Cluster Computing (2019) 22:S3213–S3230
https://doi.org/10.1007/s10586-018-2023-4

Clustering of biomedical documents using ontology-based TF-IGM enriched semantic smoothing model for telemedicine applications

R. Sandhiya[1] · M. Sundarambal[2]

Received: 26 December 2017 / Revised: 31 January 2018 / Accepted: 2 February 2018 / Published online: 20 March 2018

Abstract

Clustering of biomedical documents has become a vital research concept due to its importance in the clinical and telemedicine applications. The clustering of the medical documents is being considered as a major issue because of its unstructured nature. This paper focuses on developing an efficient document clustering approach for the medical documents to be utilized in telemedicine applications. Most existing models utilize n-gram techniques for phrase identification and term, concept or semantic based models for clustering applications. However n-gram does not perform well when the original document has been modified while only hybrid models provide relatively improved clustering. The proposed document clustering approach is named as enriched semantic smoothing model which has been developed on the concept of Mesh ontology. As the semantic smoothing model is not effective in handling the density of general words, an improved model with term frequency and inverse gravity moment (TF-IGM) factor and improved background elimination is used. Unlike term frequency and inverse document frequency), TF-IGM precisely measure the class distinguishing power of a term by making use of the fine-grained term distribution across different classes of text in documents. The modified n-gram technique, which detects the cases of substitution and deletion in the documents and averts them, improves the phrases identification. The clustering efficiency of the k-means clustering and hierarchical clustering algorithms is improved by utilizing the proposed model. The experiments are made on Mesh ontology based PubMed documents with similarity measures and cluster validity indexes used for comparisons. The results show that the proposed approach of medical document clustering is highly accurate and thus improves the concepts of clinical practices and telemedicine.

Keywords Document clustering · Telemedicine · n-gram · Mesh ontology · Semantic smoothing · Term frequency · k-means · Hierarchical clustering

1 Introduction

Document clustering has been broadly utilized in different scientific fields for supporting web indexes, content mining, and Information Retrieval [1]. It has been also utilized as a post-retrieval instrument for sorting query results into topical subjects. These composed outcomes can be intuitively browsed, envisioned, and explored by the clients. Hierarchical clustering, specifically, is used to create subject hierarchies. Depending upon the basic algorithmic procedure, hierarchical clustering algorithms can be classified into agglomerative (base up) and divisive (top-down) [2,3]. Divisive methodologies begin with all documents in a similar root cluster and work by iteratively split each cluster into various littler ones until the point when an end rule is met for each cluster. Partitional clustering algorithms are appropriate for clustering vast informational collections because of their low computational requirements (linear in the quantity of documents) [2]. Moreover, often partitional strategies have been found to prompt preferable clustering solutions over agglomerative algorithms [4]. Partitional clustering techniques can likewise be utilized to get a hierarchical clustering arrangement by means of rehashed

✉ R. Sandhiya
rsandhiya@cit.edu.in

M. Sundarambal
msundarambal@cit.edu.in

[1] Department of Information Technology, Coimbatore Institute of Technology, Coimbatore, Tamil Nadu 641014, India

[2] Department of Electrical and Electronics Engineering, Coimbatore Institute of Technology, Coimbatore, Tamil Nadu 641014, India

use of the level partitional algorithm such as Bisecting K-Means.

Document clustering has been explored for utilization in various ranges, for example, processing collections of document [5], enhancing the precision and review in data recovery frameworks [5], naturally creating hierarchical clusters of documents [6] and so forth. Document clustering is a principal method for content summarization [7], cluster based data recovery [8] and programmed point extraction [9]. Doctors' interpretations of images, signals, or other clinical information, are composed as unstructured free-content reports or documents. Such documents are extremely hard to mine; even specialists from a same field cannot agree on unambiguous terms to be utilized as a part of depicting a patient's condition [10]. Creating strategies or methods to sorting out large unstructured clinical documents into few important groups will help clients to discover their searches, separate significant data and finding patterns and examples covered up inside these documents more efficiently. As only the cluster that will contain relevant documents is considered, it will enhance viability and effectiveness [11]. The delivered clusters contain collections of documents that are more similar [12]. Along these lines, the objective of discovering high-quality document clustering algorithms is to decide an arrangement of clusters with the aim of inter-cluster similarities is minimized and intra-cluster similarity is maximized. Since knowledge extraction and information mining will be connected to the created clusters, accomplishing excellent clustering arrangement is critical. As the clinical notes have an incredible use in pharmacy store, it is necessary to lessen falsification and avoid drug abuse. Document clustering for clinical notes has been to research for clustering them into significant groups. This is done principally to find imperative examples. This has demonstrated by expanding rate, effectiveness and accuracy of managing data in the area of medical diagnoses.

Recent research has been focused on how to incorporate domain ontology as basic knowledge to document clustering procedure. It demonstrates that ontology can enhance document clustering execution with its concept hierarchy knowledge [13–15]. In [16], ontology based method has been developed for biomedical document clustering. However the developed model has certain issues of general word density handling, low performance during document modification, etc. This paper focuses on resolving those issues and developing an enhanced model of biomedical document clustering to improve the clinical practices. The remainder of the article is organized as follows: Sect. 2 describes recent research works related to the biomedical document clustering. Section 3 presents the clustering methods. Section 4 explains the proposed clustering model while Sect. 5 evaluates its performance. Section 6 makes a conclusion of the proposed research model.

2 Related works

Clustering has been utilized as a part of many fields and by various methodologies [17–19]. Many examinations have been completed on medicinal document clustering. Customarily, local-content (LC) data of documents from the informational collection to be clustered has been used for clustering [20], where each document is spoken to by "bag of words," bringing about a weighted vector as per the vector space model, and after that, clustering is completed on weighted vectors. In any case, MEDLINE documents have some distinct elements that could be used for upgrading the clustering execution. To start with, for every MEDLINE document, PubMed gives an arrangement of related articles in the entire MEDLINE gathering, which is pre-processed by looking at words from the title, the theoretical, and the medical subject heading (MeSH) utilizing a word-weighting algorithm [21]. Theodosiou et al. [22] have influenced utilization of this sort of worldwide to content (GC) data for clustering MEDLINE documents. Second, a large portion of MEDLINE documents have been commented on by the MeSH (http://www.nlm.nih.gov/work/). The MeSH is a controlled vocabulary thesaurus with an arrangement of depiction terms sorted out in a hierarchical structure where general ideas show up at the best and particular ideas show up at the base [23]. Rich semantic data in MeSHs can enhance the execution of clustering MEDLINE documents. Yoo et al. [24] adjusted terms in documents into MeSH ideas as indicated by the MeSH thesaurus, demonstrating the change in clustering execution under different techniques, for example, k-means, bisecting k-means, and suffix tree clustering. A comparative procedure was additionally utilized as a part of term reweighting of document clustering [25]. In any case, this strategy never again utilizes unique writings, causing an issue that critical content data in unique documents might be lost. Generally, existing methodologies on biomedical document clustering have two genuine limitations: (i) utilizing just a single or two sorts of data and (ii) lacking successful algorithms to incorporate diverse sorts of data.

Recently, an approach of directly consolidating both the LC and MeSH-semantic (MS) similarities has been proposed, exactly demonstrating the performance advantage over that utilizing just one of the two similarities [26]. The linear combination technique has been likewise utilized as a part of different bioinformatics issues, for example, gene clustering with various information (or requirements), including Gene Ontology, metabolic systems, and quality articulation. For this case, once the informational indexes are incorporated, we can utilize an assortment of clustering models, e.g., hierarchical clustering [27], Gaussian mixture model [28], k-medoids [29], and Markov random fields [30]. But, this methodology has approximately three basic disadvantages in document clustering. Initially, the genuine closeness is not really a straightforward direct connection between various sorts of similarities (sources). Second, the nature of comparability in an informational index may not be notwithstanding for all document sets. A few sets are more dependable and ought to be given careful consideration. Third, it is hard to pick a reasonable weighting arrangement to adjust at least three unique sorts of similarities in coordinating them.

Semisupervised clustering algorithms consolidate past knowledge to enhance the clustering performance. Constrained k-means (SS-K-means) is a prior semisupervised clustering algorithm, which was straightforwardly created from k-means [31]. SS-K-means tries to relegate each example to the cluster with the most comparable centroid, unless, in the same time, the imperatives are disregarded. Spectral clustering is a very much acknowledged strategy for clustering hubs over a diagram (or a nearness network), where clustering is a graph cut issue that can be resolved by lattice follow streamlining. Normalized cut (NCut) [32] is a common one, which limits the cost of inter-cluster edges under the limitation of the volume (the entirety of node degrees) in clusters. Ji et al. [33] proposed a constrained normalized cut technique, which consolidates ML limitations into the info contiguousness grid yet does not consider CL imperatives, which must be critical for enhancing clustering performance.

