Document Clustering Using Semantic Relationship Between Target Documents and Related Documents

Minoru Sasaki
Dept. of Computer and Information Sciences
Faculty of Engineering, Ibaraki University
4-12-1, Nakanarusawa, Hitachi, Ibaraki, 316-8511, Japan
Email: msasaki@mx.ibaraki.ac.jp

Hiroyuki Shinnou
Dept. of Computer and Information Sciences
Faculty of Engineering, Ibaraki University
4-12-1, Nakanarusawa, Hitachi, Ibaraki, 316-8511, Japan
Email: shinnou@mx.ibaraki.ac.jp

***Abstract*—Document clustering is one of the most major techniques to group documents automatically. This technique is to divide a given set of documents into a certain number of clusters automatically. In this technique, the first step is 'feature extraction' from documents. As a feature used in the conventional methods, we frequently use a set of words that contains nouns and verbs. Although words are used as features in a generic clustering framework, some previous research proposes the clustering method using the other features based on vector space model such as kernel methods and adaptive sprinkling. However, in previous research of document clustering, the method of appending new feature vectors obtained by using relationship between the existing documents and other documents has not been reported yet. So, we propose a new method for clustering documents using the relationship between the existing documents and other documents to acquire the more useful clusters for users. Our method can expand features of document similarities as semantic relationships by using relevant documents that user is interested in, like semi-supervised clustering. To evaluate the efficiency of this system, we made experiments on clustering newsgroup documents by using our method and by using the dimension reduction method based on the singular value decomposition. As the results of these experiments, we found that (i) it is effective for document clustering to combine the similarity matrix with the original matrix, and (ii) low similarity values cause adverse effect to the clustering performance when we use all the similarity value. Moreover, the proposed method is more effective for the document clustering in comparison with the clustering through the dimensionality reduction.**

***Keywords*-document clustering; semi-supervised clustering; semantic feature expansion;**

I. INTRODUCTION

Document clustering is one of the most major techniques to group documents automatically. This technique is to divide a given set of documents into a certain number of clusters automatically. Each cluster obtained by this technique represents a topic, which is different from the other topics. Thus, it enables a user to have an overall view of the topics contained in the documents so that this technique is often applied to the analysis of web data [13], news articles [12], patents and research papers [1] and so on.

In the document clustering, the first step of preprocessing is term extraction from a set of documents. After the term extraction process, various clustering methods can be applied by utilizing these extracted characteristics of terms. As a feature used in the conventional methods, we frequently use a set of words that contains noun and verb words obtained by using a morphological analyzer from the documents. For the set of these terms, the weight of each word in each document is calculated by using term weighting methods such as term frequency (TF), inverse document frequency (IDF) or log-likelihood ratio to construct a term-document matrix.

Although words are used as features in a generic clustering framework, some previous research proposes the clustering method using the other semantic features based on vector space model. For example, co-clustering methods [5] [8], which is the simultaneous clustering of both words and documents, partitions the documents using word cluster as the feature. The kernel trick [6], which is used to measure the similarity (or distance) of vectors, enables the computation of inner product in a space of possibly very high dimension by some linear combination of words as the feature. Moreover, adaptive sprinkling [4] is effective method to obtain feature vectors by appending some principal component vectors to the term-document matrix by using the singular value decomposition (SVD).

As mentioned above, the co-clustering method and the kernel trick method produce new features obtained by using the relationship between existing words. However, in previous research of document clustering, the method of appending new feature vectors obtained by using relationship between the existing documents and other documents has not been reported yet. The similarity between relevant documents increases with additional feature of other documents so that we consider these features to be efficient for users to obtain useful clustering results. For this reason, we propose a new method for clustering documents using the relationship between the target documents and the other documents. To evaluate the efficiency of this system, we make experiments on clustering newsgroup documents by

 ISBN: 978-1-61208-104-5

using our method. Moreover, as comparative experiment we make an experiment by using the dimension reductio method based on the singular value decomposition.

