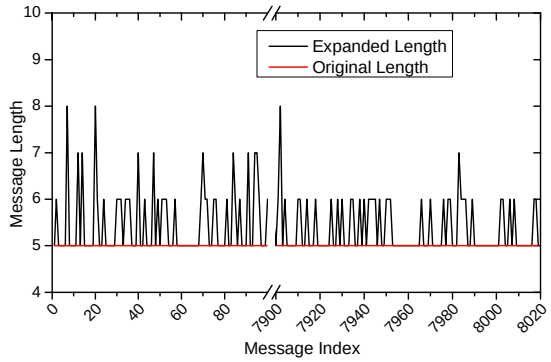

Figure 4. Message Length after Expansion.

Because there is no standard data set to assess the clustering results of our human study. If an instant messaging with others in the same cluster is consistent, then we believe that this result is correct, otherwise false. Figure 5 (a) shows the 15 iterations for different k-means clustering performance and (b) the production of 15 clusters for different iterations.

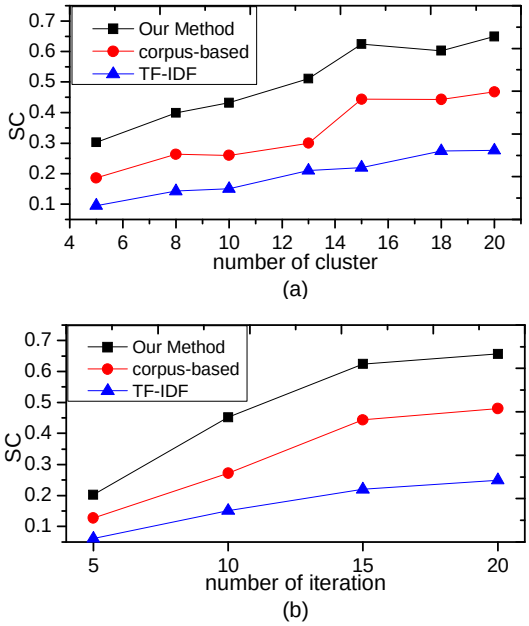

Figure 5. The SC of clustering results

SC said divisible cluster. From Figure 5, we can find that our model can be relatively clear partition corpus, which is illuminated by SC in FIG 5 (a) (SC = 0.65, K = 20). This is a reasonable result. We can find instant messaging with a poor performance, due to lack of sufficient context, the traditional TF-IDF method. In addition, TF-IDF method cannot benefit from more iteration, because the line is relatively flat, as in Figure 5 (b) below. Corpus-based approach has a better performance than the TF-IDF.

V. CONCLUSION AND FUTRUE WORKS

In this paper, we propose a new mixed-mode similarity measure instant messaging clustering found that the views of the public. The model combines semantic knowledge corpus callosum the vocabulary classification and statistical information, and integrated graphics structural relationship information. By logical extension, in the drawings, the relationship is added to the instant message of the carrier. After that, the message can be clustered based on such an expansion, said. Experimental results show that our model, SC, about 10%, and nearly 40% SC than corpus-based approach than the traditional TF-IDF method of instant messaging in the cluster effect.

REFERENCES

[1] Yongheng Wang, Yan Jia, and Shuqiang Yang. Ontology-based Short Conversations Clustering in Very Large Text Database. In The 2nd VLDB Workshop on Ontologies-based techniques for DataBases and Information Systems (ODBIS'06)(Seoul, Korea, September 11, 2006). Springer Press, Berlin, LNCS 4256, 2006, 83-93.

[2] Le Wang, Yan Jia, and Weihong Han. Instant Message Clustering Based on Extended Vector Space Model. In Proceeding of The 2nd International Symposium on Intelligence Computation and Applications (ISICA'07)(Wuhan, China, Sept. 21-23, 2007). Springer Press, Berlin, LNCS 4683, 2007, 435-443.

[3] Khaled M. Hammouda, and Mohamed S. Kamel. Document Similarity Using a Phrase Indexing Graph Model. Springer Knowledge and Information Systems, 6, 6 (Dec, 2004), 710-727.

[4] Christopher D. Manning, and Hinrich Schuetze. Foundations of Statistical Natural Language Processing. The MIT Press, Cambridge, MA, 1999.

[5] Sahami, M., and Heilman, T. A Web-based Kernel Function for Measuring the Similarity of Short Text Snippets. In Proceedings of the 15th International Conference on World Wide Web (WWW'06) (Edinburgh, Scotland, May 23 - 26, 2006). ACM Press, New York, NY, 2006, 377-386.

[6] Metzler, D., Dumais, S., and Meek, C. Similarity measures for short segments of text. In Proceeding of 29th European Conference on Information Retrieval (ECIR'07)(Rome, April 2-5, 2007). Springer Press, Berlin, LNCS 4425, 2007, 16–27.

[7] Xiaojun Wan and Yuxin Peng. A Measure Based on Optimal Matching in Graph Theory for Document Similarity. In Proceeding of Asia Information Retrieval Symposium (AIRS'04)(Beijing, China, October 14-16, 2004). Springer Press, Berlin, LNCS 3411, 227–238

[8] Ana G. Maguitman, Filippo Menczer. Algorithmic detection of semantic similarity. In Proceedings of the 14th international conference on World Wide Web (WWW'05) (Chiba, Japan May 10 - 14, 2005). ACM Press, New York, NY, 2005, 107–116.

[9] Mihalcea R, Corley C, and Strapparava C. Corpus-Based and knowledge-based measures of text semantic similarity. In Proceeding of the 21st American Association for Artificial Intelligence (AAAI'06)(Boston, USA, July 16–20, 2006). AAAI Press, Menlo Park, California, 2006.

[10] T. H. Cormen, et al. Introduction to Algorithms (Second Edition). The MIT Press, Cambridge, MA, 2001..

[11] J. MacQueen. Some methods for classification and analysis of multivariate observations. In Proceeding of the 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, 1967, 281-297.