CLUSTERING DOCUMENTS WITH LARGE OVERLAP OF TERMS INTO DIFFERENT CLUSTERS BASED ON SIMILARITY ROUGH SET MODEL

Nguyen Chi Thanh, Koichi Yamada and Muneyuki Unehara
Department of Management and Information System Science, Nagaoka University of Technology
Nagaoka-shi, Niigata-ken 940-2188 Japan

Keywords: Document clustering, Document representation, Rough sets, Text mining.

Abstract: Similarity rough set model for document clustering (SRSM) uses a generalized rough set model based on similarity relation and term co-occurrence to group documents in the collection into clusters. The model is extended from tolerance rough set model (TRSM) (Ho and Funakoshi, 1997). The SRSM methods have been evaluated and the results showed that it perform better than TRSM. However, in document collections where there are words overlapped in different document classes, the effect of SRSM is rather small. In this paper we propose a method to improve the performance of SRSM method in such document collections.

1 INTRODUCTION

Document clustering is a process of grouping similar documents into clusters so that they would exhibit high intracluster similarity and low intercluster similarity. Document clustering has become an important text mining technique to explore useful information from text collections.

In recent years various document clustering methods have been proposed, including hierarchical clustering algorithms using result from a k-way partitional clustering solution (Zhao and Karypis, 2005), spherical k-means (Dhillon and Modha, bisecting k-means (Steinbach, Karypis and Kumar, 2000), frequent word meaning sequences based method (Li, Chung and Holt, 2008), k-means with Harmony Search optimization (Mahdavi and Abolhassani, 2008).

TRSM, a document presentation model based on vector space model, was proposed by Ho and Funakoshi (1997). The model extended the vector space mode by using rough set theory. TRSM has been successfully applied to document clustering.

We proposed a new presentation model for document clustering which uses similarity relation instead of tolerance relation in TRSM. The new model is SRSM for documents (Nguyen, Yamada and Unehara, 2010). The experiment results show that the SRSM method performed better than TRSM and other methods in two experiment data sets. However, with the second data set which has documents in near fields, the effect of improvement by SRSM is rather small. The reason is that the terms used in different classes of document are overlapped. The upper approximation of a document incorporates terms used in other fields, and as the results, it is natural that the upper approximation is not effective. To improve the performance of SRSM when documents have a large overlap of terms among document vectors, we propose a two pass approach method. The objective of the proposed method is to remove terms overlapped in document classes before applying the SRSM algorithm.

The paper is organized as follows. Section 2 describes the SRSM document clustering algorithm and results of our experiments on document collections. Section 3 presents an approach trying to improve the SRSM method by removing overlapped terms. Finally, Section 4 concludes with a summary and discussion about future research.

2 SRSM DOCUMENT CLUSTERING ALGORITHM

TRSM is a model extended from Pawlak's rough set model (Pawlak, 1982) by using tolerance relation

Chi Thanh N., Yamada K. and Unehara M..
CLUSTERING DOCUMENTS WITH LARGE OVERLAP OF TERMS INTO DIFFERENT CLUSTERS BASED ON SIMILARITY ROUGH SET MODEL .
DOI: 10.5220/0003068803960399
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2010), pages 396-399
ISBN: 978-989-8425-28-7

instead of equivalence relation. TRSM is used as a basis to model documents and terms for information retrieval, text mining, etc (Ho and Funakoshi, 1997). It was then applied to document clustering to improve the quality of similarity measure between documents (Ho and Nguyen, 2002). It is also applied in clustering of web search results (Meng, Chen and Wang, 2009). However, tolerance relation does not reflect exactly the relationship between meanings of words when there are large differences between frequencies of words.

To resolve the problem of TRSM, we use similarity relation instead of tolerance relation to define the model (Nguyen et al., 2010). We propose a new function I_α, which uses a relative value α of word frequency instead of a constant threshold, to define the similarity class of words. The new function depends on a parameter α $(0 < \alpha < 1)$ as

$$I_\alpha(t_i) = \{t_j \mid f_D(t_i, t_j) \geq \alpha.f_D(t_i)\} \cup \{t_i\}, \quad (1)$$

where $f_D(t_i, t_j)$ is the number of documents in D in which term t_i and t_j co-occur, $f_D(t_i)$ is the number of documents in D in which term t_i occurs. $I_\alpha(t_i)$ is a set of terms that co-occur with t_i at a probability more than or equal to α.

SRSM document clustering algorithm is extended from spherical k-means algorithm (Dhillon and Modha, 2001) using SRSM. In the algorithm, we used a document vector with terms in the upper approximation of terms in document instead of an ordinary document vector.

