Keyword-based Document Clustering

Seung-Shik Kang
School of Computer Science, Kookmin University & AITrc
Chungnung-dong, Songbuk-gu, Seoul 136-702, Korea
sskang@kookmin.ac.kr

Abstract

Document clustering is an aggregation of related documents to a cluster based on the similarity evaluation task between documents and the representatives of clusters. Terms and their discriminating features of terms are the clue to the clustering and the discriminating features are based on the term and document frequencies. Feature selection method on the basis of frequency statistics has a limitation to the enhancement of the clustering algorithm because it does not consider the contents of the cluster objects. In this paper, we adopt a content-based analytic approach to refine the similarity computation and propose a keyword-based clustering algorithm. Experimental results show that content-based keyword weighting outperforms frequency-based weighting method.

Keywords: Document Clustering, Weighting Scheme, Feature Selection

1 Introduction

Document clustering is an aggregation of documents by discriminating the relevant documents from the irrelevant documents. The relevance determination criteria of any two documents is a similarity measure and the representatives of the documents [1,2,3,4]. There are some similarity measures such as Dice coefficient, Jaccard's coefficient, and cosine measure. These similarity measures require that the documents are represented in document vectors and the similarity of two documents is calculated from the operation of document vectors.

In general, the representatives of a document or a cluster are document vectors that consist of <term, weight> pairs and the document similarities are determined by the terms and their weighting values that are extracted from the document [7,9]. In the previous studies on the document clustering, we focused on the clustering algorithm, but the document representation methodology was not the important issue. Document vectors are simply constructed from the term frequency (TF) and the inverted document frequency (IDF). This representation of term weighting method starts from the precondition that terms or keywords representing the document are calculated by TF-IDF. Term weighting method by TF-IDF is generally used to construct a document vector, but we cannot say that it is the best way of representing a document. So, we suppose that there is a limitation to improve the accuracy of the clustering system only by improving the clustering algorithm without changing the document/cluster representation method.

Also, document clustering requires a large amount of memory spaces to keep the representatives of documents/clusters and the similarity measures [6, 8, 10]. Given N documents to be clustered, $N \times N$ similarity matrix is needed to store document similarity measures. Also, the recursive iteration of similarity calculation and reconstructing the representative of the clusters need a huge number of computations.

In this paper, we propose a new clustering method that is based on the keyword weighting approach. The clustering algorithm starts from the seed documents and the cluster is expanded by the keyword relationship. The evolution of the cluster stops when no more documents are added to the cluster and irrelevant documents are removed from the cluster candidates.

2 Keyword-based Weighting Scheme

In general, the construction of a document vector depends on the term frequency and document frequency. If keywords are determined by frequency information of the document, we are apt to generate an error that nouns are often used regardless of substance of the document and the words of a high frequency are extracted. The clustering method, which is focused on similarity calculation considers the whole words except stopwords as the representative of the document, and constitutes a document vector that is calculated by the weight value from the term frequency and document frequency.

It is common that terms and their weight values represent a document and <term, weight> pairs are the unique elements of the document vector. When we construct a document vector, term frequency and document frequency are the most important features to calculate the weight of a term. As for the terms and

This work was supported by the Korea Science and Engineering Foundation(KOSEF) through the Advanced Information Technology Research Center(AITrc).

their weight values, the weight value of a term means a ranking score just as an importance factor to the document. So, the term weighting can be seen as an evaluation of the term as a keyword or a stopword to the document. The weighting function *w(t)* from a term to its weight is described in expression (1).

$$w: \text{term} \rightarrow \text{weight} \qquad (1)$$

$$w(t) = \begin{cases} 0, & \text{if } t \text{ is a stopword} \\ 1, & \text{if } t \text{ is a keyword} \\ a, & \text{otherwise} \quad 0 \le a \le 1 \end{cases}$$

For the weighting scheme of terms, there are two points of views as the representation of a document:

(1) a discriminative value that distinguishes or characterizes the document from others;
(2) an importance measure as a keyword or a stopword.

Frequency-based term weighting (FBW) is a statistical measure of terms in an inter-document relationship. This weighting scheme is a very efficient method for distinguishing and characterizing a document from others, and it performs well for the applications of document classification or clustering in the information retrieval system. The only evaluation measure to characterize a document in frequency-based weighting scheme is a frequency statistics, but term frequencies are not the best measures to characterize the document by terms.

