Advanced Science and Technology Letters
Vol.99 (ITCS 2015), pp.87-92
http://dx.doi.org/10.14257/astl.2015.99.22

A Level-wise Hierarchical Document Clustering method for Categorization

Kil Hong Joo[1] and Nam Hun Park[2]

[1] Dept. Of Computer Education, Gyeongin National University of Education, San 6-8
Seoksudong Manangu Anyangsi, Gyeonggi, Korea, 430-040
khjoo@ginue.ac.kr

[2] Dept. Of Computer Science, Anyang University, 102 Samsungli, Buleunmyun, Ganghwagun,
Incheon, Korea, 417-833
nmhnpark@anyang.ac.kr

Abstract. For document categorization, numerous words appearing in similar documents are divided into stopwords and keywords and to precisely describe documentary characteristics, documents are expressed by keywords without stopwords. For enhanced clustering precision, this paper proposed SHODC algorithm, a seed cluster-based hierarchical document clustering method, and DHODC method through domain stopwrod removal and tree structure expansion for document categorization. Through several experiments, it was found that the deeper the domain levels, the more precise results were produced by the suggested method compared to other algorithm. The suggested algorithm.

Keywords: a hierarchical clustering, categorization

1 Introduction

Document categorization is to sort mass amount of documents into their corresponding themes. Information retrieval (IR) refers to a series of process to extract necessary information from large-scale document pools. Most IR targets text documents. Such documents are characterized by following clustering, document classification, and document summarization. Regarding document clustering, in particular, many studies were performed for the purpose of search result browsing or classification itself for easier IR. For example, Scatter/Gather [1,2,3] used the clustering method to make people find out their desired documents easily by browsing database details.

In this research, to classify documents in an automated manner, SHODC (Seed HODC) algorithm based on the HODC algorithm [7], a hierarchical document clustering method, is presented along with the level-wise domain stop-word removing (LSR) method which is to remove stopwords based on word frequency and deviation within a domain where relevant documents belong. For the expression of document-

ISSN: 2287-1233 ASTL

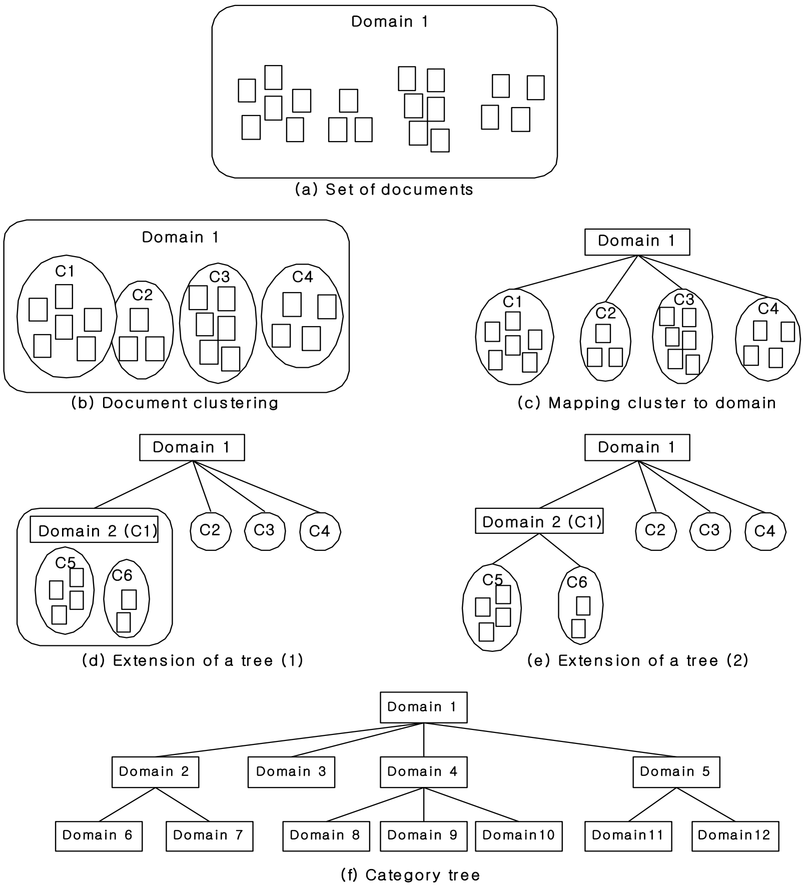
set tree relationships, clusters are generated and tree expansion is repeatedly applied to the generated clusters to form a tree structure expressing inter-document set relationships. Such a tree structure is defined as category tree.

