Text Clustering using Semantic Terms

Sun Park[1] and Seong Ro Lee[2]

*[1]Institute Research of Information Science and Engineering,
Mokpo National University, South Korea
[2]Department of Information and Electronic, Mokpo Naitional University,
South Korea
[1,2]{sunpark,srlee}@mokpo.ac.kr*

Abstract

In traditional text clustering, documents appear terms frequency without considering the semantic information of each document (i.e., vector model). The property of vector model may be incorrectly classified documents into different clusters when documents of same cluster lack the shared terms. Recently, to overcome this problem uses knowledge based approaches. However, these approaches have an influence of structure of document set and a cost problem of constructing ontology. In this paper, we propose a text clustering method using semantic terms for clustering label and term weights. The semantic terms of clustering label can well express the internal structure of document clusters using non-negative matrix factorization (NMF). It can also improve the quality of text clustering which uses the term weights by WordNet. The experimental results demonstrate that the proposed method achieves better performance than other text clustering methods.

Keywords: *document clustering, NMF, semantic terms, term weight, WordNet*

1 Introduction

Traditional text 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. The clustering performance of the BOW model is dependent on a distance measure of document pairs. But the distance measure cannot reflect the real distance between two documents because the documents are composed of the high dimension terms with relation to the complicated document topics. In addition, the results of clustering documents are influenced by the properties of documents or the desired cluster forms by user [1]. Recently, to overcome the problems of the vector model-based text clustering, internal and external knowledge approaches are applied.

Internal knowledge-based text clustering approaches use the internal structure of the document set by means of a factorization technique. The factorization techniques for document clustering including non-negative matrix factorization [2, 3, 4], adaptive subspace iteration [5], and clustering with local and global regularization [6] have been proposed, which can accurately identify the topics of document set from their semantic features. However, the successful construction of a semantic features from the original document set remains limited regarding the organization of very different documents or the composition of similar documents [7]. External knowledge-based text clustering uses external knowledge database with respect to ontology. Recently, the ontology approaches are proposed such as term mutual information with conceptual knowledge by concept weighting from domain ontology [8], and fuzzy associations with WordNet [9], etc. These techniques can improve the BOW term representation of text clustering. However, it is often difficult to locate a comprehensive ontology that covers all concepts mentioned in the documents collection, which is a cause of loss of information [1].

To solve the disadvantages of the knowledge-based approaches, this paper proposes a text clustering method that uses sematic terms by NMF and WordNet. The proposed method combines the advantages of the internal and external knowledge-based methods.

2. Non-negative Matrix Factorization

This section reviews non-negative matrix factorization (NMF) theory with algorithm. In this paper, we define the matrix notation as follows: Let X_{*j} be j'th column vector of matrix X, X_{i*} be i'th row vector, and X_{ij} be the element of i'th row and j'th column. NMF is to decompose a given $m\times n$ matrix A into a non-negative semantic feature matrix W and a non-negative semantic variable matrix H as shown in Equation (1) [7].

$$A \approx WH \tag{1}$$

where W is a $m \times r$ non-negative matrix and H is a $r \times n$ non-negative matrix. Usually r is chosen to be smaller than m or n, so that the total sizes of W and H are smaller than that of the original matrix A.

The objective function is used minimizing the Euclidean distance between each column of A and its' approximation $\widetilde{A} = WH$, which was proposed by Lee and Seung [7]. As an objective function, the Frobenius norm is used:

$$\Theta_E(W,H) \equiv \sum_{i=1}^{m}\sum_{j=1}^{n}\left(A_{ij} - \sum_{l=1}^{r} W_{il}H_{lj} \right)^2 \tag{2}$$

Updating W and H is kept until $\Theta_E(W,H)$ converges under the predefined threshold or exceeds the number of repetition. The update rules are as follows:

$$H_{\alpha\mu} \leftarrow H_{\alpha\mu}\frac{(W^T A)_{\alpha\mu}}{(W^T WH)_{\alpha\mu}}, \qquad W_{i\alpha} \leftarrow W_{i\alpha}\frac{(AH^T)_{i\alpha}}{(WHH^T)_{i\alpha}} \tag{3}$$

3. Proposed Text Clustering Method

This paper proposes a text clustering method using NMF and WordNet. The proposed method consists of three phases: preprocessing, extracting semantic terms, and clustering text document. In the subsections below, each phase is explained in full.