Another weighting scheme is a keyword-based term weighting (KBW) method that is based on the keyword importance factors in a document. It is an analytic approach that analyses the contents of a document to get a keyword list from the document. The weight value of a word is calculated by the importance factors as a keyword in a document. The weight value of a word is a combination value of keyword-weighting factors and the terms are ordered by the keyword ranking score. The ranking scores in this weighting scheme are calculated from the analysis results of the document. Keyword-based term weighting will be a good solution to overcome the limitation of the frequency-based weighting scheme.

Keywords in a text are the terms that represent a document and the candidate keywords are extracted from the analysis results of the document. Keyword ranking method depends on several factors of a term such as the type of a document, the location and the role of words in a sentence or a paragraph [5].

Thematic words of a document are representative terms for the document. Thematic words are extracted from a text by analysing the contents of the text, but keyword extraction depends on the type of text. Keywords are easily found in the title or an abstract in a research paper that consists of a title, abstract, body, experiment, and conclusion. Also, newspaper article contains a keyword in the title or the first part of the text. There are some clues of determining a keyword and we may classify them as word level, sentence level, paragraph level, and text level features. Word-level features are the type of part-of-speech and case-role information. The part-of-speech of Korean noun is divided into common noun, compound noun, proper noun, and numeral.

Syntactic or sentence-level features are the type of a phrase or a clause, sentence location, and sentence type. From the rhetoric word in a sentence, the importance of the sentence is computed and the terms in a sentence are affected by the type of a sentence. Also, the weighting scheme of a term in the subjective clause is not the equal to the same term that appeared in an auxiliary clause or in a modifying clause. Basic term weight is assigned by the type of a term and recomputed by the features that it accompanies in the text. That is, the weight value of a term is also determined by the characteristics of word, sentence, phrase, and clause where the term is extracted.

3 Keyword-based Document Clustering

Keyword-based document clustering creates a cluster by the keywords of each document. Suppose that C is a set of clusters that is finally created by the clustering algorithm. If *n* is the number of clusters in C, then C is a set of clusters C_1, C_2, $\cdots$, C_n.

$$C = \{ C_1, C_2, \cdots, C_n \}$$

Each cluster C_i is initialised by document *d* that is not assigned to the existing clusters, and *d* is a seed document of C_i. When a new cluster is created, expansion and reduction steps are repeated until it reaches a stable state from the start state. In each evolution steps for cluster C_i, C_i^j is the *j*-th state of C_i.

C_i^j : the *j*-th state of a cluster C_i

The characteristic vector of a cluster is a set of <keyword, weight value> pairs that represents the cluster. If K_D is a keyword set of a document D and K_{C_i} is a keyword set of cluster C_i, then $K_{C_i}^j$ is the *j*-th state of cluster C_i. Figure 1 shows a keyword-based clustering algorithm for the cluster C_i. Given the keyword sets for each document, cluster C_i is created by the self-expanding algorithm.

3.1 Cluster Initialisation

The first step of the clustering algorithm is a creation and initialisation of a new cluster. A document D is selected that does not belong to any other cluster, and it is assigned to a new cluster C_i^0 that is an initial state

of cluster C_i.

$$C_i^0 = \{ D \}$$

At this time, a document D that is the first document in the new cluster is called a seed document (or an initialisation document). The seed document is randomly selected among the documents that do not belong to the clusters C_1 ~ C_{i-1}. Keyword set K_D of a document *D* is a set of keywords k_1, k_2,⋯, k_n that are extracted from document D. The initial state of keyword set $K_{C_i}^0$ is initialised by K_D.

$$K_{C_i}^0 = K_D$$

K_D = { k | k is a keyword that is extracted from D }

$C_i^0 = \{D\}$
$K_{C_i}^0 = K_D$
$C_i^1 =$ { D_x | document D_x, where $k \in K_{D_x}$
for $\forall k$ such that $k \in K_{C_i}^0$ }
$j = 1$
do {
$K_{C_i}^j = \cup K_{D_x}$, where $D_x \in C_i^j$
$C_i^{j+1} = C_i^j$
for all $D_x \in C_i^j$ begin
$s = sim(D_x, K_{C_i}^j)$
if $(s < threshold)$
$C_i^{j+1} = C_i^{j+1} - \{D_x\}$
end for
$j = j + 1$
} *while*(*isDeleteDocument*())
$C_i = C_i^j$

Figure 1. Keyword-based clustering algorithm

3.2 Expanding the Cluster

In the initialisation step of the cluster, a new cluster C_i^0, an initial state of cluster C_i, is established as the seed document, and the keyword set $K_{C_i}^0$ is initialised by the key word set of the seed document. In the expanding step of the cluster, the cluster is expanded by adding more related documents to the cluster, that include the keywords of the seed document as the related documents of the seed document. That is, adding the total documents that appear each keyword of $K_{C_i}^0$ (the keyword extracted from the seed document) to the cluster C_i^1 that is the next state of cluster C_i expands the cluster.