II. Related Works

We consider this proposed method to be positioned as on of the semi-supervised clustering [3] [9]. Our objective o the method is to improve the efficiency of clustering results by providing related information. In the standard semi-supervised clustering, it is hard to find similar (or dissimilar) document pairs. However, the proposed method use related data as additional features like must-link constraints so it is easy to provide the constraints by comparison with the semi-supervised clustering. In the co-clustering algorithm [5] [8], features are first grouped to perform document clustering. In contrast, the proposed method uses the features that consist of both the original bag of words and the group of words. The kernel method [6] computes the inner product in a space of possibly very high dimension by some linear combination of words. It is similar to the proposed method in the use of additional features. However, the kernel method is difficult to find efficient combination of words from the high dimensional space. Moreover, adaptive sprinkling method [4] appends some principal component vectors of the term-document matrix to obtain the effective features. In contrast, the proposed method appends new feature vectors obtained by using relationship between the existing documents and other documents.

III. Clustering Method Based on Relationship Feature Expansion

A. Motivation

In previous researches, there are some methods that insert additional features obtained by using the relationship between existing words. For example, there is a method that creates combinations of features using kernel methods, and another method that learns similarity metric by information that consists of a set of similar(dissimilar) pair such as semi-supervised clustering. However, sometimes it is hard to construct the additional information of pairs, even though the semi-supervised methods use only a small number of pairwise relations. For this reason, we propose a new method for clustering documents using the relationship between the existing documents and other documents.

The similarity between relevant documents increases with additional feature of other documents. We show an example to explain the reason of this efficiency. In the Figure 1, we consider that there are four documents A, B, C and D in the target document set. The similarity between the document and the other two is nearly equal (e.g., A and B, A and D) so that it is difficult to cluster these data (in the Figure 1 a). Then, we consider another document X in which a user needs relevant information. The similarities between the X and the target document set are calculated and added to their

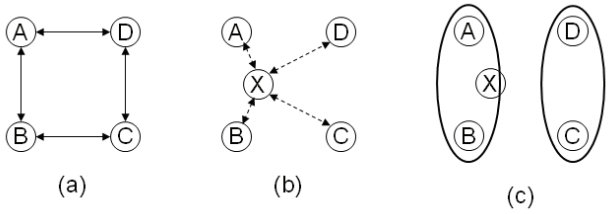

Figure 1. Description of the efficiency by expanding correlation features

document vectors as an additional features (in the Figure 1 b) so that the similarity between the target documents changes with the relationship between the target and the additional documents. Therefore, the document A and B are more similar than the C and D and the C and D are more similar than the A and B so that we are able to acquire the more useful clusters for users from the target document set (in the Figure 1 c).

Our proposed method resembles the kernel methods in the point of dimensionality expansion. The kernel methods calculate the weights of additional features that are combined by the existing features. Then, some of these additional features work well for learning appropriate document vectors. However, Our method is possible to expand features of document similarities by using relevant documents that user is interested in like semi-supervised clustering.

B. Keywords Extraction

As features of the clustering, we extract words from documents from the 20 Newsgroups data set [11]. We first preprocessed all documents in the documents to remove all the stop words using a stop list of common English words such as “a” or “about”. For the obtained words, we calculate the relevancy of each word with respect to each document by using TF-IDF weighting scheme [14]. Therefore, a term-document matrix is generated by normalizing document vectors as shown in the upper part in the Figure 2 .

C. Method of Feature Expansion

For this term-document matrix, we combine a similarity matrix between target documents and other documents to construct an expanded matrix. As a first step, we provide other relevant documents which are different to the target documents. Next, we extract words from the additional documents in the same way as the word extraction from the target documents. Then, we construct a term-document matrix of the additional documents as described in lower part of the Figure 2, where the terms of this matrix are same as the terms appearing in the target documents. To make the correlation matrix, we calculate document similarity matrix based on the cosine similarity measure. Finally, we combine the term-document matrix of the target documents and this similarity matrix by rows to construct the expanded matrix.

