Development of an efficient hierarchical clustering analysis using an agglomerative clustering algorithm

Arshia Naeem[1], Mariam Rehman[2,*], Maria Anjum[1] and Muhammad Asif[3]

[1]Department of Computer Science, Lahore College for Women University, Lahore 54000, Pakistan
[2]Department of Information Technology, Government College University Faisalabad 38000, Pakistan
[3]Department of Computer Science, National Textile University, Faisalabad 37610, Pakistan

Clustering algorithms are used to generate clusters of elements having similar characteristics. Among the different groups of clustering algorithms, agglomerative algorithm is widely used in the document clustering domain. This study aimed to examine the effectiveness of agglomerative clustering algorithm in document clustering by enhancing its efficiency and evaluating it through implementation. The resulting values, precision = 0.8571, recall = 0.8571 and *F*-measure = 0.857076 indicate the highest level of accuracy and efficiency compared to existing algorithm.

Keywords: Cosine similarity measure, document clustering, *F*-measure, hierarchical agglomerative clustering, preprocessing, TF-IDF.

THE navigation mechanisms are dependent on efficient and high-quality document clustering algorithms as these algorithms use small number of clusters to handle and manage huge amount of information. Moreover, they enhance retrieval performance in many ways such as cluster driven dimensionality reduction, term weighting or query expansion. The development of new clustering criteria functions and novelty in several algorithms is due to the increase in the importance of document clustering and continuous expansion in its applications[1]. Therefore, it is imperative to enhance the performance of various clustering algorithms. Clustering is a technique used to group similar documents[2]. The data tuples are considered as objects in clustering techniques. A clustering technique partitions the objects into clusters, or groups, and so in this way, the objects that belong to the same cluster are 'similar' and 'dissimilar' to objects of another cluster. The closeness of objects in space, depending on the 'distance' function, is a common way to define the concept of similarity. Moreover, the term 'quality' of a cluster can be representable by its diameter, which specifies the maximum distance between any two objects in the cluster[2].

The clustering technique can be of specific types such as: partitioning, hierarchical, density-based and grid-based. There are three groups of clustering methods namely model, grid and density-based clustering methods[2].

In this study, focus is on hierarchical clustering algorithms. Hierarchical clustering algorithms create nested clusters by continuously splitting the instances in agglomerative mode or divisive mode. In the agglomerative[3] mode, bottom-up approach is followed, where initially every data point is considered as a cluster. However, in divisive mode, all data points initially consist of one cluster and then further smaller clusters are built. The divisive mode follows the top-down approach. Similarity measures such as 'summing the squares' are used to merge the clusters, as these measures are chosen to optimize the criterion. The hierarchical clustering methods are further classified based on the different ways of calculating the similarity measure. They are: single-linkage, complete-linkage and average-linkage clustering[4]. Hierarchical agglomerative clustering analysis merges similar clusters based on selected distance and linkage measure[4]. The clustering taxonomy is shown in Figure 1.

Partition-based algorithms do not create any hierarchy. These algorithms simultaneously consider each cluster as a partition of data. The core aim of these algorithms is to break down the set of objects into an already assigned number of disjoint clusters[4]. The data set is broken down into clusters in such a way that every cluster has at least one single data point, as well as each data point will have one cluster. In these algorithms, the number of clusters is preset by the user. The global optimality can be achieved by proper counting process of all partitions. The local minima issue occurs due to distinct partitions and a limited number of data points. This issue can be overcome by using exhaustive search methods. However, finding a global optimal partition is an NP (non-deterministic polynomial-time hardness)-hard problem. So it can only be true in theory and in practice the exhaustive methods are not beneficial[4]. Some of the well-known algorithms that belong to the partition-based clustering category are *k*-means, *k*-medoids, Clustering Large Applications (CLARA) and Clustering Large Applications based on Random Search (CLARANS)[5].

*For correspondence. (e-mail: dr.mrehman13@gmail.com)

Review of literature

Clustering is a completely unsupervised learning method which has flexibility and adaptability. The grouping of documents based on their similarity is known as document clustering[6]. The comprehensive data clustering problem is document clustering, where every object is in the form of document. The grouping of similar documents is the core aim of the clustering process, while similarity can be based on document type or the availability of contents in document. However, similar documents will create a single group (cluster)[6]. The documents are categorized into different groups. All such documents that are part of the same group are considered similar to each other.

To produce effective search results, several document clustering algorithms are mentioned in the literature[7]. Among these algorithms, partitioned clustering and hierarchical clustering are two core techniques of document clustering. Hierarchical clustering methods are used to break down the collection of documents into hierarchies to develop the hierarchical structure. Hierarchical clustering consists of two categories, viz. agglomerative (AHC) and divisive (DHC) hierarchical clustering[7].

