Integrating Feature-Based Document Summarization as Feature Reduction in Document Clustering

Catur Supriyanto, Abu Salam, Abdul Syukur
Dept. of Postgraduate Computer Science
Dian Nuswantoro University
Semarang, Indonesia
caturs@research.dinus.ac.id, masaboe@dinustech.com, abdul.syukur@research.dinus.ac.id

***Abstract*—Document clustering used term-document matrix to represents the collection of document. Number of documents leads high dimensionality of term-document matrix. This paper proposed document summarization as feature reduction to reduce the dimensionality of term-document matrix. We evaluated document summarization in document clustering, compare to feature selection and feature transformation as feature reduction. By comparing the document summarization and other feature reduction, it was found that document summarization improved the accuracy and reduced the time computational of document clustering.**

Keywords-component; feature based document summarization; document clustering; feature reduction

I. INTRODUCTION

Automatic document clustering is the task of grouping text documents into several different clusters. Automatic document clustering used vector space model (VSM) to represent the collection of documents. Unfortunately, the problem of VSM is the high dimensionality of term-document matrix [1]. This problem reduces the performance of automatic document clustering.

High dimensionality of tem-document matrix can be reduced by stopword removal and stemming. Srematthy and Balamurugan [2] stated the using stopword removal and stemming can improve the accuracy of classifier. Another approach to solve the problem of term-document matrix is feature reduction. Basically, feature reduction can be classified into feature selection and feature transformation. Feature selection is selecting the important terms to be used in clustering and feature transformation is the transformation of high dimension matrix into small dimension matrix. The purpose of both is to reduce the high dimensionality of term-document matrix. By implementing feature selection can speed up the computation of document clustering [3].

In order to overcome the high dimensionality of term-document matrix, this paper proposed feature-based automatic document summarization. This approach reduces the dimensionality of VSM by selecting the important sentences of a document before the collection of document is preprocessed.

The outline of this paper is as follows: section 2 describes the related work. Section 3 describes the methodology of research. Section 4 describes the dataset and shows the performance analysis of proposed approach. Section 5 presents the conclusion and future work.

II. RELATED WORK

The aim of feature reduction is to speed up the time processing of a system without decrease the accuracy. Liu et al. [4] have compared document frequency (DF), term contribution (TC), term variance (TV) and term variance quality (TVQ) as unsupervised feature selection on document clustering. No predefined label on document clustering is the reason of using unsupervised feature selection. The experiment shows that the unsupervised feature selection can improve the accuracy of document clustering.

Meng and Lin [5] used feature selection and Latent Semantic Indexing (LSI) via Singular Vale Decomposition (SVD) for text categorization. SVD as the second stage of feature reduction is expected to discover the semantic relationship among text in the collection of document.

Xiao-Yu et al. [6] have implemented automatic document summarization on document classification. Automatic document summarization is used to reduce the dimensionality of vector space model and the complexity of categorization. Experiment is carried out on several news dataset. The result of the experiment shows the advantage of automatic document summarization as feature reduction in document classification.

III. METHODOLOGY

In this section we describe our approach to cluster the documents. Firstly, we used feature-based document summarization to select the important sentences that represent the topic. Next, the selected sentences would be preprocessed by using tokenization, stopword removal and stemming to construct term-document matrix. Feature selection and feature transformation also used to reduce the dimensionality of term-document matrix. Fig. 1 shows our proposed document clustering. Finally, clustering algorithm is performed. Detailed description of each stage is explained in the next subsection.

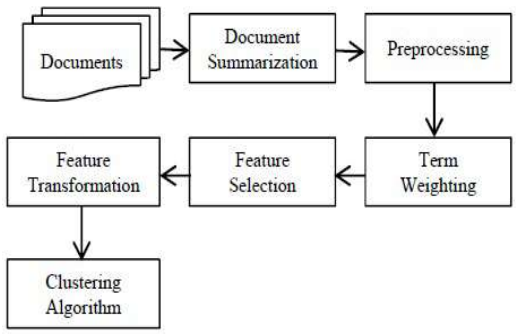

Fig. 1. Clustering Process

A. Feature-Based Document Summarization

Document summarization can be classified into extractive and abstractive summarization [7]. Extractive summarization gives a score for each sentence in a document and selects the important sentences that have high score. Abstractive summarization attempts to understand the concept of each sentence and used natural language processing to change each sentence. We focused on feature-based document summarization.

This paper used eight feature namely title feature, sentence length, term weight, sentence position, sentence to sentence similarity, proper noun, thematic word and numerical data. Detailed explanation of these features shown in Suanmali et al. [8].

B. Preprocessing

Text preprocessing used in this paper are tokenization, stopword removal and stemming. We stemmed the terms using Porter stemming algorithm.

