Big Data Clustering

Supervisor:
Dr J. Shanbehzadeh
Shanbehzadeh@gmail.com
By:
Azam Shakori
a.shakori.com@gmail.com

Big Data Clustering

Supervisor:
Dr J. Shanbehzadeh
Shanbehzadeh@gmail.com
By:
Azam Shakori
a.shakori.com@gmail.com

BIG
DATA

Outline

- Introdution
- Clustering
- Characteristic of Big Data
- Big Data Clustering Algorithms
- Open Issues
- Conclusion
- References

flickr from YAHOO!

~4.5 million photos uploaded/day

Google News

Articles from over 10,000 sources in real time

YouTube

48 hours of video uploaded/min; more than 1 trillion video views

Over 225 million users generating over 800 tweets per second

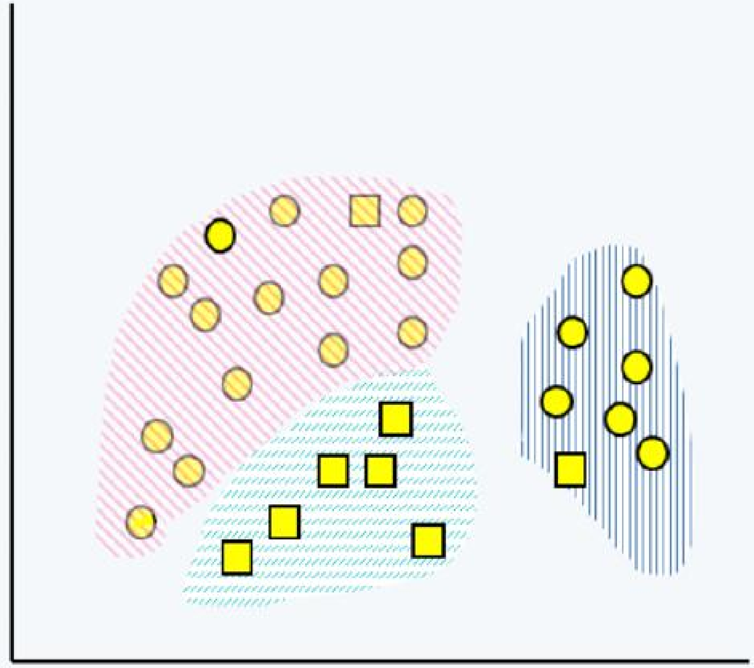
twitter

Over 50 billion pages indexed and more than 2 million queries/min

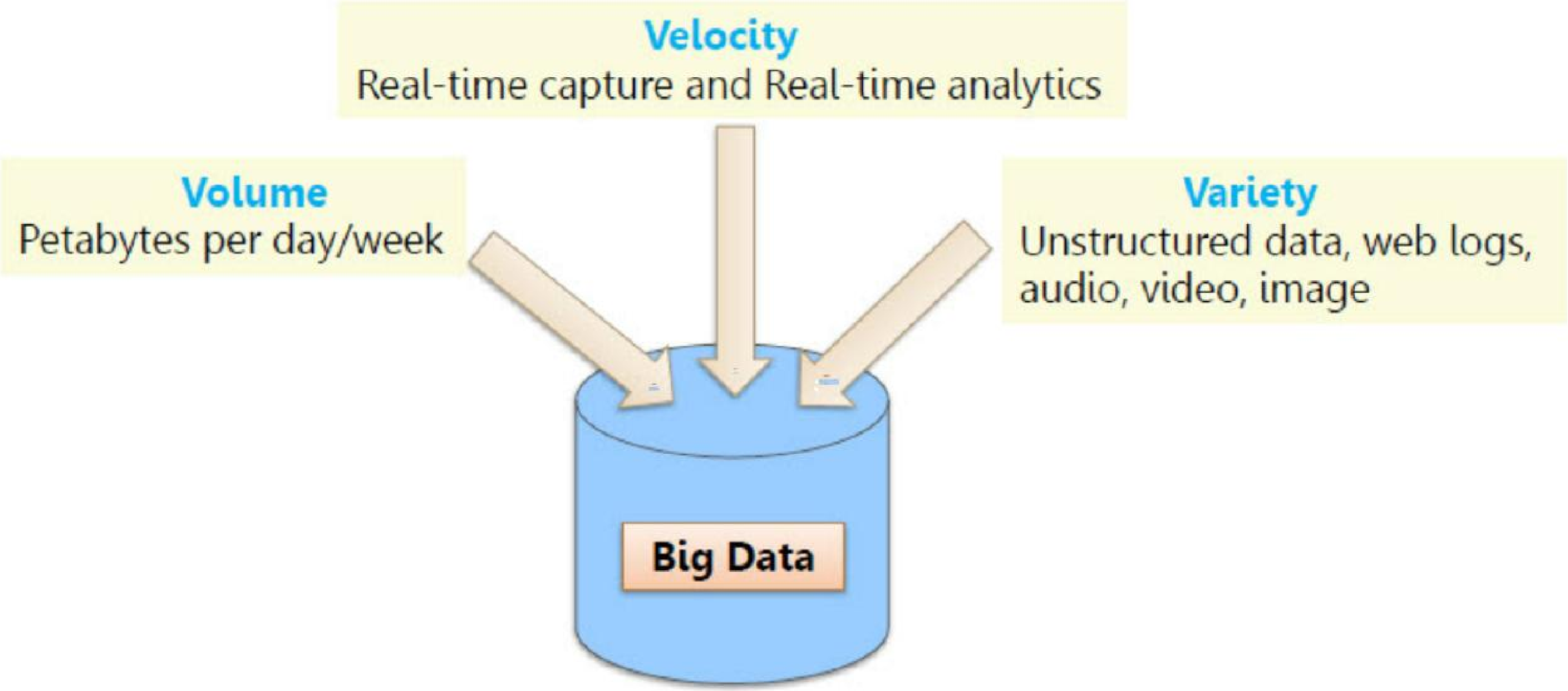
STRUCTURED

UNSTRUCTURED

What is Clustering



Is a method in which, data are divided into groups in a way that objects in each group share more similarity than with other objects in other groups



Velocity
Real-time capture and Real-time analytics

Volume
Petabytes per day/week

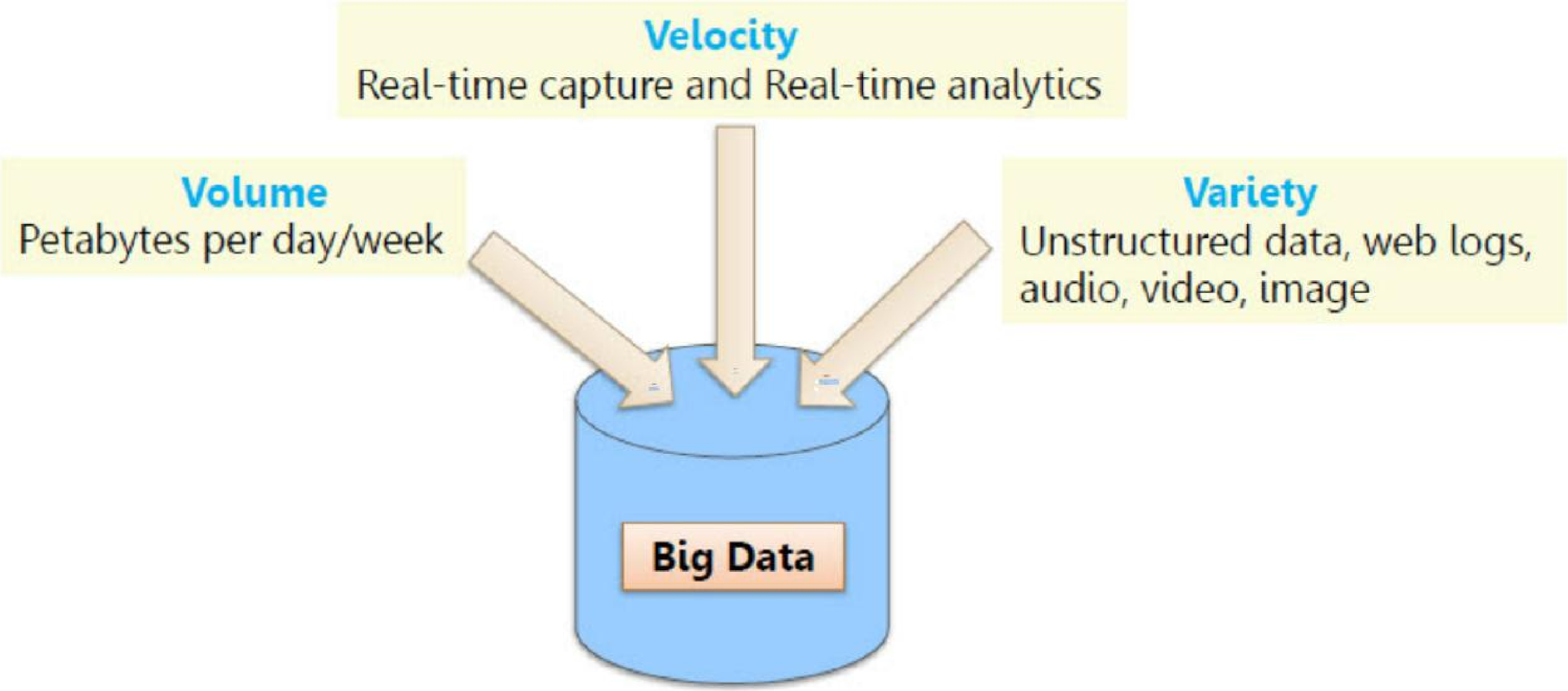
Variety
Unstructured data, web logs, audio, video, image

