American Journal of Applied Sciences

Literature Reviews

Advances in Document Clustering with Evolutionary-Based Algorithms

[1,2]Sarmad Makki, [1]Razali Yaakob, [1]Norwati Mustapha and [1]Hamidah Ibrahim

[1]Faculty of Computer Science and Information Technology, University Putra Malaysia, 43400 Selangor, Malaysia
[2]College of Science, University of Baghdad, 10071 Baghdad, Iraq

Article history
Received: 20-01-2015
Revised: 22-06-2015
Accepted: 14-09-2015

Corresponding Author:
Razali Yaakob
Faculty of Computer Science and
Information Technology,
University Putra Malaysia, 43400
Selangor, Malaysia
E-mail: razaliy@upm.edu.my

Abstract: Document clustering is the process of organizing a particular electronic corpus of documents into subgroups of similar text features. Formerly, a number of conventional algorithms had been applied to perform document clustering. There are current endeavors to enhance clustering performance by employing evolutionary algorithms. Thus, such endeavors became an emerging topic gaining more attention in recent years. The aim of this paper is to present an up-to-date and self-contained review fully devoted to document clustering via evolutionary algorithms. It firstly provides a comprehensive inspection to the document clustering model revealing its various components with its related concepts. Then it shows and analyzes the principle research work in this topic. Finally, it compiles and classifies various objective functions, the core of the evolutionary algorithms, from the related collection of research papers. The paper ends up by addressing some important issues and challenges that can be subject of future work.

Keywords: Text Document Clustering, Hypertext Clustering, Evolutionary Algorithms, Genetic Algorithms, Text Dimensional Reduction

Introduction

With the rapid tendency towards the usage of information systems along the world, more and more data have been stored in electronic form. Approximately 80% of these data are stored in text format (IndiraPriya and Ghosh, 2013; Xiao, 2010). Hence, there is a need for organizing and categorizing these data in such a way satisfying the needs for more mining information. One of these text mining techniques is the document clustering or the unsupervised document classification process. With unsupervised it meant the attempt to automatically construct groups (clusters or partitions) of documents without having a prior knowledge or domain expertise alongside the given data, such as the class label. The resulting groups should possess: (1) *homogeneity* within the cluster, i.e., documents belongs to the same partition should be as similar as possible and (2) *heterogeneity* among the clusters, i.e., documents belongs to different partitions should be as different as possible.

Document clustering can be useful in a number of applications, such as the query term routing, cluster-based browsing, result set clustering or expansion and query suggestion refinement. Hence, it becomes a vital research area in text mining with a contemporary trend towards applying the machine learning techniques especially the evolutionary algorithms to enhance the performance of the clustering algorithm (Sheikh *et al*., 2008).

Mathematically, the clustering problem can be modeled as follows:

Assume D is the given document set (corpus) $D=\left\{\vec{d}_1, \vec{d}_2,\ldots, \vec{d}_N\right\} \textit{ where } \vec{d}_i \in R^n$ and $\vec{d}_i$ represents a single document.

The aim of document clustering is to find K partitions $P_1, P_2, \ldots, P_K$ such that:

- $p_i \neq 0 \quad \forall_i = 1,2,\ldots,k$ (not empty set)
- $p_i \cap p_j = \varphi \quad \forall_i = 1,2,\ldots,k \textit{ and } i \neq j$ (non-overlapped partitions)
- $\bigcup_{i=1}^{k} p_i = D$ (The compositions of all partitions represent the data set)

All of the clustering algorithms based on the cluster hypothesis (van-Rijsbergen, 1979), which states that: Related text documents tend to be more coherent to each other than to non related documents (separation).

The remaining of this paper is organized as follows: Section 2 will be devoted to explain the general model for a typical document clustering system. Section 3 will be dedicated to summarize other surveys on document clustering in general. Section 4 will focus specifically on the recent proposed algorithms and approaches of document clustering from the evolutionary algorithms point of view. Section 5 will be dedicated entirely for presenting and discussing the objective functions for the reviewed researches. We close our work in section 6 with conclusions and suggestions for future work.