We use tf×idf weighting scheme to calculate the weights of terms in upper approximations of document vectors. The term weighting method is extended to define weights of terms in the upper approximation. It ensures that each term in the upper approximation, but not contained in the document, has a weight smaller than the weight of any term in the document. The weight a_{ij} of term t_i in the upper approximation of document d_j is defined as follows.

$$a_{ij} = \begin{cases} f_{ij} \times \log\left(\frac{N}{f_{D(t_i)}}\right) & \text{if } t_i \in d_j \\ \min_{t_h \in d_j} a_{hj} \times \frac{\log\left(\frac{N}{f_{D(t_i)}}\right)}{1+\log\left(\frac{N}{f_{D(t_i)}}\right)} & \text{if } t_i \in \overline{apr}(d_j) \setminus d_j \\ 0 & \text{if } t_i \notin \overline{apr}(d_j) \end{cases} \quad (1)$$

where f_{ij} is the frequency of term i in document j, $f_{D(t_i)}$ is the number of documents containing the term i, N is number of documents and t_h is the term with the smallest weight in the document j. This equation is derived from the TRSM algorithm (Ho and Nguyen, 2002). Then normalization is applied to the upper approximations of document vectors. The cosine similarity measure is used to calculate the similarity between two vectors.

The algorithm is described as follows (Nguyen et al., 2010).

1. Preprocessing (word stemming, stopwords removal).
2. Create document vectors.
   2.a. Obtain sets of words appearing in documents.
   2.b. Create document vectors using tf×idf.
   2.b. Generate similarity classes of words.
   2.c. Create vectors of upper approximations of documents and normalize the vectors.
3. Apply the clustering algorithm

The method was evaluated with two document collections. The results showed that SRSM method provide better results than TRSM and the clustering quality with SRSM is less affected by the value of parameter (θ and α) than TRSM. With the second document collection, the result with SRSM is slightly better than the result when we used ordinary document vector instead of upper approximation of terms. The effect of improvement by SRSM is rather small. The reason for the small effect of SRSM with this data set is that the terms used in four classes are overlapped, because the classes' fields are rather near fields. Therefore, the upper approximation of a document X incorporates terms used in other classes, and as a result the proposed approach is not so effective.

Table 1: Evaluation of clustering results with SRSM for the second data set (Nguyen et al., 2010).

SRSM			
α	Entropy	Mutual information	F measure
0.30	0.420	0.312	0.825
0.35	0.396	0.321	0.853
0.40	0.375	0.328	0.859
0.45	0.348	0.337	0.877
0.50	0.327	0.345	0.892
0.55	0.309	0.352	0.900
0.60	0.309	0.352	0.905
0.65	0.306	0.353	0.907
0.70	0.308	0.353	0.908
0.75	0.311	0.351	0.907
0.80	0.310	0.352	0.908

3 EXPERIMENT OF OVERLAPPED TERM REMOVAL

As discussed in the previous section, there are cases in real applications where documents that should be classified into different clusters have a large overlap of terms among document classes. The task of classifying these documents into different clusters is very challenging. It expands applicability of documents clustering; i.e. classifying research documents in near fields, classifying research papers and instruction papers in the same field, etc.

To improve the performance of SRSM in these cases, we try to remove terms overlapped in different classes before apply the SRSM algorithm. We proposed a two pass approach for overlapped term removal. The method is described as follows.

(1) At the first pass, the current approach with upper approximation is applied.

(2) We check each word how many clusters each word is related to. If a word is included in multiple word vectors each of which represents a cluster, we remove the word from word vectors.

(3) Then, we apply the current clustering again as the second pass.

A word vector representing a cluster (cluster vector) in (2) is just a word vector that consists of all words in the cluster.

The idea of the approach is to remove overlapped words between clusters. The overlapped words make the upper approximation not effective because it could incorporate words used in other classes. After the first pass, we can determine which words occurred in multiple clusters and remove them in the step (2).

To implement the idea, first we have to obtain the word vectors of clusters, then in each document, we remove the words appear in multiple word vectors representing clusters.

In the experiment, we try the two pass approach, with step (2) remove all words included in more than one word vector of clusters, which means words simultaneously appear in two clusters. We call this method as two pass appoach for two clusters. First, we build a word vector representing a cluster which is just a word vector that consists of all words in the cluster. Then we calculate a word vector consist of words appear in two cluster word vector, for example, words appear in cluster 1 word vector and cluster 2 word vector. For each document word vector, we remove words appear in the calculated vectors. The total number of words to be removed when $\alpha = 50$ is 4322 and when $\alpha = 60$ is 4298.

Table 2 shows the result of the experiment with two pass approach for two clusters.

Table 2: Results with two pass approach for two clusters.

α	Mutual information	F measure
0.35	0.196	0.853
0.40	0.196	0.715
0.45	0.193	0. 720
0.50	0.195	0.721
0.55	0.131	0.645
0.60	0.124	0.635
0.65	0.126	0.637

As we can compare between original method (Table 1) and the two pass approach for two clusters (Table 2), the result of the two pass approach is worse. This is can be explained by the large number of words removed. We remove so many words from document so the document vectors can not represent the content of the document.

After the two pass approach for two clusters, we do the experiment, with two pass appoach for three clusters, which in step (2) removes all words simultaneously appear in three clusters. Similar to the two pass approach for two clusters, we calculate a word vector consisting of words appear in three cluster word vectors. Then for each document word vector, we remove words appear in the calculated vectors.

Table 3 shows the result of the experiment with two pass approach for three clusters. This result is better than the result in Table 2 but still worse than the result of the original method.