Big Data

Volume

Refers to large amount of data. Criteria:

- size of data
- Handling high dimensionality
- Handling outliers / noisy data



Velocity
Real-time capture and Real-time analytics

Volume
Petabytes per day/week

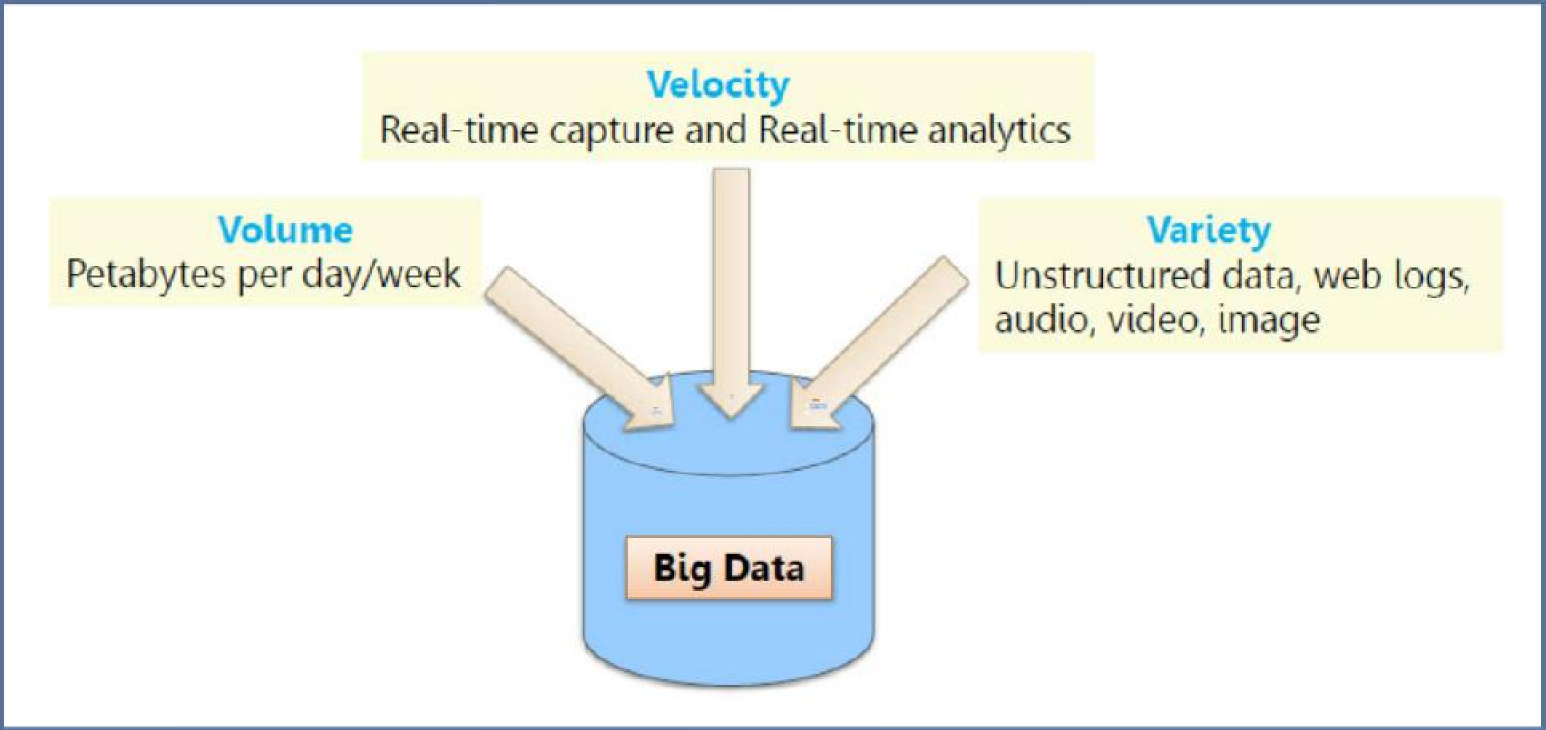
Variety
Unstructured data, web logs, audio, video, image

Big Data

Velocity

Refers speed of processing data. Criteria:

- Complexity of algorithm
- The run time performances



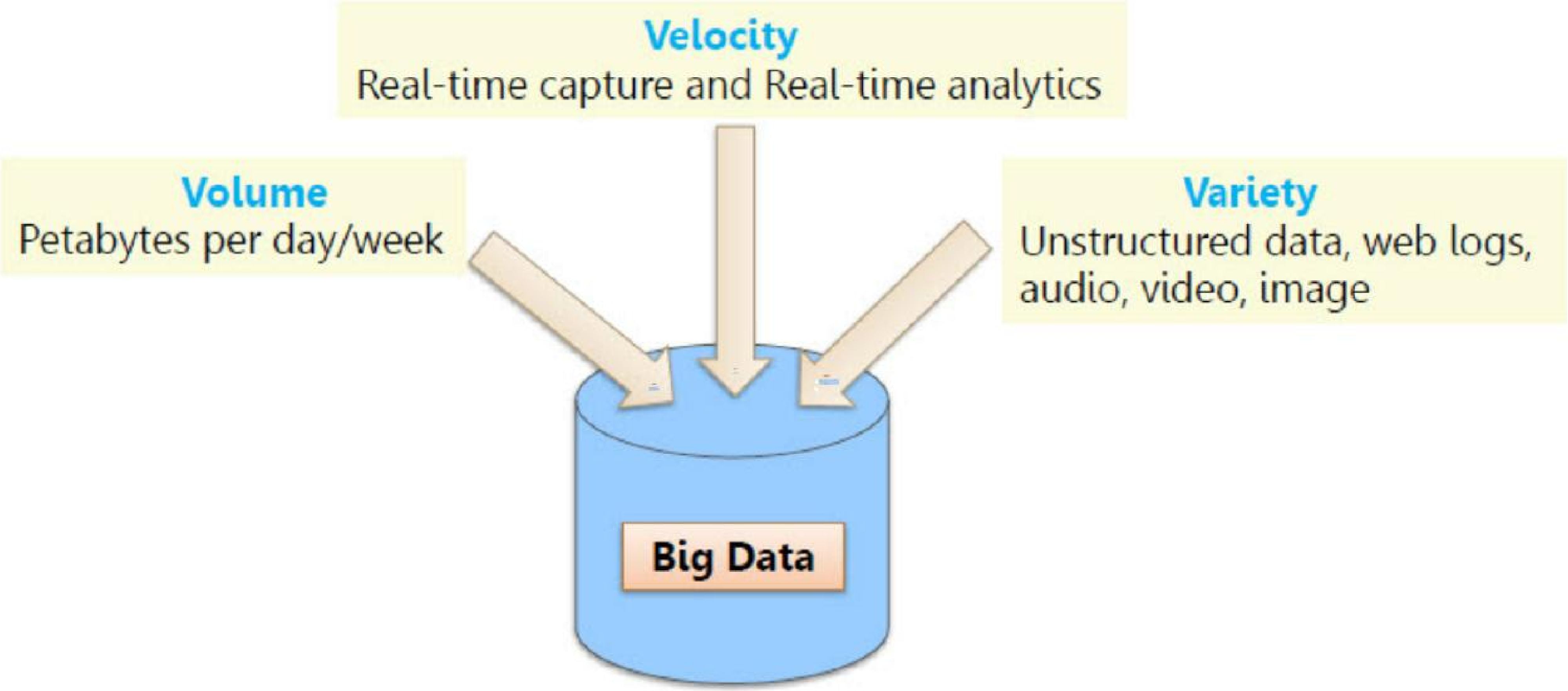
Velocity
Real-time capture and Real-time analytics

