Advanced Science and Technology Letters
Vol.30 (ICCA 2013), pp.6-9
http://dx.doi.org/10.14257/astl.2013.30.02

Document Clustering using Weighting and Labels based on Inherent Structure of Document

Yong-Il Kim[1], Yoo-Kang Ji[2], Sun Park[3*]

[1]Honam University, South Korea,
[2]Dept. of Information & Communication Engineering, Dongshin Univ., Korea
[3]Mokpo National University, South Korea
[1]yikim@honam.ac.kr, [2]neobacje@gmail.com, [3]sunpark@mokpo.ac.kr

Abstract. In classic document 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 inherent structure of documents on clustering and a cost problem of constructing ontology. This paper proposes a new document clustering method using terms of class label and term weights based on inherent structure of documents by NMF. The experimental results demonstrate that the proposed method achieves better performance than other document clustering methods.

Keywords: document clustering, NMF, semantic features, term weight, WordNet, term mutual information, cosine similarity

1 Introduction

Clustering of class labels can be generated automatically however there are different from labels specified by humans. The automatic class label is much lower quality than a maunal class label. If the specified class labels for clustering are provided with no human intervention, the clustering is more effective [1, 2, 3]. 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. 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 document clustering, internal and external knowledge approaches are applied.

* Corresponding author

ISSN: 2287-1233 ASTL

In order to resolve the limitations of the knowledge-based approaches, this paper proposes a document clustering method that uses terms of class label by semantic features of NMF and term weights by term mutual information (TMI) in connection with WordNet. The proposed method combines the advantages of the internal and external knowledge-based methods. In the proposed method, first, meaningful terms of class label for describing cluster topics of documents are extracted using NMF. The extracted terms well represents the class label of document clusters by means of semantic features (i.e., internal knowledge) having inherent structure of documents. Second, the term weights of documents are calculated using the TMI based on the synonyms of WordNet (i.e., external knowledge) with respect to documents terms. The term weights can easily classify documents into an appropriate class label by extending the coverage of document with respect to class label.

2 Non-negative Matrix Factorization

This section reviews NMF theory. 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*×*n* matrix *A* into a non-negative semantic feature matrix *W* and a non-negative semantic variable matrix *H* as shown in Equation (1) [4].

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

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 $\tilde{A} = WH$, which was proposed by Lee and Seung [10]. As an objective function, the Frobenius norm is used:

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

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 Document Clustering Method

This paper proposes a document clustering method using class label terms by NMF and term weights based on TMI with WordNet. The proposed method consists of three phases: preprocessing, extracting class label terms, and clustering document, as shown in Figure 1. In the subsections below, each phase is explained in full.

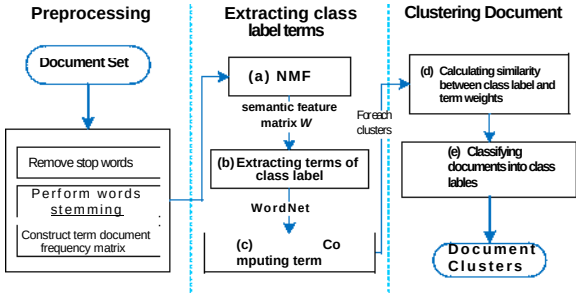

Fig. 1. Document clustering method using class label terms and term weights

In the preprocessing phase of Figure 1, Rijsbergen's stop words list is used to remove all stop words, and word stemming is removed using Porter's stemming algorithm [5]. Then, the term document frequency matrix A is constructed from the document set.

The extracts class label terms in regard to the properties of the document clusters using NMF as in Figure 1(b). The terms of class label that can well explain the topic of document cluster are derived by the semantic features of the NMF. The extracting method is described as follows. First, term document frequency matrix A is constructed by performing the preprocessing phase. Second, let the number of cluster (i.e., the number of semantic feature r) be set, and then NMF is performed on the matrix A to decompose the two sematic feature matrices W and H. Finally, matrix W is used to extract class label terms. The column vector of matrix W corresponds to class label of cluster and the row vector of matrix W refers to terms of document, which the element of matrix W (i.e., the semantic feature value) indicates how much the term reflects the cluster class labels. The equation of extracting terms of class labels is as follows.

The term weights are calculated by TMI (term mutual information) based on the synonyms of WordNet as in Figure 1(c). 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 [6]. Class label terms may be restricted from properties of document cluster and document composition. To resolve 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 [7].

The clustering documents use cosine similarity between class label terms and term weights of documents. The proposed method in Figure 1(d) and 1(e) is described as follows. First, the cosine similarity between class label terms and term weights is

calculated. And then a document having a highest similarity value with respect to the class label is clustered into cluster label in connection with the document clusters [3, 5].

4 Conclusion

This paper proposes the enhancing document clustering method using class label terms and term weights. The proposed method uses the semantic features by internal knowledge of NMF to extract the class label terms, which are well represented within the important class labels of the documents cluster. To resolve the limitation of the semantic features 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. In addition, it uses a similarity between the class label terms and term weights to improve the quality of the document clustering. It was demonstrated that the normalized mutual information is higher than other document clustering methods for 20 Newsgroups test collections using the proposed method.

Acknowledgement. This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(No. 2009-0093828), “This research was supported by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the ITRC (Information Technology Research Center) support program (NIPA-2013-H0301-13-2005) supervised by the NIPA (National IT Industry Promotion Agency)”

References

1. J. Hu, L. Fang, Y. Cao, H. J. Zeng, H. Li, Q. Yang, Z. Chen, "Enhancing Text Clustering by Leveraging Wikipedia Semantics," In proceeding of SIGIR’08, 179-186 Singapore, Jul. (2008)
2. S. Chakrabarti, “mining the web: Discovering Knowledge from Hypertext Data”, Morgan Kaufmann Publishers, (2003)
   B. Y. Ricardo, R. N. Berthier, “Moden Information Retrieval: the concepts and technology behind search Second edition , ACM Press, (2011)
3. D. D. Lee, H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” Nature, 401, pp. 788-791, Oct. (1999)
4. W. B. Frankes, B. Y. Ricardo, “Information Retrieval: Data Structure & Algorithms”, Prentice-Hall, (1992)
5. G. Miller "WordNet: A lexical databassed for english", CACM, vol. 38(11), 39-41, (1995)
6. L. Jing, L. Zhou, M. K. Ng, J. Z. Huang, “Ontology-based Distance Measure for Text Clustering”, In proceeding of SIAM International conference on Text Data Mining, Bethesda, MD. (2006)