2 Document Clustering Methods

The hierarchical agglomerative document clustering-based document categorization method [4,5,6] utilizes a tree-shaped figure to perform categorization. This method does not require extra time for document categorization and provides high document search efficiency. However, as its sub-tree structure is only binary, its results are largely different from the manually-classified actual category structure. For this reason, the method can hardly provide precisely desired outcome to users. In this paper, since domain stopword removal and document clustering method are utilized together to form semantically similar document sets, clustering is performed first then categories are created through domain mapping. This clustering and mapping process is repeated to create clusters. Then such clusters are added to the category tree and category tree is completed. Category tree creation is repeated in automated procedures until domains are no longer semantically distinguishable. The result expresses the hierarchical relationships among clusters.

The SHODC algorithm proposed in this study bases on the HODC [7] algorithm, one of the hierarchical agglomerative clustering allowing overlaps. The SHODC algorithm conducts document clustering in two divided phases of seed cluster generation and document combination to such seed clusters. The seed cluster generation phase is where cluster similarities are compared based on the cohesion and participation used in the HODC algorithm. This two-step approach is to overcome the weaknesses of HODC algorithm, such as creating only little number of small clusters and leaving too many documents outside clusters. Category tree generation and expansion is done in the following steps;

[Phase 1] Generate clusters from document sets in the SHODC method.
[Phase 2] Add the generated clusters to the sub-domain of category tree route.
[Phase 3] Remove stopwords from the domains added to the category tree and perform clustering.
[Phase 4] Repeat Phases 2 and 3 until there is no more change in the category tree structure.



(a) Set of documents

(b) Document clustering

(c) Mapping cluster to domain

(d) Extension of a tree (1)

(e) Extension of a tree (2)

(f) Category tree

Fig. 1. Category tree generation process

The figure 1 shows document categorization process through document clustering and category tree generation. As in the figure 1(a), a domain is consisted of multiple documents. The figure 1(b) groups Domain 1 documents into multiple clusters by using the SHODC algorithm. The produced clusters are filled with semantically similar documents and domains are structured with multiple semantically distinguished clusters. In the figure 1(c), the generated clusters are set as domains and mapped into Domain 1. Here, the newly generated domains all become the child domain of Domain 1 and category tree level rises. In the figures 1(d) and 1(e), the above procedures are repeated continuously to finally create the category tree as shown in the figure 1(f).

Through category tree expansion, all of the documents are expressed in a suitable structure for IR as displayed in the figure 1. However, automatically-produced category trees are different in its shape from manually-classified categories. If the whole documents are classified automatically, resulting clusters' characteristics are largely affected. So, compared to manual classification results, such automated classification results have lower accuracy in their hierarchical order and relationships. Also, full automation in classification tends to produce a larger number of small clusters. To overcome such weaknesses, this research proposes a semi-automatic document categorization (SDC). SDC means to do both the automated method and manual method. In the category tree, a lot of documents are included in the upper part. So the method involves manual classification for the upper part while leaving the lower part to the automated process to form the whole category tree. This method provides more accurate results than the full automation method. Here, the scope of manual classification is determined by partition criterion, δ, and by adjusting the partition criterion, δ, clustering efficiency could be expanded.