Volume
Petabytes per day/week

Variety
Unstructured data, web logs, audio, video, image

Variety

- Refers to the ability to handle different type of data. Criteria:
    - Type of data
    - Clusters shape

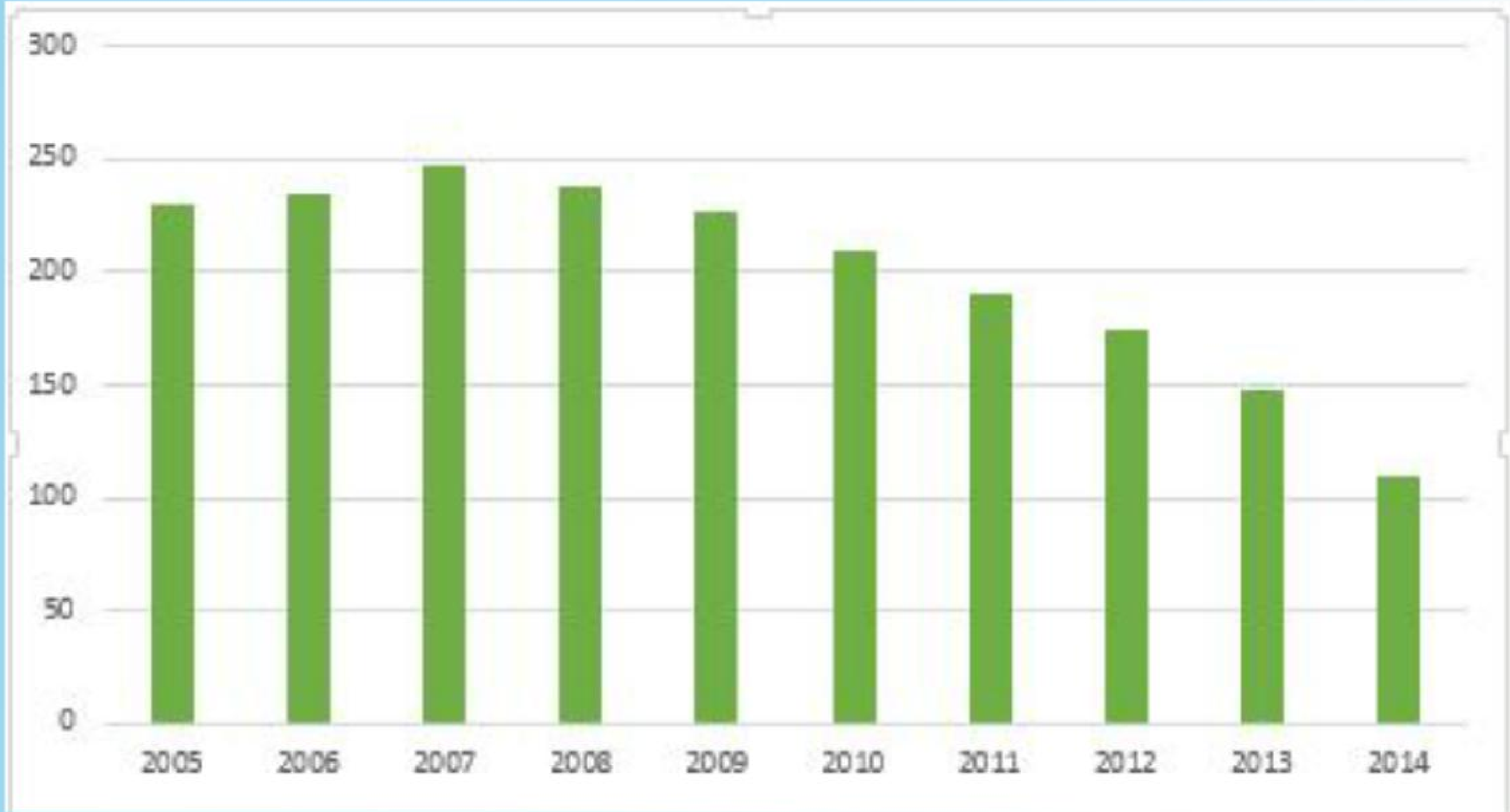


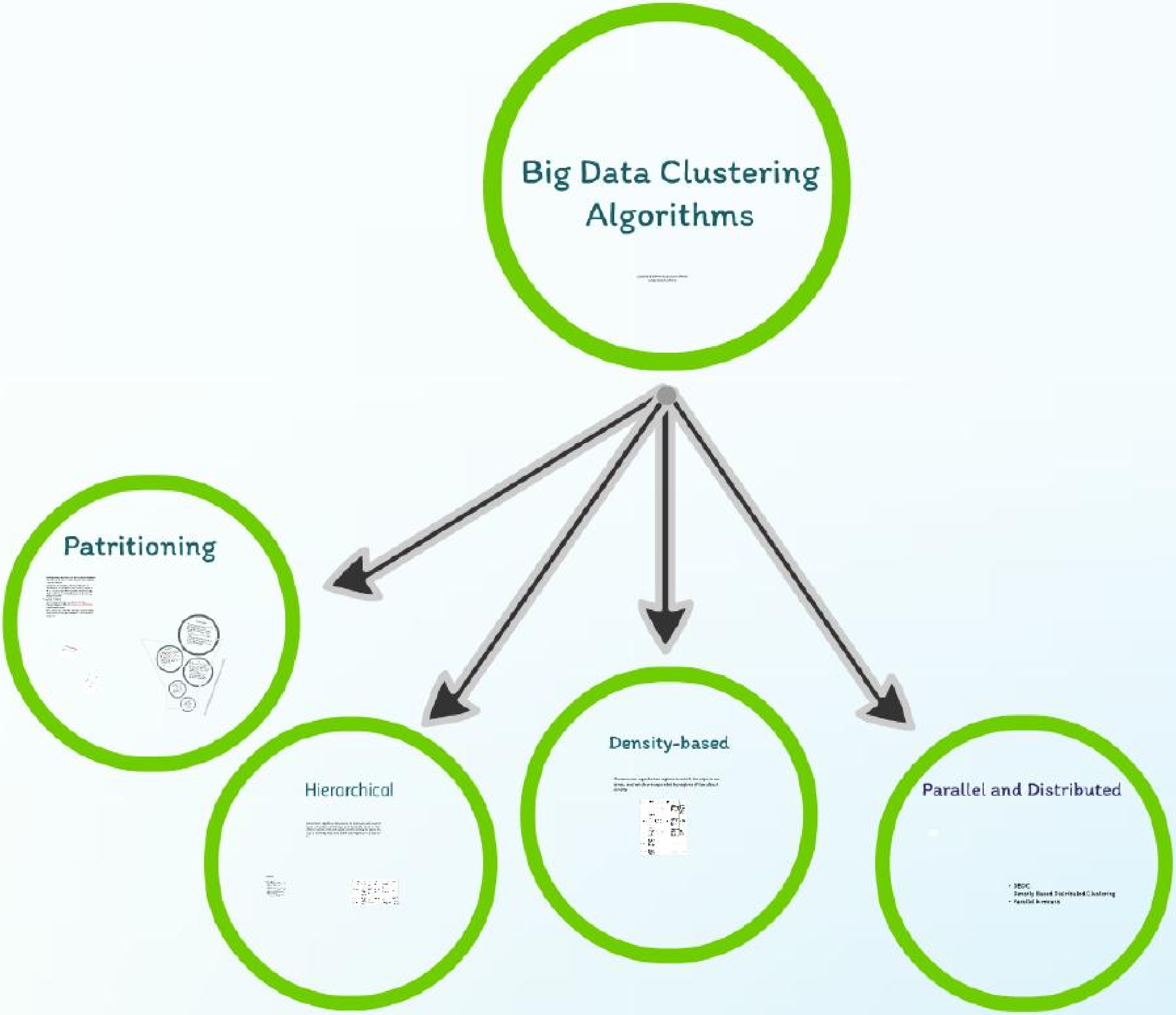
Velocity
Real-time capture and Real-time analytics

