Document Clustering using Reweighted Term

Sun Park[1], Ronesh Asnil Sharma[2], Jin Gwan Park[3], Min A Jeong[1], Jun Woo Shin[4], Seong Ro Lee[1]

[1,2,3] Mokpo National University, South Korea, [4]National IT Industy Promotion Agency, South Korea
[1]{sunpark, majung, srlee}@mokpo.ac.kr, [2]sharmaronesh@yahoo.com, [3]chrispj@nate.com, [4]sjw@nipa.kr

Abstract. This paper proposes a new document clustering method using the reweighted term based on semantic features for enhancing document clustering. The proposed method uses document samples of cluster by user to reduce the semantic gap between the user's requirement and clustering results by machine.

Keywords: Document clustering, reweighting term, sematic feature.

1 Introduction

Traditional document clustering methods are based on bag of words (BOW) model, which represents documents with features such as weighted term frequencies (i.e., vector model). However, these methods ignore semantic relationship between the terms within a document set. Recently, to overcome the problems of the vector model-based document clustering, knowledge based approaches are applied [1].

Knowledge based approaches can be either internal knowledge based or external knowledge based document clustering. Internal knowledge-based document clustering uses the inherent structure of the document set by means of a factorization technique [1, 2, 3, 4, 5, 6]. External knowledge-based document clustering exploits the constructed term ontology from external knowledge database with regard to ontology as WordNet and Wikipedia [1, 2, 3].

In order to enhance the internal knowledge-based approaches, this paper proposes a document clustering method that uses the reweighted terms by semantic features of NMF and the selected sample document of cluster by user.

2 Proposed Method

The method of reweighting term is described as follows. First, let the number of cluster be set (it also can use to set the number of semantic feature r with connection to NMF), and then the sample documents regarding the clusters are selected by user. Second, preprocessing is performed. Finally, the reweighting term g_a^{new} is calculated by using equation (1). However, we cannot directly calculate a new weight of a'th term. In order to solve this limitation, it calculates the average weight of *a*'th row vector with regard to semantic features of document set by NMF a corresponding *a*'th term of the selected sample document .

$$g_a^{new} = g_a^{old} + \Delta g_a \qquad (1)$$

Where g_a^{new} is a new weight of *a*'th term, g_a^{old} is a weight of a'th term (i.e., initial value is 1.), Δg_a is variance in average weight of *a*'th row vector.

This section presents the clustering document using *k*means clustering method and reweighting terms of document set. The reweighting terms are calculated by using equation (2).

$$\tilde{A} = GA \qquad (2)$$

Where $\tilde{A}$ is reweighting term document frequency matrix, *G* is weight matrix, *A* is term document frequency matrix with relation to document set.

The *k*mean algorithm takes the input parameter, *k*, and partitions a set of n objects into *k* clusters so that the resulting intra-cluster similarity is high but inter-cluster similarity is low [7]. In this paper, we use cosine similarity for cluster distance measure with association to *k*means.

3 Experiments and Evaluation

This paper uses 20 Newsgroups data set for performance evaluation [8]. To evaluate the proposed method, mixed documents were randomly chosen from the 20 Newsgroups documents. Normalized mutual information metric used to measure the document clustering performance [1-7].

In this paper, the eight different document clustering methods are implemented. The RT, KM [13, 14], NMF [4], CF [5], ASI [6] methods are document clustering methods based on internal knowledge. The TR denotes the proposed method described within this paper. The average normalized metric of RT is 20.8% higher than that of KM, 17.58% higher than that of NMF, 14.48% higher than that of CF, 12.88% higher than that of ASI.

4 Conclusion

This paper presents a document clustering method using the reweighted term based on semantic features for enhancing document clustering. The proposed method uses document samples of cluster by user to reduce the semantic gap between the user's requirement and clustering results by machine. The method can enhance the document clustering because it uses the reweighted term which can well represent an inherent structure of document set relevant to a user's requirement.

Acknowledgement

This work was supported by Priority Research Centers Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2009-0093828). “This research was supported by the MKE (The Ministry of Knowledge Economy), Korea, under the ITRC (Information Technology Research Center) support program supervised by the NIPA (National IT Industry Promotion Agency)”(NIPA-2012-H0301-12-2005).

References

1. Hu, X. Zhang, X. Lu, C. Park, E. K. Zhou, X.: Exploiting Wikipedia as External Knowledge for Document Clustering. In proceeding of the 15th ACM SIGKDD Conference On Knowledge Discovery and Data Mining (KDD'09), Paris, France, 389-396 (2009)
2. Hu, T. Xiong, H. Zhou, W. Sung, S. Luo, Y. H.: Hypergraph Partitioning for Document Clustering: A Unified Clique Perspective. In proceeding of the ACM SIGIR conference on research and development in information retrieval (SIGIR’08), Singapore, 871-872 (2008)
3. Park, S. Kim, K. J.: Document Clustering using Non-negative Matrix Factorization and Fuzzy Relationship. The Journal of Korea Navigation Institute, vol. 14(2), 239-246 (2010)
4. Xu, W. Liu, X. and Gong, Y.: Document clustering based on non-negative matrix factorization. In proceeding of the ACM SIGIR conference on research and development in information retrieval (SIGIR’03), Toronto, Canada, (2003)
5. Xu, W. Gong, Y.: Document Clustering by Concept Factorization. In proceeding of the ACM SIGIR conference on research and development in information retrieval (SIGIR’04), UK, 202-209 (2004)
6. Li, T. Ma, S. Ogihara, M.: Document Clustering via Adaptive Subspace Iteration. In proceeding of the ACM SIGIR conference on research and development in information retrieval (SIGIR’04), UK, 218-225 (2004)
7. Frankes, W. B. Ricardo, B. Y.: Information Retrieval, Data Structure & Algorithms. Prentice-Hall, (1992)
8. The 20 newsgroups data set. http://people.csail.mit.edu/jrennie/20Newsgroups/, (2012)