In agglomerative hierarchical clustering (AHC) algorithms[8], documents are classified according to their similarity measure referred to as inter-cluster similarity measure[9]. These algorithms are further classified into three categories based on the distance measurement between clusters. These categories are single-linkage, complete-linkage and average-linkage[10]. In single-linkage, there exists minimum distance within the pair of clusters, in complete-linkage, maximum distance exists within the pair of clusters and in average-linkage, there exists an average distance within the pair of clusters. Our study concentrated on single-linkage clustering. In single-linkage clustering, similarity of two clusters is calculated by considering similarity in the closest pair of data points which are available in different clusters. Therefore, until all the objects are merged into a single cluster, the process of merging objects will be repeated[11]. The complete flow of the study is shown in Figure 1. The chaining issue of single-linkage clustering algorithm is discussed here to improve document clustering[12,13].

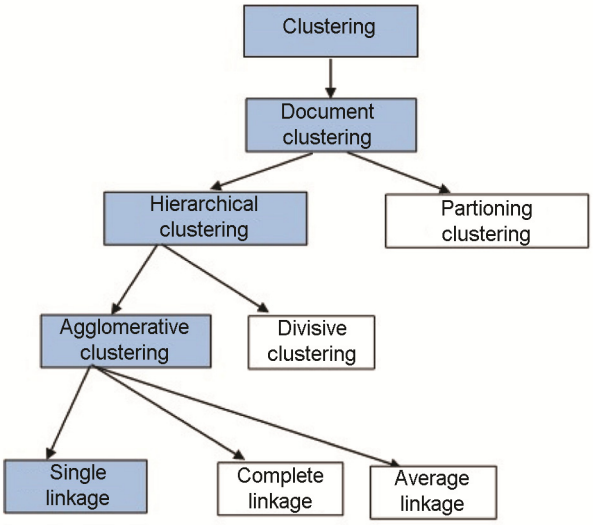

Figure 1. Clustering taxonomy flow of the study[4].

Document clustering techniques are used to make the clustering process efficient[14]. The process to find similar documents using different clustering techniques consists of comparison and selection of clustering techniques as well as the similarity measures. Al-Anazi *et al.*[14] discussed the effective and efficient combination of clustering techniques with similarity measures which could produce good quality of clustering. The platform of Rapid Miner was used for experimentation. Preprocessing of data, data mining, validation of model and visualization of results were the major elements for knowledge discovery.

Another way to cluster the documents is Malay document clustering using complete-linkage clustering technique with cosine coefficient[15]. Samat *et al.*[15] conducted experiments that made use of recall (R) and precision (P) as effective measures. The main outcome of this experiment involved retrieval from corpus of significant Malay text documents. The Malay Islamic documents were retrieved to help users extract their required data of hadith. The experiments were performed for 20, 50 and 100 clusters where clustered and non-clustered documents were compared. The results indicated that when the clusters size was 20, the effectiveness (E) score was highest. Thus, this type of clustering technique is quite efficient for Malay document clustering. However, further improvement is required to make the searching process of Malay retrieval system more efficient and effective.

Rajavat and Gupta[16] compared the different clustering approaches in their study. Initially, the documents will be unstructured; therefore, preprocessing is applied to convert data into numeric vectors. Vector space model calculates similarity measures as the clustering methods are not directly applicable to cluster documents. To evaluate cluster quality and goodness, there are different measures such as purity–purity, entropy and F-measure. To compare hierarchical with partitioning algorithms, hierarchical algorithms are more efficient than partitioning algorithms. Large databases use BIRCH technique because of its efficiency, in case of hierarchical clustering. BIRCH scans a database only once and is not sensitive to noise. The main limitation partitioning clustering process is that the *k*-means is expensive to deal with large datasets. However, BIRCH has limited memory and takes more time than *k*-means in processing. Parameters such as complexity, efficiency and sensitivity outliers are adequate for comparing other clustering algorithms.

Bsoul *et al.*[17] reviewed the different steps of document clustering such as features extraction and clustering

algorithms. They focused on detecting the crime news stories by using clustering methods. Here, for crime document clustering, 2400 documents were considered as a dataset for testing. To measure the external standards, F-measure and purity measure were considered. However, few challenges such as detecting, identifying and tracking of crime documents clustering were identified. The weaknesses of extraction terms and *k*-means algorithm's wrong detections and identifications are mentioned in the study. The selection of initial centroid and the number of clusters are the main limitations of *k*-means clustering algorithm. The other limitation is the issue of extraction features. Aggarwal and Zhai[18] explained the problem of text clustering in detail. Similarity function is used to measure the similarity within objects. Text domain consists of objects like documents, paragraphs, sentences or terms. Clustering is beneficial for documents to enhance their retrieval and browsing. Distance-based clustering algorithms are most efficient for document clustering procedures. To perform clustering, two clustering methods are used namely agglomerative clustering and *k*-means clustering. Aggarwal and Zhai[18] used two types of data that included data related to dynamic applications and data of heterogeneous applications.

Yim and Ramdeen[19] explained hierarchical cluster analysis by comparing three linkage measures using Statistical Package for Social Sciences (SPSS) by considering psychological data. Cluster analysis is an efficient technique that indicates homogeneous groups within the cluster and heterogeneous groups among the clusters. The hierarchical agglomerative cluster analysis consisted of two parts, i.e. type of linkage methods and measuring distance between cases. This study[19] focused on the theoretical background of hierarchical clustering. There were some limitations such as, the squared Euclidean distance calculated only one case per cluster. To overcome this issue, linkage measures were considered.