Volume
Petabytes per day/week

Variety
Unstructured data, web logs, audio, video, image

Big Data


300
250
200
150
100
50
0
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014



Big Data Clustering Algorithms

Patritioning

Hierarchical

Density-based

Parallel and Distributed

- DBDC
  Density Based Distributed Clustering
- Parallel k-means

Big Data Clustering Algorithms

clustering algorithms according to different
categorization schemes

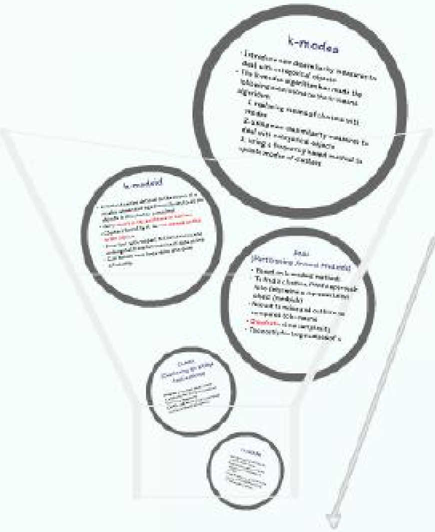
clustering algorithms according to different categorization schemes

Patritioning

- Partitioning algorithms use the distances between the objects directly in order to optimize a global cluster criterion
- Construct a flat (single level) partition of a database D of n objects into a set of k clusters
- The k-means algorithm is best suited for large data set because of its efficiency in clustering large data sets

K-means Problem:

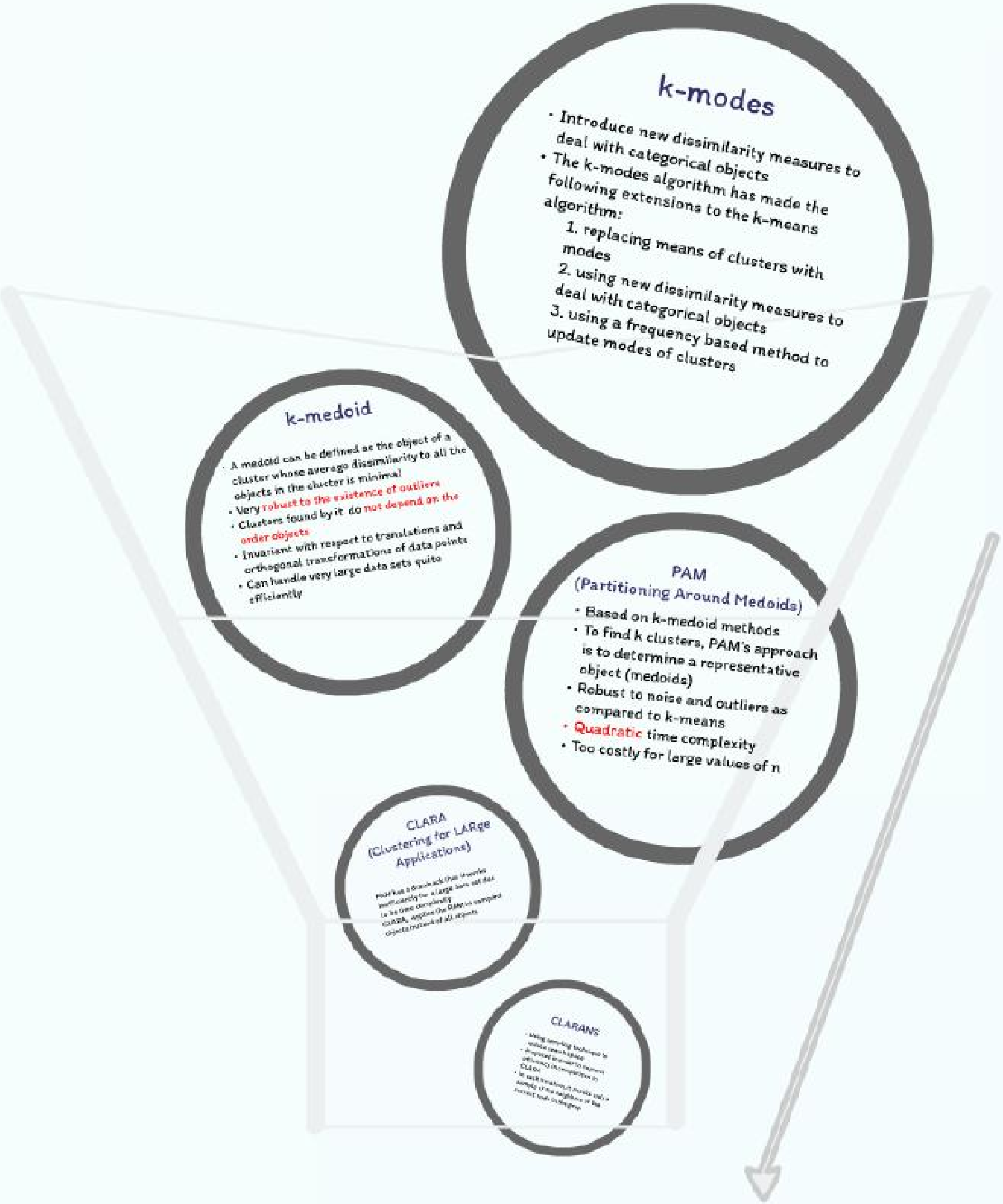
- Can't determine the number of cluster
- Handle dataset with only numerical attributes
- Can't handle noise
- Its performance depends strongly on the initial centroids and may get trapped in local optimal solutions



- Partitioning algorithms use the distances between the objects directly in order to optimize a global cluster criterion
- Construct a flat (single level) partition of a database D of n objects into a set of k clusters
- The k-means algorithm is best suited for large data set because of its efficiency in clustering large data sets

K-means Problem:

- Can’t determine the number of cluster
- Handle dataset with only numerical attributes
- Can't handle noise
- Its performance depends strongly on the initial centroids and may get trapped in local optimal



k-modes

- Introduce new dissimilarity measures to deal with categorical objects
- The k-modes algorithm has made the following extensions to the k-means algorithm:
  1. replacing means of clusters with modes
  2. using new dissimilarity measures to deal with categorical objects
  3. using a frequency based method to update modes of clusters

k-medoid

- A medoid can be defined as the object of a cluster whose average dissimilarity to all the objects in the cluster is minimal
- Very robust to the existence of outliers
- Clusters found by it do not depend on the order objects
- Invariant with respect to translations and orthogonal transformations of data points
- Can handle very large data sets quite efficiently

PAM
(Partitioning Around Medoids)

- Based on k-medoid methods
- To find k clusters, PAM's approach is to determine a representative object (medoids)
- Robust to noise and outliers as compared to k-means
- Quadratic time complexity
- Too costly for large values of n

CLARA
(Clustering for LARge Applications)

[illegible]

CLARANS

- [illegible]
- [illegible]
- [illegible]

k-modes

- Introduce new dissimilarity measures to deal with categorical objects
- The k-modes algorithm has made the following extensions to the k-means algorithm:
  1. replacing means of clusters with modes
  2. using new dissimilarity measures to deal with categorical objects
  3. using a frequency based method to update modes of clusters

medoid

the object of a
to all the

k-medoid

- A medoid can be defined as the object of a cluster whose average dissimilarity to all the objects in the cluster is minimal
- Very robust to the existence of outliers
- Clusters found by it do not depend on the order objects
- Invariant with respect to translations and orthogonal transformations of data points
- Can handle very large data sets quite efficiently

d
nts

PAM
(Partitioning Around Medoids)

- Based on k-medoid methods
- To find k clusters, PAM's approach is to determine a representative object (medoids)
- Robust to noise and outliers as compared to k-means
- Quadratic time complexity
- Too costly for large values of n

ARA
LARge
ng for
es)

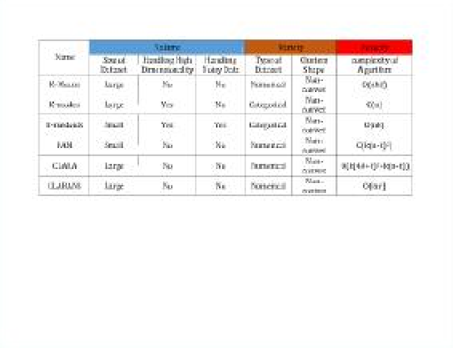
CLARA (Clustering for LARge Applications)