Health care data by means of Clinical source has seen a noteworthy increment in both volume and assortment. In [34], authors have proposed a strategy for information extraction from online medical forums. Lexico-syntactic pattern from annotated data with seed vocabularies is utilized to separate two element sorts, specifically, medications and medications. This proposed framework extracts symptom names and the medicines, which are truant from unique vocabulary lexicon. In [35], organized information is removed from clinical data utilizing rule based strategy alongside machine learning system and feature engineering. In [36], authors have proposed a plan to extricate entity extraction utilizing local grammar. In this technique, medicinal related data from French clinical notes is separated utilizing principle based local syntax. The drawback of this strategy is that an incredible human exertion and additional time are required. In [37], authors have taken a shot at a strategy called computerized de-distinguishing proof and extensive scale assessment of clinical notes. A NLP apparatus is utilized for automatic identification of large set of various clinical sets. This proposed approach helps in de-recognizable proof of millions of clinical documents.

In [38], the authors manufacture a coordinating framework for extricating prescription names and manifestation names from clinical notes; at that point the nonnegative matrix factorization (NMF) and multi-see NMF are connected to cluster clinical notes into significant clusters in view of test include lattices. However clustering these clinical documents depends only on the side effect/pharmaceutical names yet factors like patients' demographical data are excluded for clustering process. In [39], the authors proposed a mixed integer model where clustering clinical divisions to level out related bed necessities after some time, thus fusing occasional impacts of individual offices. In [40], the authors developed a joined clustering technique utilizing measurement diminishment and K-means clustering based on vector clustering and Silhouette measure. This approach conquers the sparseness condition in clustering however expert knowledge is not considered for extricating term particular substance. In [41], another approach is proposed for clustering MEDLINE abstracts in light of an augmentation of a transformative algorithm which is the hereditary algorithm consolidated with a Vector Space Model and an agglomerative algorithm. In Telemedicine applications, the major source of transmission is through wireless body area networks however transmitting real-time multimedia data of the patients using geographic multipath routing models [42] will be more efficient and will be insightful for future. This type of research works can improve the medical document clustering through novel methods. Though there have extensive research works on document clustering, the exploitation of the same for clinical and biomedical documents faces performance hindrances due to various factors. This limitation forms the motivation for this proposed clustering model.

3 Clustering methods

By utilizing the proposed model of enriched semantic smoothing, the document clustering can be improved. The performance of partitioning [43] and hierarchical techniques [2] will be explored utilizing this model. K-means is a partitional clustering strategy that takes as input a dataset and makes a level, non-hierarchical clustering solution that comprises of K clusters. At first, the algorithm picks K documents as initial centroids and calculates the similarity between each document and all cluster centroids. Each document is allotted to its nearest cluster centroid. The following stage is the "centroid re-calculation": All documents assigned to a similar centroid are averaged to compute another centroid. The clusters (and their centroids) are balanced iteratively by the algorithm until convergence (i.e., the centroids do not change fundamentally). In most cases, K-means algorithm keeps repeating until the point that the centroids do not change fundamentally between iterations. In any case, because of the way that the centroids seldom quit moving completely and additional time is required to check for minimal movement (or greatest cluster cohesion), the algorithm keeps running at most iterations. Cluster cohesion (overall cluster similarity) is characterized as the sum of the average pairwise similarities between all documents allotted to a cluster and equivalents to $\|c\|^2$. The overall cohesion of a clustering solution is then estimated as:

$$Overall\ Clustering\ Cohession = \sum\nolimits_{r=1}^{K} \|\mathbf{c}\|^2 \qquad (1)$$

Hierarchical document clustering is observed to be superior to the partitional clustering techniques. The primary purpose of hierarchical document clustering is to construct a hierarchical tree of clusters whose leaf nodes represent the subset of a document collection. In addition, this strategy can be additionally arranged into agglomerative and divisive methodologies, which work in a bottom up and top-down form, separately. An agglomerative clustering iteratively blends two most comparable clusters until the point when a terminative condition is fulfilled. A divisive technique begins with one cluster, which comprises of all documents, and recursively parts one cluster into smaller sub-clusters until the point that termination is satisfied. There are three fundamental stages in hierarchical clustering. In the principal stage, the key terms will be separated from the document set, and each document is pre-processed into the designated representation for the accompanying mining process. In this stage, a feature selection technique will be utilized to viably minimize the irrelevant terms for each document. In the second stage, to find a set of relevant terms effectively, a mining algorithm is utilized for text. By utilizing this algorithm, similarity is evaluated for each term in view of its recurrence to separate the level of significance of the term inside a document in the mining procedure. The derived items contain key terms to be viewed as the labels of candidate clusters. In the last stage, the documents will be clustered into a hierarchical cluster tree depending upon these candidate clusters. The cluster tree will be built in a top-down manner to recursively choose the parent clusters at level $k-1$ for separating the documents into its reasonable child clusters at level k.

4 Ontology based enriched semantic smoothing approach

In this section, the proposed model of enriched semantic smoothing is illustrated as in Fig. 1. Most existing semantic smoothing models are not compelling for partitional clustering. The key reason is that the current semantic smoothing models are viable for dealing with the sparsity of core words which affects the agglomerative clustering. But, due to the density of general words, the sparsity of core words is not effective for "reducing" general words. The capacity of the semantic smoothing model for "reducing" general words and in the meantime keeps the capacity to appoint sensible counts to concealed core words can be upgraded based on the guideline of term frequency weighting model.

TF-IDF is the most commonly utilized term frequency weighting model. However TF-IDF factor is not completely effective in all text document classification and influences the overall precision. Consequently the proposed model uses TF-IGM factor [44] for precisely measure the class distinguishing power of a term. It makes utilization of the

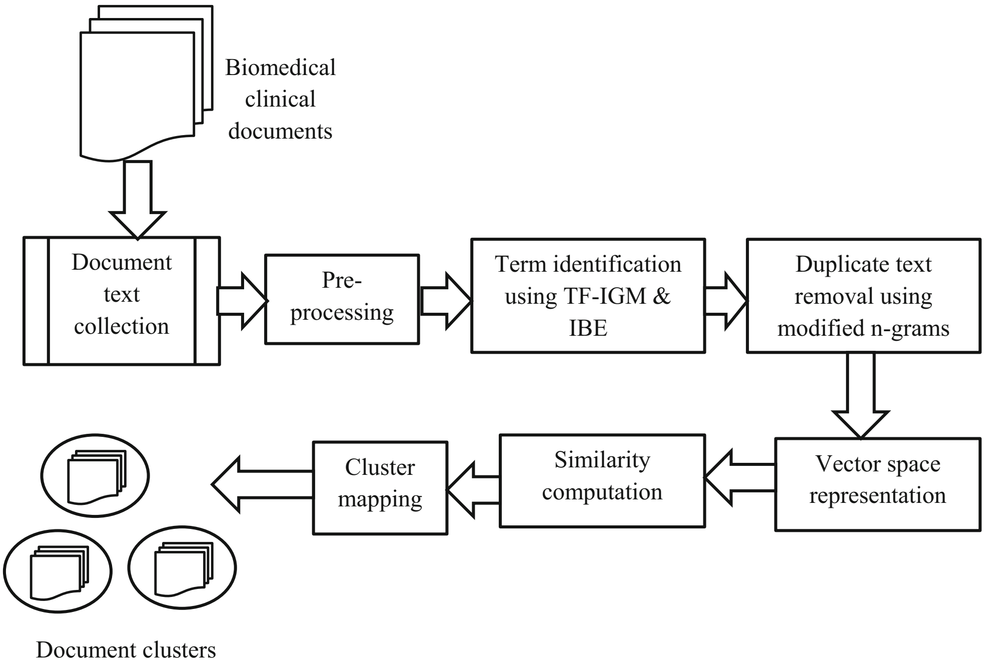
Biomedical clinical documents
Document text collection
Pre-processing
Term identification using TF-IGM & IBE
Duplicate text removal using modified n-grams
Vector space representation
Similarity computation
Cluster mapping
Document clusters

Fig. 1 Ontology based document clustering using enriched semantic smoothing model

fine-grained term distribution over various classes of text in documents. Initially the improved background elimination is used for undesirable background data removal followed by the TF-IGM in the proposed model of ontology based enriched semantic smoothing model.

4.1 Improved background elimination (IBE)

Generally the background model signified as $p(w|C)$ is computed by the frequency count of word w background corpus $(c\ (w|C))$. As a matter of fact, the background model $p(w|C)$ to a great extent impacts the probability of the general word created by a given document or cluster. Since the general words regularly have high frequency in the background corpus, and that influences its model estimation to very huge. Thus an enhanced background model is presented in which the coefficient α is set as zero as to remove the background model in the simple language model. In the semantic smoothing model without background model, the words' probability relies only upon the frequency count of the words and the translation model. The estimation equations of the simple language model without background are given in (2) for document model and (3) for cluster model.

$$p_b'\ (w|d) = p_{ml}(w|d)\delta \tag{2}$$

$$p_b'\left(w|c_j\right) = p_{ml}(w|c_j)\delta \tag{3}$$

where c_j is the cluster, d is the document, w is the word, p denotes likelihood with b as simple language model and ml as the multiword phrase, δ is the term factor.