Another study[20] included partitioning patent dataset using AHC technique. The experiment was conducted by using Tanagra tool and accessing dataset from the EPO Worldwide Patent Statistical Database, known as 'PATSTAT'. The core objective of Tanagra tool was to provide complete and wide range algorithms related to machine learning as well as tools consisting of preprocessing of data for researchers and practitioners guidance. Agglomerative clustering technique is applied on the dataset by considering different values of *k*. The values of *k* consisted of high BSS (between sum of squares) ratio and less gap ratio. The clustering results were visualized and validated. Further study could be made by performing family patent analysis and time series analysis by using the dataset analysed in the above study[20].

Mishra *et al.*[21] conducted performance analysis of single- and complete-linkage measures by tagging questions from question papers during agglomerative clustering. The Euclidean distance measure was used for calculating distances. The dataset was taken from UGC-NET question paper II, which had 50 questions belonging to several topics of computer science and application areas. The tagging of area names such as artificial intelligence (AI), data structure (DS) is useful for clustering the questions. The papers clustered in several iterations were demonstrated. One important finding of this study was that single-linkage was found to be much more time consuming than complete-linkage.

Mohbey[22] conducted an experimental survey on single-linkage on two-dimensional space. Space consisted of several objects which were combined by the Euclidean distance where the distance of two objects was calculated and distance matrix was developed using MATLAB. The aim of this study[22] was to provide complete and systematic information about clustering algorithms. The resultant clusters and dendrogram were used as minimum spanning tree for searching or other purposes. Takumi and Miyamoto[23] compared different methods of AHC with pairwise constraints. The pairwise constraints included penalty method as well as dissimilarity modification method. The initial technique used was to change the distance between two objects by using a kernel function, however, the condition was that the objects should belong to different clusters. Another technique used was the penalty term to measure similarity. In this study, two measures namely asymmetric similarity or dissimilarity were used. Thus the above study[23] found that penalty method worked better for single-linkage, centroid and asymmetric single-linkage, but the dissimilarity modification approach worked only for the centroid method.

Wu *et al.*[24] implemented an efficient algorithm known as linear text segmentation using AHC without user corporation, parameter setting and auxiliary knowledge base. Linear text segmentation splits large text into several contemporary chunks. The task of segmenting large document into separate topics is beneficial for users, as in this way they retrieve only the topical segments of their need. The computational complexity and segmentation accuracy are considered by Text Segmentation based on Hierarchical Agglomerative Clustering (TSHAC). TSHAC consists of four steps such as text preprocessing, representation of text in vectors, sentence similarity matrix and identification of optimal topic boundaries.

Single linkage is the clustering process that works on bottom-up technique to cluster the data. This clustering method is used to create a cluster by finding minimum distance within any two points of two different clusters. Initially, each and every cluster is in the form of a singleton cluster by applying the single-linkage technique. The points with the shortest distance are combined to produce the clusters.

Each linkage method requires distance measure to proceed with its clustering process. Hence, single-linkage clustering uses Euclidean distance measure for proper and efficient clustering process[25]. The single-linkage process

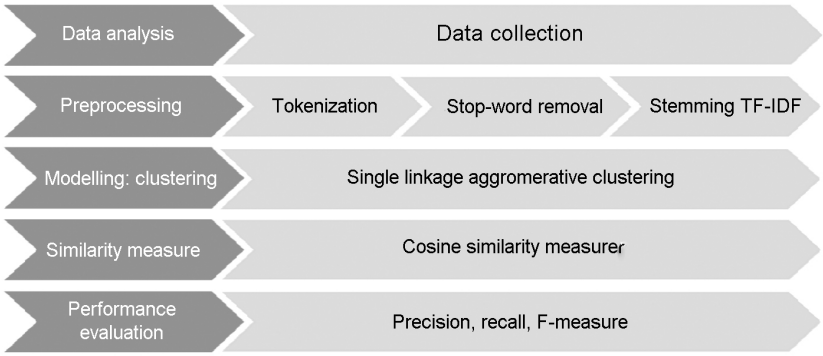

Figure 2. Research process.