 ISBN: 978-1-61208-104-5

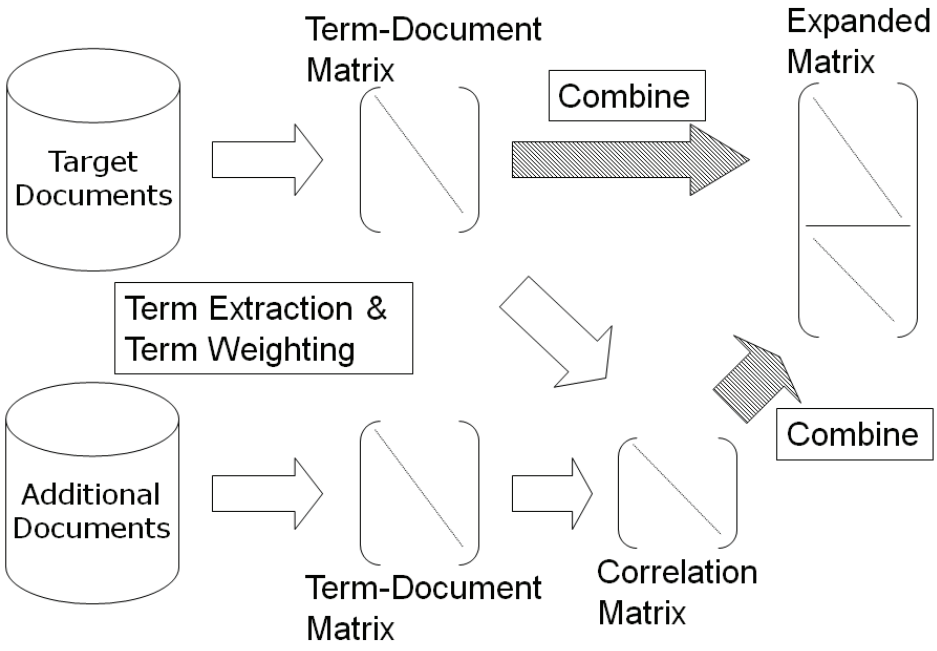

Figure 2. Process of constructing an expanded matrix

However, when we use all values in the similarity matrix, our system tends to have a higher sensitivity to statistical noise due to low values of this matrix. To solve this problem, for all the elements of the document vector, we use the top k similarity values and set all other similarity values to 0. k is defined as the number of the similarity scores. This process enables to reduce the noise in the similarity matrix.

D. Clustering Method

For the expanded matrix generated as mentioned above, we apply a clustering method to group similar documents in the target documents. Let the expanded matrix M, which is obtained above to be the $(m+l)\times n$ matrix, consisting of the number of terms m, the number of additional documents l and the number of the documents n. Then our system groups the documents to generate clusters using the matrix M. In our system, we use CLUTO [10] as the tool for clustering documents. For the purpose of evaluation of the efficiency of document clustering for the expanded matrix, we apply the same clustering method of CLUTO in all our experiments.

IV. Experiments

To evaluate the efficiency of the proposed method, we make some experiments on clustering newsgroup documents. In this experiment, we use the 20 newsgroups as a document set. The 20 newsgroups data set is a collection of approximately 20000 articles from 20 Usenet newsgroups. We extract 50 articles from each 10 newsgroups and construct the subset which consists of the total 500 articles. We also extract 100 articles from each 10 newsgroups and construct another subset which consists of the total 1000 articles. As additional documents, we extract 50 articles from each 20 newsgroups and construct the subset which consists of the total 1000 articles. For the similarity matrix, the number of top-ranked similarity values which are used in each document vector is varied from 100 (set 900 values to 0) to 1000 (use all values) incremented by 100. Then, we compare the performance using the proposed method with that using only the target documents as a baseline method for the evaluation of our method.

