The reason is that the extended vector space model richer semantic information than traditional TF-IDF model, and to strengthen the vector representation of the text content, the real theme. The carrier used, adding, in which only the relevant terms to calculate the similarity. This method brilliant avoiding the warp yarns from the sparse keywords, measuring the similarity of the text, thereby achieving a better efficiency than the original TF-IDF model method.

Experimental study of K NG20 Reuters-21578 on the effect of the SC, the results shown in Figure 1. SC Clusters divisible. WR-KMEANS point of the original class included in the data set, you can get a clear partition corpus can induce SC Figure 1 (NG200.67 when K = 20, Reuters -215780.69, K = 80). This is a reasonable result.

Extended vector space model, combined with the long-term mutual information, more knowledge of the language than the TF-IDF model, we can conclude that: From the above results. Context information needed to distinguish classes of documents.

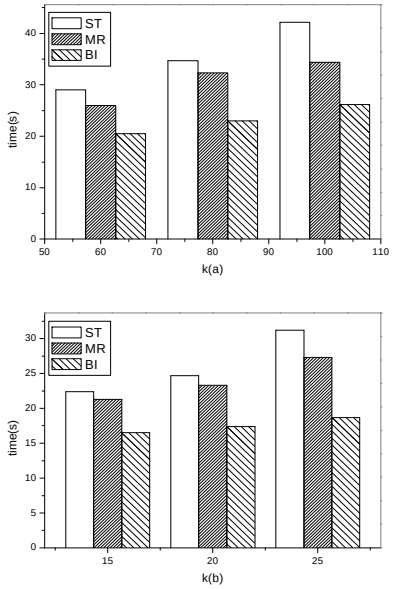

Figure 3. Comparing the best SC results on NG20(a) and on Reuters-21578(b)

Figure 3 illustrates the running time of the two data sets three algorithms. We can see, WR-KMEANS relative need more time than bisecting k-means, but a little faster than the standard k-means. The reason is, WR-kmeans is optimized in the pretreatment and text only, said not included in the clustering process. WR-KMEANS better benefits than the other two algorithms, efficiency are not much advantage.

V. CONCLUSION AND FUTURE WORKS

In this article, we focus on instant messaging cluster, and put forward the average WR-K method to solve sparse keywords, followed by the. WR-KMEANS automatic synthesis of instant messaging conversations, it has more keywords and more complete context information than simple message, and expand the traditional TF-IDF model dialogue assistance HowNet words. Experimental evidence indicates, WR-kmeans significantly outperformed traditional TF-IDF model based on the other two methods.

We plan to WR-K mean clustering, the initial partition optimization to improve speed. In addition, we have to analyze the IM networks, social network analysis, in future work.

REFERENCES

[1] J. Resig and A. Teredesai.: A framework for mining instant messaging services. In Proceedings of the 2004 SIAM Lake Buena Vista, Florida (2004)

[2] J. MacQueen.: Some methods for classification and analysis of multivariate observations. In proceedings of 5th berkeley SMSP, pp. (1967) 281–297

[3] Yi Guan, etc.: Quantifying Semantic Similarity of Chinese Words from Hownet. IEEE Proceedings of ICMLC02, Volumn 1. Beijing (2002) 234-239

[4] Sack, etc.: A Content-Based Usenet Newsgroup Browser. Proceedings of the international conference on Intelligent user interfaces. 233 –240. New Orleans, Louisianna, 2000.

[5] Faisal M. Khan, Todd A. Fisher, Lori Shuler, Tianhao Wu, and William M. Pottenger.: Mining chat-room conversations for social and semantic interactions (2002)

[6] Hearst, Marti A. TextTiling: A Quantitative Approach to Discourse Segmentation, Technical Report UCB: S2K-93-24, 1993

[7] Scott Deerwester, etc. Indexing by latent semantic analysis. Journal of the American Society of Information Science, vol. 41, issue 6, 1990, 391-407.

[8] Ding, C. H. Q. A probabilistic model for dimensionality reduction in information retrieval and filtering. In Proc. of the 1st SIAM, Raleigh, NC, 2000.

[9] Ikehara, S., etc. Vector space model based on semantic attributes of words. In Proc. of the Pacific Association for Computational Linguistics (PACLING), Kitakyushu, Japan, 2001.

[10] A.Daemi, etc. From Ontologies to Trust through Entropy, Proceedings of the International Conference on Advances in Intelligent System, Luxembourg (2004)

[11] Andreas Hotho, etc.: Ontology-based Text Document Clustering. KI 16(4) (2002) 48-54

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

[13] M. F. Porter. An algorithm for suffix stripping. Program, 14(3):130–137, 1980.