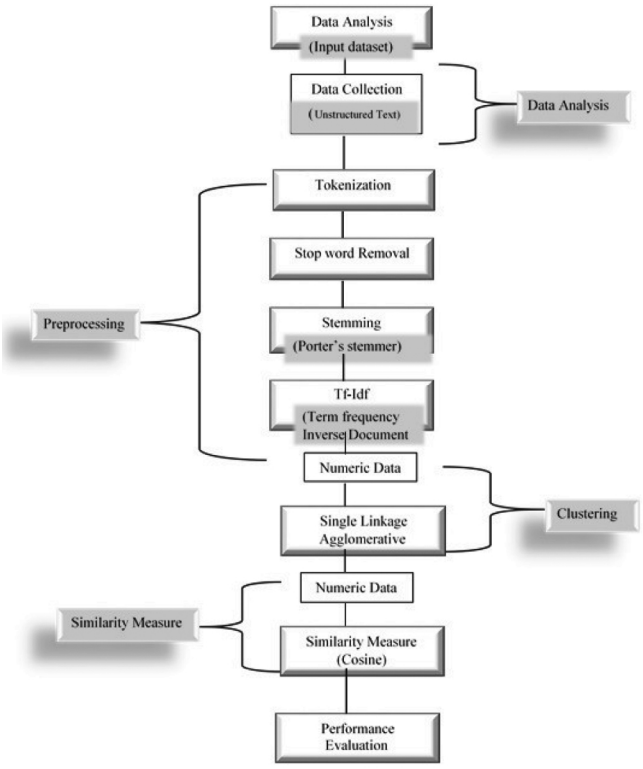

Figure 3. Detailed flow of research process.

is able to cluster limited dataset because it produces chaining issue while creating linkage clusters. This leads to the $O(n^2)$ time complexity for the clustering process. Furthermore, chaining issue slows down the clustering process which is the cause of low algorithm performance[26]. Hence, to make agglomerative clustering process efficient, it is essential to keep the chaining issue of single-linkage method at its minimum so that the efficiency of clustering process is increased.

Research methodology

In our study we have employed validation research method to enhance efficiency in single-linkage clustering algorithm. Therefore, the research methodology is composed of five core processes which are: data analysis, preprocessing, clustering, similarity measure and performance evaluation (Figure 2). In Figure 2, an abstract view of these core processes and their further implementation techniques are provided.

First process of our research is data analysis which is carried out by creating the dataset and then preprocessing is applied by using different methods to convert data into numerical form. Later, the proposed clustering technique is applied on the created dataset. Finally, validation is performed to evaluate the proposed technique and compare it with existing solutions. The steps involved in the five core processes are shown in Figure 3.

The data analysis includes data collection and description of the collected dataset used for document clustering. The dataset was created by considering research publications of faculty members of the Computer Science Department, Lahore College for Women University, Lahore, Pakistan. The dataset created at this stage is called unstructured text which is converted into a numerical form to obtain results from document clustering. The preprocessing technique used in this study consists of various steps[27]. These steps are performed to transform unstructured data into numerical form. The steps involved during preprocessing are:

Tokenization: the unstructured text document is divided into tokens or words.

Stop word removal: in this level of preprocessing, stop words are removed[28–30] using dictionary based algorithm, where, the given text in the document is compared with predefined stop words list.

The dictionary based algorithm is implemented with the following steps[30]: Step 1: Initially each word is stored in an array when the whole document is tokenized; Step 2: Each time from the list of stop-words, single stop-word is read; Step 3: By using sequential search technique, each word of text stored in array is compared with the list of stop-words; Step 4: The matched word in an array will be removed, and the comparison will continue till the whole tokenized words are scanned; Step 5: Once first stop-word is removed, another stop-word is read and

algorithm will continue starting from Step 2 and this will work recursively until all stop-words are removed.

Stemming: The process of converting the words into their root format is known as stemming. In this study, stemming task is done by Porter's stemming algorithm[31].

Term frequency-inverse document frequency (TF-IDF): During the final step of preprocessing, all the words will be in root form; therefore, TF-IDF weighting is applied to the whole bag of words[32] to transfer documents into numeric vectors. Each word is considered to be a dimension and every document is in high dimensional space.

We have used, single-linkage technique to cluster documents. The algorithm is improved in terms of its efficiency by handling the chaining issue. Moreover, the results generated by this algorithm are effective.

Once the data is converted into numeric form by applying TF-IDF measure/weighting, the similarity measure is applied on the input dataset. Similarity measure provides valid results related to the closeness or separation of targeted objects[33,34]. A similarity measure known as cosine is used in this study. In cosine similarity measure, the documents which are in the form of vectors are correlated to show the similarity within two documents. Therefore, relationship between two sets is represented by this measure. In information retrieval literature, various studies have used cosine measure to compare text documents[34]. Moreover, this measure consists of inner product of two vectors that are further divisible by the product of their length. Hence, in terms of geometry, this is called cosine of angle for two vectors[34]. The formula for measuring cosine similarity between two documents is a general formula to check similarity of any two documents from dataset

$$\cos(a, b) = \frac{a * b}{|a||b|}. \tag{1}$$

The documents a and b are considered as vectors of m-dimension having T as term set where $T = (t_1, t_2, \ldots, t_m)$. In the document, every dimension is considered as a term and each term has a certain weight. Hence, the weight of each term is non-negative. The cosine similarity lies in the range of 0 to 1. The cosine similarity computes similarity of two documents d_i and d_j, which are defined as

$$\cos(d_i, d_j) = \frac{d_i^t d_j}{\| d_i \| \| d_j \|}. \tag{2}$$

In the clustering procedure, the formula $\cos(d_i, d_j) = d_i^t d_j$, indicates the simplification of cosine formula, where unit length indicates the document vectors. The similarity of document will be indicated by 1 and non-similarity by 0. When cosine function used for similarity measure, then the properties of centroid and composite vector of set of the documents are considered. The set S_i and S_j are unit length documents which have n_i and n_j as documents respectively, while D_i, D_j, C_i and C_j are considered as composite and centroid vectors[35] respectively.