$$C_i^1 = \left\{ D_x \mid k \in D_x,\ k \in K_{C_i}^0 \right\}$$

The cluster expansion is performed by the iteration of keyword expansion and cluster expansion. More documents are added to a cluster by the similarity evaluation between the keyword set and the document. If a new document is added to a cluster, then the keywords in the added document are also added to the keyword set of the cluster. The first expansion is performed by the keyword set extracted from the seed document. The second expansion is performed by new keywords that are added to a cluster as a result of the first expansion. And the *i*-th expansion is performed by the *(i-1)*-th state of the keyword set.

The number of iterations is decided through the experiment. When a cluster is expanded from C_i^0 to C_i^1, the keyword set $K_{C_i}^0$ is also expanded to a new keyword set $K_{C_i}^1$ that appears in the total documents of the cluster C_i^1. The keyword set $K_{C_i}^j$ of C_i^j is a union of the total keyword sets of C_i^j.

$$K_{C_i}^i = \cup K_{D_x}, \text{where } D_x \in C_i^j$$

The keyword set $K_{C_i}^j$ of the cluster C_i^j is used to calculate the characteristic vector of each cluster. The characteristic vector is constituted the weight value calculated by term frequency (TF) and inverted document frequency (IDF) of the keywords and this is used to calculate the similarity measure between a document and the cluster.

3.3 Cluster Reduction and Completion

This step is to produce a complete cluster by removing the documents that are not related to the cluster. For the cluster C_i^j, documents of a low similarity to the cluster are removed, that are not related to a cluster C_i^j through the similarity computation with the cluster C_i^j. The result of cluster reduction is a filtering of documents that are not related to the cluster, and the cluster C_i^{j+1} is generated as a next step of the cluster C_i^j. Ultimately, the cluster C_i is completed that consists of the related documents after filtering the non-related documents. If a cluster C_i is completed, the next cluster C_{i+1} is created through the same process. Clustering is terminated if all the documents are clustered or no more clusters are created.

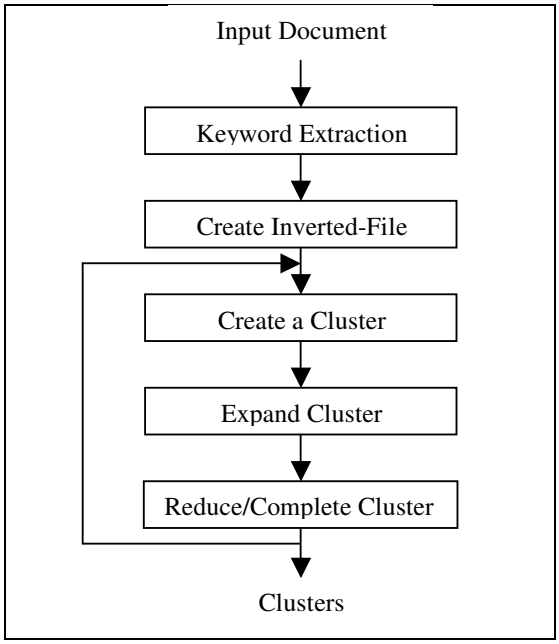

Figure 2. Overall architecture of keyword-based clustering

4 Design and Implementation

The structure of a keyword-based clustering system is shown in Figure 2. At first, keywords are extracted from each input document and the weight values of them are computed. Keywords and their scores are stored in an inverted-file structure. Inverted-file structure is a good for the expansion of the cluster and adding the documents that includes a keyword to the initial cluster. Figure 3 shows an example of the operation of the document clustering system: initialization, expansion, reduction, and completion of clusters.

A new cluster is created and it includes a seed document *D*. An initial set of keywords for the initial state of a cluster is a keyword set K_D of document *D*.

$$K_D = \{ T_{1,D}, T_{2,D}, \ldots, T_{i,D \ldots}, T_{n,D} \}$$

For the terms in K_D, documents that contain the same term are added as a candidate document in the cluster. Let the candidate documents be D_{1a}, D_{1b}, … , D_{2a}, D_{2b}, …, D_{na}, D_{nb}, …. then D_{xy}, is a document that is expanded by term T_x. Keyword set of the cluster is reconstructed by new set of documents.

In each step of the cluster expansion, the number of keywords that are used for the expansion, and the threshold of the weight value are decided through experiments considering the maximum number of document candidates in a cluster. Also, <keyword, weight> pairs as an intermediate representative of the cluster are much important factor of the cluster expansion.