A. Evaluation Measures

In this paper, we use entropy and purity to evaluate the clustering quality [2]. The purity is defined as the degree to which each cluster contains documents primary from a single class. The purity of a clustering result is obtained as a weighted sum of the purity of individual clusters as follows,

$$Purity = \sum_{i=1}^{C} \frac{1}{N} \times \max_{j}(n_{ij}), \quad (1)$$

where N is the total number of documents, C is the number of clusters and n_{ij} is the number of documents of the

Table I
CLUSTERING RESULTS 1 WITH 500 DOCUMENTS FOR THE EACH NUMBER OF SIMILARITY VALUES

The number of top-ranked similarity values	Entropy	Purity
None	0.400	0.686
1000	0.415	0.648
900	0.415	0.620
800	0.415	0.620
700	0.415	0.620
600	0.406	0.676
500	0.406	0.676
400	0.415	0.620
300	0.415	0.620
200	0.410	0.672
100	0.410	0.672

Table II
CLUSTERING RESULTS 2 WITH 500 DOCUMENTS FOR THE EACH NUMBER OF SIMILARITY VALUES

The number of top-ranked similarity values	Entropy	Purity
None	0.371	0.698
1000	0.402	0.664
900	0.360	0.690
800	0.306	0.768
700	0.306	0.768
600	0.312	0.760
500	0.312	0.760
400	0.312	0.760
300	0.312	0.760
200	0.312	0.760
100	0.306	0.768

Table III
CLUSTERING RESULTS 1 WITH 1000 DOCUMENTS FOR THE EACH NUMBER OF SIMILARITY VALUES

The number of top-ranked similarity values	Entropy	Purity
None	0.379	0.755
1000	0.378	0.696
900	0.327	0.779
800	0.328	0.774
700	0.363	0.762
600	0.328	0.774
500	0.328	0.774
400	0.328	0.774
300	0.328	0.774
200	0.356	0.731
100	0.327	0.779

Table IV
CLUSTERING RESULTS 2 WITH 1000 DOCUMENTS FOR THE EACH NUMBER OF SIMILARITY VALUES

The number of top-ranked similarity values	Entropy	Purity
None	0.315	0.750
1000	0.300	0.764
900	0.308	0.750
800	0.308	0.750
700	0.295	0.767
600	0.284	0.780
500	0.284	0.780
400	0.295	0.767
300	0.322	0.752
200	0.284	0.780
100	0.293	0.777

category j in the cluster C_i. In general, the larger the purity value are obtained, the clustering algorithm is the better.

The entropy of a clustering result is defined as the weighted sum of cluster entropies as follows,

$$Entropy = -\sum_{i=1}^{C} \frac{n_i}{N} \sum_{j=1}^{K} \frac{n_{ij}}{n_i} \log \frac{n_{ij}}{n_i}, \quad (2)$$

where n_i is the number of documents in the cluster C_i. A good clustering algorithm should have low cluster entropy.

V. EXPERIMENTAL RESULTS

A. Clustering Results with 500 Documents

Table I and Table II show the results of the experiments with two sets of 500 articles, which consist 50 articles from each 10 newsgroups as a mentioned above, respectively. In the Table I, the precision is the approximately same as the result using the original matrix. If the additional documents have little association with the target documents about these contents, the clustering performance is less affected by the similarity matrix. However, in the Table I, the purity score represents a 6.2% increase by the addition of the similarity matrix. When the additional documents are relevant to the original documents, we found that it is effective for document clustering to combine the similarity matrix with the original matrix.

Additionally, the system provides the highest accuracy when we use the top 500-600 similarity values in the Table I and the top 700-800 similarity values in the Table II. We found that low similarity values cause adverse effect to the clustering performance when we use all the similarity value.

B. Clustering Results with 1000 Documents

Table III and Table IV show the results of the experiments with two sets of 1000 articles, which consist 100 articles from each 10 newsgroups as a mentioned above, respectively. Though these results are smaller accuracy than that with the 500 documents, the clustering performance is improved by the addition of similarity matrix. This shows that the addition of the similarity matrix is effective for the clustering performance even when we change the number of documents. When we change the number of top similarity value, we obtain the highest accuracy by using the top 500 values.