(1) Documents S_i and S_j sum of pairwise similarity is equal to D_i^t and D_j.

$$\sum_{dq \in D_i,\, dr \in D_j} \cos(\mathrm{d}q, \mathrm{d}r) = D_i^t d_j. \tag{3}$$

(2) The combined sum of pairwise similarities between the documents in S_i is equal to $||D_i||^2$ such as

$$\sum_{dq,\, dr \in D_i} \cos(Dq, Dr) = ||D_i||^2. \tag{4}$$

Proposed single-linkage agglomerative hierarchical clustering algorithm

The proposed single-linkage AHC algorithm is discussed in this section. The algorithm consists of the following modules.

Data collection module

This module consists of research publications that are retrieved from the databases. In this module, at a time three documents can be retrieved. These documents can be of two combinations: They can be of same research domain and different scholars; and they can be of same scholars and different research domain.

Input component: pdf format of the research documents are the input components in this collection module.

Output component: Simple text documents are the output components after the conversion of all pdf format documents.

Preprocessing of data module

In this module, effective preprocessing techniques are applied on all text documents. The core purpose of this module is to refine the documents before clustering.

Input component: The output of data collection module will be input for this module.

Output component: The resulting set of research documents after application of tokenization, stop-word removal, stemming and TF-IDF.

Clustering module

In this module, the proposed single-linkage agglomerative clustering algorithm is applied on text documents. This module divides the whole collection of documents into two major clusters – the documents that are most

similar will be in one cluster while the non-similar documents will be part of the second cluster. The clustering procedure depends upon the combination of documents. However, development of clusters will be different for both combination of documents, i.e. same research domain and different scholars; and same scholar and different research domains.

Input component: The resulting set of research documents after application of tokenization, stop-word removal, stemming and TF-IDF.

Output component: Two sets of clusters depending on input combination.

Similarity measure of data module

Here, the similarity score will be measured between pair of documents. Hence, the threshold of similarity will be from 0 to 1. When the cosine similarity measure is applied non-similar documents will have score to 0, while the highest level of similarity between the pair of documents will be 1. All the similarity measure results that lie between 0 to 1 threshold values will be considered for average measure of similarity.

Input component: The input components are the clusters after application of agglomerative clustering algorithm.

Output component: The output shows the calculation of similarity measure between pair of documents.

Results generation module

In this module, statistical measures are shown with the help of a graph. Hence *x*-axis represents the clustered documents, while *y*-axis represents the similarity measure among the clustered documents.

In our study, the accuracy and efficiency of the 'proposed single-linkage AHC algorithm' was measured on the basis of evaluation parameters such as precision, recall and F-measure. The results concluded with

Algorithm 1. Proposed single-linkage agglomerative clustering algorithm

Input: Three set of documents of research scholars.
$S = \{d1, d2, d3\}$
Output: Similar clustered documents
Comments:

Symbol	Description
$(d1, d2, d3)$	Each independent document
(S)	Denotes the set of documents

Require: Set of $S = \{d1, d2, d3\}$
For each: Set of $S = \{d1, d2, d3\}$
Compute: Tokenization and words will be generated and each word is stored in array.
Repeat until tokens generated.
$\{d1, d2, d3\} \rightarrow \{d1', d2', d3'\}$
Compute Stop-word removal
Repeat until the resultant text of documents after stop-words will be displayed.
$\{d1', d2', d3' - \{\text{set of stop words }\} = \{di, dj, dk\}$
Compute Stemming by Porter's stemming algorithm.
Applying TF-IDF where each document is considered as term
Repeat until:
$\{di, dj, dk\}$ is considered as $\{t1, t2, t3\}$
Comments:

Symbol	Description
T	as term set
$t1, t2, t3$	each independent term

Compute Cosine similarity measure (as mentioned in methodology section)
End for
Stored into similarity matrix $M = |di|*|dj|*|dk|$
Initial each $di|dj|dk$ as assigned to singleton cluster
Repeat
Merge the two closet clusters with minimal distance
For each cluster
Compute the similarities between clusters
End for
Update distance matrix
Until all clusters are merged in one cluster

Table 1. Evaluation results of dataset

Evaluation Parameters	Values
True positive (TP)	6
False positive (FP)	1
False negative (FN)	1
Precision	0.8571
Recall	0.8571
F-score	0.857076

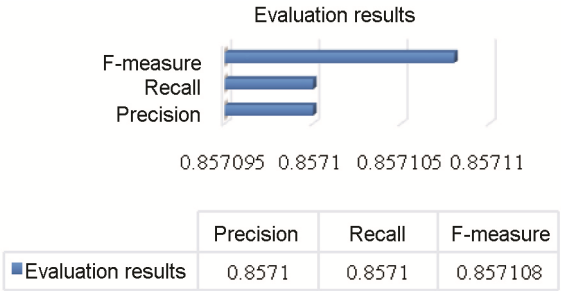

Figure 4. Representation of evaluation results.