Table 3: Results with two pass approach for three clusters.

α	Entropy	Mutual Information	F measure
0.35	0.523	0.275	0.830
0.40	0.511	0.279	0.839
0.45	0.504	0.282	0.841
0.50	0.485	0.288	0.854
0.55	0.455	0.299	0.855
0.60	0.434	0.307	0.862
0.65	0.443	0.304	0.855

Table 4: Results with two pass approach for four clusters.

α	Entropy	Mutual information	F measure
0.35	0.406	0.317	0.873
0.40	0.400	0.319	0.873
0.45	0.371	0.330	0.874
0.50	0.369	0.330	0.891
0.55	0.330	0.344	0.906
0.60	0.336	0.342	0.901
0.65	0.334	0.343	0.903

Next, we do the experiment with two pass appoach for four clusters, which in step (2) removes all words simultaneously appear in four clusters.

Table 4 shows the result of the experiment with two pass approach for four clusters. This result is better than the result in Table 2 and Table 3 but still worse than the result in Table 1.

The proposed method is not good as the original SRSM method. With the two pass approach for two clusters, there are many words appear in two clusters so when we remove them, many document vectors only have few words, the result becomes worse. In the case of removing words appearing in three and four clusters, the results are better but still worse than the original approach. The reason can be the words removed are not the "right" words, which are the words we try to remove. We try to remove the words overlapped in "actual class" of documents, but here we removed words overlapped in the algorithm's clustering results. These words may not the same because our results are not exactly similar to the "actual class" of documents.

4 CONCLUSIONS

A Similarity Rough Set Model for document clustering has been introduced and used in representation of document in clustering process. Experiment results (Nguyen et al., 2010) show that the quality of the clustering with SRSM is better than the one with TRSM. However, it has been shown that the effectiveness of our approach is not so large when terms used in different classes overlap.

To improve the performance of SRSM clustering when terms used in different classes are overlapped, a two pass approach has been proposed. However, it works not good as expected. The reason is that the proposed method does not remove the overlapped words in "actual class" of documents, it only removes words overlapped in clusters generated from the clustering method. In future work, we will continue to improve the efficiency of the algorithm by trying to remove the overlapped words using different approaches. This is a challenging study and it can be applied to real applications where documents have a large overlap of terms among classes such as research papers in near fields or research papers and instruction papers in the same field.

REFERENCES

Dhillon, I. S. and Modha, D. S. (2001). Concept decompositions for large sparse text data using clustering. *Machine Learning*, 42 (1-2), pp. 143-175.

Ho, T. B. and Funakoshi, K. (1997). Information retrieval using rough sets. *Journal of Japanese Society for Aritificial Intelligence*, 13 (3), pp. 424-433.

Ho, T. B. and Nguyen, N. B. (2002), Nonhierarchical document clustering based on a tolerance rough set model. *International Journal of Intelligent Systems*, 17 (2), pp. 199-212.

Li, Y., Chung, S. M. and Holt, J. D. (2008). Text document clustering based on frequent word meaning sequences. *Data and Knowledge Engineering*, 64 (1), pp. 381-404.

Luce, R. D. (1956). Semiorders and a Theory of Utility Discrimination. *Econometrica*, Vol. 24, No. 2, pp. 178-191.

Mahdavi, M. and Abolhassani, H. (2008). Harmony K-means algorithm for document clustering. *Data Mining and Knowledge Discovery*, pp. 1-22.

Meng, X.-J., Chen, Q.-C. and Wang, X.-L. (2009). A tolerance rough set based semantic clustering method for web search results. *Information Technology Journal*, 8 (4), pp. 453-464.

Nguyen, C. T., Yamada, K. and Unehara, M. (2010), A similarity rough set model for document representation and document clustering. *IEICE Transactions on Information and Systems* (submitted).

Pawlak, Z. (1982). Rough sets. *Int. J. of Information and Computer Sciences*, 11 (5), pp. 341-356.

Salton, G. and McGill, M. J. (1983). *Introduction to modern information retrieval*. MCGraw-Hill Book Company.

Stefanowski, J. and Tsoukias, A. (2001). Incomplete Information Tables and Rough Classification. *Computational Intelligence*, 17(3), pp.545-566.

Steinbach, M., Karypis, G. and Kumar, V. (2000). A comparison of document clustering techniques. *Proceedings of the KDD Workshop on Text Mining*.

Strehl, A., Ghosh, J. and Mooney, R. (2000). Impact of similarity measures on web-page clustering, *Proceedings of the 17th National Conference on Artificial Intelligence: Workshop of Artificial Intelligence for Web search (AAAI 2000)*, Austin, TX, pp. 58–64.

Yao, Y. Y. Wong, S. K. M. and Lin, T. Y. (1997). A review of rough set models. *Rough Sets and Data Mining: Analysis for Imprecise Data*, pp. 47-73.

Zhao, Y. and Karypis, G. (2005). Hierarchical clustering algorithms for document datasets. *Data Mining and Knowledge Discovery*, 10 (2), pp. 141 - 168.