4.2 TF-IGM

Most supervised term weighting schemes fuse the measurements of the inter-class distribution (over various classes) and intra-class distribution (inside a class) of a term. In any case, it is questioned that the intra-class distribution factor has any clearly constructive outcome on term weighting. In this view, the weight of a term ought to be determined principally by its class distinguishing power, which is embodied basically by its uneven distribution over various classes. Usually, the all the more uniformly a term disseminates in the text dataset, the weaker its class distinguishing power is. The examples are some basic words, which have no peculiarity or representativeness for any class. Despite this, a term with class representativeness or uniqueness frequently appear in only a single class or few classes of text, and clearly it has solid class distinguishing power. Obviously, some class-particular terms may periodically show up in different classes of text. In other words, however these terms may show up in various classes of text, yet they happen habitually in only a couple of classes or even a solitary class of text. Due to the high non-consistency of their distribution in general dataset, these class-particular terms can be utilized to recognize text among various classes and ought to be assigned more noteworthy weights. Generally, a term with more concentrated inter-class distribution has a tendency to have more grounded class distinguishing power than others. In this way, a term can be weighted by its interclass distribution focus or non-consistency. The non-consistency of sample distribution is measured usually by the entropy in data theory or variance in arithmetic. Hence, in this investigation another factual model called "inverse gravity moment" (IGM) is proposed to quantify the non-consistency or concentration level of a term's inter-class distribution, which reflects the term's class distinguishing power. On this premise, an appropriate weight is assigned to the term.

To quantify the inter-class distribution convergence of term t_k, as a matter of first importance, one needs to sort every one of the frequencies of t_k's happening in the individual classes of text in diving request. The resulting sorted list is $f_{k1} \geq f_{k2} \geq \cdots \geq f_{km}$, where f_{kr} $(r = 1, 2, \ldots, m)$ is the frequency of t_k's happening in the r-th class of text in the wake of being arranged, and m is the quantity of classes. In the event that f_{kr} is viewed as a class-specific "gravity", at that point for the sorted list, the head on the left is heavier than the tail on the right, and the focal point of gravity of the general inter-class distribution is inclination to one side. Clearly, the fewer classes the events of a term gathers in, the nearer the focal point of gravity is to one side head. Particularly, when the term happens in just a single class of text, the focal point of gravity is at the beginning position (r = 1). Just when the inter-class distribution is uniform, that is, $f_{k1} = f_{k2} = \cdots = f_{km}$, the focal point of gravity is situated at the inside position (m/2). In this manner, the inter-class distribution centralization of a term can be reflected by the position of the gravity focus. However, we do not straightforwardly embrace the position of the gravity focus, yet rather another metric related with it, to gauge the inter-class distribution concentration. For the class-specific "gravity" f_{kr}, if its rank r is viewed as the separation to the root 0, the result of $f_{kr}.r$ is called "gravity moment" (GM) in material science. For an aggregate frequency of a term happening in the text corpus, the more concentrated the term inter-class distribution is, the shorter the separation of the gravity focus is to the origin and the less the sum of all the class-specific gravity moments is, and in the meantime the higher the most maximum class-specific frequency (f_{k1}) is, too. The above realities demonstrate that the concentration level of the term inter-class distribution is corresponding to the proportional estimation of the total gravity moment. In this way, a new statistical model called "inverse gravity moment" (IGM) is proposed to gauge the measure the inter-class distribution concentration of a term, which is defined as

$$igm\left(t_k\right) = \frac{f_{k1}}{\sum_{r=1}^{m} f_{kr}.r} \tag{4}$$

where $igm\left(t_k\right)$ denotes the inverse gravity moment of the inter- class distribution of term t_k, and f_{kr} $(r = 1, 2, \ldots, m)$ are the frequencies of t_k's occurring in different classes, which are sorted in descending order with r being the rank. Usually, the frequency, f_{kr}, refers to the class-specific document frequency (DF), i.e., the number of documents containing the term tk in the r-th class, denoted as df_{kr}.

The inverse gravity moment of term inter-class distribution ranges from $2/\left(\left(1+m\right).m\right)$ to 1.0. Since the first element, f_{k1}, is the maximum in the list of $\{f_{kr}|r = 1, 2, \ldots, m\}$ in descending order, it can be rewritten as

$$igm\left(t_k\right) = \frac{1}{\sum_{r=1}^{m} \frac{f_{kr}}{\max_{1 \le i \le m}\left(f_{ki}\right)}.r} \tag{5}$$

This equation demonstrates that the IGM of a term is the inverse of the aggregate gravity moment computed from the normalized frequencies of the term's event in all the individual classes. The term weighting in a document by TF-IGM ought to be determined by its significance in the document and its contribution to text classification, which compare separately to the local and global weighting factors in term weighting. A term's contribution t to text classification relies upon its class distinguishing power which is reflected by its inter-class distribution focus. The higher the focus level is the more prominent weight ought to be allotted to the term. The previous can be measured by the IGM model. Thus, rather than the customary IDF factor, another worldwide factor in term weighting is characterized in view of the IGM metric of the term

$$w_g\left(t_k\right) = 1 + \lambda.igm\left(t_k\right) \tag{6}$$

where $w_g\left(t_k\right)$ denotes the IGM-based global weighting factor of term t_k, and λ is an adjustable coefficient. The purpose of introducing the coefficient λ is to keep the relative balance between the global and local factors in the weight of a term.

The TF-IGM weight of term t_k in document d is the product of the TF-based local weighting factor and the IGM-based global weighting factor, i.e., $w\left(t_k, d\right) = w_l\left(t_k, d\right).w_g\left(t_k\right)$, which is expressed as

$$w\left(t_k, d\right) = tf_{kd}.\left(1 + \lambda.\frac{f_{k1}}{\sum_{r=1}^{m} f_{kr}.r}\right) \tag{7}$$

$$w\left(t_k, d\right) = \sqrt{tf_{kd}}.\left(1 + \lambda.\frac{f_{k1}}{\sum_{r=1}^{m} f_{kr}.r}\right) \tag{8}$$

where the TF of t_k in d, $tf_{kd} > 0$. Otherwise, $w\left(t_k, d\right) = 0$ if $tf_{kd} = 0$. The frequency f_{kr} $(r = 1, 2, \ldots, m)$ usually refers to the class-specific DF of the term.

4.3 Modified n-gram technique

An n-gram is a string of n-adjacent words that occur inside a text. Contrasting n-grams has demonstrated to be a powerful strategy for identifying duplicate or similar documents that has been connected to a scope of issues, including the distinguishing proof of text reuse in journalism and plagiarism detection. A limitation of utilizing n-gram overlap to determine document similarity is that it does not perform well when the first document has been modified (Example, by paraphrasing) [45]. Actually, systems in view of n-gram overlap can be tricked utilizing extremely simple changes to a document. Addition, deletion, or substitution of even a solitary token in a text brings about the mismatch of at least one n-gram. If every nth word in a text is altered in some way, at that point the two documents would have no n-grams of length n in common and measurements namely the containment measure would neglect to distinguish the similarity between them. To maintain a strategic distance from this issue modified n-grams is proposed to be utilized. These are n-grams which are gotten from those found in one of the documents and are intended to reflect the progressions that may happen if a document was modified in order to disguise the fact that it is a duplicate. Two techniques for changing the n-grams were investigated: substitution and deletion.

The main type of modified n-grams is made by substituting one of the words in the n-gram with one of its equivalent words from the Unified Medical Language System (UMLS) Metathesaurus. The reference is first go through MetaMap and each term mapped onto an arrangement of potential Concept Unique Identifiers (CUIs). The MRCONSO table in the UMLS Metathesaurus records different routes in which each CUI is depicted in the different vocabularies that are utilized to make the UMLS Metathesaurus. This table is utilized to create a set of alternate terms that can be substituted for each term in the n-gram. A list of modified n-grams is then made by picking one of the terms in the n-gram and substituting it with one of the alternative terms from the table.

A second type of modified n-grams is created by essentially deleting words from the n-gram. If $w_1, w_2, \ldots, w_n$ is a n-gram, then n-2 modified n-grams can be made by expelling one of $w_2, w_3, \ldots, w_{n-1}$. Modified n-grams are not made by eliminating the first or last word since doing as such would just copy existing n-grams of order n− 1. Unigrams only comprise of one word and are too short for deleted n-grams to be generated from them.

Let A and B be two documents. The overlap between these two documents can be determined using the containment measure

$$score_n\left(A, B\right) = \frac{\sum_{ngram \in B} count\left(ngram, A\right)}{\sum_{ngram \in B} count\left(ngram, B\right)} \tag{9}$$

where $count\ (ngram, A)$ is the number of times n-gram appears in A

To recognize duplicate citations, modified n-grams are generated for the document that is doubted as a duplicate. Correlation between the documents is then completed by determining the extent of n-grams in B which likewise happen as n-grams in A or as modified n-grams produced from A. For every n-gram in A, the list of conceivable modified n-grams is made, signified as $mod\ (ngram)$. The first n-gram $ngram$ is likewise incorporated into $mod\ (ngram)$. The modified count for the number of events of a n-gram in A, $mod\_count\ (ngram, A)$, is then calculated as the number of times it shows up in $mod\ (ngrams)$, that is,

$$mod\_count\ (ngram, A) = \sum\nolimits_{ngram' \in mod(ngram)} count\ \left(ngram', A\right) \tag{10}$$

The deletion and substitution approaches produce huge quantities of modified n-grams. This implies the quantity of shared n-grams can surpass the aggregate number of n-grams in B, prompting a score more noteworthy than 1. To maintain a strategic distance from this, the overlap counts are limited by the quantity of times that n-gram shows up in B. Consequently, the text reuse detection score, $score_n\ (A, B)$, is computed as:

$$score_n\ (A, B) = \frac{\sum_{ngram \in B} \min\left(mod.count\ (ngram, A), count\ (ngram, B)\right)}{\sum_{ngram \in B} count\ (ngram, B)} \tag{11}$$

N-grams do not happen with equal frequency in Medline/PubMed: the fact that a couple of references has an uncommon n-gram in common is substantially more stronger proof that they are copies than if they shared a n-gram that occurs in numerous references. In the proposed model, the likelihood of every n-gram is evaluated in the correlation of reference sets to determine the duplicate detection.