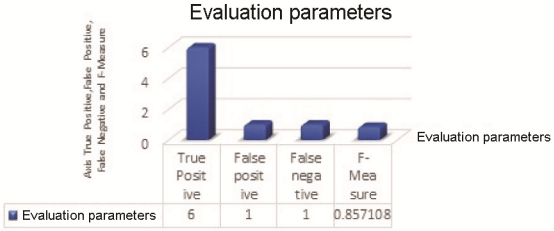

Figure 5. Representation of evaluation parameters.

Table 2. Research synthesis on the basis of performance

Research studies	Proposed algorithm	Adopted similarity	Measure evaluation mechanism
Using the complete-linkage technique along with the cosine coefficient over the document known as Malay document[36].	Complete-linkage algorithm	Cosine coefficient similarity measure	Effectiveness = 0.59
Document clustering by *k*-mean algorithm using vector space model[37].	Clustering algorithm *k*-means	Cosine similarity means	• Clustering algorithm *k*-means: • *F*-measure = 0.6 • Genetic algorithm: • *F*-measure = 0.8 • Proposed algorithm: • *F*-measure = 0.81
The text clustering improved algorithm as well as improved algorithm for text clustering application in microblogging public opinion analysis[38].	Agglomerative *k*-means clustering algorithm	Jaccard Coefficient Law	*F*-measure = 0.680
Clustering by the efficient phrase-based document similarity[39].	Agglomerative *k*-means clustering algorithm	Jaccard Coefficient Law	*F*-measure = 0.8
A clustering method for text mining[40].	*k*-means clustering algorithm	N/A	*F*-measure = 0.7
Development of an efficient hierarchical clustering analysis using agglomerative clustering algorithm.	Proposed single-linkage agglomerative hierarchical clustering algorithm	Cosine similarity measure	*F*-measure = 0.857

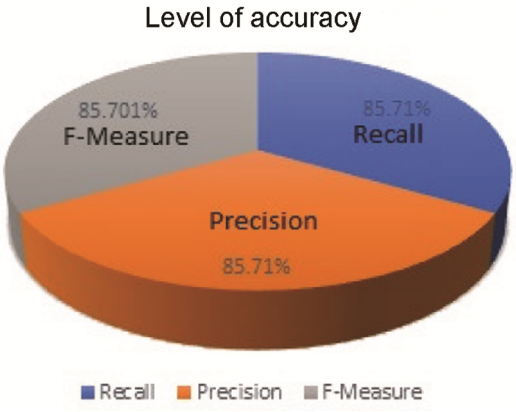

Figure 6. Accuracy of results.

precision value of 85.71%, recall value of 85.71% and F-measure also equal to 85.701% which clearly showed highest level of accuracy of results generated.

Evaluation results of dataset

The results derived from the evaluation of documents dataset are presented in this section. The evaluation parameters are applied on the self-developed documents dataset to achieve the results. The results containing the values of precision, recall and F-measure are shown in Table 1.

The precision, recall and F-measure are represented in graphical format as shown in Figure 4. The graphical representation of true positive, false positive and false negative are shown in Figure 5.

Figures 4 and 5 indicate that the dataset generated by the proposed system using Microsoft Visual Studio tool has the highest level of true positive records. Moreover, false negatives and false positive records are less than the true positive records. Figure 5 shows that in documents dataset there exist six true positives, while the ratio of false positive records and false negative records are 1 and 1 respectively. The values of precision, recall and F-measure are 85.71%, 85.71% and 85.701% respectively, which show highest level of accuracy of results.

From Figure 6, it is concluded that the results are satisfactory as 85.71% of both precision and recall, while 85.701% F-measure are achieved. Moreover, the efficiency and goodness of dataset results show that true positives are maximum in number while false positives and false negatives are minimum in number. The F-measure of 85.70% points towards the measurement of error percentage. The error percentage is calculated using the following formula

$$\frac{\left(\begin{array}{c}\text{Number of same}\\ \text{domain and same}\\ \text{documents existed}\end{array}\right)-\left(\begin{array}{c}\text{Number of same}\\ \text{domain and same scholar}\\ \text{documents retrieved}\end{array}\right)}{\text{Number of total documents in dataset}}\times 100.$$

$$\text{Percentage of error} = \frac{16-2}{24}\times 100 = 16.7\%. \quad (5)$$

Thus, the percentage of error is low. Therefore, it is concluded that the evaluation results generated on the dataset by applying the proposed technique are accurate.

The evaluation of performance of results is of vital importance. In this section, some representative research publications of document clustering are compared with our proposed document clustering technique. Table 2 provides performance evaluation of studies discussed in the literature and their comparison with the proposed algorithm.

The main challenging task of our study was enhancing and improving accuracy as well as efficiency of agglomerative clustering algorithm during document clustering. For this, a research process consisting of five phases was devised. After data analysis, preprocessing was applied on the dataset. The results obtained from preprocessing were used to apply the proposed single-linkage AHC algorithm. In the next step, cosine similarity measure was applied. In the last phase, performance evaluation was carried out. The end results showed highest level of precision, recall and F-measure when compared to all other evaluation mechanisms adopted for different document clustering techniques.

