Volume 6, Issue 11, November 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Various Document Clustering Algorithms and their Key Parameters

Sonia Saini
Department of CSE, Kurukshetra,
Haryana, India

Abstract: ***Document clustering is the application of cluster analysis. It has applications in automatic document organization, topic extraction and fast information retrieval or filtering. KNN-classifier, K-means, Improved K-means, Simhash, SAHKC and various other algorithms and their keyparameters are discussed in the presented paper. This paper is quite helpful for document clustering point of view.***

Keyword: SAHKC(simple agglomerative hierarchical k-means), MFVSM(multiple feature vector space model), KNN(K- Nearest Negibhor

I. INTRODUCTION

Text Clustering: Grouping of text documents into meaningful clusters in an unsupervised manner. Relevant documents tend to be more similar than to non-relevant documents.

Clustering is an automatic learning technique aimed at grouping a set of objects into subsets and clusters.

The goal is to create a cluster where objects are similar as much as possible with in the cluster and dissimilar as much as possible from objects in other clusters. [1]

Document clustering is the application of cluster analysis to textual documents. It has applications in automatic document organization, topic extraction and fast information retrieval or filtering.

II. ALGORITHMS

K-NN classifier: K-NN improves the effectiveness of categorization system. Term association is main key parameter to be included to improve the performance of K-NN classifier.[2]

K-means :is an algorithm that helps in text clustering. *k*-means clustering aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean.

Improved k-means: algorithm based on modified cosine distance measure for document clustering. Minimum intra-cluster distance and maximum intra cluster distance by reducing cluster size. Term frequency is main parameter that was used in this approach.[9]

Improved transfer learning algorithm for document categorization , that was based on data set reconstruction. Results of this algorithm was compared with other algorithms and results showed that this better than others in some extent.[7]

Simhash algorithm: simhash reduces text size and computes hamming distance b/w fingerprints as the vector distance, on the basis of that can conclude that to which cluster data belongs to. term weight is main key parameter.[8]

An approach for text clustering based on k-means. Response time is reduced and accuracy is increased as compare to well known approach for clustering like hierarchal and partional clustering.[11]

An approach based on improved partical swarn optimization and k-means .This approach is better than previous work in this field in terms of accuracy, robustness and high compact clustering.[6]

An algorithm for web document clustering based on hierarchical k-means algorithm, this algorithm is better than others in terms of quality. A model MFVSM(multiple feature vector space model) is proposed.[10]

III. ALGORITHMS AND THEIR KEYPARAMETERS

Algorithm	parameters
K-means	Retrieval time, fast convergence
Suffix tree clustering	Common phrases, accuracy

Simhash	Speed, term weight
Particle Swarn Optimization	Globalised searching
Improved Particle Swarn Optimization	High compact clustering
SAHKC(simple agglomerative hierarchical k-means)	Running time
MFVSM(multiple feature vector space model)	Quality of clustering

IV. CONCLUSION

- Some algorithms reduces running time but quality of clustering reduces in that case.
- No work done on term probability factor in the previous work.
- There is no method defined for center prediction in the previous work.
- Optimal cost, minimum running time and higher accuracy are not there in a single algorithm.
- K-NN, centroid based, neighborhood based and statistical approaches and compared all these approaches for effective document categorization and concluded that KNN is better than others.[3]

REFERENCES

[1] L. Yanjun, L. Congan, C. M. soon,"Text Clustering with Feature Selection by Using Statistical Data",IEEE Trans,pp. 1-11,2007,DOI 10.1109\TKDE.2007

[2] K. Hauizhong, G. Georges, "Similarity model and term association for document categorization",IEEE , ISBN 1529-4188, pp. 3-7,2002.

[3] T. Vincent, S. Ardi and S. Rudy,"A Comparative Study of Centroid-Based , Neighborhood-Based and Statistical Approaches for Effective Document Categorization",IEEE, pp. 235-238, 2002.

[4] L. Yongxin, L. Zhijng,"An Improved Hierarchical K-Means Algorithm for Web Document Clustering",IEEE,pp. 606-610,2008,DOI 10.1109/ICCSIT.2008.

[5] C. Hamounda K. and C. David ,W. "Feature Subset Selection for Arabic Document Categorization using BPSO-KNN", IEEE,ISBN 978-1-4577-1124-4,pp. 546-551,2011.

[6] R.M. A. Jaya, P. Latha, "Clustering Analysis by Improved Particle Swarm Optimization and K- Means Algorithm",IEEE,2012.

[7] S. Wei and X. Qian, "An Improved Transfer Learning Algorithm for Document Categorization Based on Data Sets Reconstruction" , IEEE, ISBN 978-1-4673-1398-8,pp. 575-578,2012.

[8] W. Guohua,L. Hairong,F. Erashuai, W. Liuyang,"An Improved K-means Algorithm for Document Clustering",ICSMA(IEEE),ISBN 978-1-4673-9166-5/15, pp. 65-69,2015,DOI 10.1109/CSMA.2015.20.

[9] S. Lokesh,Biju and M. R.,"An improved K-means algorithm using modified cosine distance measure for document clustering using Mahout with Hadoop",IEEE,ISBN 1479964999,pp. 1-5,2014.

[10] L. Yongxin, L. Zhijng,"An Improved Hierarchical K-Means Algorithm for Web Document Clustering",IEEE,pp. 606-610,2008,DOI 10.1109/ICCSIT.2008.

[11] O.H. ODukoya,G. A. Aderounmu, E.R. Adagunodo,"An Improved Data Clustering Algorithm for Mining Web Documents",IEEE,ISSN 978-1-4244-5392-4,2010

