Comparision of Keyword based Clustering of Web Documents by using OPENSTACK 4J and by Traditional Method

International Journal of Computer Applications
Foundation of Computer Science (FCS), NY, USA

Volume 156
Number 9

Year of Publication: 2016

Authors:
Shiza Anand, Pradeep Pant, Mukesh Rawat

10.5120/ijca2016912583

Abstract

As the number of hypertext documents are increasing continuously day by day on world wide web. Therefore, clustering methods will be required to bind documents into the clusters (repositories) according to the similarity lying between the documents. Various clustering methods exist such as: Hierarchical Based, K-means, Fuzzy Logic Based, Centroid Based etc. These keyword based clustering methods takes much more amount of time for creating containers and putting documents in their respective containers. These traditional methods use File Handling techniques of different programming languages for creating repositories and transferring web documents into these containers. In contrast, openstack4j SDK is a new technique for creating containers and shifting web documents into these containers according to the similarity in much more less amount of time as compared to the traditional methods. Another benefit of this technique is that this SDK understands and reads all types of files such as jpg, html, pdf, doc etc. This paper compares the time required for clustering of documents by using openstack4j and by traditional methods and suggests various search engines to
adopt this technique for clustering so that they give result to the user queries in less amount of time.

References

2. Tom Fifield, Diane Fleming & Joe Topjian, “OpenStack Operations Guide” by O'Reilly Publications, @itarchitectkev, openstack.prov12n.com
10. IBM-Object Storage, www. IBM.com/object-storage/, Date: 18.05.2016, Time: 11.30 am
13. Wei Xu, Xin Liu, Yihong Gong, “Document Clustering Based On Non-negative Matrix Factorization”,

Index Terms

Computer Science Information Sciences
Keywords

Clustering, openstack4j, K-Means, centroid based, document-matching