4.4 Building enriched semantic smoothing model

The proposed model of semantic smoothing is improved by the inclusion of TF-IGM and modified n-grams for the concept extraction and analysis. The phrase model has been efficiently modelled using the modified n-gram. The model can be organized as:

$$Concept\ model = \lambda.w\ (t_k, d) + score_n\ (A, B)\ mod\ (ngram) \tag{12}$$

Thus the enriched semantic smoothing model can be organized and utilized for enhanced clustering of the biomedical documents. The experimental results can justify the theoretical concepts.

5 Experimental results and evaluation

5.1 Datasets and validity measures

The evaluation of the proposed model has been performed in MATLAB using the 350 medical documents collected from PubMed based on seven categories namely neoplasms, viral diseases, cardiovascular diseases, Intestinal diseases, eye diseases, respiratory diseases and skin diseases. Each disease related documents are of three sub-types of documents; hence amounting to 21 document classes. The accuracy and the quality of the clustering results of the proposed model can be examined by using the cluster validity measures [46] namely Silhouette Index, Fowlkes–Mallows (FM) Index, Jaccard Index, Dunn Index and Davies–Bouldin Index.

5.1.1 Silhouette Validity Index

It is a measure utilized for confirming the precision in task of data points into the suitable clusters. This strategy processes the silhouette width for every data point, average silhouette width for each cluster and overall average silhouette width for the total dataset. The silhouettes width S_i of i-th data point is given as

$$S_i = \frac{b_i - a_i}{\max\left(a_i, b_i\right)} \tag{13}$$

where a_i is average dissimilarity of i-th data point to every other point in a similar cluster; b_i is least of average dissimilarity of i-th data point to all data focuses in other cluster. An estimation of S_i near 1 demonstrates that the data point is allotted to an exceptionally suitable cluster. On the off chance that S_i is near zero, it implies that that data half quart could be dole out to another nearest cluster also on the grounds that it is equidistant from both the clusters. On the off chance that S_i is near – 1, it implies that data is misclassified and lies somewhere in the middle of the clusters. The overall average silhouette width for the whole data set is the average S_i for all data focuses in the entire dataset. The largest overall average silhouette shows the best clustering. In this manner, the quantity of cluster with the most extreme general average silhouette width is taken as the ideal number of the clusters.

5.1.2 FM Index

The FM Index evaluates the similarity between the clusters returned by the clustering algorithm and the benchmark classifications. The estimation of the FM list is in the vicinity of

Table 1 Term, concept and semantic smoothing weights of sample documents

Domain	Query term	Sample documents	Term-based weight TF-IGM	Concept weight	Semantic smoothing weight
Neoplasm	Carcinoma	N1	0.1025	0.0154	0.2727
		N2	0.0384	0.0154	0.2032
		N3	0.1739	0.032	0.1793
Respiratory diseases	Pneumonia	RD1	0.5454	0.0923	0.1006
		RD2	0.0336	0.2298	0.083
		RD3	0.0487	0.3311	0.106
Cardiac diseases	heart failure	CVD1	0	0.0143	0.075
		CVD2	0.4093	0.32	0.368
		CVD3	0.2518	0.2876	0.2
Eye infection	Glaucoma	ED1	0	0.121	0
		ED2	0.0215	0.2	0.0176
		ED3	0.202	0.067	0.0165
Viral diseases	Small pox	VD1	0	0.023	0
		VD2	0.0416	0.022	0.04
		VD3	0.0194	0.134	0.019
Intestinal diseases	Appendicitis	A1	0.2289	0.3321	0.2
		A2	0.2051	0.043	0.195
		A3	0.3137	0.043	0.2746
Skin diseases	Scabies	S1	0.2857	0.223	0.2511
		S2	0.2587	0.012	0.253
		S3	0.2631	0.004	0.2259

0 and 1, and a high esteem implies better exactness. It can be processed as

$$FM = \sqrt{\frac{TP}{TP + FP} \cdot \frac{TP}{TP + FN}} \tag{14}$$

where TP is the number of true positives, FP is the number of false positives, and FN is the number of false negatives.

5.1.3 Jaccard Index

Jaccard Index is utilized to evaluate the similarity between two datasets. Jaccard Index, which measures dissimilarity between test sets, is integral to the Jaccard coefficient and is acquired by subtracting the Jaccard coefficient from 1 (meaning identical datasets), or, identically, by isolating the distinction of the sizes of the union and the intersection of two sets by the span of the union. The Jaccard Index for the datasets A and B is defined by

$$J\left(A \cup B\right) = \frac{|A \cap B|}{|A \cup B|} \tag{15}$$

5.1.4 Dunn Index

Dunn Index is a metric for assessing clustering algorithms. The point is to distinguish sets of clusters that are minimal, with a small variance between individuals from the cluster, and well isolated, where the means of various clusters are adequately far separated, when contrasted within the cluster variance. For a given task of clusters, a higher Dunn Index indicates better clustering. Dunn Index is computed as

$$D = \min_{i=1,..n_c} \times \left\{ \min_{j=i+1,\ldots n_c} \left(\frac{d\left(c_i, c_j\right)}{\max_{k=1,\ldots n_c}\left(diam\left(c_k\right)\right)} \right) \right\} \tag{16}$$

where n_c is the number of clusters, $d\left(c_i, c_j\right)$ is the inter-cluster distance metric and $diam\left(c_k\right)$ distance of all points from mean

5.1.5 Davies–Bouldin Index

The Davies–Bouldin Index depends on similarity measure of clusters (R_{ij}) whose bases are the scattering measure of a cluster (dm_i) and the cluster dissimilarity measure (d_{ij}). The Davies–Boludin Index measures the average of simi-

Table 2 Performance result analysis for seven clusters

Model	Technique	Silhouette Index	FM Index	Jaccard Index	Dunn Index	Davies–Boludin Index
Using Euclidean measure						
Ontology based semantic smoothing	K-means	0.5867	0.4765	0.8734	0.5432	0.34
	Hierarchical single	0.6765	0.4654	0.6987	0.5545	0.3356
	Hierarchical complete	0.4876	0.4654	0.6323	0.5543	0.3675
	Hierarchical centroid	0.712	0.4654	0.6231	0.5567	0.354
Ontology based enriched semantic smoothing	K-means	0.612	0.5232	0.912	0.5812	0.4565
	Hierarchical single	0.6878	0.5543	0.768	0.5765	0.4644
	Hierarchical complete	0.5768	0.5543	0.7765	0.5887	0.4644
	Hierarchical centroid	0.7288	0.55	0.7567	0.566	0.4644
Using Pearson correlation						
Ontology based semantic smoothing	K-means	0.7989	0.8545	0.7432	0.6543	0.7112
	Hierarchical single	1	1	1	1	1
	Hierarchical complete	1	1	1	1	1
	Hierarchical centroid	1	1	1	1	1
Ontology based enriched semantic smoothing	K-means	0.8342	0.8876	0.7987	0.7121	0.7338
	Hierarchical single	1	1	1	1	1
	Hierarchical complete	1	1	1	1	1
	Hierarchical centroid	1	1	1	1	1
Using Cosine Similarity						
Ontology based semantic smoothing	K-means	0.7112	0.8456	0.7675	0.7087	0.6912
	Hierarchical single	1	1	1	1	1
	Hierarchical complete	1	1	1	1	1
	Hierarchical centroid	1	1	1	1	1
Ontology based enriched semantic smoothing	K-means	0.7986	0.8763	0.7745	0.6894	0.7288
	Hierarchical single	1	1	1	1	1
	Hierarchical complete	1	1	1	1	1
	Hierarchical centroid	1	1	1	1	1

larity between each cluster and its most comparable one. As the clusters must be reduced and isolated the lower Davies–Bouldin Index implies better cluster arrangement. The Davies–Bouldin Index is characterized as

$$DB = \frac{1}{n_c}\sum_{i=1}^{n_c} R_i \qquad (17)$$

where $R_i = \max_{j=1,\ldots n_c, i \neq j}\left(R_{ij}\right), i = 1, \ldots n_c$

The experimental analysis of the proposed model is done on both hierarchical and partitional clustering algorithms with the Euclidean distance, Pearson correlation and Cosine similarity. For each query term, weight is assigned based on term, concept and semantic smoothing from randomly selected three documents from the seven document corpuses. The weights are assigned to evaluate the similarity between the query term and documents. The weight values are shown in Table 1.

5.2 Performance analysis

The performance of the enriched semantic smoothing approach is compared with that of the ontology based semantic smoothing approach are compared using Euclidean distance, Pearson correlation and Cosine similarity. Table2 shows the result analysis for 7 clusters using the three measures for the enriched semantic smoothing model and existing semantic smoothing model.

5.2.1 Semantic smoothing based clustering

The performance of the ontology based semantic smoothing and enriched semantic smoothing based clustering models are evaluated and illustrated to compare the differences.