3.1. Preprocessing

In preprocessing phase, Rijsbergen's stop words list is used to remove all stop words, and word stemming is removed using Porter's stemming algorithm [10]. Then, the term document frequency matrix *A* is constructed from the document set.

3.2. Extracting Semantic Terms

This section extracts semantic terms to connection with the properties of the document clusters. Extracting semantic terms phase consists of constructing cluster labels and computing term weights. In constructing cluster labels phase, the method is described as follows. First, term document frequency matrix *A* is constructed by performing the preprocessing phase. Second, let the number of cluster be set, and then NMF is performed on the matrix *A* to decompose the two sematic feature matrices *W* and *H*. Finally, matrix *W* and Equation (4) are used to extract terms of clustering labels.

$$L^k \leftarrow A_{ij}\ if\ k = \arg\max_{1 \le j \le r} W_{ij} \tag{4}$$

Where L^k is the term set of *k*'th clustering label of cluster, A_{ij} is the term corresponding to the semantic feature of *i*'th row and the *j*'th column in the matrix *W*. In computing term weights phase, the weights are calculated by TMI (term mutual information) based on the synonyms of WordNet. WordNet is a lexical database for the English language where words (i.e., terms) are grouped in synsets consisting of synonyms and thus representing a specific meaning of a given term [11]. Clustering label terms may be restricted from properties of document cluster and document composition. To solve this problem, this paper uses term weight of documents by using the TMI on synonyms of WordNet. Term weights of the document are calculated by jing's TMI as in Equation (5) [12]. The Jing's TMI is as follows. In Equation (5), δ_{il} is to indicate semantic information between two terms. If term A_{lj} appears in the synonyms of A_{ij} by means of WordNet, δ_{il} will be treated in a same level for different A_{ij} and A_{lj}, otherwise, δ_{il} will be set zero.

$$\tilde{A}_{ij} = A_{ij} + \sum_{\substack{l=1 \\ i \ne l}}^{m} \delta_{il} A_{lj} \tag{5}$$

3.4. Clustering text document

This section presents the clustering text documents using cosine similarity between clustering labels and term weights of documents. The proposed method is described as follows. First, the cosine similarity of Equation (6) between clustering labels and term weights is calculated. And then a document having a highest similarity value with respect to the clustering label is clustered into cluster label in connection with the document clusters [10].

$$sim(A_{*j}, \tilde{A}_{*j}) = \frac{A_{*j} \cdot \tilde{A}_{*j}}{|A_{*j}| \times |\tilde{A}_{*j}|} = \frac{\sum_{i=1}^{m} A_{ij} \times \tilde{A}_{ij}}{\sqrt{\sum_{i=1}^{m} A_{ij}^2} \times \sqrt{\sum_{i=1}^{m} \tilde{A}_{ij}^2}} \qquad (6)$$

Where A_{*j} denotes the terms vector of *j*th clustering label, $\tilde{A}_{*j}$ denotes the term weights vector of *j*th clustering label, and *m* denotes the number of terms.

4. Experiments

This paper uses 20 Newsgroups corpus for performance evaluation [13]. Normalized mutual information metric used to measure the text clustering performance [2, 3, 4, 5, 6, 7]. In this paper, the seven different document clustering methods are implemented. The NMF, ASI, CLGR, RNMF, and FPCA methods are document clustering methods based on internal knowledge. The FAWDN and STNW methods are clustering methods based on combining the internal and external knowledge. STNW denotes the proposed method described within this paper. FAWDN denotes the previously proposed method using the WordNet and fuzzy theory [9]. FPCA is the previously proposed method using PCA (principal component analysis) and fuzzy relationship [4], and RNMF is the method proposed previously using NMF and cluster refinement [3]. NMF denotes Xu's method using non-negative matrix factorization [2]. ASI is Li's method using adaptive subspace iteration [5]. Lastly, CLGR denotes Wang's method using local and global regularization [6]. The average normalized metric of STNW is 12.1% higher than that of NMF, 10.12% higher than that of ASI, 6.22% higher than that of CLGR, 4.24% higher than that of RNMF, 2.37% higher than that of FPCA, and 2.18% higher than that of FAWDN.

5. Conclusion

This paper proposes the text clustering method using semantic terms with respect to clustering label and term weights. The proposed method uses the semantic terms by internal knowledge of NMF to extract the clustering labels, which are well represented within the important clustering labels of the text documents. To solve the limitation of the clustering labels with respect to be influenced by internal structure of documents, the method uses TMI (term mutual information) to calculate term weights of documents based on external knowledge of WordNet.