PAM has a drawback that it works inefficiently for a large data set due to its time complexity

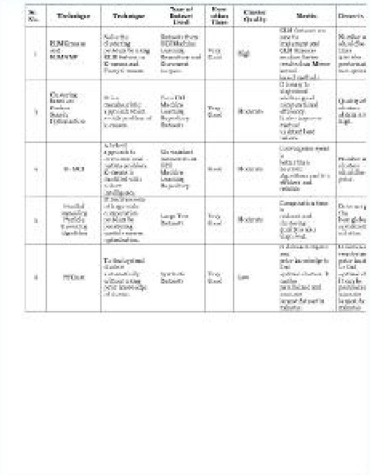
CLARA, applies the PAM to sampled objects instead of all objects

CLARANS

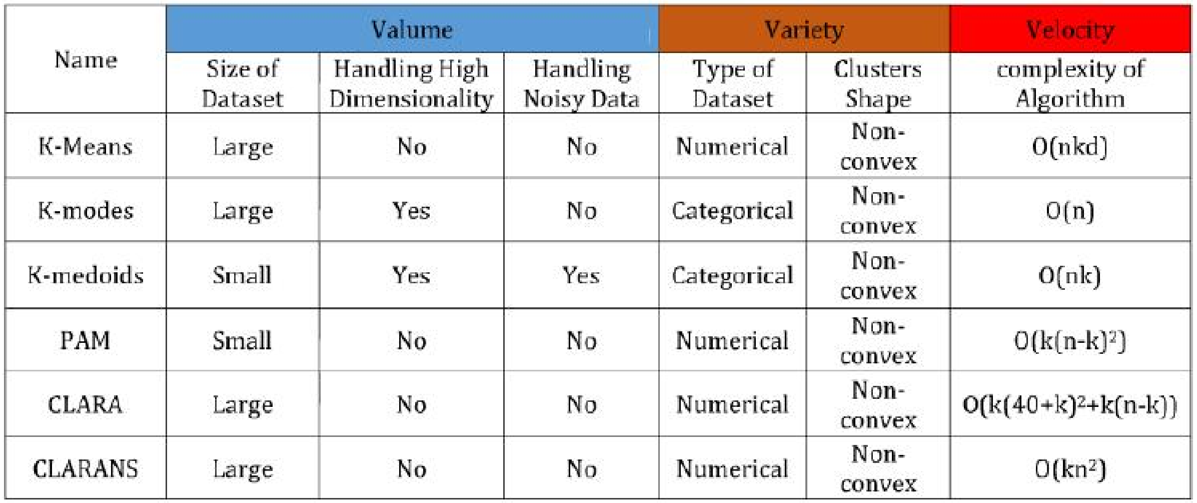
- Using sampling technique to reduce search space
- Proposed in order to improve efficiency in comparison to CLARA
- In each iteration, it checks only a sample of the neighbors of the current node in the grap



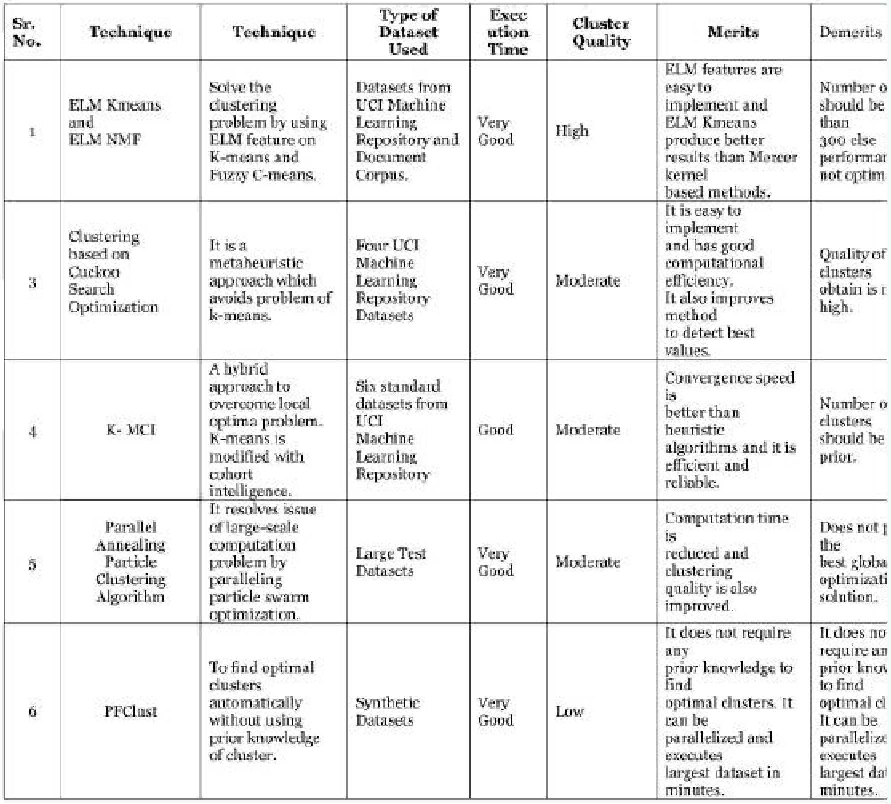
Name	Volume			Variety		Velocity
	Size of Dataset	Handling High Dimensionality	Handling Noisy Data	Type of Dataset	Clusters Shape	Complexity of Algorithm
K-Means	Large	No	No	Numerical	Non-convex	O(nkd)
K-modes	Large	Yes	No	Categorical	Non-convex	O(n)
K-medoids	Small	Yes	Yes	Categorical	Non-convex	[illegible]
PAM	Small	No	No	Numerical	Non-convex	O(k(n-k)²)
CLARA	Large	No	No	Numerical	Non-convex	O(k(40+k)²+k(n-k))
CLARANS	Large	No	No	Numerical	Non-convex	O(kn²)



Sr. No.	Technique	Technique	Type of Dataset Used	Execution Time	Cluster Quality	Merits	Demerits
1	[illegible]	[illegible]	[illegible]	Very Good	High	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	Very Good	Moderate	[illegible]	[illegible]
4	[illegible]	[illegible]	[illegible]	Good	Moderate	[illegible]	[illegible]
5	[illegible]	[illegible]	Large Test Datasets	Very Good	Moderate	[illegible]	[illegible]
6	[illegible]	To find optimal clusters automatically without using prior knowledge of clusters.	Synthetic Datasets	Very Good	Low	[illegible]	[illegible]



Name	Valume			Variety		Velocity
	Size of Dataset	Handling High Dimensionality	Handling Noisy Data	Type of Dataset	Clusters Shape	complexity of Algorithm
K-Means	Large	No	No	Numerical	Non-convex	O(nkd)
K-modes	Large	Yes	No	Categorical	Non-convex	O(n)
K-medoids	Small	Yes	Yes	Categorical	Non-convex	O(nk)
PAM	Small	No	No	Numerical	Non-convex	$O(k(n-k)^2)$
CLARA	Large	No	No	Numerical	Non-convex	$O(k(40+k)^2+k(n-k))$
CLARANS	Large	No	No	Numerical	Non-convex	$O(kn^2)$



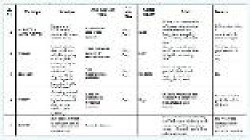
Sr. No.	Technique	Technique	Type of Dataset Used	Execution Time	Cluster Quality	Merits	Demerits
1	ELM Kmeans and ELM NMF	Solve the clustering problem by using ELM feature on K-means and Fuzzy C-means.	Datasets from UCI Machine Learning Repository and Document Corpus.	Very Good	High	ELM features are easy to implement and ELM Kmeans produce better results than Mercer kernel based methods.	Number o should be than 300 else performa not optim
3	Clustering based on Cuckoo Search Optimization	It is a metaheuristic approach which avoids problem of k-means.	Four UCI Machine Learning Repository Datasets	Very Good	Moderate	It is easy to implement and has good computational efficiency. It also improves method to detect best values.	Quality of clusters obtain is high.
4	K- MCI	A hybrid approach to overcome local optima problem. K-means is modified with cohort intelligence.	Six standard datasets from UCI Machine Learning Repository	Good	Moderate	Convergence speed is better than heuristic algorithms and it is efficient and reliable.	Number o clusters should be prior.
5	Parallel Annealing Particle Clustering Algorithm	It resolves issue of large-scale computation problem by paralleling particle swarm optimization.	Large Test Datasets	Very Good	Moderate	Computation time is reduced and clustering quality is also improved.	Does not the best globa optimizati solution.
6	PFClust	To find optimal clusters automatically without using prior knowledge of cluster.	Synthetic Datasets	Very Good	Low	It does not require any prior knowledge to find optimal clusters. It can be parallelized and executes largest dataset in minutes.	It does no require an prior kno to find optimal cl It can be paralleliz executes largest da minutes.