Thus, from the results, it can be concluded that the proposed agglomerative clustering algorithm has higher efficiency and effectiveness for document clustering. The efficiency of the proposed single-linkage agglomerative clustering algorithm is further elaborated by comparing it with pre-document clustering algorithms as shown in Table 2. Apart from this, the achieved evaluation results from dataset also indicate that the proposed agglomerative clustering algorithm is generating highest F-measure compared to other existing F-measure results.

This study is limited to clustering document dataset by agglomerative clustering only. However, other clustering techniques can also be considered. For evaluating the efficiency and accuracy of the algorithm, external as well as internal quality measures can be adopted. In this research, 100% precision and recall are not achieved. Nevertheless, the results are justified in the form of research synthesis which provides a detailed comparison with existing literature studies. Moreover, since self-developed document dataset is limited in size and domains of publications of research scholars, this implies that the dataset needs to be extended.

1. Mishra, R. K., Saini, K. and Bagri, S., Text Document Clustering on the basis of Inter passage approach by using K-means. In International Conference on Computing, Communication and Automation (ICCCA), Greater Noida, India, 2015, pp. 110–113.
2. Sathiyakumari, K., Manimekalai, G., Preamsudha, V. and Scholar, M. P., A survey on various approaches in document clustering. *Int. J. Comput. Technol. Appl*., 2011, **2**(5), 1534–1539.
3. Sunanda, P. and Vineela, A., An agglomerative hierarchical clustering for hybrid recommender systems. In International Conference on Power, Control, Communication and Computational Technologies for Sustainable Growth (PCCCTSG), Karnool, India, 11–12 December 2015, pp. 283–288.
4. Halkidi, M., Batistakis, Y. and Vazirgiannis, M., Clustering algorithms and validity measures. In Proceedings of the Thirteenth International Conference on Scientific and Statistical Database Management (SSDBM), Fairfax, USA, 18–20 July 2001, pp. 3–22.
5. Bhagat, A., Kshirsagar, N., Khodke, P., Dongre, K. and Ali, S., Penalty parameter selection for hierarchical data stream clustering. *Proc. Comput. Sci.*, 2016, **79**, 24–31.
6. Rafi, M., Maujood, M., Fazal, M. M. and Ali, S. M., A comparison of two suffix tree-based document clustering algorithms. In International Conference on Information and Emerging Technologies, Karachi, Pakistan, 14–16 June 2010.
7. Sun, H. S. H., Liu, Z. L. Z. and Kong, L. K. L., A document clustering method based on hierarchical algorithm with model clustering. In Proceedings of the 22nd International Conference on Advanced Information Networking and Applications, Okinawa, Japan, 25–28 March 2008, pp. 1229–1233.
8. Lu, Y. and Wan, Y., PHA: a fast potential-based hierarchical agglomerative clustering method. *Pattern Recognit.*, 2013, **46**(5), 1227–1239.
9. Liu, F., Wei, Y., Ren, M., Hou, X. and Liu, Y., An agglomerative hierarchical clustering algorithm based on global distance measurement. In Seventeenth International Conference on Information Technology in Medicine and Education (ITME), Huangshan, China, 13–15 November 2015, pp. 363–367.
10. García-Lapresta, J. L. and Pérez-Román, D., Consensus-based hierarchical agglomerative clustering in the context of weak orders. In IFSA World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS), Edmonton, Canada, 24–28 June 2013, pp. 1010–1015.
11. Zamir, O. and Etzioni, O., Web document clustering: a feasibility demonstration. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, 24–28 August 1998, pp. 46–54.
12. Zhang, G. Z. G., Liu, Y. L. Y., Tan, S. T. S. and Cheng, X. C. X., A novel method for hierarchical clustering of search results. In International Conference on Web Intelligence and Intelligent Agent Technology, Fremont, USA, 2–5 November 2007, pp. 181–184.
13. Zhou, S., Xu, Z. and Liu, F., Method for determining the optimal number of clusters based on agglomerative. *IEEE Trans. Neur. Net. Learn. Syst.*, 2016, **28**(12), 3007–3017.
14. Al-Anazi, S., Almahmoud, H. and Al-Turaiki, I., Finding similar documents using different clustering techniques. *Proc. Comput. Sci.*, 2016, **82**, 28–34.
15. Samat, N., Murad, M. A., Abdullah, M. T. and Atan, R., Malay documents clustering algorithm based on singular value decomposition. *J. Theor. Appl. Infor. Technol.*, 2005, **8**(2), 180–186.
16. Rajavat, A. and Gupta, M., Comparison of algorithms for document clustering. In International Conference on Computational Intelligence and Communication Networks (CICN), Bhopal, India, 14–16 November 2014, pp. 542–546.
17. Bsoul, Q., Salim, J. and Zakaria, L. Q., An intelligent document clustering approach to detect crime patterns. *Proc. Technol.*, 2013, **11**, 1181–1187.
18. Aggarwal, C. C. and Zhai, C. (eds), A survey of text clustering algorithms. In *Mining Text Data*, Springer, Boston, USA, 2012, pp. 77–128.
19. Yim, O. and Ramdeen, K. T., Hierarchical cluster analysis: comparison of three linkage measures and application to psychological data. *Quant. Meth. Psychol.*, 2015, **11**(1), 8–21.
20. Mattas, N., Kalra, P. and Mehrotra, D., Agglomerative hierarchical clustering technique for partitioning patent dataset. In Fourth International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), Noida, India, 2–4 September 2015, pp. 2–5.
21. Mishra, R. B., Modi, N. K. and Shah, R. R., Performance analysis of single and complete link during agglomerative clustering of question papers by tagging the questions and trend analysis using single link. In International Conference on Advanced Communications, Control and Computing Technologies (ICACCCT), Ramanathapuram, India, 8–10 May 2014, vol. 978, pp. 616–618.
22. Mohbey, K. K., An experimental survey on single linkage clustering. *Int. J. Comput. Appl.*, 2013, **76**(17), 6–10.
23. Takumi, S. and Miyamoto, S., Comparing different methods of agglomerative hierarchical clustering with pairwise constraints. In Joint Sixth International Conference on Soft Computing and Intelligent Systems (SCIS*)* and 13th International Symposium on