Figure 2 shows the performance of semantic smoothing based clustering using Euclidean distance. The measures given in Fig. 2a, Silhouette Index; b, FM Index; c, Jaccard

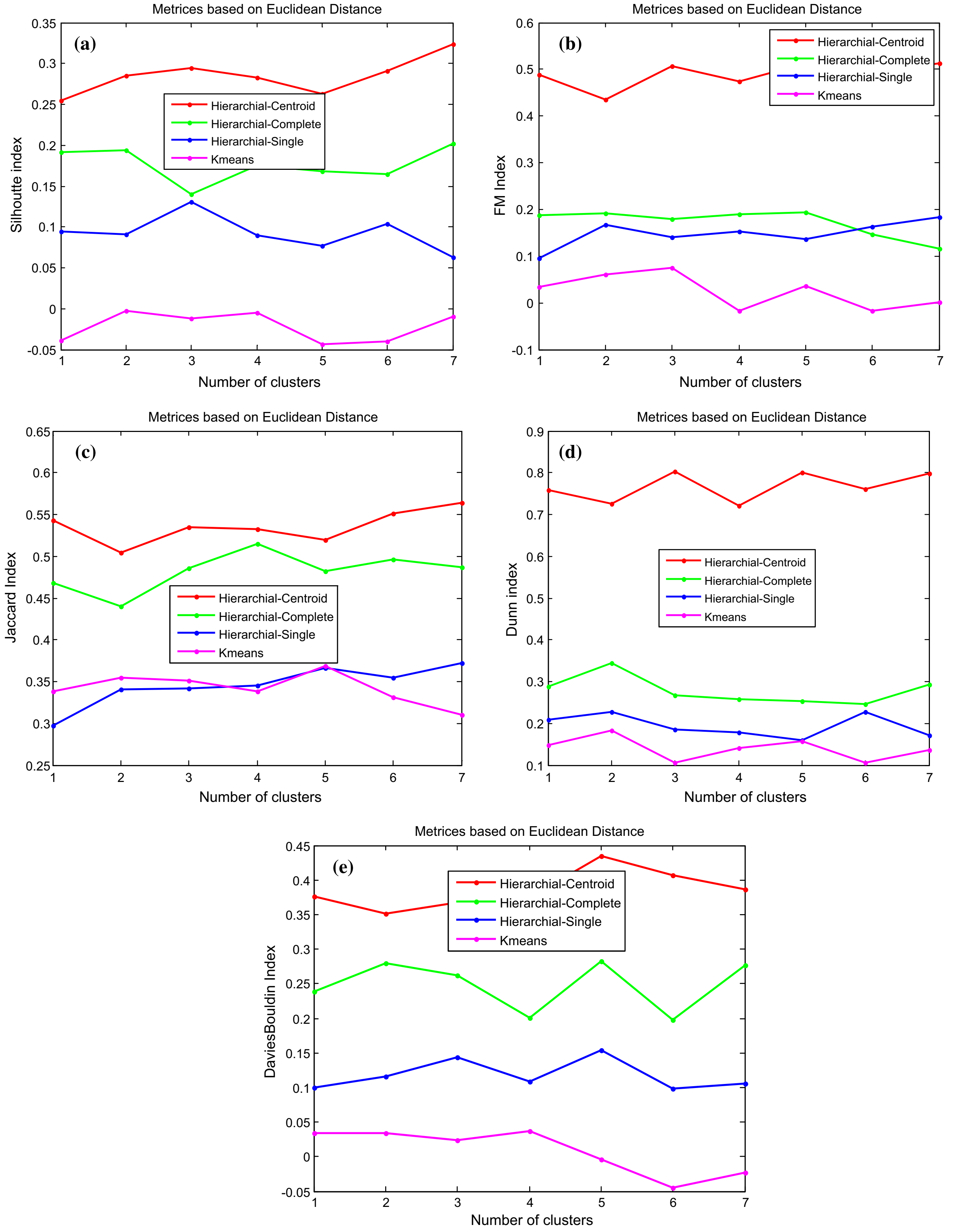

Fig. 2 Semantic smoothing based clustering using Euclidean distance. **a** Silhouette Index, **b** FM Index, **c** Jaccard Index, **d** Dunn Index, **e** Davies–Bouldin Index

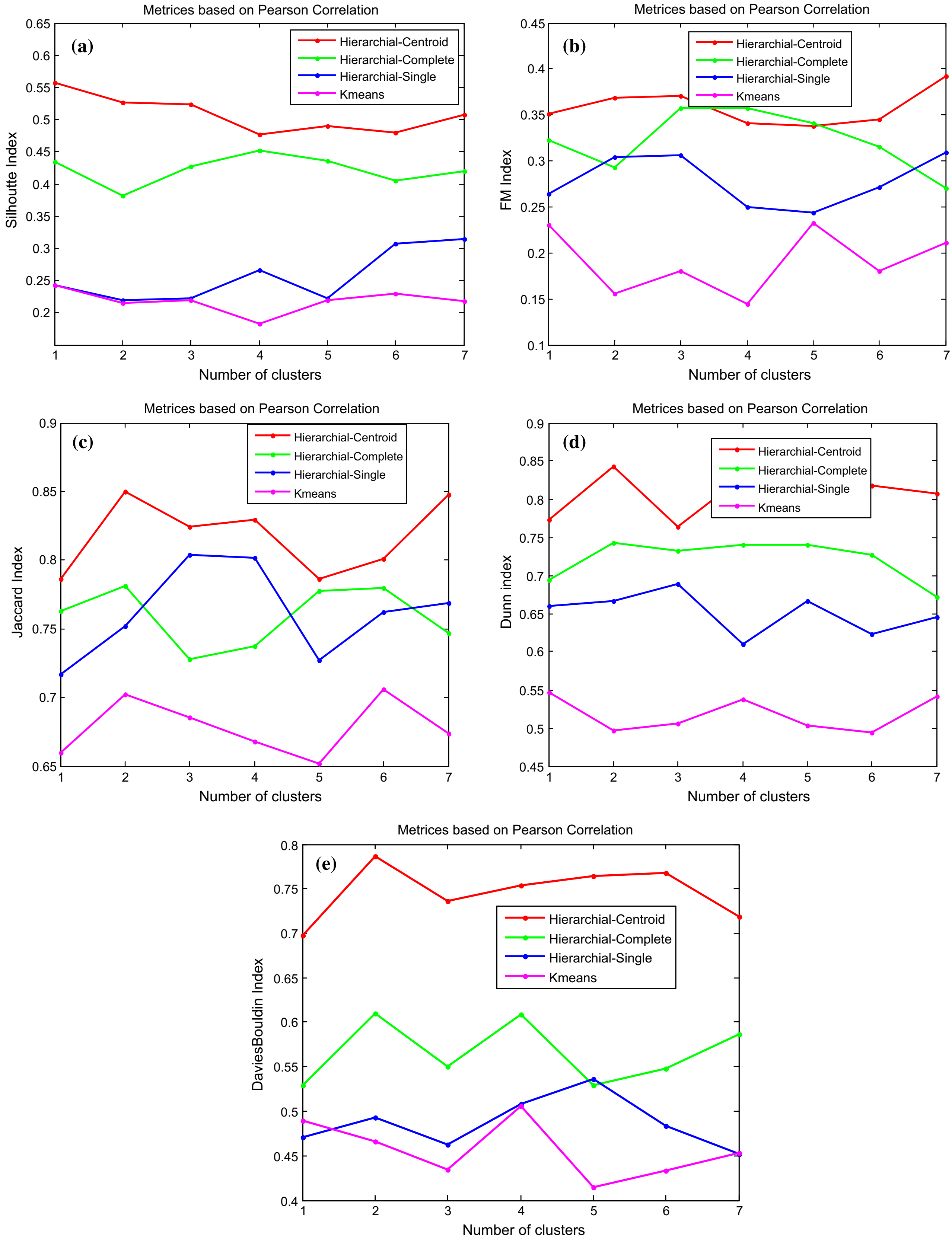

Fig. 3 Semantic smoothing based clustering using Pearson correlation. **a** Silhouette Index, **b** FM Index, **c** Jaccard Index, **d** Dunn Index, **e** Davies–Bouldin Index