Hierarchical

Hierarchical algorithms decompose the data base into several levels of nested partitionings and iteratively splits D into smaller subsets until each subset consists of only one object. In such a hierarchy, each node of the tree represents a cluster of D.

BIRCH

[illegible]

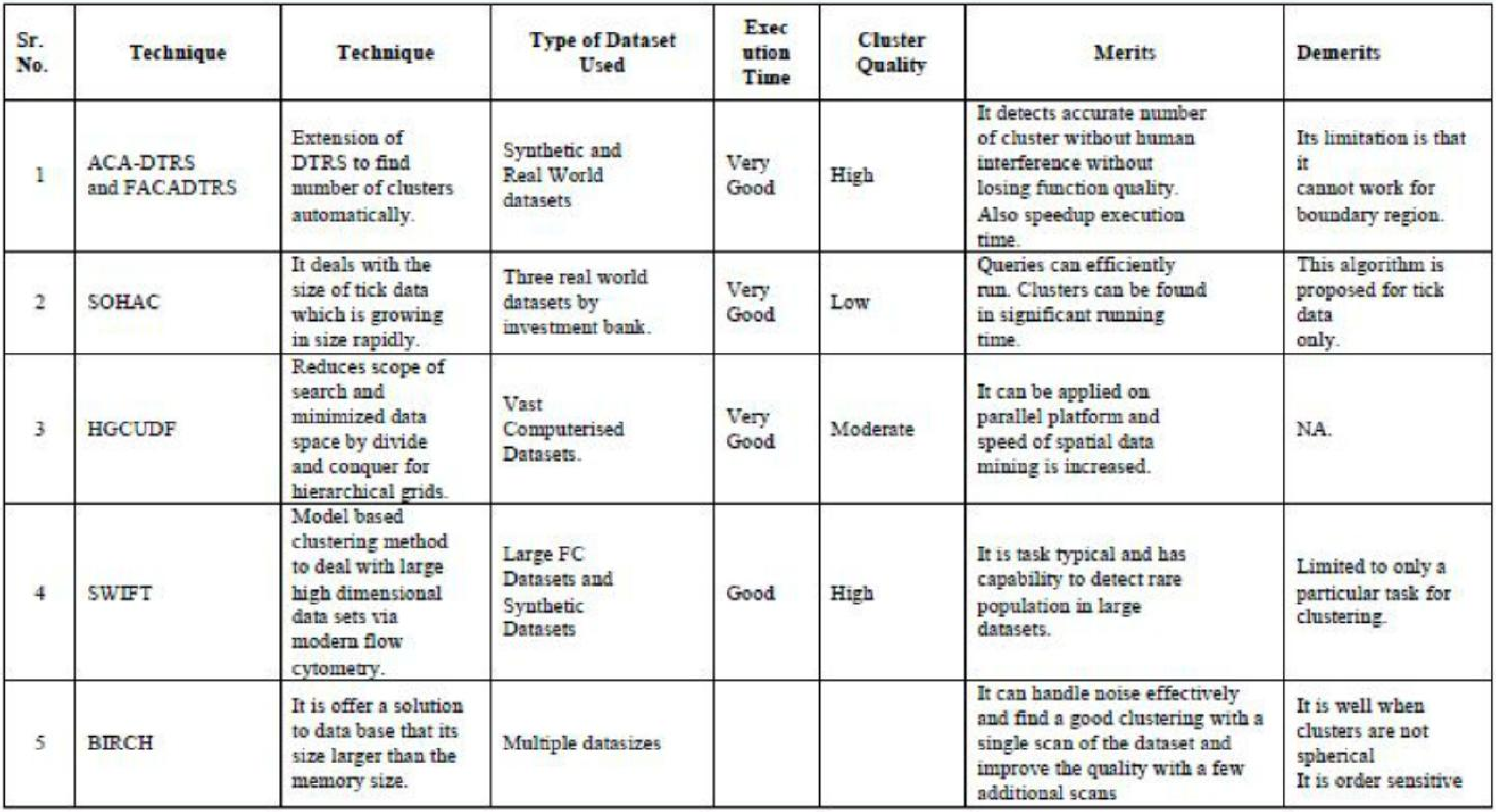


Hierarchical algorithms decompose the data base into several levels of nested partitionings and iteratively splits D into smaller subsets until each subset consists of only one object. In such a hierarchy, each node of the tree represents a cluster of D.

BIRCH

Properties of algorithm

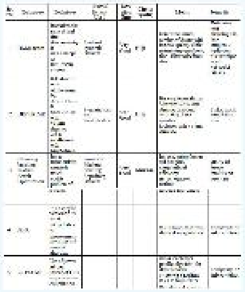
- Handles mixed types of attributes
- Automatically determine the best number of clusters(A desirable feature in clustering is to determine the number of clusters utomatically)
- Identifies outlier or noise data records (95%)
- Linear scalability by increasing data set (We test the scalability of our algorithm by increasing number of data records and number of attributes.)
- Generates better quality clusters than the traditional k-means algorithms



Sr. No.	Technique	Technique	Type of Dataset Used	Execution Time	Cluster Quality	Merits	Demerits
1	ACA-DTRS and FACADTRS	Extension of DTRS to find number of clusters automatically.	Synthetic and Real World datasets	Very Good	High	It detects accurate number of cluster without human interference without losing function quality. Also speedup execution time.	Its limitation is that it cannot work for boundary region.
2	SOHAC	It deals with the size of tick data which is growing in size rapidly.	Three real world datasets by investment bank.	Very Good	Low	Queries can efficiently run. Clusters can be found in significant running time.	This algorithm is proposed for tick data only.
3	HGCUDF	Reduces scope of search and minimized data space by divide and conquer for hierarchical grids.	Vast Computerised Datasets.	Very Good	Moderate	It can be applied on parallel platform and speed of spatial data mining is increased.	NA.
4	SWIFT	Model based clustering method to deal with large high dimensional data sets via modern flow cytometry.	Large FC Datasets and Synthetic Datasets	Good	High	It is task typical and has capability to detect rare population in large datasets.	Limited to only a particular task for clustering.
5	BIRCH	It is offer a solution to data base that its size larger than the memory size.	Multiple datasizes			It can handle noise effectively and find a good clustering with a single scan of the dataset and improve the quality with a few additional scans	It is well when clusters are not spherical It is order sensitive

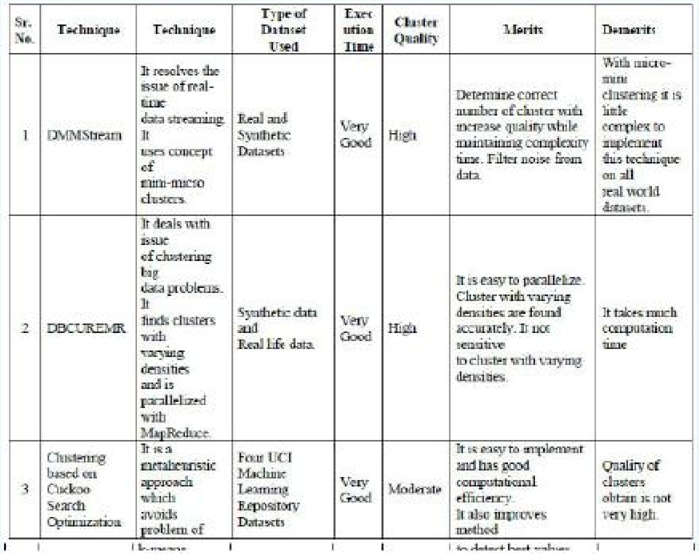
Density-based

Clusters are regarded as regions in which the objects are dense, and which are separated by regions of low object density

