{tag} {/tag}International Journal of Computer Applications
Foundation of Computer Science (FCS), NY, USA

Volume 181
-
Number 18

Year of Publication: 2018

Authors:

Shyam Mohan J. S., Shanmugapriya P.

10.5120/ijca2018917856

{bibtex}2018917856.bib{/bibtex}

## Abstract

Cluster identification is useful for finding insights into the huge datasets for finding out the attributes, characteristics of a particular dataset. Today, many organizations have started to use their own data analytic tools for finding clusters.

This paper focuses on various algorithms for finding clusters for huge and different datasets. We have used different datasets and applied MapReduce algorithms for achieving the results. The experimental results obtained in substantial algorithmic computations provide clusters that are used for quick decision making. We present the results performed over various datasets that scales well with respect to both data set size and data set dimensionality.

## References

1. Steinbach M, Ertöz L, Kumar V (2004) The challenges of clustering high dimensional data. In: New directions in statistical physics. Springer, Berlin Heidelberg. pp 273–309.

2. Fan J, Han F, Liu H (2014) Challenges of big data analysis. National Science Review 1(2):293–314.

3. Parsons L, Haque E, Liu H (2004) Subspace clustering for high dimensional data: a review. ACM SIGKDD Explor News l6(1):90–105.

4. Babu MM (2004) Introduction to microarray data analysis. In: Grant RP (ed). Computational genomics: Theory and application. Horizon Press, UK. pp 225–249

5. Jiang D, Tang C, Zhang A (2004) Cluster analysis for gene expression data: a survey. IEEE Trans Knowl Data Eng 16(11):1370–1386

6. Cheng Y, Church GM (2000) Biclustering of expression data. Proc Int Conf Intell Syst Mol Biol 8:93–103.

7. Jongeneel CV (2000) Searching the expressed sequence tag (est) databases: panning for genes. Bioinformatics 1:76–92.

8. Liu B, Wang X, Zou Q, Dong Q, Chen Q (2013) Protein remote homology detection by combining chous pseudo amino acid composition and profile-based protein representation. Mol Inf 32(9–10):775–782.

9. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215(3):403–410.

10. Halko, N.H., Martinsson, P.-G., Shkolnisky, Y., Tygert, M.: An algorithm for the principal component analysis of large data sets. SIAM J. Sci. Comput. 33(5), 2580–2594 (2011).

11. Carlos Ordonez Et.Al.” PCA for large data sets with parallel data summarization” Distrib Parallel Databases (2014) 32:377–403, Springer, DOI 10.1007/s10619-013-7134-6.

12. Alexios Kotsifakos et.al.” DRESS: dimensionality reduction for efficient sequence search”, Data Min Knowl Disc (2015) 29:1280–1311, Springer, DOI 10.1007/s10618-015-0413-2.

13. Ester M, Kriegel H, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. Int Conf Knowl Discov Data Min 96(34):226–231

14. Dongkuan Xu et.al,” A Comprehensive Survey of Clustering Algorithms”. Ann. Data. Sci. Springer-Verlag Berlin Heidelberg 2015.

15. Dean, Jeffrey, and Sanjay Ghemawat. "MapReduce: simplified data processing on large clusters." Communications of the ACM 51, no. 1 (2008): 107-113.

16. Nivranshu Hans et.al,” Big Data Clustering Using Genetic Algorithm On Hadoop Mapreduce”. INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 4, ISSUE 04, APRIL 2015: 58-62.

17. J. Dean and S. Ghemawat. Mapreduce: Simplified data processing on large clusters. OSDI, 2004.

18. Jeffrey Dean and Sanjay Ghemawat, “MapReduce: Simplified Data Processing on Large Clusters” Google, Inc.USENIX Association OSDI ’04: 6th Symposium on Operating Systems Design and Implementation.

19. Altschul S, Madden T, Schffer R, Zhang J, Zhang Z, MillerW, Lipman D (1997) Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402.

20. Korf I, Gish W (2000) Mpblast : improved blast performance with multiplexed queries. Bioinformatics 16:1052–1053.

21. Bhadra R, Sandhya S, Abhinandan KR, Chakrabarti S, Sowdhamini R, Srinivasan N (2006) Cascade psiblast web server: a remote homology search tool for relating protein domains. Nucleic Acids Res 34(Web–Server–Issue):143–146.

22. Zhang Z, Schwartz S,Wagner L,MillerW(2000) A greedy algorithm for aligning dna sequences. J Comput Biol 7:203–214.

23. Kent WJ (2002) Resource BLAT-The BLAST-like alignment tool. Genome Res.

24. SunY,Han J,YanX,Yu PS,Wu T (2011) Pathsim: meta path-based top-k similarity search in heterogeneous information networks. PVLDB 4(11):992–1003.

25. Hastie, T., Tibshirani, R., Friedman, J.H.: The Elements of Statistical Learning, 1st edn. Springer, New York (2001).

26. Ming Ji Et.al," Mining strong relevance between heterogeneous entities from unstructured biomedical data", Data Min Knowl Disc (2015) ,Springer,29:976–998. DOI 10.1007/s10618-014-0396-4.

27. Max Bodoia , "MapReduce Algorithms for k-means Clustering".

28. Weizhong Zhao et.al," Parallel K-Means Clustering Based on MapReduce". CloudCom , Springer-Verlag Berlin Heidelberg 2009. LNCS 5931, pp. 674–679.

## Index Terms

Computer Science Artificial Intelligence

## Keywords

Machine Intelligence, Dimensionality reduction.