Scopus

Documents

Mallik, M.A.[a b] , Zulkurnain, N.F.[a] , Nizamuddin, M.K.[c] , Sarkar, R.[d] , Chalil, A.K.[e]

A SPARK-BASED PARALLEL FUZZY C MEDIAN ALGORITHM FOR WEB LOG BIG DATA

(2022) *International Journal on Technical and Physical Problems of Engineering*, 14 (3), pp. 212-220.

[a] International Islamic University Malaysia, Kuala Lumpur, Malaysia
[b] VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, India
[c] Shadan Women's College of Engineering and Technology, Jawaharlal Nehru Technological University, Hyderabad, India
[d] University of Science and Technology, Meghalaya, India
[e] Malla Reddy College of Enginnering, Hyderabad, India

Abstract
Now-a-days, the World Wide Web (WWW) is regarded as an exceptionally large data storehouse. The WWW is becoming more complicated and substantive every day. At the moment, the situation is such that we are starved for knowledge while drowning in data. Due to these factors, the data mining clustering technique is one of the most crucial tools for collecting useful data from the web. Clustering techniques for small datasets have led to the development of numerous successful clustering techniques. Nevertheless, these techniques do not provide adequate results when trading with extensive data sets. The most important problems are excessive computational difficulty and lengthy evaluating time, which is not acceptable for real-time context. It is very prime to process this enormous information on time. This paper proposes an efficient parallel Fuzzy C median solution based on Spark for large-scale web log data. Based on the Rand Index and SSE (sum of squared error), the parallel Fuzzy C median algorithm's performance is evaluated in the PySpark platform. According to the experimental findings, the parallel Fuzzy C median method built on Spark performs better. © 2022, International Organization on 'Technical and Physical Problems of Engineering'. All rights reserved.

Author Keywords
Apache Spark; Fuzzy Clustering; Parallel Computing; Web Log Big Data

References

- Sardar, T.H., Ansari, Z.
  MapReduce-Based Fuzzy C-Means Algorithm for Distributed Document Clustering
  (2022) *The Institution of Engineers*, 103 (3), pp. 131-142.
  Kolkata, India

- Guliyev, H.B.
  Fuzzy Probabilistic Model for Managing the Modes of Networks with Renewable Energy Sources
  (2021) *International Journal on Technical and Physical Problems of Engineering (IJTPE)*, 13 (1), pp. 46-50.
  46 March

- Mallik, M.A.
  An Efficient Fuzzy C-Least Median Clustering Algorithm
  (2021) *IOP Conf. Ser., Mater. Sci. Eng*, 1070 (1), p. 012050.
  Tamil Nadu, India

- Bendechachea, M., Tarib, A.K., Kechadiaa, M.T.
  Insight Centre for Data Analytics", University College Dublin, Ireland University A-Mira of Bejaia, Algeria "Parallel and Distributed Clustering Framework for Big Spatial Data Mining
  (2018) *Article in International Journal of Parallel Emergent and Distributed Systems*, 34 (6), pp. 671-689.
  March

- Cooley, R., Mobasher, B., Srivastava, J.
  Web Mining: Information Andpattern Discovery on the World Wide Web
  (1997) *The Ninth IEEE International Conference*, pp. 558-567.
  London, UK, November

- Mallik, M.A.
  A Survey on Parallel Clustering Techniques for Big Data Framework
  (2022) *The 2nd Global Conference on Artificial Intelligence and Applications (GCAIA 2021)*, pp. 49-56.
  CRC Press, Taylor and Francis, Jaipur, India

- Sinha, A., Jana, P.K.
A Novel K-Means Based Clustering Algorithm for Big Data
(2016) *IEEE International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 1875-1879.

- Sreedhar, C., Kasiviswanath, N., Reddy, P.C.
Clustering Large Datasets using K-Means Modified Inter and Intra Clustering (KM-I2C) in Hadoop
(2017) *Journal of Big Data*, 4 (1), p. 27.

- Akthar, N., Ahamad, M.V., Khan, S.
Clustering on Big Data Using Hadoop MapReduce
(2015) *The 2015 IEEE International Conference on Computational Intelligence and Communication Networks (CICN)*, pp. 789-795.

- Sinha, A., Jana, P.K.
A Novel K-Means Based Clustering Algorithm for Big Data
(2016) *The 2016 IEEE International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 1875-1879.

- Ben Haj Kacem, M.A., Ben N'cir, C.E., Essoussi, N.
One-Pass MapReduce-Based Clustering Method for Mixed Large-Scale Data
(2017) *Journal of Intelligent Information Systems*, pp. 1-18.

- Shafiq, M.O., Torunski, E.
A Parallel K-Medoids Algorithm for Clustering Based on MapReduce
(2016) *The 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 502-507.

- Ben Haj Kacem, M.A., Ben N'cir, C.E., Essoussi, N.
MapReduce-Based K-Prototypes Clustering Method for Big Data
(2015) *The 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 1-7.

- Ludwig, S.A.
MapReduce-Based Fuzzy C-Means Clustering Algorithm: Implementation and Scalability
(2015) *International Journal of Machine Learning and Cybernetics*, 6 (6), pp. 923-934.