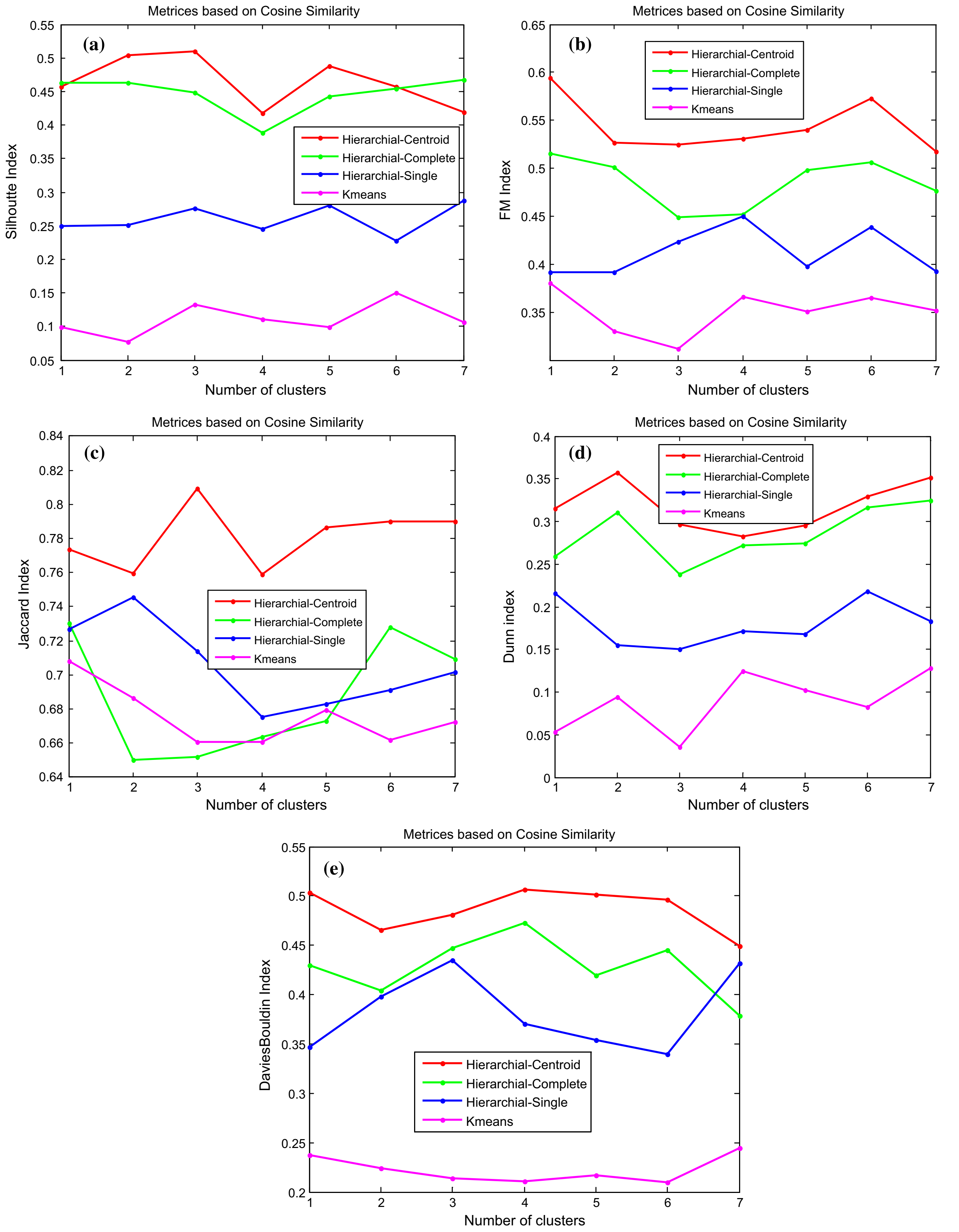

Fig. 4 Semantic smoothing based clustering using Cosine similarity. **a** Silhouette Index, **b** FM Index, **c** Jaccard Index, **d** Dunn Index, **e** Davies–Bouldin Index

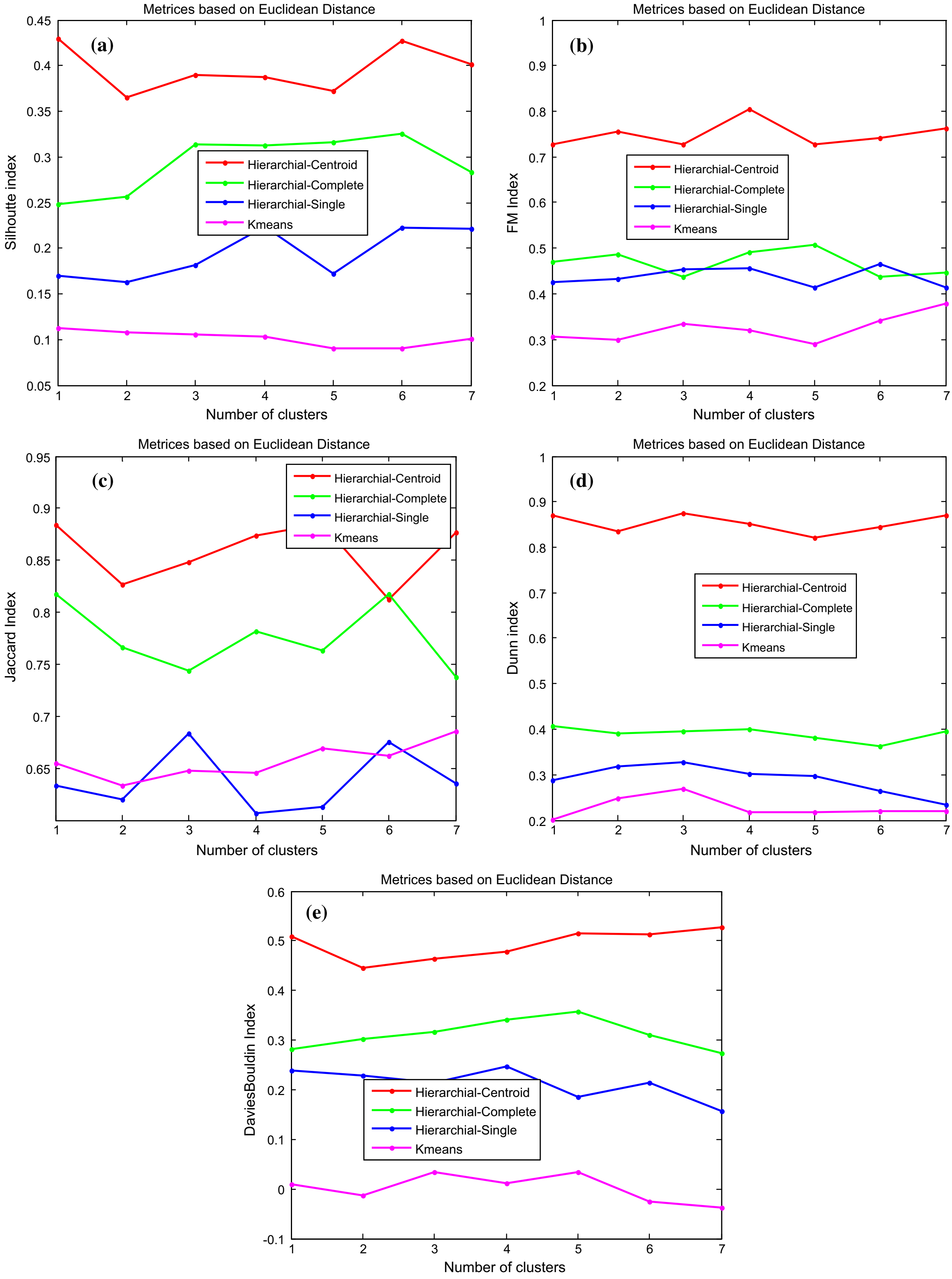

Fig. 5 Enriched semantic smoothing based clustering using Euclidean distance. **a** Silhouette Index, **b** FM Index, **c** Jaccard Index, **d** Dunn Index, **e** Davies–Bouldin Index