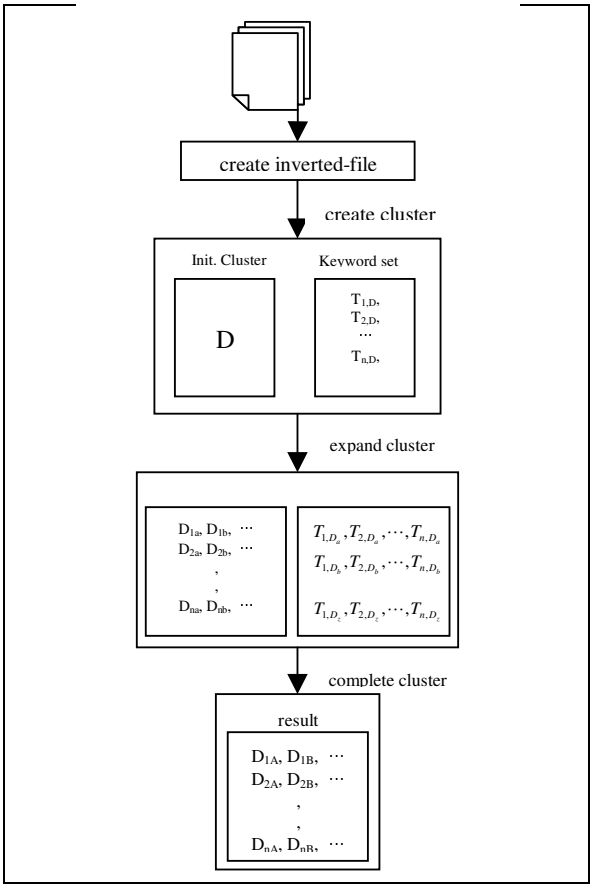

Figure 3. Example of keyword-based clustering

Now, a new keyword set that is limited to the cluster candidates is constructed to get cluster documents. Through the similarity calculation between the document and the candidate centroid of the cluster, relevant documents are selected to be a member of the cluster. Through the iterations on keyword selection and the reconstruction of the related documents, a new cluster is completed that reaches in a stable status with a strong relationship between keyword set and document set.

5 The Experiments

We implemented our clustering algorithm and applied it to the clustering of similar documents. The test documents for the experiment are collected from the three days of newspaper articles. The total number of articles is 383 and average 132 terms are extracted from the articles. We performed a document clustering by applying the difference criteria for term selection: 1) frequency-based term selection; 2) percentage-based keyword selection; and 3) keyword selection by absolute number of keywords. Figure 4 shows the result of similarity clustering by frequency-based term selection. In this experiment, three types of term selection are performed.

- all terms are used to the clustering
- terms with more than frequency 2
- terms with more than frequency 3

In each experiment, we varied the similarity decision ratio by the percentage of term matches. Figure 4 shows that term selection by frequency 2 or 3 is not good for the representation of a document.