- Hidri, M.S., Zoghlami, M.A., Ben Ayed, R.
Speeding up the Large-Scale Consensus Fuzzy Clustering for Handling Big Data
(2017) *Fuzzy Sets and Systems*,

- Zhang, Q., Chen, Z.
A Weighted Kernel Possibilistic C-Means Algorithm Based on Cloud Computing for Clustering Big Data
(2014) *International Journal of Communication Systems*, 27 (9), pp. 1378-1391.

- Zhang, Q., Yang, L.T., Chen, Z., Li, P.
PPHOPCM: Privacy-Preserving High-Order Possibilistic C-Means Algorithm for Big Data Clustering with Cloud Computing
(2017) *IEEE Transactions on Big Data*, (99), pp. 1-11.
May

- Hu, R., Dou, W., Liu, J.
ClubCF: A Clustering-Based Collaborative Filtering Approach for Big Data Application
(2014) *IEEE Transactions on Emerging Topics in Computing*, 2 (3), pp. 302-313.

- Subramaniyaswamy, V., Vijayakumar, V., Logesh, R., Indragandhi, V.
Unstructured Data Analysis on Big Data Using MapReduce

(2015) *Procedia Computer Science*, 50, pp. 456-465.

- Sachar, P., Khullar, V.
  Social Media Generated Big Data Clustering Using Genetic Algorithm
  (2017) *The 2017 IEEE International Conference on Computer Communication and Informatics (ICCCI)*, pp. 1-6.
  Coimbatore

- Karimov, J., Ozbayoglu, M.
  High Quality Clustering of Big Data and Solving Empty-Clustering Problem with an Evolutionary Hybrid Algorithm
  (2015) *The 2015 IEEE International Conference on Big Data (Big Data)*, pp. 1473-1478.

- Yang, Y., Teng, F., Li, T., Wang, H., Wang, H., Zhang, Q.
  Parallel Semi-Supervised Multi-Ant Colonies Clustering Ensemble Based on MapReduce Methodology
  (2015) *IEEE Transactions on Cloud Computing*, 6 (1), pp. 1-12.

- Sais, M., Rafalia, N., Abouchabaka, J.
  Enhancements and Intelligent Approach to Optimize Big data Storage and Management: Random Enhanced HDFS (REHDFS) and DNA Storage
  (2022) *International Journal on Technical and Physical Problems of Engineering (IJTPE)*, 14 (1), pp. 196-203.
  50 March

- Ansari, Z., Azeem, M.F, Babu, A.V., Ahmed, W.
  A Fuzzy Clustering based Approach for Mining Usage Profiles from Web Log Data
  (2011) *International Journal of Computer Science and Information Security (IJC-SIS)*, 9 (6), pp. 70-79.
  9, June

- Castellano, G., Mesto, F., Minunno, M., Torsello, M.A.
  Web User Profiling Using Fuzzy Clustering
  (2007) *Lecture Notes in Computer Science*, 4578, pp. 94-101.
  WILF (F. Masulli, S. Mitra, G. Pasi, eds), Springer

- Nasraoui, O., Frigui, H., Krishnapuram, R., Joshi, A.
  Extracting Web User Profiles Using Relational Competitive Fuzzy Clustering
  (2000) *International Journal on Artificial Intelligence Tools*, 9 (4), pp. 509-526.

- Castellano, G., Fanelli, A.M., Torsello, M.A.
  Mining Usage Profiles Fromaccess Data Using Fuzzy Clustering
  (2006) *The 6th WSEAS International Conference on Simulation, Modelling and Optimization (SMO)*, pp. 157-160.
  Stevens Point, Wisconsin, USA

- Ansari, Z., Azeem, M.F., Babu, A.V., Waseem, A.
  A Fuzzy Approach for Feature Evaluation and Dimensionality Reduction to Improve the Quality of Web Usage Mining Results
  (2012) *International Journal on Advanced Science, Engineering and Information Technology*, 2 (6), pp. 67-73.

- Ansari, Z., Babuy, A., Ahmed, W., Azeemz, M.
  A Fuzzy Set Theoretic Approach to Discover User Sessions from Web Navigational Data
  (2011) *Recent Advances in Intelligent Computational Systems (RAICS)*, pp. 879-884.
  September

- Gallego, S.R., Garcia, S., Benitez, J.M., Herrera, F.
  A Distributed Evolutionary Multivariate Discretizer for Big Data Processing on Apache Spark

(2018) *Swarm Evol. Comput*, 38, pp. 240-250.
February

Publisher: International Organization on 'Technical and Physical Problems of Engineering'

ISSN: 20773528
Language of Original Document: English
Abbreviated Source Title: Int. J. Tech. Phys. Probl. Eng.
2-s2.0-85139463678
Document Type: Article
Publication Stage: Final
Source: Scopus

ELSEVIER

Copyright © 2022 Elsevier B.V. All rights reserved. Scopus® is a registered trademark of Elsevier B.V.

RELX Group™