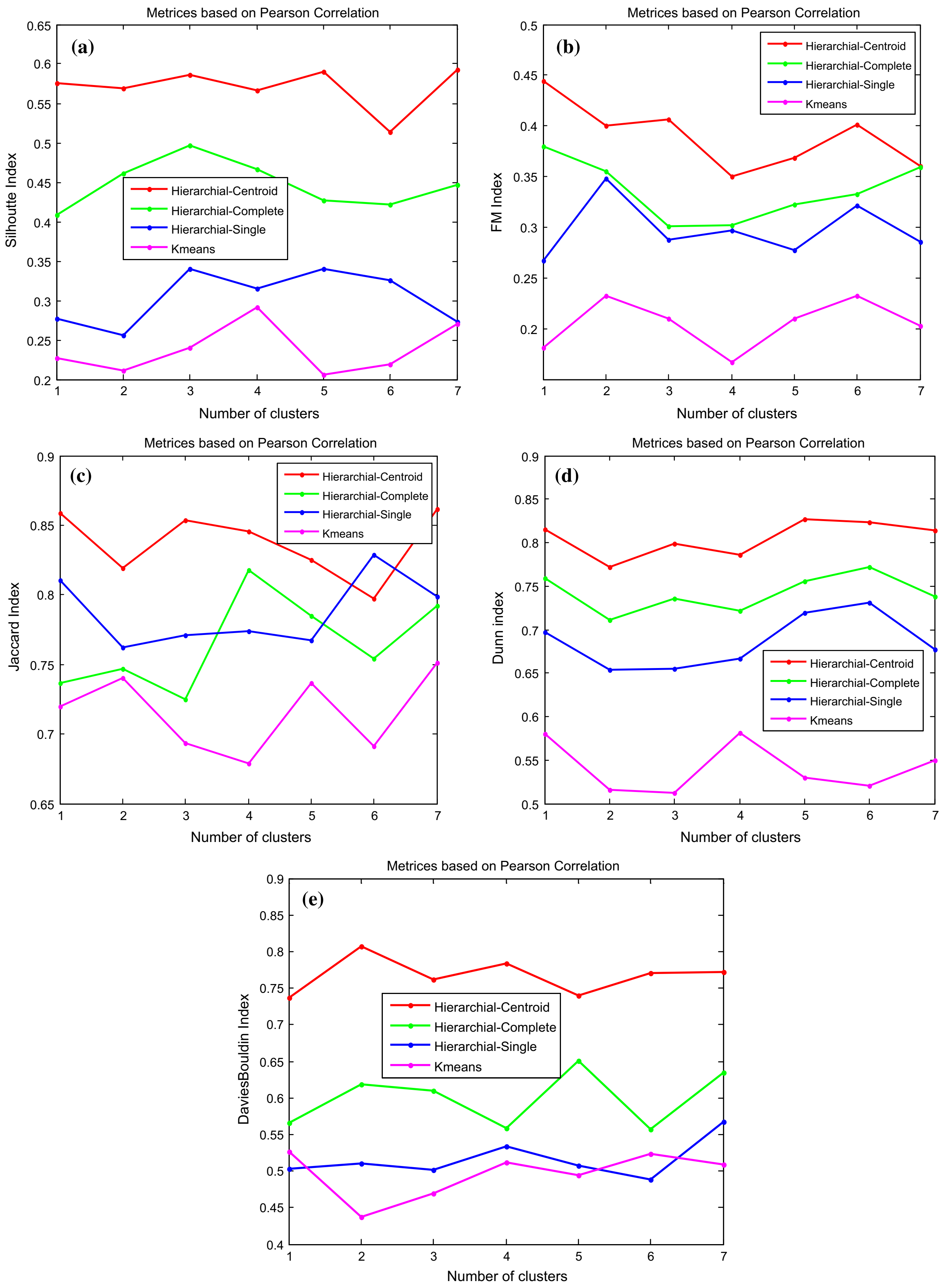

Fig. 6 Enriched semantic smoothing based clustering using Pearson correlation. **a** Silhouette Index, **b** FM Index, **c** Jaccard Index, **d** Dunn Index, **e** Davies–Bouldin Index

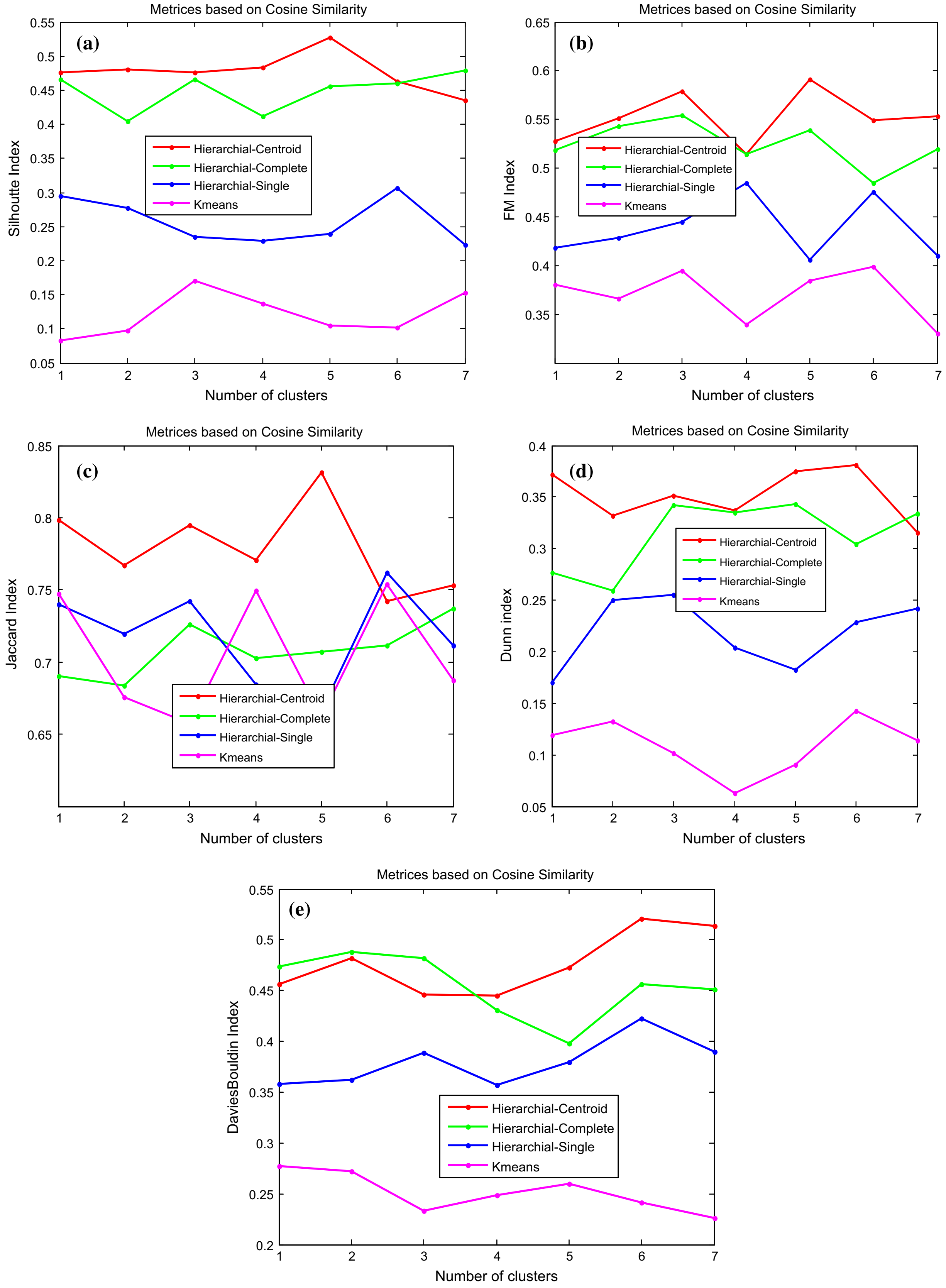

Fig. 7 Enriched semantic smoothing based clustering using Cosine similarity. **a** Silhouette Index, **b** FM Index, **c** Jaccard Index, **d** Dunn Index, **e** Davies–Bouldin Index

Index; d, Dunn Index; and e, Davies–Bouldin Index helps in analysing the performance of K-means clustering and hierarchical clustering with the semantic smoothing based clustering. From Fig. 1, it can be seen that the hierarchical clustering produces efficient clusters for Silhouette Index, Dunn Index and Davies–Bouldin Index while k-means clustering outperforms hierarchical clustering for FM Index and Jaccard Index.

Figure 3 shows the performance of semantic smoothing based clustering using Pearson correlation measure. It is seen that the performance of k-means clustering is better in Davies–Bouldin Index while hierarchical clustering outperforms k-means clustering for all other index measures.

Figure 4 shows the performance of semantic smoothing based clustering using Cosine similarity measure. It can be inferred from the figures that the k-means clustering provides best clusters for Jaccard Index while hierarchical clustering outperforms k-means clustering in terms of other measures.

5.2.2 Enriched semantic smoothing based clustering

The performance of the enriched model of semantic smoothing based clustering model is utilized for partitional as well as hierarchical clustering methods to determine its proficiency than the existing models.

Figure 5 shows the performance of enriched semantic smoothing based clustering using Euclidean distance. From the figure, it can be seen that the hierarchical clustering produces efficient clusters for Silhouette Index, FM Index, Dunn Index and Davies–Bouldin Index while k-means clustering outperforms hierarchical clustering for Jaccard Index.

Figure 6 shows the performance of enriched semantic smoothing based clustering using Pearson correlation measure. It is seen that the performance of k-means clustering is better in Davies–Bouldin Index while hierarchical clustering outperforms k-means clustering for all other index measures.

Figure 7 shows the performance of enriched semantic smoothing based clustering using Cosine similarity measure. It can be inferred from the figures that the k-means clustering provides best clusters for Jaccard Index while hierarchical clustering outperforms k-means clustering in terms of other measures. Thus better clustering can be obtained while utilizing the proposed model.

6 Conclusion

Biomedical document clustering has been one of the most sought research areas in recent times and this paper proposed enriched semantic smoothing enabled clustering model based on domain knowledge through ontology. The model utilized modified n-grams and TF-IGM concepts for improving the clustering experience. The performance of the model is analysed using partitional and hierarchical clustering methods with Silhouette Index, Jaccard Index, FM Index, Dunn Index and Davies–Bouldin Index used as cluster quality measures. Both the methods perform significantly better with the proposed model; k-means performs better in Euclidean distance while hierarchical clustering performs better for Pearson correlation and cosine similarity measures. The proposed model outperforms the semantic smoothing model in about 80% of the quality measures, thus proving its efficiency. From these analysis results, it is evident that the proposed model of enriched semantic smoothing improves the clustering performance and is of high proficiency in the biomedical document clustering for telemedicine applications.

