GREEDY APPROACH FOR DOCUMENT CLUSTERING

Lim Cheon Choi and Soon Cheol Park
Division of Electronics and Information Engineering, Chonbuk National University, Jeonju, Korea

Keywords: Document clustering, Cluster validity indices, Greedy algorithm, Average similarity.

Abstract: A Greedy Algorithm for Document Clustering (Greedy Clustering) is proposed in this paper. Various cluster validity indices (DB, CH, SD, AS) are used to find the most appropriate optimization function for Greedy Clustering. The clustering algorithms are tested and compared on Reuter-21578. The results show that AS Index shows the best performance and the fastest running time among cluster indices in various experiments. Also Greedy Clustering with AS Index has 15~20% better performance than traditional clustering algorithms (K-means, Group Average).

1 INTRODUCTION

The document clustering is to group the documents which are similar in a set of documents without prior information(Christopher, 2008, Croft, 2009). The researches of document clustering are actively progressing for analysis amount of information in internet.

One of the main issues in document clustering is cluster validity index(Halkidi, 2001). Cluster validity index measure the clustering result is good or not. Document clustering is unsupervised classification technique, thus cluster validity index is very important.

Document clustering is optimization problem about the cluster validity index(Maulik, 2002). In recent years, some optimization algorithms such as the genetic algorithm(Song and Park, 2009) and the particle swarm optimization algorithm(Cui and Potok, 2005) are applied document clustering. These algorithms show good performance but have long running time.

Greedy algorithm is one of the fast optimization algorithms(Richard, 2011). And greedy algorithm sometimes works well for optimization problems. In this paper, we proposed the greedy algorithm for document clustering to make a fast and good performance clustering algorithm.

This paper is organized as follows. The next section introduces some cluster indices. Section 3, presents the principle of proposed algorithm. Section 4, explains experiment setting, results, evaluation approaches and analysis. Section 5, concludes and discusses future work.

2 CLUSTER VALIDITY INDICES

Cluster validity indices are used for measuring a result of clustering(D.L and D.W, 1979). In this section, the four cluster validity indices have been introduced. Table 1 shows the used notation in cluster validity indices.

Table 1: Meaning of Notation.

Notation	Meaning
n_c	number of cluster
nc_i	number of document in ith cluster
c_i	ith cluster
cv_i	centroid vector of ith cluster
N	number of document
x_i	ith document
$d(x_i, x_j)$	similarity between x_i, x_j (cosine similarity)

2.1 Davies - Bouldin (DB) Index

DB Index (D.L and D.W, 1979) based on the similarity of two clusters(R_{ij}). The R_{ij} calculated by the within cluster scatter and the between cluster separation.

Usually R_{ij} is defined as

Choi L. and Park S. (2012).
GREEDY APPROACH FOR DOCUMENT CLUSTERING.
In *Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods*, pages 597-600
DOI: 10.5220/0003836605970600

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}$$
$$d_{ij} = d(cv_i, cv_j),\ s_i = \sum_{x \in c_i} d(x, cv_i) \quad (1)$$

where s_i means scatter within *i*th cluster, d_{ij} means similarity between *i*th cluster and *j*th cluster.

Then the DB Index is defined as

$$DB = \frac{1}{n_c}\sum_{i=1}^{n_c} R_i$$
$$R_i = \max_{i=1,\ldots n_c, i \neq j} R_{ij}, i = 1,\ldots, n_c \quad (2)$$

The lower value of DB Index means better result of clustering.

2.2 Calinski Harabasz (CH) Index

The CH Index(Maulik, 2002) based on the between cluster scatter matrices($trace\ B$) and within cluster scatter matrices($trace\ W$).

The $trace\ B$ defined as

$$trace\ B = \sum_{k=1}^{n_c} nc_k \left\| nc_k - z \right\|^2 \quad (3)$$

where z means centroid of the entire data set.

The $trace\ W$ defined as

$$trace\ W = \sum_{k=1}^{n_c}\sum_{i=1}^{n_k} \left\| x_i - nc_k \right\|^2 \quad (4)$$

The CH Index is defined as follow Equation 3 and Equation 4.

$$CH = \frac{trace\ B / (n_c - 1)}{trace\ W / (N - n_c)} \quad (5)$$

The upper value of CH Index means better result of clustering.

2.3 SD Validity Index

The SD Index(Halkidi, 2001) based on the average scattering of clusters and total separation of clusters..

The scattering is calculated by variance of the clusters and variance of the dataset.

The average scattering for clusters is defined as

$$Scatt = \frac{1}{n_c}\sum_{i=1}^{n_c} \frac{\left\| \sigma(c_i) \right\|}{\left\| \sigma(X) \right\|} \quad (6)$$

where $\sigma(X)$ means the variance of the dataset, $\sigma(c_i)$ means the variance of the ith cluster. So the average scattering calculated by variance of the dataset and variance of the each cluster.

The total separation of cluster is calculated by distance between each cluster.