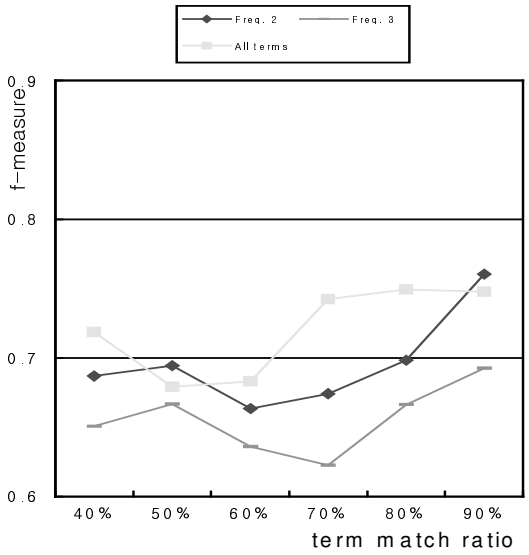

Figure 4. Frequency-based keyword selection

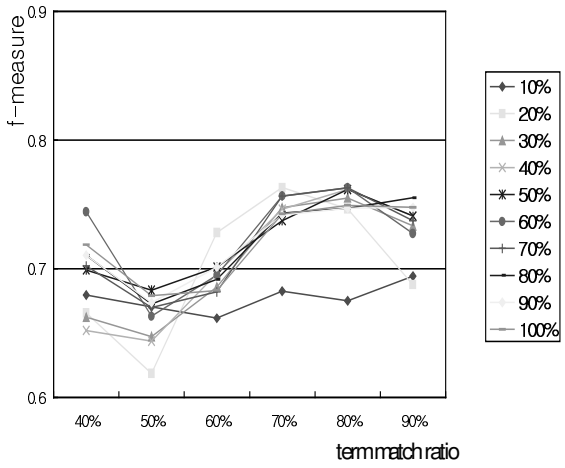

Figure 5. Percentage-based keyword selection

In the experiment of percentage-based keyword selection, terms of high weight values are selected for the similarity calculation of the document. All the curves in Figure 5 are a similar shape, except for 10 % selection. In case of 10% selection, we guess that less than 10% of keywords are not sufficient for the similarity decision and auxiliary keywords are also needed for the accuracy. Another point in this experiment is that 30%~60% keyword selection resulted better than the selection of all terms.

We compared the F_1-measure for the selection of maximum keywords. All the experiments in Figure 6 resulted better than the experiment of using all the terms in the document. Also, 30~70 keywords with 60%~70% match ratio resulted a good performance for the comparison of document similarity.

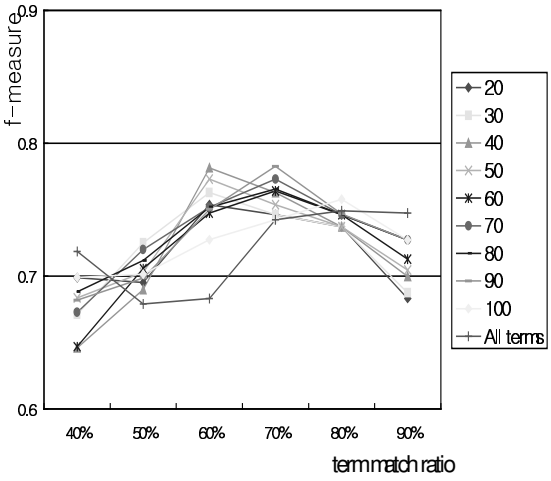

Figure 6. Keyword selection by maximum

6 Conclusion

It is common that clustering algorithm is based on the similarity computation by frequency-based statistics to aggregate the related documents. This metric is an important factor for term weighting. We proposed a term weighting method that is based on the keyword features and we tried to complement the drawback of frequency-based metric. Based on the keyword weighting scheme, documents of the same keywords are grouped into a cluster candidate and a new cluster is created by removing irrelevant documents. We performed an experiment for the clustering of similar documents and the results showed that keyword-based weighting scheme is better than the frequency-based method.

Our keyword-based algorithm is using 30%~60% of terms for a clustering and the similarity matrix is not a necessity that it will be good for the clustering of a huge number of documents. We also expect that this algorithm will be good for the topic tracking of special events. In the experiment, we randomly selected a seed document and it is a bit sensitive for the seed document. So, our next research will be focused on minimizing the effect of the seed document by getting representative keywords before starting the clustering.

References

[1] Anderberg, M. R., "Cluster Analysis for Applications", New York: Academic, 1973.

[2] Can, F., and E. A. Ozkarahan, "Dynamic Cluster Maintenance", Information Processing & Management, Vol. 25, pp.275-291, 1989.

[3] Dubes, R., and A. K. Jain, "Clustering Methodologies in Exploratory Data Analysis", Advances in Computers, Vol. 19, pp.113-227, 1980.

[4] Frakes, W. B. and R. Baeza-Yates, Information Retrieval, Prentice Hall, 1992.

[5] Kang, S. S., H. G. Lee, S. H. Son, G. C. Hong, and B. J. Moon, "Term Weighting Method by Postposition and Compound Noun Recognition", Proceedings of 13th Conference on Korean Language Computing, pp.196-198, 2001.

[6] Murtagh, F., "Complexities of Hierarchic Clustering Algorithms: State of the Art", Computational Statistics Quarterly, Vol. 1, pp.101-113, 1984.

[7] Perry, S. A., and P. Willett, "A Review of the Use of Inverted Files for Best Match Searching in Information Retrieval Systems", Journal of Information Science, Vol. 6, pp.59-66, 1983.

[8] Sibson, R. "SLINK: an Optimally Efficient Algorithm for the Single-Link Cluster Method", Computer Journal, Vol. 16, pp.328-342, 1973.

[9] Willett, P., "Document Clustering Using an Inverted File Approach", Journal of Information Science, Vol. 2, pp.223-231, 1980.

[10] Willett, P., "Recent Trends in Hierarchic Document Clustering: A Critical Review", Information Processing and Management, Vol. 24, No.5, pp.577- 597, 1988.