Advanced Intelligent Systems (ISIS), Kobe, Japan, 20–24 November 2012, pp. 1545–1550.
24. Wu, J. W., Tseng, J. C. and Tsai, W. N., An efficient linear text segmentation algorithm using hierarchical agglomerative clustering. In Seventh International Conference on Computational Intelligence and Security (CIS), Hainan, China, 3–4 December 2011, pp. 1081–1085.
25. Wong, K. C., A short survey on data clustering algorithms. In Second International Conference on Soft Computing and Machine Intelligence, 2015, pp. 64–68.
26. Zamir, O. and Etzioni, O., Web document clustering: a feasibility demonstration. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, 24–28 August 1998, pp. 46–54.
27. Gupta, M. and Rajavat, A., Comparison of algorithms for document clustering. In IEEE International Conference on Computational Intelligence and Communication Networks, Bhopal, India, 14–16 November 2014, pp. 541–545.
28. Sharma, D. and Jain, S., Evaluation of stemming and stop word techniques on text classification problem. *Int. J. Sci. Res. Comput. Sci. Eng.*, 2015, **3**(2), 1–4.
29. Xue, X. and Zhou, Z., Distributional features for text categorization. *IEEE Trans. Knowl. Data Eng.*, 2009, **21**(3), 428–442.
30. Raulji, J. and Saini, J., Stop-word removal algorithm and its implementation for Sanskrit language. *Int. J. Comput. Appl.*, 2016, **150**(2), 15–17.
31. Porter, M. F., An algorithm for suffix stripping. *Program. Electron. Lib. Inf. Syst.*, 1980, **14**(3), 130–137.
32. Agnihotri, D., Verma, K. and Tripathi, P., Pattern and cluster mining on text data. In Fourth International Conference on Communication Systems and Network Technologies Bhopal, India, 7–9 April 2014, pp. 428–432.
33. Subhashini, R. and Kumar, V. J. S., Evaluating the performance of similarity measures used in document clustering and information retrieval. In First International Conference on Integrated Intelligent Computing (ICIIC), Bangalore, India, 5–7 August 2010, pp. 27–31.
34. Liao, H. and Xu, Z., Approaches to manage hesitant fuzzy linguistic information based on the cosine distance and similarity measures for HFLTSs and their application in qualitative decision making. *Expert Syst. Appl.*, 2015, **42**(12), 5328–5336.
35. Zhao, Y. and Karypis, G., Evaluation of hierarchical clustering algorithms for document dataset. In Proceedings of the 11th International Conference on Information and Knowledge Management, Virginia, USA, 4–9 November 2002, pp. 515–524.
36. Nurazzah, A. R. *et al.*, Malay document clustering using complete linkage clustering technique with cosine coefficient. In IEEE International Conference on Open Systems (ICOS), Bandar Melaka, Malaysia, 24–26 August 2015, pp. 103–107.
37. Ravindran, R. M. and Thanamani, A. S., K-means document clustering using vector space model. *Int. J. Data Min.*, 2015, **5**(2), 10–14.
38. Wang, Y. *et al.*, Improved text clustering algorithm and application in microblogging public opinion analysis. In Proceedings of the Fourth World Congress on Software Engineering (WCSE), Hong Kong, China, 3–4 December 2013, pp. 27–31.
39. Chim, H., Deng, X. and Member, S., Efficient phrase-based document similarity for clustering. *IEEE Trans. Knowl. Data Eng.*, 2008, **20**(9), 1217–1229.
40. Sun, H., Liu, Z. and Kong, L., A document clustering method based on hierarchical algorithm with model clustering. In 22nd International Conference on Advanced Information Networking and Applications, Okinawa, Japan, 25–28 March 2008, pp. 1229–1233.

Received 24 September 2018; revised accepted 10 April 2019

doi: 10.18520/cs/v117/i6/1045-1053