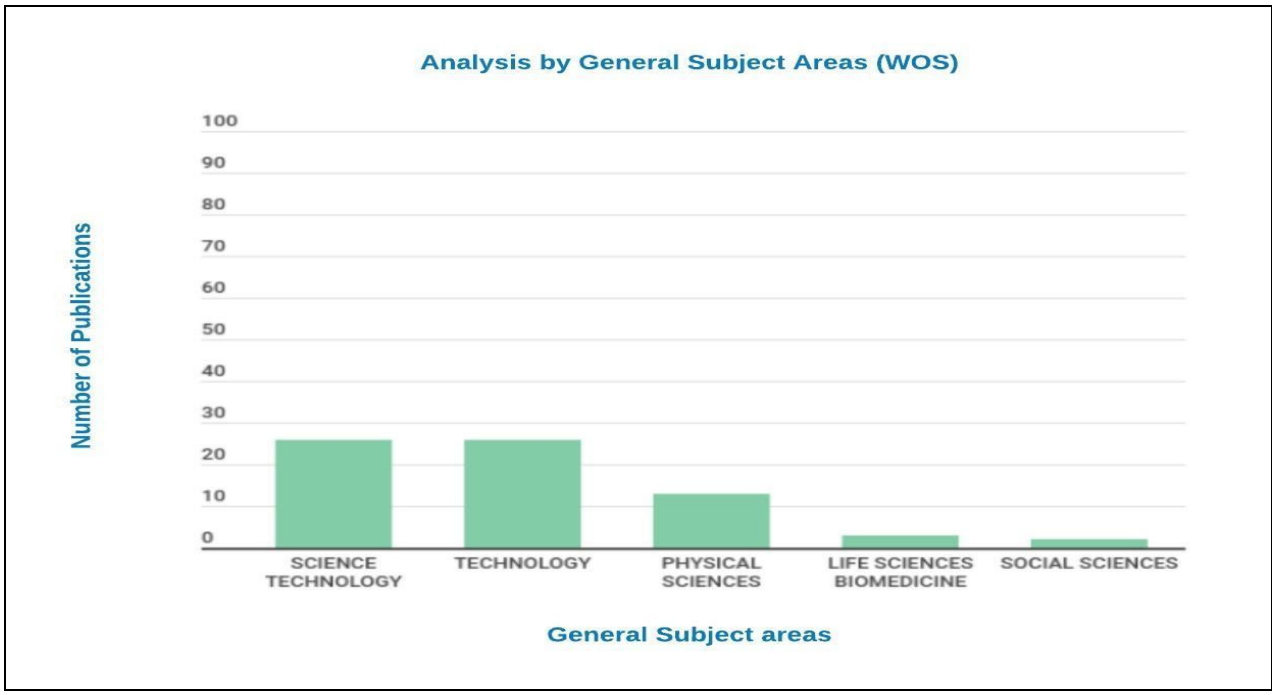

Figure 12 : General subject areas for DICA, text clustering and data clustering (WOS)

Funding Agencies

Figure 13 . shows the top 10 funding agencies for DICA ,text clustering and data clustering from Scopus. From the analysis we can see that the National Natural Science Foundation of China is the biggest funding sponsor for DICA ,text clustering and data clustering.