The total separation of cluster is defined as

$$Dis = \frac{D_{\max}}{D_{\min}} \sum_{k=1}^{n_c} \left(\sum_{z=1}^{n_c} \left\| cv_k - cv_z \right\| \right)^{-1} \quad (7)$$

where the D_{max} means the maximum distance between cluster centroids, the D_{min} means the minimum distance between cluster centroids.

Then SD index defined as follow Equation 6 and Equation 7.

$$SD = \alpha \times Scat + Dis \quad (8)$$

where α is weighting factor that is equal to parameter in case of maximum number of cluster (Csaba, 2006).

2.4 Average Similarity (AS) Index

The AS Index(Cutting, 1992) based on the average similarity. Average similarity of AS Index calculate similarity all documents in each cluster.

The AS Index defined as

$$AS = \frac{1}{n_c}\sum_{i=1}^{n_c} S_i$$
$$S_i = \sum_{j=1}^{nc_i - 1}\sum_{k=j+1}^{nc_i} d(x_j x_k) \quad (9)$$

The upper value of AS Index means better clustering result.

3 GREEDY APPROACH FOR DOCUMENT CLUSTERING

A greedy algorithm is any algorithm that follows the problem solving heuristic of making the locally optimal choice at each stage with the hope of finding the global optimum(Zhang, 2000).

Document clustering is optimization problem about cluster index. An optimization problem is one in which you want to find, not just a solution, but the best solution. And greedy algorithm sometimes works well for optimization problems.

In this paper, we proposed a greedy algorithm for document clustering. In this paper our proposed algorithm is called Greedy Clustering(GC).

3.1 Data Structure

First of all we need to define a data structure for greedy algorithm applied to document clustering. The data structure of GC can express the

information of cluster number all documents belongs to. Figure 1 shows the data structure of GC.

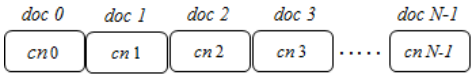

Figure 1: Data Structure of GC.

As shown the Figure 1, GC has N data. Index of data means a document number and value of data means a cluster number that document belongs to.

3.2 Greedy Clustering Operations

GC operations consist of Initialization and Greedy Searching. Figure 2 shows the progress of GC.

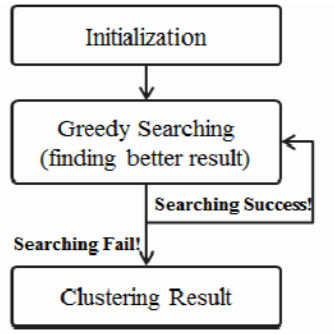

Figure 2: Progress of GC.

As shown Figure 2, the progress of GC is simple. First initialize data in Initialization operation and then repeat the Greedy Searching operation until the algorithm can't find better result.

3.2.1 Initialization

In Initialization operation, all documents in data-set are randomly assigned to a cluster.

3.2.2 Greedy Searching

In this operation, just one cluster number of document is changed. Therefore GC selects the cluster number of document that gives the biggest improvement in performance. The performance is measured by cluster validity indices.

In this paper, the GC use various cluster validity indices, so that GC is represented by

- GC(DB) : Greedy Clustering with DB Index
- GC(CH) : Greedy Clustering with CH Index
- GC(SD) : Greedy Clustering with SD Index
- GC(AS) : Greedy Clustering with AS Index

4 EXPERIMENT AND RESULT

This paper proposed a greedy approach for document clustering. For estimating its performance, the Reuter-21578 text collection set is used. Two Topic-Sets are experimented and four subjects were allocated to each Topic-Set. Each subject has 50 doc1uments, so that a Topic-Set has 200 documents. Table 2 shows the subjects of Topic-Sets.

Table 2: Subjects of Topic-Sets.

	Subjects
Topic-Set1	coffee, acq, trade, interest
Topic-Set2	earn, grain crude, ship

Documents in Topic-Set are represented by VSM. The term weight as follow(Christopher, 2008)

$$w_{ij} = tf_{ij} \times \log \frac{N}{df_j} \quad (10)$$

w_{ij} means jth term frequency in ith document, tf_{ij} means a document frequency of jth term, N means the number of documents in Topic-Set.

To evaluate the clustering performances, F-measure is used. The F-measure defined as (Croft, 2009)

$$F-measure = \frac{2 \times precision \times recall}{precision + recall} \quad (11)$$

Traditional clustering algorithms such as K-means(Shokri,1984) and Group Average(Cutting, 1992) are compare with the GCs.

Figure 3 shows the performance of traditional clustering algorithms and GCs in Topic-Set1.