C. Term Weighting

Term weighting is performed to transform text document into numeric. Term Frequency Inverse Document Frequency (TFIDF) is used to measure the important of term in document collection [9]. TFIDF is computed as (1).

$$TFIDF = TF \times IDF \tag{1}$$

$$IDF = \log(\frac{N}{DF}) \tag{2}$$

where TF is the number of term i in a document, N is the number of document, DF is the number of document containing term i and IDF is Invers Document Frequency.

D. Feature Selection

The aim of feature selection is to select the important features or terms that can be used to document clustering. By applying the appropriate feature selection, the using of small features can improve the performance of document clustering, especially time performance. There two types of feature selection, namely supervised and unsupervised feature selection. This paper focused on unsupervised feature selection, since there is no predefined label in document clustering.

The unsupervised feature selection methods include Document Frequency (DF), Term Contribution (TC), Term Variance (TV) and Term Variance Quality (TVQ). We used TC to select our feature, since TC has good performance than others [4]. TC can be computed by using (3).

$$TC_{(t)} = \sum_{i,j \cap i \neq j} f(t_k, D_i) \times f(t_k, D_j) \tag{3}$$

where

$$f(t_k, D_i) = TF_j \times \log(\frac{N}{DF_j}) \tag{4}$$

TF is term frequency of a term j, N is the number of documents, DF is the number of document contains term j.

E. Latent Semantic Indexing

Latent Sematic Indexing (LSI) via Singular Value Decomposition (SVD) is used to transform high dimension term-document matrix into small dimension term-document matrix. SVD attempts to find semantic corresponding between term and document in the collection of document [5]. Let A is the term-document matrix of size $m \times n$ where m is the number of term and n is the number of document. The singular value decomposition of term-document matrix A can be defined as (5).

$$A = U\Sigma V^T \tag{5}$$

where U is term vector, Σ is the diagonal matrix of singular value and V^T is document vector. Next, matrix V^T is used to cluster document collection.

F. Clustering Algorithm

K-means is a clustering algorithm to cluster the document collection. K-means attempts to classify document into k-cluster [10]. K-means will randomly picking K document as centroid. Then compute the similarity distance between each document and centroid. Documents that have higher distance will be placed in the same cluster. New centroid will be determined when all data has been placed in the nearest cluster. The process of determining the centroid and the placement of data within the cluster is repeated until the centroids converge. Table I shows the pseudocode of K-means algorithm [11].

Cosines similarity of two documents is defined using (6).

$$Cosines(d_A, d_B) = \frac{\Sigma w_A \times w_B}{\sqrt{\Sigma(w_A)^2} \times \sqrt{\Sigma(w_B)^2}} \tag{6}$$

Where $w(d_A)$ and $w(d_B)$ is word in document A and document B, w_A and w_B is the $TFIDF$ value of each term in document A and document B.

TABLE I
K-MEANS ALGORITHM

Input	Document collection $D = \{d_1, d_2, d_3, , d_n\}$ Number of cluster // k
Output	k cluster
Process	1. Choosing documents to be k initial centroid (cluster center) randomly 2. Calculating the distance of each document to each centroid using cosines similarity, document that has the closest distance to the centroid is placed in the same cluster with the centroid. 3. Determining the new centroids 4. Return to step 2 if the new centroid is different from the previous centroid.

IV. EXPERIMENTS

This paper used Lucene as java library to implement the methodology. Lucene has provided standard text preprocessing such as tokenization, stopword removal and stemming.

A. Dataset

We have used total 300 documents collected from yahoo news. The collection documents belonging five different classes (sport, economy, politic, entertainment and business). Each class contains 30 documents. These documents clustered by using K-means clustering algorithm. Random selection is used to select the initial centroid of the clustering algorithm. We executed 5 times to obtain the average result of performance.

B. Evaluation Measure

In order to evaluate the quality of document clustering, we employ F-measure as the standard evaluation measurement widely used in document clustering. F-measure is the combination between recall and precision. The recall, precision and F-measure is defined as (7), (8) and (9) respectively.

$$Recall_{(i,j)} = \frac{N_{i,j}}{N_i} \tag{7}$$

$$Precision_{(i,j)} = \frac{N_{i,j}}{N_j} \tag{8}$$

$$F - measure = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{9}$$

where N_{ij} is the number of document of class i in cluster j, N_i is the number of document of class i and N_j is the number of document of cluster j.

C. Experiment Result

Our experiment attempts to evaluate the influence among feature selection (FS), feature transformation (LSI) and document summarization as feature reduction in document clustering. In our experiment, we summarized documents into 30%, 50% and 80%. For feature selection, we used 20% and 40% of a number of terms. The minimum of the number document and the number of term is used to set the parameter of k in LSI via SVD. The influence of each feature reduction is compared to the original k-means. There is no feature reduction in original k-means. Fig. 2 presents the F-measure of different feature reduction in document clustering. As we have shown in this evaluation, the using of feature-based document summarization has improved the accuracy of original K-means when we summarized the documents into 30%, FS 20% and LSI. Otherwise, the accuracy of document clustering was decreased when we summarized documents into 80%, FS 20% and LSI.