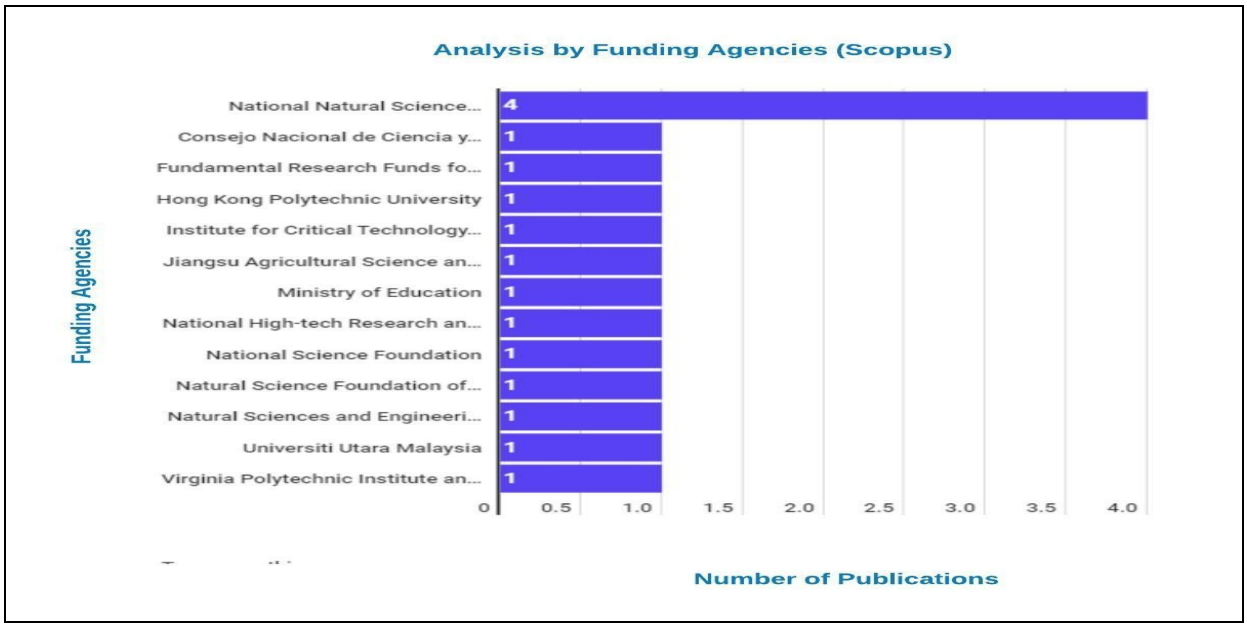

Figure 13: Top 15 Funding agencies in DICA,text clustering and data clustering (Scopus)

Most Cited Articles

Till date there are 236 articles for DICA and 57 articles for DICA and Image Clustering in Scopus. Out of these 57 articles the most cited article in DICA and Image Processing is Resource-aware distributed scheduling strategies for large-scale computational cluster/grid systems by Viswanathan, S., Veeravalli, B., Robertazzi, T.G. (2007) cited by total 58 times.

Proposed Methodology

Feature Extraction

Feature extraction is the process in which the required features within an image are detected and represented for further processing. The feature is defined as a function of one or more measures, each specifying an object's measurable property. Text in an image can vary based on the following properties:

1. Geometry: This includes the text's size, the text's alignment and geometric distortion
2. Colour: This is one of the crucial properties, making it possible to locate and detect text.
3. Edge: Most captions in an image are designed in such a way that they can be easily read and result in strong edges at text and background boundaries.

A Text Information Extraction (TIE) system accepts an input either in a sequence of images or pdfs. The input further undergoes the process of text detection, text location, tracking extraction

and further enhancement.
Initially, the text or digits in a given data frame is determined; this process is generally referred to as Text detection. Text localization is further used to determine the text's location in images and generate bounding boxes. Further, text tracking is also done to reduce the processing time for localization. This additionally helps to maintain position across frames. Even though bounding boxes can indicate the location of text in an image, the text has to be segmented from the background for accurate recognition. Then the extracted text is converted into a binary image fed into an OCR engine or Open-cv2. The text sections are then partitioned from the background. Thereafter, the extracted text must be enhanced as the text area typically has poor resolution and is sensitive to noise. Finally,the text images extracted are translated into plain text using OCR or Open-cv2 technology.

Feature Selection

Feature selection is the process in which we select a subset of some relevant features to maximize the learner's ability. For example, if we have a set of features F = {f1…..fn….fm} and after feature selection the subset is F1, then F1 would increase some scoring function.
The Feature Selection in a text dataset could be made using TF-IDF (term frequency-inverse document frequency). This tells us how relevant a word is to a document in a bunch of papers. Two metrics are used to make this selection - (1) how many times the word appeared in the document. (2) inverse document frequency of the word in a bunch of documents.