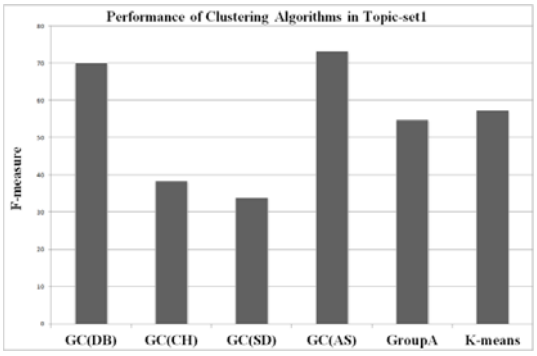

Figure 3: F-measure of Clustering Algorithms(Topic-Set1).

Figure 4 shows the performance of traditional clustering algorithms and GCs in Topic-Set2

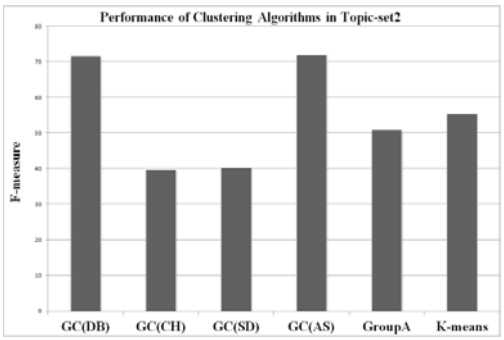

Figure 4: F-measure of Clustering Algorithms(Topic-Set2).

As shown Figure 3 and Figure 4, AS index and DB index have better performance than traditional clustering algorithms(K-means, group average) in all topic-Sets. But CH index and SA index show lower performance than traditional clustering algorithms.

Table 3: Running time of GC with Cluster Index.

Cluster Index	time(s)
GC(DB Index)	2,549
GC(CH Index)	3,612
GC(SD Index)	3,892
GC(AS Index)	15

Table 3 shows a running time of GC with Cluster Index. As shown Table 3, AS Index has faster running time than other cluster indices.

Consequently AS Index has the best performance and the fastest running time for Greedy Clustering.

5 CONCLUSIONS

In this paper, we propose the greedy algorithm for document clustering(Greedy Clustering). Main goal of this paper is find optimal function for Greedy Clustering(high performance, fast running time).

So various cluster indices are used to optimal function for Greedy Clustering. As the results of experiments in this paper, the Average Similarity index is the most suitable for the Greedy Clustering among cluster indices(DB, CH, SD, AS). Moreover Greedy Clustering with AS Index has 15~20% better performance than traditional clustering algorithms (K-means, Group Average Clustering).

But Greedy Clustering has weakness that is a long running time due to the complexity of calculation of cluster index compare with traditional clustering algorithms. We will fix this problem through the optimization of Greedy Clustering with AS Index.

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 and Technology(No. 2011-0004389) And second stage of Brain Korea 21 Project in 2011.

REFERENCES

Csaba, Legany., Sandor, Juhasz., Attila, Babos., 2006, Cluster validity measurement techniques. *Knowledge Engineering and Data Bases*

Christopher D., Manning, Prabhakar, Raghavan., Hinrich, Schütze., 2008, *Introduction to Information Retrieval*, Cambridge University Press.

Croft, W. B., Metzler, D., Strohman, T., 2009, *Search engines information retrieval in practice*. Addison Wesley.

Cui, X., Potok, T.E., Palathingal, P., 2005, Document clustering using particle swarm optimization. Swarm Intelligence Symposium 185 - 191

Cutting, D. R., Pedersen, J. O., Karger, D. R., Tukey, J.W., 1992, Scatter/Gather: a cluster-based approach to browsing large document collections. *SIGIR*, 318-329

D, L, Davies., D, W, Bouldin., 1979, A cluster separation measure. *IEEE Trans. Pattern Anal. Intell. 224-227*

Halkidi, M., Batistakis, Y., Vazirgiannis, M., 2001, On Clustering Validation Techniques. *J. Intell. Inf. Syst*. 107-145

Likas, A., Vlassis, N.A., Verbeek, J.J., 2003, The global k-means clustering algorithm. *Pattern Recognition* 451-461

Maulik, U., Bandyopadhyay, S., 2000, Genetic algorithm-based clustering technique. Pattern Recognition 1455-1465

Maulik, U., Bandyopadhyay, S., 2002, Performance Evaluation of Some Clustering Algorithms and Validity Indices. *IEEE Trans. Pattern Anal. Intell* 1650-1654

Richard, Neapolitan., Kumarss, NaimipourSmith. 2011. Foundations of Algorithms, Jones & Bartlett 4th edition.

Shokri, Z., Selim, M., A, Ismail., 1984, K-Means-Type Algorithms: A Generalized Convergence Theorem and Characterization of Local Optimality. *IEEE Trans. Pattern Anal. Mach. Intell.* 81~87

Song, W., Park, S. C., 2009, Genetic algorithm for text clustering based on latent semantic indexing. Computers & Mathematics with Applications 1901-1907

Zhang, Z., Schwartz, S., Wagner, L., Miller, W., 2000, A Greedy Algorithm for Aligning DNA Sequences. *Journal of Computational Biology* 203-214