C. Comparison with Clustering Through Dimensionality Reduction

To evaluate the efficiency of the proposed method, we make another experiment using the clustering through di-

 ISBN: 978-1-61208-104-5

Table V
CLUSTERING RESULTS FOR EACH REDUCED DIMENSIONS (1000 DOCUMENTS)

dimension	Entropy	Purity
900	0.314	0.730
800	0.309	0.728
700	0.252	0.786
600	0.326	0.693
500	0.264	0.769
400	0.325	0.729
300	0.314	0.730
200	0.310	0.736
100	0.328	0.712
50	0.355	0.675
10	0.389	0.626
5	0.478	0.544

mensionality reduction. We compute the singular value decomposition for the term-document matrix generated from the above 1000 documents [7]. The documents are projected in a lower dimensional space spanned by the leading l left singular vectors to obtain dimension reduced vectors. Then our system groups these vectors to generate clusters by the same clustering algorithm.

Table V shows the results of this experiment with the 1000 documents. In this Table V, the clustering accuracy drops continuously as the number of dimensions grows. The projection transforms a document's vector in n-dimensional word space into a vector in the k-dimensional reduced space. Because the characteristics of words are reduced by this projection, it is difficult to make clear distinction between words.

VI. CONCLUSION AND FUTURE WORKS

In this paper, we proposed a new method for clustering documents using the relationship between the existing documents and other documents. To evaluate the efficiency of this system, we make experiments on clustering newsgroup documents by using our method and by using the dimension reduction method based on the singular value decomposition. As the results of these experiments, we found that it is effective for document clustering to combine the similarity matrix with the original matrix and low similarity values cause adverse effect to the clustering performance when we use all the similarity value. Moreover, the proposed method is more effective for the document clustering in comparison with the clustering through the dimensionality reduction.

Further work would be required to compare the other semi-supervised clustering methods by the many kinds of document data.

REFERENCES

[1] B. Aljaber, N. Stokes, J. Bailey, and J. Pei, "Document clustering of scientific texts using citation contexts," *Information Retrieval*, vol. 13, no. 2, pp. 101–131, 2010.

[2] E. Amigó, J. Gonzalo, J. Artiles, and F. Verdejo, "A comparison of extrinsic clustering evaluation metrics based on formal constraints," *Information Retrieval*, vol. 12, no. 4, pp. 461–486, 2009.

[3] S. Basu, A. Banerjee, and R. J. Mooney, "Semi-supervised clustering by seeding," in *ICML '02: Proceedings of the Nineteenth International Conference on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2002, pp. 27–34.

[4] S. Chakraborti, R. Mukras, R. Lothian, N. Wiratunga, S. Watt, and D. Harper, "Supervised latent semantic indexing using adaptive sprinkling," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI07)*, 2007, pp. 1582–1587.

[5] H. Cho, I. Dhillon, Y. Guan, and S. Sra, "Minimum sum squared residue co-clustering of gene expression data," in *Proceedings of the Fourth SIAM International Conference on Data Mining*, 2004, pp. 114–125.

[6] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[7] S. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41(6), pp. 391–407, 1990.

[8] I. S. Dhillon, S. Mallela, and D. S. Modha, "Information-theoretic co-clustering," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD03)*. New York, NY, USA: ACM, 2003, pp. 89–98.

[9] N. Grira, M. Crucianu, and N. Boujemaa, "Unsupervised and semi-supervised clustering: a brief survey," in *'A Review of Machine Learning Techniques for Processing Multimedia Content', Report of the MUSCLE European Network of Excellence (FP6)*, 2005.

[10] G. Karypis, *CLUTO - Software for Clustering High-Dimensional Datasets : A Clustering Toolkit, Release 2.1.1*, http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview, 2003.

[11] K. Lang, "Newsweeder: Learning to filter netnews," in *Proceedings of the Twelfth International Conference on Machine Learning*, 1995, pp. 331–339.

[12] M. Naughton, N. Kushmerick, and J. Carthy, "Clustering sentences for discovering events in news articles," in *Advances in Information Retrieval, 28th European Conference on IR Research*, 2006, pp. 535–538.

[13] N. T. S. Sambasivam, "Advanced data clustering methods of mining web documents," *The Journal of Issues in Informing Science and Information Technology*, vol. 3, pp. 563–579, 2006.

[14] I. H. Witten, A. Moffat, and T. C. Bell, *Managing Gigabytes: Compressing and Indexing Documents and Images*. Van Nostrand Reinhold, New York, 1994.

 ISBN: 978-1-61208-104-5