Acknowledgements

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 (2010-0028295). "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-2011-C1090-1121-0007).

References

[1] J. Hu, L. Fang, Y. Cao, H. J. Zeng, H. Li, Q. Yang and Z. Chen, "Enhancing Text Clustering by Leveraging Wikipedia Semantics", Proceedings of the ACM SIGIR conference on research and development in information retrieval (SIGIR'08), pp.179-186, **(2008)** July 20-24; Singapore.

[2] W. Xu, X. Liu and Y. Gon, "Document Clustering Based On Non-negative Matrix Factorization", Proceedings of the ACM SIGIR conference on research and development in information retrieval (SIGIR'03), pp.267-274, **(2003)** July 28-August 1; Toronto Canada.

[3] S. Park, D. U. An, B. R. Cha and C. W. Kim, "Document Clustering with Cluster Refinement and Non-negative Matrix Factorization", Proceeding of the 16th International Conference on Neural Information Processing (ICONIP'09), pp.281-288, **(2009)** December 1-5; Bangkok, Thailand.

[4] S. Park and K. J. Kim, "Document Clustering using Non-negative Matrix Factorization and Fuzzy Relationship", The Journal of Korea Navigation Institute, Vol. 14(2), pp. 239-246, **(2010)**.

[5] T. Li, S. Ma and M. Ogihara, "Document Clustering via Adaptive Subspace Iteration", Proceedings of the ACM SIGIR conference on research and development in information retrieval (SIGIR'04), pp. 218-225, UK, **(2004)** July 25-39; The University of Sheffield, UK.

[6] F. Wang and C. Zhang, "Regularized Clustering for Documents", Proceedings of the ACM SIGIR conference on research and development in information retrieval (SIGIR'07), pp. 95-102, **(2007)** July 23-27; Amsterdam.

[7] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization", Nature, 401, pp. 788-791, **(1999)**.

[8] H. H. Tar and T. T. S. Nyaunt, "Ontology-based Concept Weighting for Text Documents", World Academy of Science, Engineering and Technology 81, pp. 249-253, **(2011)**.

[9] S. Park and S. R. Lee, "Enhancing Document Clustering Using Condensing Cluster Terms and Fuzzy Association", Journal of IEICE TRANS, Information and System. Vol. E94-D, No.6, pp.1227-1234, June, **(2011)**.

[10] W. B. Frankes and B. Y. Ricardo, "Information Retrieval: Data Structure & Algorithms", Prentice-Hall, **(1992)**.

[11] G. Miller, "WordNet: A lexical databassed for English", CACM, vol. 38(11), pp.39-41, **(1995)**.

[12] L. Jing, L. Zhou, M. K. Ng and J. Z. Huang, "Ontology-based Distance Measure for Text Clustering", Proceeding of SIAM International conference on Text Data Mining, Bethesda, **(2006)** April 20-22; Maryland.

[13] The 20 newsgroups data set. http://people.csail.mit.edu/jrennie/20Newsgroups/, **(2012)**.

Authors

Sun Park

He is a research professor at Institute Research of Information Science and Engineering, Mokpo National University, South Korea. He received the Ph.D degree in Computer & Information Engineering from Inha University, Korea, in 2007, the M.S. degree in Information & Communication Engineering from Hannam University, Korea, in 2001, and the B.S. degree in Computer Engineering from Jeonju University, Korea, in 1996. Prior to becoming a researcher at Mokpo National University, he has worked as a postdoctoral at Chonbuk National University, and professor in Dept. of Computer Engineering, Honam University, South Korea. His research interests include Data Mining, Information Retrieval, and Information Summarization, Convergence IT and Marine.

Seong Ro Lee

He received th B.S. degree in electronics engineering from Korea University, Seoul, Korea, in 1987, respectively, and the M.S., and Ph.D. degrees in electrical engineering from Korea Advanced Institute of Science and Technology, Daejeon, Korea, 1990 and 1996, respectively. In September 1997, he joined the Division of Electronics Engineering, Mokpo National University, Jeonnam, Korea. His research interests include digital communication system, mobile and satellite communications system, applications of telematics, USN and embedded system. He serves as chairman of detection and estimation committee for the Korea Information and Communications Society.