3 Experimental results

Diverse experiments were conducted to evaluate the suggested algorithm from various different perspectives. Precision is the number of appropriate documents among the classified documents. It means the ability not to include inappropriate documents. Recall means the number of correctly classified documents among the entire appropriate documents, representing the ability to include correct documents. Precision and recall are calculated as table 1. Precision and recall are mutually complementary. Depending upon their values, generated clusters are evaluated. Thus, cluster evaluation based on the two values has the break even points (BEP) of precision and recall, or *F-measure* [7]. BEP is calculated as follows;

$$BEP = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

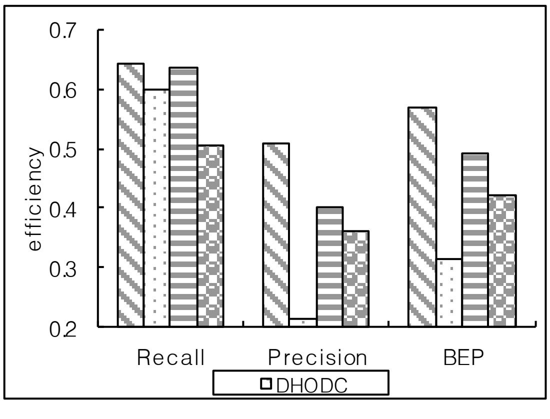


Fig. 2. Performance comparison

The figure 2 shows performance comparison between the DHODC method proposed herein and the three clustering methods – single linkage (SL) method, agglomerative linkage (AL) method and ward method (Ward) according to similarity comparison methods - under the same condition. As in the figure 8, the SL method showed the lowest performance while the AL method that represents the hierarchical agglomerative algorithms gave the best results among the three methods above. This results proves that cohesion-based methods are more excellent. But, by employing not only the cohesion but also the participation, the DHODC method proposed in this research, delivers better performance than the AL method. In this sense, it is noted that clustering efficiency improves by utilizing both cohesion and participation.

4 Conclusion

In this paper, for document categorization, numerous words appearing in similar documents are divided into stopwords and keywords and to precisely describe documentary characteristics, documents are expressed by keywords without stopwords. For enhanced clustering precision, this paper proposed SHODC algorithm, a seed cluster-based hierarchical document clustering method, and DHODC method through domain stopwrod removal and tree structure expansion for document categorization. An experiment was conducted herein for any change in precision according to domain support and document support changes to test the DHODC algorithm performance. The generally-utilized document clustering algorithm was compared with the suggested method. As a result, it was found that the deeper the domain levels, the more precise results were produced by the suggested method compared to other algorithm. The suggested algorithm, however, should re-conduct document clustering if there is any huge change in document sets. In this sense, study will be necessary on a gradual document clustering method capable of effectively dealing with increasing number of documents.

Acknowledgements. This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science Technology (NRF-2014R1A1A1008524).

References

1. Yunjae Jung, Haesun Park and Ding-Zhu Du, "A Balanced Term-Weighting scheme for Improved Document Comparions and Classification", NEC 2000
2. I.Aalbersberg, " A Document Retrieval Model based on Term Frequency Ranks", 17th International ACM SIGIR Conference on Research and Development in Information Retrieval, 163-172, 1994

3. http://www.yahoo.com
4. Michael Steinbach and George Karypis " A Comparison of Document Clustering Techniques" 5th Pacific Asia Conference on Knowledge Discㅐvery And Data Mining,2001
5. A.El-Hamdouchi and P.Willet, "Hierarchical document clustering using Ward's method," in Proceedings of the 9th ACM SIGIR, 1986
6. R. M. Cormack. "A review of classification". Journal of the Royal Statistical Society, 134:321-367, 1971
7. Paul Bradley and Usama Fayyad, Refining Initial Points for K-Means Clustering" Proceedings of fifteenth International Conference on Machine Learning ICML 98, pages91-99.