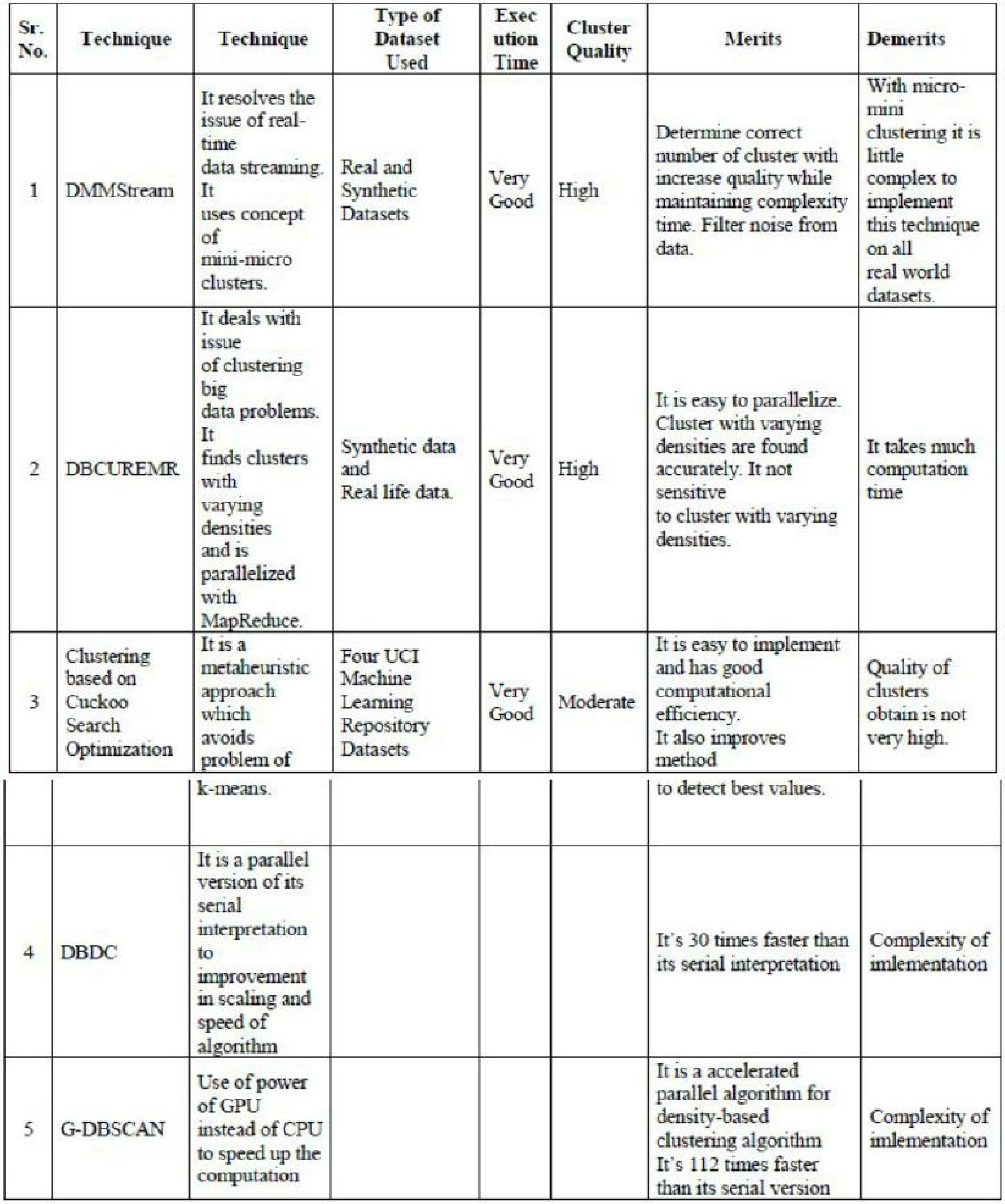


Density-based

Clusters are regarded as regions in which the objects are dense, and which are separated by regions of low object density



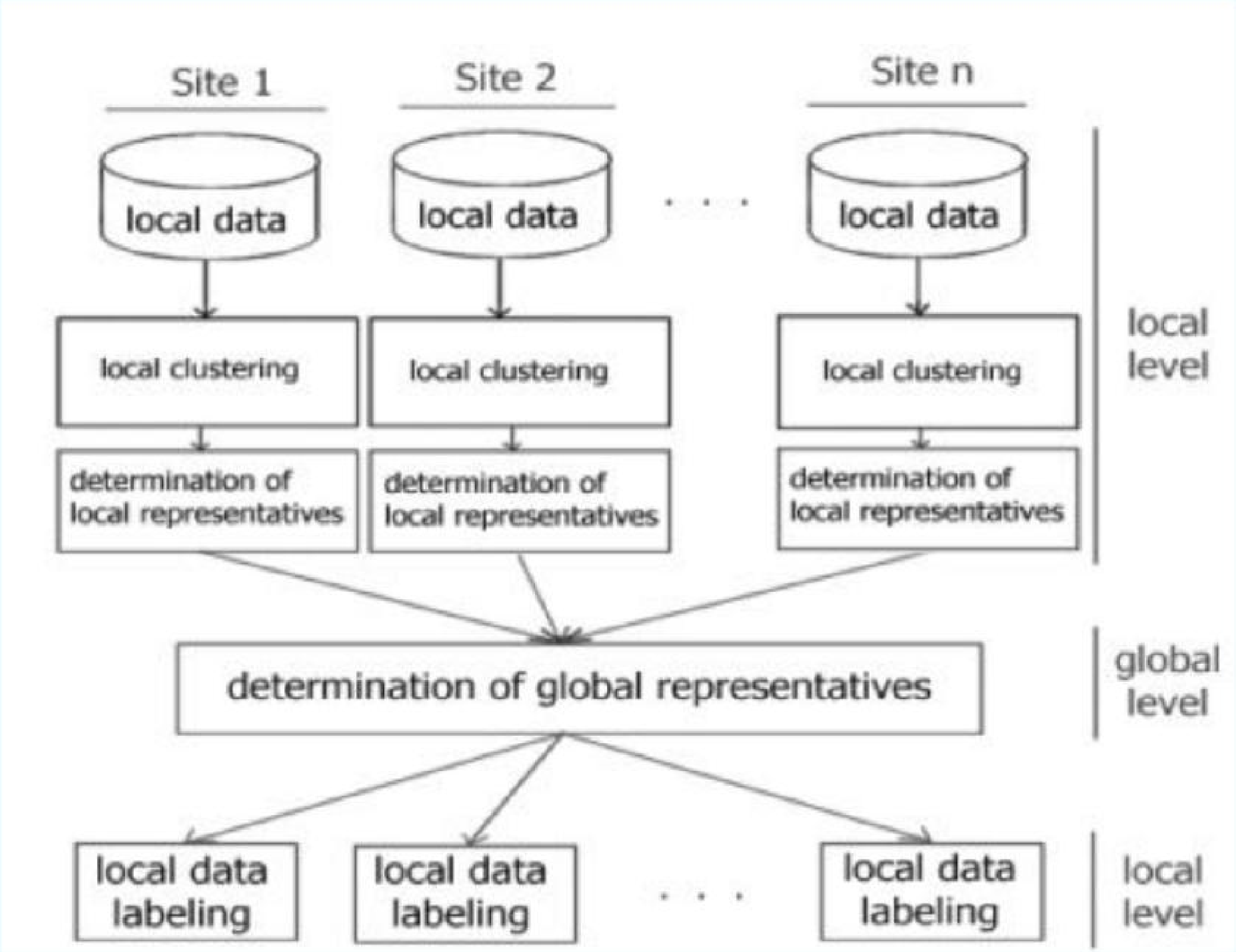
Sr. No.	Technique	Technique	Type of Dataset Used	Execution Time	Cluster Quality	Merits	Demerits
1	DMMStream	It resolves the issue of real-time data streaming. It uses concept of mini-micro clusters.	Real and Synthetic Datasets	Very Good	High	Determine correct number of cluster with increase quality while maintaining complexity time. Filter noise from data.	With micro-mini clustering it is little complex to implement this technique on all real world datasets.
2	DBCUREMR	It deals with issue of clustering big data problems. It finds clusters with varying densities and is parallelized with MapReduce.	Synthetic data and Real life data.	Very Good	High	It is easy to parallelize. Cluster with varying densities are found accurately. It not sensitive to cluster with varying densities.	It takes much computation time
3	Clustering based on Cuckoo Search Optimization	It is a metaheuristic approach which avoids problem of	Four UCI Machine Learning Repository Datasets	Very Good	Moderate	It is easy to implement and has good computational efficiency. It also improves method	Quality of clusters obtain is not very high.



