NG20 and Reuters -21 578 different categories. From the results, we can see that the SHDC better benefits than other variants of k-means, k-means and random initial partition of the poorest, which will lead to sensitive k-means initial partition. The k-means to provide the best conditions, can achieve better results than those of random initial partition.

In order to study the effect of k, we use different K experimental NG20 and RTR -21,578. If in the space limitations, only the main point is plotted in Figure 2. NMI, the best result is close to the true number of classes, respectively, is similar to the results of the two sets of data.

Figure 1 irradiation, the average performance for the running time of 10 iterations. We show only the main points on two data sets k.

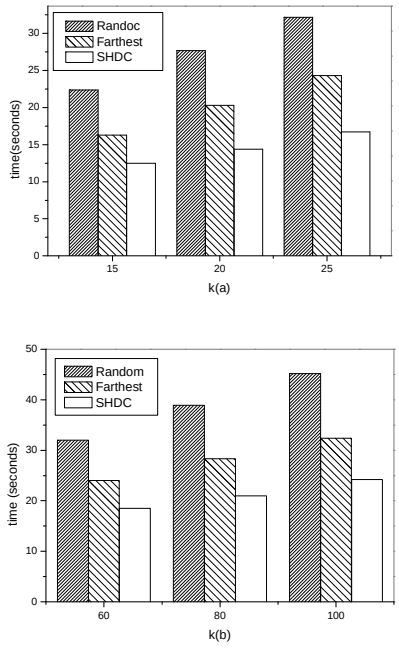

Figure 1. Comparing the average time performance on NG20(a) and Reuters-21578(b)

We can see that SHDC can accommodate a variety of clusters. Its running time a little more, when faced with the rise of the cluster. This characteristic facilitates the handling of the large capacity scalability corpus. In addition, the heuristic initial cluster, SHDC consume less running time than the other two algorithms. K-means furthest better than other algorithms from random initial partition. This result can be explained as follows: the more optimal initial partition fewer files are re-assigned to the cluster, wherein the retrenches of running time to re-calculate the similarities between the device and file cluster means and unchanged.

Besides, we have studied the effect of min_length for clustering effectiveness and efficiency. We found that the clustering quality is better when min_length is about 10. The quality would decline when the value of min_length beyond or below 10. From current first step analysis, the reason is the impact of min_length over parTFI. Furthermore, this impact would also interface the running time of SHDC.

REFERENCES

[1] Ling Zhuang, Honghua Dai.: A Maximal Frequent Itemset Approach for Web Document Clustering. Proceedings of the Fourth International Conference on Computer and Information Technology (CIT'04)

[2] Wang Yongheng, Jia Yan and Yang Shuqiang.: Parallel Mining of Top-K Frequent Items in Very Large Text Database. WAIM (2005)

[3] Oren Zamir, Oren Etzioni.: Web document clustering: A feasibility demonstration. In Melbourne, Australia Proceedings of SIGIR' 98.

[4] Oren Zamir, Oren Etzioni 99 Grouper.: A Dynamic Clustering Interface to Web SearchResults. In Proceedings of the 8th WWW Conference, Toronto Canada, 1999

[5] Jiawei Han, Micheline Kamber.: Data Mining: Concepts and Techniques, Second Edition.: Morgan Kaufmann Press, 2006..

[6] Andreas Hotho, Andreas Nürnberger, and Gerhard Paaß.: A Brief Survey of Text Mining. LDV FORUM – Vol. 20 – 2005, p.19-62

[7] Beil F., Ester M., Xu X.: Frequent Term-Based Text Clustering, Proc. 8th Int. Conf. on Knowledge Discovery and Data Mining (KDD '2002), Edmonton, Alberta, Canada, 2002.

[8] Benjamin C. M. Fung, Ke Wang, Martin Ester.: Hierarchical Document Clustering using Frequent Itemsets. SDM 2003.

[9] J.B. MacQueen. Some methods for classification and analysis of multivariate observations. In Proceeding of the 5th Berkeley symposium in mathematics and probability, 1967.

[10] Steinbach M., Karypis G., Kumar V.: A Comparison of Document Clustering Techniques, Proc. TextMining Workshop, KDD 2000, 2000.

[11] Oracle Text technical white paper. Oracle Corporation, January 2004.

[12] Andreas Hotho, Alexander Maedche, Steffen Staab.: Ontology-based Text Document Clustering. KI 16(4) (2002) 48-54

[13] Strehl, A., & Ghosh, J. (2002). Cluster ensembles - a knowledge reuse framework for combining partitions. Journal of Machine Learning Research, 3, 583–617.