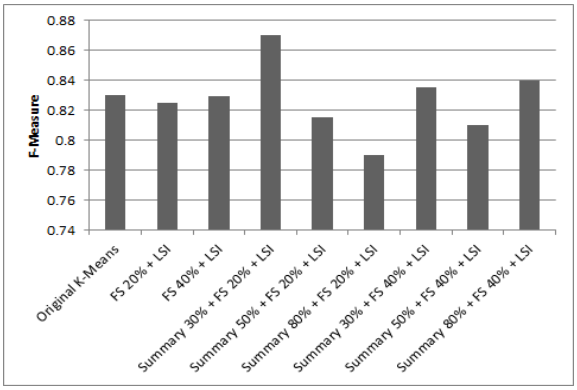

Fig. 2. Accuracy of different feature reduction

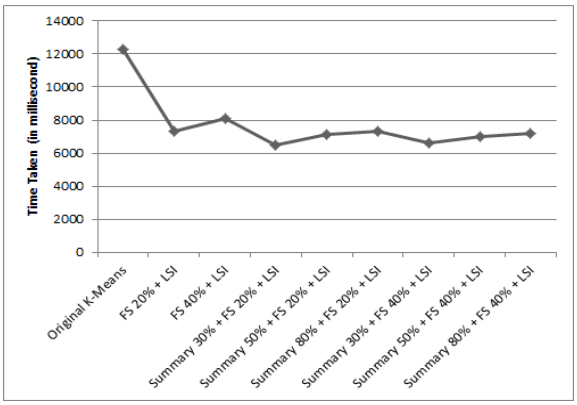

Fig. 3. Time taken of different feature reduction

Result presented in Figure 4 demonstrated that the using of document summarization as feature reduction can reduce the time computation of document clustering. By using 30% document summarization, FS 20% and LSI, the computation of original k-means clustering (12.300 milliseconds) can be decrease into 47% (6.500 milliseconds).

V. CONCLUSION

This paper studied and evaluated the influence of feature-based document clustering as feature reduction in document clustering. The experimental shows the effectiveness of document summarization as feature reduction. Compared to original k-means, the accuracy of document

clustering can be increased and the time computation can be decreased by implementing feature-based document summarization in document clustering. As future work, we plan to evaluate and compare another document summarization approach as feature reduction in document clustering.

REFERENCES

[1] M. Thangamani and P. Thangaraj, "Integrated clustering and feature selection scheme for text documents," *Journal of Computer Science*, vol. 6, no. 5, pp. 536–541, 2010.

[2] J. Sreemathy and P. S. Balamurugan, "An efficient text classification using knn and naive bayesian," *International Journal on Computer Science and Engineering*, vol. 4, no. 3, pp. 392–396, 2012.

[3] M. K.Mugunthadevi, M. S. Punitha, and D. Punithavalli, "Survey on feature selection in document clustering," *International Journal on Computer Science and Engineering*, vol. 3, no. 3, pp. 1240–1244, 2011.

[4] L. Liu, J. Kang, J. Yu, and Z. Wang, "Comparative study on unsupervised feature selection methods for text clustering," in *Proceeding of NLP-KE' 05*, pp. 597–601, 2005.

[5] J. Meng and H. Lin, "A two-stage feature selection method for text categorization," in *Proceeding of Seventh International Conference on Fuzzy Systems and Knowledge Discovery*, pp. 1492–1496, 2010.

[6] J. Xiao-Yu, F. Xiao-Zhong, W. Zhi-Fei, and J. Ke-Liang, "Improving the performance of text categorization using automatic summarization," in *Proceeding of International Conference on Computer Modeling and Simulation*, pp. 347–351, 2009.

[7] V. Gupta and G. S. Lehal, "A survey of text summarization extractive techniques," *Journal Of Emerging Technologies in Web Intelligence*, vol. 2, no. 3, pp. 258–268, 2010.

[8] L. Suanmali, N. Salim, and M. S. Binwahlan, "Fuzzy logic based method for improving text summarization," *International Journal of Computer Science and Information Security*, vol. 2, no. 1, 2009.

[9] S. K. and N. N., "Semantically enhanced document clustering based on pso algorithm," *European Journal of Scientific Research*, vol. 57, no. 3, pp. 485–493, 2011.

[10] M.-U.-S. Shameem and R. Ferdous, "An efficient k-means algorithm integrated with jaccard distance measure for document clustering," in *Proceeding of First Asian Himalayas International Conference on Internet*, 2009.

[11] M. H. Dunham, *Data Mining Introductory and Advanced Concepts*. Pearson Education, 2006.