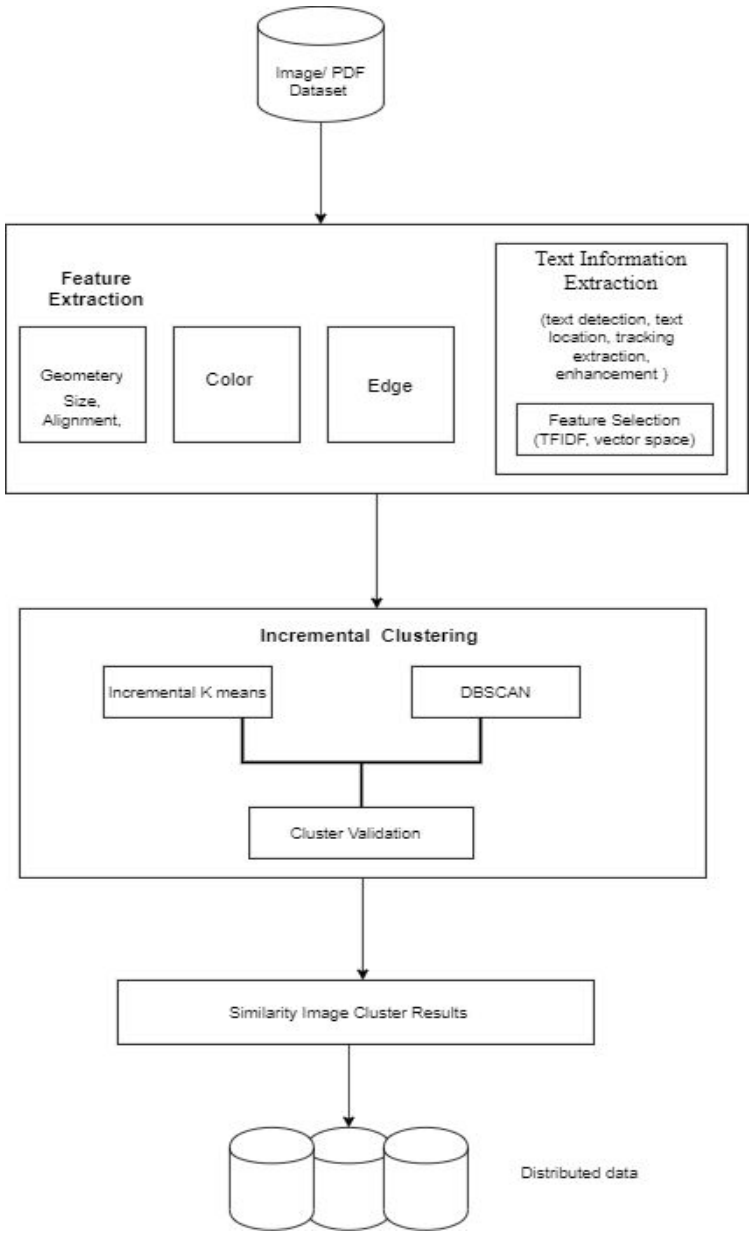

Figure 14: Figure 14 shows the flow of code i.e. we need to extract features from the given image dataset and export it in a .csv format, apply incremental clustering algorithm on the extracted data and finally deploy it on Azure.

Clustering Technique:

Algorithm 1:

Input: Data set (DS) //{t 1, t 2, ..., t 2000 } 2000 images.
Output: Clusters(K)

1. The smart electricity meter dataset that will be analysed is provided by **UFPR-AMR.** This dataset consists of 2000 images taken from inside a warehouse of the Energy company of Paraná (Copel).
2. The input DS images are **cropped** to get the reading part of the meter images.
3. The digits are extracted using **Open-Cv2 and Pytesseract** and exported in a .csv file be it our DS1.
4. Incremental **K Means** or **DBSCAN** is applied on DS1 and K are formed.

Algorithm 2:

Input: Data Set (DS) //{t1, t2....t 1000} 1000 Pdf/Images
Output: Clusters

1. The text within the pdfs are extracted using Open-Cv2 and Pytesseract and exported in a .csv file.
2. Remove commas, line breaks, slashes, and hyphens using regex expressions.
3. Using TfIdfVectorization pull out the unique words that can be used for Clustering.
4. Use DBSCAN or Incremental Kmeans to form clusters for unique values found.
5. Plot the values and compare the results.

Conclusion

We presented a bibliometric analysis of the work accomplished in the area of Distributed Incremental Clustering on Images. Specifically, for journal papers published in the field of incremental clustering on images using distributed systems, we provided an overview for the period 2000-2020. It is advent from the above research that the areas of Computer Science, Engineering, Mathematics and Medicine are leading in this field of study.It is also evident that China, the US, and India publish in this area as the world's leading nations and Shen C, Son L.H., Zheng Y., etc., are prevailing writers.Famous universities are the Chinese Academy of Sciences, Ministry of Education of China and University of Chinese Academy of Sciences.

This paper briefly describes the Distributed Incremental Clustering Algorithm on two of the use cases, one is to segregate the FDP Certificates based on the institute, and another one is to implement Electricity Smart Meter Data Extraction. We have also proposed a Common System Architecture, which explains the various steps involved in implementing the above use cases. The steps include feature extraction, Text Information extraction, Incremental clustering algorithm etc. For simplicity, we have also structured a stepwise algorithm for the implementation of the above mentioned. The purpose of the work done in this survey paper is to give valuable insight into the research on Distributed Incremental Clustering on Images in the future.

References

[1] Aljalbout, Elie & Golkov, Vladimir & Siddiqui, Yawar & Cremers, Daniel. (2018) Clustering with Deep Learning: Taxonomy and New Methods.