References

1. Leuski, A.: Evaluating document clustering for interactive information retrieval. In: Proceedings of the Tenth International Conference on Information and Knowledge Management, pp. 33–40. ACM (2001)
2. Steinbach, M., Karypis, G., Kumar, V.: A comparison of document clustering techniques. In: KDD Workshop on Text Mining vol. 400(1), pp. 525–526 (2000)
3. Ding, C.H., He, X., Zha, H., Gu, M., Simon, H.D.: A min-max cut algorithm for graph partitioning and data clustering. In: Proceedings IEEE International Conference on Data Mining, 2001. ICDM 2001, pp. 107–114. IEEE (2001)
4. Zhao, Y., Karypis, G.: Evaluation of hierarchical clustering algorithms for document datasets. In: Proceedings of the Eleventh International Conference on Information and Knowledge Management, pp. 515–524. ACM (2002)
5. Chim, H., Deng, X.: Efficient phrase-based document similarity for clustering. IEEE Trans. Knowl. Data Eng. **20**(9), 1217–1229 (2008)
6. Saad, F.H., de la Iglesia, B., Bell, D.G.: A comparison of two document clustering approaches for clustering medical documents. In: DMIN, pp. 425–431 (2006)
7. Wan, X., Yang, J.: Multi-document summarization using cluster-based link analysis. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 299–306. ACM (2008)
8. Liu, X., Croft, W.B.: Cluster-based retrieval using language models. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 186–193. ACM (2004)
9. Silva, J., Mexia, J., Coelho, A., Lopes, G.: Document clustering and cluster topic extraction in multilingual corpora. In: Proceedings IEEE International Conference on Data Mining, 2001. ICDM 2001, pp. 513–520. IEEE (2001)
10. Cios, K.J., Moore, G.W.: Uniqueness of medical data mining. Artif. Intell. Med. **26**(1), 1–24 (2002)
11. Prather, J.C., Lobach, D.F., Goodwin, L.K., Hales, J.W., Hage, M.L., Hammond, W.E.: Medical data mining: knowledge discovery in a clinical data warehouse. In: Proceedings of the AMIA Annual Fall Symposium, p. 101. American Medical Informatics Association (1997)
12. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review. ACM Comput. Surv. (CSUR) **31**(3), 264–323 (1999)
13. Hotho, A., Staab, S., Stumme, G.: Ontologies improve text document clustering. In: Third IEEE International Conference on Data Mining, 2003. ICDM 2003, pp. 541–544. IEEE (2003)

14. Jing, L., Zhou, L., Ng, M.K., Huang, J.Z.: Ontology-based distance measure for text clustering. In Proceedings of SIAM SDM Workshop on Text Mining, Bethesda, MD (2006)
15. Yoo, I., Hu, X., Song, I.Y.: Integration of semantic-based bipartite graph representation and mutual refinement strategy for biomedical literature clustering. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 791–796. ACM (2006)
16. Logeswari, S., Premalatha, K.: Ontology-based semantic smoothing model for biomedical document clustering. Int. J. Telemed. Clin. Pract. **1**(1), 94–110 (2015)
17. Ding, C., He, X.: K-means clustering via principal component analysis. In: Proceedings of the Twenty-First International Conference on Machine Learning, p. 29. ACM (2004)
18. Pan, J.Y., Zhang, J.S.: Relationship matrix nonnegative decomposition for clustering. Math. Probl. Eng. **2011**, 842325 (2011)
19. Zhong, Y., Zhang, L.: A new fuzzy clustering algorithm based on clonal selection for land cover classification. Math. Probl. Eng. **2011**(2), 253–266 (2011)
20. Lee, M., Wang, W., Yu, H.: Exploring supervised and unsupervised methods to detect topics in biomedical text. BMC Bioinform. **7**(1), 140 (2006)
21. Lin, J., Wilbur, W.J.: PubMed related articles: a probabilistic topic-based model for content similarity. BMC Bioinform. **8**(1), 423 (2007)
22. Theodosiou, T., Darzentas, N., Angelis, L., Ouzounis, C.A.: PuReD-MCL: a graph-based PubMed document clustering methodology. Bioinformatics **24**(17), 1935–1941 (2008)
23. Nelson, S.J., Schopen, M., Savage, A.G., Schulman, J.L., Arluk, N.: The MeSH translation maintenance system: structure, interface design, and implementation. Stud. Health Technol. Inf. **11**(Pt 1), 67–69 (2004)
24. Yoo, I., Hu, X., Song, I.Y.: Biomedical ontology improves biomedical literature clustering performance: a comparison study. Int. J. Bioinform. Res. Appl. **3**(3), 414–428 (2007)
25. Zhang, X., Jing, L., Hu, X., Ng, M., Zhou, X.: A comparative study of ontology based term similarity measures on PubMed document clustering. In: Concepts, Systems and Applications, Advances in Databases, pp. 115–126 (2007)
26. Zhu, S., Zeng, J., Mamitsuka, H.: Enhancing MEDLINE document clustering by incorporating MeSH semantic similarity. Bioinformatics **25**(15), 1944–1951 (2009)
27. Hanisch, D., Zien, A., Zimmer, R., Lengauer, T.: Co-clustering of biological networks and gene expression data. Bioinformatics **18**(suppl 1), S145–S154 (2002)
28. Pan, W.: Incorporating gene functions as priors in model-based clustering of microarray gene expression data. Bioinformatics **22**(7), 795–801 (2006)
29. Huang, D., Pan, W.: Incorporating biological knowledge into distance-based clustering analysis of microarray gene expression data. Bioinformatics **22**(10), 1259–1268 (2006)
30. Shiga, M., Takigawa, I., Mamitsuka, H.: Annotating gene function by combining expression data with a modular gene network. Bioinformatics **23**(13), i468–i478 (2007)
31. Wagstaff, K., Cardie, C., Rogers, S., Schrödl, S.: Constrained k-means clustering with background knowledge. ICML **1**, 577–584 (2001)
32. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **22**(8), 888–905 (2000)
33. Ji, X., Xu, W.: Document clustering with prior knowledge. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 405–412. ACM (2006)
34. Gupta, S., MacLean, D.L., Heer, J., Manning, C.D.: Induced lexico-syntactic patterns improve information extraction from online medical forums. J. Am. Med. Inf. Assoc. **21**(5), 902–909 (2014)
35. Xu, Y., Hong, K., Tsujii, J., Chang, E.I.C.: Feature engineering combined with machine learning and rule-based methods for structured information extraction from narrative clinical discharge summaries. J. Am. Med. Inf. Assoc. **19**(5), 824–832 (2012)
36. Ghoulam, A., Barigou, F., Belalem, G., Meziane, F.: Using local grammar for entity extraction from clinical reports. IJIMAI **3**(3), 16–24 (2015)
37. Deleger, L., Molnar, K., Savova, G., Xia, F., Lingren, T., Li, Q., Solti, I.: Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. J. Am. Med. Inf. Assoc. **20**(1), 84–94 (2013)
38. Ling, Y., Pan, X., Li, G., Hu, X.: Clinical documents clustering based on medication/symptom names using multi-view nonnegative matrix factorization. IEEE Trans. Nanobiosci. **14**(5), 500–504 (2015)
39. Hübner, A., Walther, M., Kuhn, H.: Approach to clustering clinical departments. In: Health Care Systems Engineering for Scientists and Practitioners, pp. 111–120. Springer (2016)
40. Jun, S., Park, S.S., Jang, D.S.: Document clustering method using dimension reduction and support vector clustering to overcome sparseness. Expert Systems with Applications **41**(7), 3204–3212 (2014)
41. Karaa, W.B.A., Ashour, A.S., Sassi, D.B., Roy, P., Kausar, N., Dey, N.: Medline text mining: an enhancement genetic algorithm based approach for document clustering. In: Applications of Intelligent Optimization in Biology and Medicine, pp. 267–287. Springer (2016)
42. Al-Ariki, H.D.E., Swamy, M.S.: A survey and analysis of multipath routing protocols in wireless multimedia sensor networks. Wirel. Netw. **23**(6), 1823–1835 (2017)
43. Celebi, M.E. (ed.).: Partitional Clustering Algorithms. Springer, Cham (2014)
44. Chen, K., Zhang, Z., Long, J., Zhang, H.: Turning from TF-IDF to TF-IGM for term weighting in text classification. Expert Syst. Appl. **66**, 245–260 (2016)
45. Barrón-Cedeño, A., Rosso, P.: On Automatic Plagiarism Detection Based on n-Grams Comparison. Advances in Information Retrieval, pp. 696-700. Springer, Berlin (2009)
46. Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J.M., Perona, I.: An extensive comparative study of cluster validity indices. Pattern Recognit. **46**(1), 243–256 (2013)

R. Sandhiya has obtained her Bachelors in Information Technology at the CSI College of Engineering from Anna University in 2011. She obtained her Master's in Information Technology at Anna University Regional Centre, Coimbatore from Anna University in 2014. She is currently working as Professor in the Department of Information Technology at Coimbatore Institute of Technology, Coimbatore, India. She is doing her Ph.D. in the area of Ontology under the guidance of Dr. M. Sundarambal.

M. Sundarambal has obtained her Bachelors in Electrical and Electronics Engineering at the Government College of Technology from Madras University in 1981. She obtained her Master's in Applied Electronics at PSG College of Technology from Bharathiyar University in 1984. Dr. M. Sundarambal is currently working as Professor in the Department of Electrical and Electronics Engineering at Coimbatore Institute of Technology, Coimbatore, India. Her specializations include MANET, Wireless Networks, Wireless Sensor Networks, Robotics; Agent based Intelligent System and High Tech Prosthetics. She has produced 4 Ph.D. Scholars and has more than 100 publications in international journals and national conferences. She is a member of professional bodies, Indian Society of Technical Education, System Society of India, and Institution of Engineers.