Sr. No.	Technique	Technique	Type of Dataset Used	Execution Time	Cluster Quality	Merits	Demerits
1	DMMStream	It resolves the issue of real-time data streaming. It uses concept of mini-micro clusters.	Real and Synthetic Datasets	Very Good	High	Determine correct number of cluster with increase quality while maintaining complexity time. Filter noise from data.	With micro-mini clustering it is little complex to implement this technique on all real world datasets.
2	DBCUREMR	It deals with issue of clustering big data problems. It finds clusters with varying densities and is parallelized with MapReduce.	Synthetic data and Real life data.	Very Good	High	It is easy to parallelize. Cluster with varying densities are found accurately. It not sensitive to cluster with varying densities.	It takes much computation time
3	Clustering based on Cuckoo Search Optimization	It is a metaheuristic approach which avoids problem of k-means.	Four UCI Machine Learning Repository Datasets	Very Good	Moderate	It is easy to implement and has good computational efficiency. It also improves method to detect best values.	Quality of clusters obtain is not very high.
4	DBDC	It is a parallel version of its serial interpretation to improvement in scaling and speed of algorithm				It's 30 times faster than its serial interpretation	Complexity of imlementation
5	G-DBSCAN	Use of power of GPU instead of CPU to speed up the computation				It is a accelerated parallel algorithm for density-based clustering algorithm It's 112 times faster than its serial version	Complexity of imlementation

Parallel and Distributed

- DBDC
  Density Based Distributed Clustering
- Parallel k-means


Site 1
Site 2
Site n
local data
local data
. . . .
local data
local clustering
local clustering
local clustering
determination of local representatives
determination of local representatives
determination of local representatives
local level
determination of global representatives
global level
local data labeling
local data labeling
. . . .
local data labeling
local level

- DBDC
  Density Based Distributed Clustering
- Parallel k-means

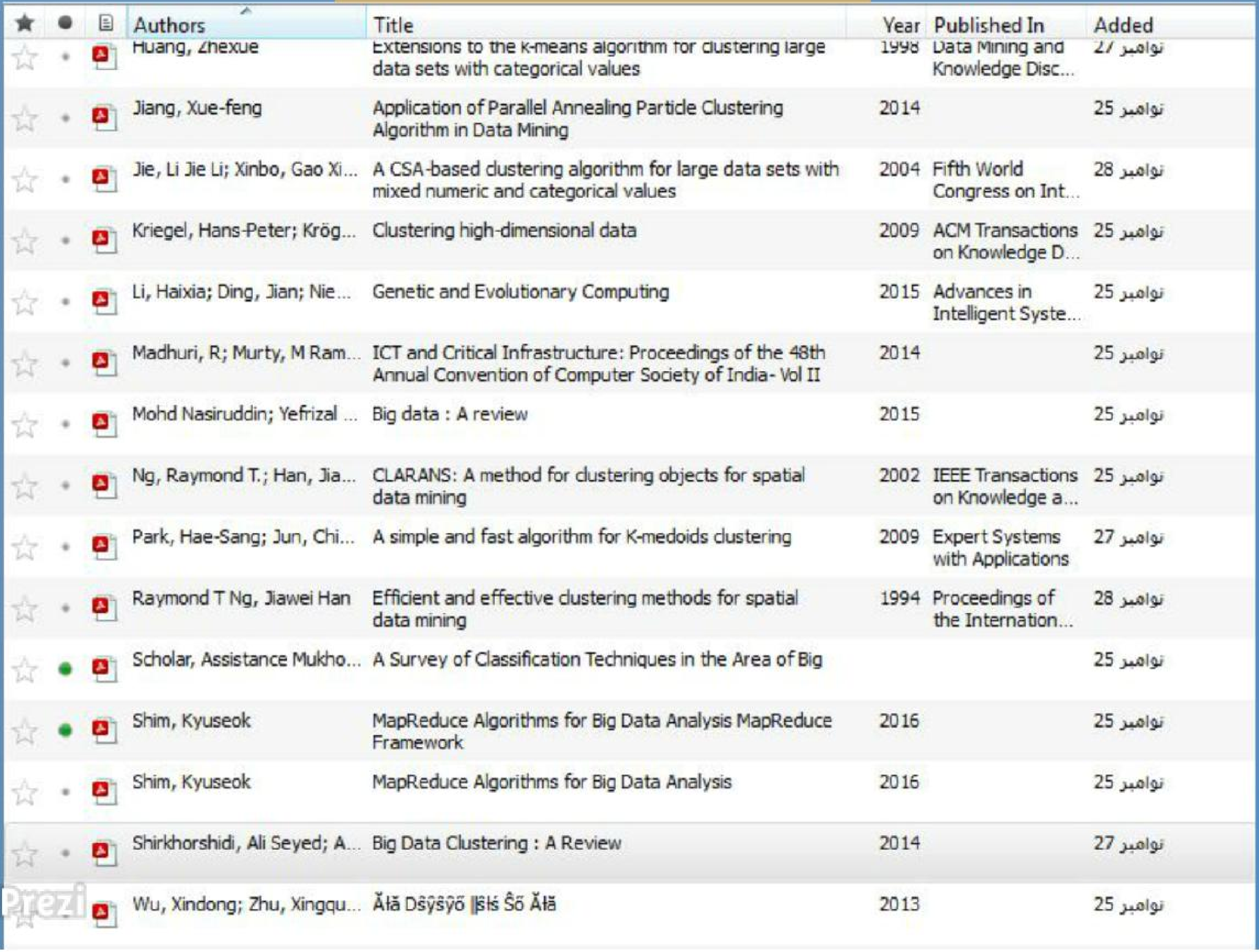
Open Issues

- Deploy clustering algorithms on GPU based MapReduce frameworks to achieve better scalability and speed
- Improvement k-means

Conclusion

- In this study the improvement trend of data clustering algorithm were discusse, the future of clustering is tied with distributed computing. Although parallel clustering is very useful for clustering, MapReduce framework provides a very satisfying base for implementing clustering algorithms.

References



★	●		Authors	Title	Year	Published In	Added
☆	•		Huang, Zhexue	Extensions to the k-means algorithm for clustering large data sets with categorical values	1998	Data Mining and Knowledge Disc...	27 نوامبر
☆	•		Jiang, Xue-feng	Application of Parallel Annealing Particle Clustering Algorithm in Data Mining	2014		25 نوامبر
☆	•		Jie, Li Jie Li; Xinbo, Gao Xi...	A CSA-based clustering algorithm for large data sets with mixed numeric and categorical values	2004	Fifth World Congress on Int...	28 نوامبر
☆	•		Kriegel, Hans-Peter; Krög...	Clustering high-dimensional data	2009	ACM Transactions on Knowledge D...	25 نوامبر
☆	•		Li, Haixia; Ding, Jian; Nie...	Genetic and Evolutionary Computing	2015	Advances in Intelligent Syste...	25 نوامبر
☆	•		Madhuri, R; Murty, M Ram...	ICT and Critical Infrastructure: Proceedings of the 48th Annual Convention of Computer Society of India- Vol II	2014		25 نوامبر
☆	•		Mohd Nasiruddin; Yefrizal ...	Big data : A review	2015		25 نوامبر
☆	•		Ng, Raymond T.; Han, Jia...	CLARANS: A method for clustering objects for spatial data mining	2002	IEEE Transactions on Knowledge a...	25 نوامبر
☆	•		Park, Hae-Sang; Jun, Chi...	A simple and fast algorithm for K-medoids clustering	2009	Expert Systems with Applications	27 نوامبر
☆	•		Raymond T Ng, Jiawei Han	Efficient and effective clustering methods for spatial data mining	1994	Proceedings of the Internation...	28 نوامبر
☆	●		Scholar, Assistance Mukho...	A Survey of Classification Techniques in the Area of Big			25 نوامبر
☆	●		Shim, Kyuseok	MapReduce Algorithms for Big Data Analysis MapReduce Framework	2016		25 نوامبر
☆	•		Shim, Kyuseok	MapReduce Algorithms for Big Data Analysis	2016		25 نوامبر
☆	•		Shirkhorshidi, Ali Seyed; A...	Big Data Clustering : A Review	2014		27 نوامبر
☆	•		Wu, Xindong; Zhu, Xingqu...	Ăłă Dŝŷŝŷő ‖ŝłś Ŝő Ăłă	2013		25 نوامبر