[2] Sahoo, Nachiketa & Callan, Jamie & Krishnan, Ramayya & Duncan, George & Padman, Rema. (2006). Incremental hierarchical clustering of text documents. International Conference on Information and Knowledge Management, Proceedings. 357-366. 10.1145/1183614.1183667.

[3] Md. Nazmul Hasan, Rafia Nishat Toma, Abdullah-Al Nahid, M M Manjurul Islam and Jong-Myon Kim. (2019) Electricity Theft Detection in Smart Grid Systems: A CNN-LSTM Based Approach.

[4] Karthick Kanagarathinam, Kavaskar Sekar. (2019)Text detection and recognition in raw image dataset of seven-segment digital energy meter display.

[5] Sajan Jaiswal, Ashish Prajapati, Niranjan Shirodkar, Pratik Tanawade, Prof. Chintal Gala Electricity Meter Reading Based on Image Processing.

[6]J. Chang, L. Wang, G. Meng, S. Xiang and C. Pan, "Deep Adaptive Image Clustering," *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, 2017, pp. 5880-5888, doi: 10.1109/ICCV.2017.626.

[7]Chaudhari, A., and P. Mulay. 2019. A bibliometric survey on incremental clustering algorithm for electricity smart meter data analysis. Iran Journal of Computer Science 2 (4):197–206. doi:10.1007/s42044-019-00043-0.

[8]Chaudhari, Archana & Joshi, Rahul & Mulay, Preeti & Kotecha, Ketan & Kulkarni, Parag. (2019). Bibliometric Survey on Incremental Clustering Algorithms.

[9]Elrefaei, Lamiaa. (2015). Automatic Electricity Meter Reading Based on Image Processing. 10.1109/AEECT.2015.7360571.

[10]Hennig S., Wurst M. (2006) Incremental Clustering of Newsgroup Articles. In: Ali M., Dapoigny R. (eds) Advances in Applied Artificial Intelligence. IEA/AIE 2006. Lecture Notes in Computer Science, vol 4031. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11779568_37

[11]Sowjanya, Mary & Shashi, M.. (2010). Cluster Feature-Based Incremental Clustering Approach (CFICA) For Numerical Data. IJCSNS Int J Comput Sci Netw Secur. 10.

[12] Joshi, Rahul & Mulay, Preeti. (2019). Closeness Factor Based Clustering Algorithm (CFBA) and Allied Implementations—Proposed IoMT Perspective. 10.1007/978-3-030-23983-1_8.

[13] Tureczek, A., Nielsen, P. S., & Madsen, H. (2018). Electricity consumption clustering using smart meter data. Energies, 11(4), [859].

[14] Zigui Jiang, Rongheng Lin and Fangchun Yang (2018). A Hybrid Machine Learning Model for Electricity Consumer Categorization Using Smart Meter Data.

[15] https://arxiv.org/pdf/1406.4751.pdf Performance Comparison of Incremental K-means and Incremental DBSCAN Algorithms Sanjay Chakraborty, N.K.Nagwani, Lopamudra Dey

[16] P. Angelov and X. Zhou, "Evolving Fuzzy Systems from Data Streams in Real-Time," 2006 International Symposium on Evolving Fuzzy Systems, Ambleside, 2006, pp. 29-35, doi: 10.1109/ISEFS.2006.251157.

[17] L. Sun and C. Guo, "Incremental Affinity Propagation Clustering Based on Message Passing," in IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 11, pp. 2731-2744, Nov. 2014, doi: 10.1109/TKDE.2014.2310215.

[18] Ricardo Marcondes Marcacini and Solange Oliveira Rezende. 2013. Incremental hierarchical text clustering with privileged information. In Proceedings of the 2013 ACM symposium on Document engineering (DocEng '13). Association for Computing Machinery, New York, NY, USA, 231–232. DOI:https://doi.org/10.1145/2494266.2494296

[19] Bi, A., Wang, S. Incremental enhanced α-expansion move for large data: a probability regularization perspective. *Int. J. Mach. Learn. & Cyber.* 8, 1615–1631 (2017).

[20] Padmalatha E., Sailekya S. (2019) Compact Clusters on Topic-Based Data Streams. In: Bapi R., Rao K., Prasad M. (eds) First International Conference on Artificial Intelligence and Cognitive Computing. Advances in Intelligent Systems and Computing, vol 815. Springer, Singapore.

[21]Nikhath, A. & Subrahmanyam, K.. (2019). Feature selection, optimization and clustering strategies of text documents. International Journal of Electrical and Computer Engineering (IJECE). 9. 1313. 10.11591/ijece.v9i2.pp1313-1320.

[22]Tian, Dongping. (2013). A review on image feature extraction and representation techniques. International Journal of Multimedia and Ubiquitous Engineering. 8. 385-395.