Stages of Document Clustering

Broadly speaking, the basic question in text processing is how to represent the unstructured natural language text as an algebraic form suitable for mathematical analysis. Salton *et al*. (1975) in their seminal paper answered this question by proposing the Vector Space Model (VSM) representation. Since VSM is one of widely applied and most popular text representation for document clustering in recent years (Sathiyakumari *et al*., 2011), therefore we'll focus our discussion on this representation and the corresponding model.

The text unit should be passes throughout a number of stages in order to be ready for analysis by the chosen/proposed algorithm (s). Figure 1 shows the main stages for a general text document clustering process.

The following subsections briefly clarify these stages.

Data Acquisition

Generally, two main data sources for the text data can be recognized. It could be obtained either from the standard data repositories such as: Reuters (Lewis, 2004), 20newsgroup (Lang, 2008), TREC (NIST, 2000), DMOZ (Cobos, 2011) and KDnuggets (Piatetsky-Shapiro, 1993). The process of obtaining the data is combined with another process called indexing. *Indexing* is the process of storing the documents and its constituent terms in a suitable representation or more specifically suitable data structure. There are five levels of representing the natural language document by means of a set of index. These are character, word, phrase, sentence or language/application specific levels (Benbrahim and Bramer, 2009). The basic and most widely-used approach for indexing is the use of word (token) level, in a process known as *tokenization*. Tokenization means segmenting the sentences into its constituent parts. In this approach the sequences of words are ignored, i.e., the document is treated as *bag-of-word.*

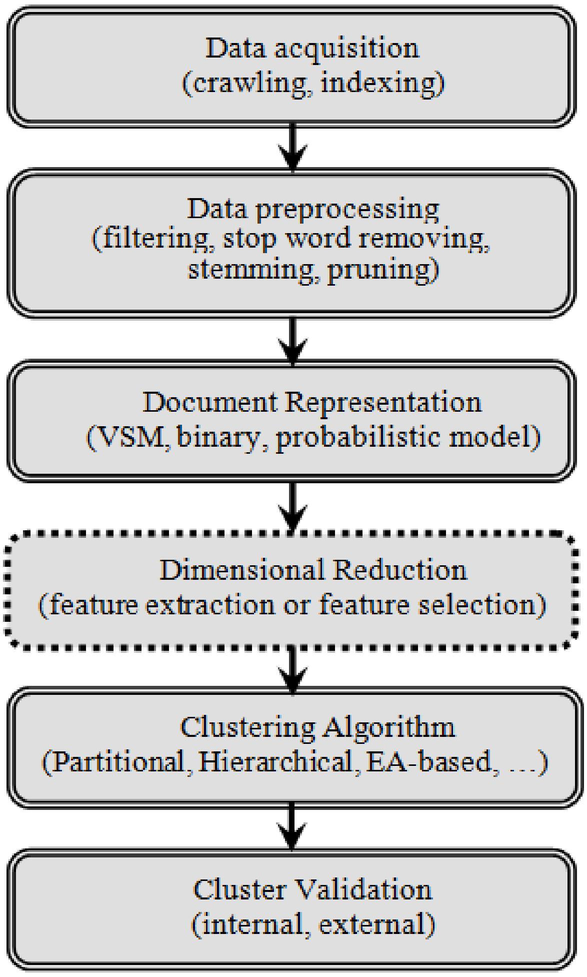

Data acquisition
(crawling, indexing)
Data preprocessing
(filtering, stop word removing, stemming, pruning)
Document Representation
(VSM, binary, probabilistic model)
Dimensional Reduction
(feature extraction or feature selection)
Clustering Algorithm
(Partitional, Hierarchical, EA-based, ...)
Cluster Validation
(internal, external)

Fig. 1. General stages of the text document clustering process

After tokens are extracted from documents, an indexing phase follows. Two significant indexing techniques exists, namely *inverted indices* and *signature files* (Han and Kamber, 2011).

An inverted index maintains two B+-tree or hash tables for doc-id and term-id. The first one consist of set of records for documents and indices to its terms, while the second one consist of a set of records for terms and its appeared-in documents. A signature file is a table for documents of fixed size columns equals to the number of terms. Initially all contents are set to zero. Whenever a term occurs in a document the corresponding bit is set to one. Careful management should be taken for multiple occurrences of term in this indexing technique.

Data Preprocessing

The preprocessing consist of a number of steps necessary to convert the natural language “web” text unit into a form (single term or n-gram) suitable to be included in the VSM. These steps typically consist of:

Filtering

Filtering is removing the characters that have little or no importance for text mining, such as numbers, punctuation symbols and special characters. It is also involved replacing tabs and other non-text characters by single space. Finally, convert all characters to upper case. In the case of formatted texts documents such as the web pages, the scripts and codes should be eliminated in this phase of preprocessing, while the tags could be either removed or a special weight could be assigned to their constituent terms.

Stopword Removal

Filtering out the terms that do not have a discriminating power, such as the function words "which", "there", "who" and etc. This process will lessen the dimensions of terms in the VSM by typically comparing each term against a list of known stopwords. Since stopword removal chooses a subset from the original feature set, it would be consider as a feature selection process. One drawback of stopword removal is that it might remove potential useful words; hence the selection must be done with care according to the intended application.

Stemming

Reducing inflected words to their root/base form. For example, the words "stemmer", "stemming", "stemmed" are all diminished to the root word "stem". Stemming (or lemmatization if part of speech is included) is a basic procedure used to minimize the dimension of the terms in the VSM model. Thus, by storing stems instead of terms, compression factors of over 50% can be achieved (Frakes and Baeza-Yates, 1992). Despite the existence of other analogue stemming algorithms such as (Hooper and Paice, 2005a; Lovins, 1968),Paice (Hooper and Paice, 2005c; Paice, 1990), S-removal (Harman, 1991), Dawson (Hooper and Paice, 2005b) and Krovetz. Nevertheless, Porter algorithm (Porter, 2006; 1980) is yet the most commonly used stemmer in English language (Frakes and Baeza-Yates, 1992). Since stemming maps the morphologically similar words into their stem, it would be consider as a feature extraction process. One drawback of stemming is that it might affect the meaning.

Pruning

Removing (stemmed) words that appear too low or too frequent throughout the corpus. The assumption is that even though these words have discriminating power, they might still form too small clusters to be useful. It's typically done by comparing the frequency of the term with pre-specified lower/upper threshold.

It should be noted that stop word removal, stemming and pruning could be an optional functions in the text preprocessing.

Document Feature Representation

Different representation models used in text processing such as the VSM, ontology-based, binary and probabilistic models. VSM is identified as the most popular representation method for text documents. In this model and after preprocessing, the next step is to represent each text document as a one dimensional vector in the multidimensional term space, consequently forming what is known as the document-term matrix as shown in Fig. 2.

In this sparse matrix each row corresponds to a document and each column correspond to the weight of unique term in the vocabulary, based on one of the term weighting schemas.

Several terms weighting schema had been applied in text processing. The schemas that specifically adopted in document clustering are as follows:

1- The first weighting schema is the classical TF/IDF (Term Frequency/Inverse Document Frequency). Simply, the $tf_{i,j}$ is the counts of occurrences (frequency) of term i in the document j. Usually this number is normalized by the number of terms in the document. While idf_j is computed as:

$$idf_j = log\left(N / n_j\right)$$

where, N is the total number of documents in the corpus and n_j is the number of documents that the term j occurs in. This factor will give a higher weight to the terms that occurs in few documents. Thus, the weight of term j in document i is computed as follows:

$$w_{i,j} = tf_{i,j} . idf_j$$

This weighting schema was used, for instance, in (Leon *et al.*, 2012; Yonghong and Wenyang, 2010). Also (Dorfer *et al.*, 2012) applied this frequency analysis and kept the terms with above-average relevance. This method achieved significant reduction, as only 29% of the terms remained afterwards.

Salton and Buckley (1988) had recommended two schemas for document weighting. These are:

$$w_{i,j} = \frac{tf_{i,j} \, . \log\left(\frac{N}{n_j}\right)}{\sqrt{\sum_{vector}\left(tf_{i,j} \, . \log\left(\frac{N}{n_j}\right)\right)^2}}$$

and:

$$w_{i,j} = \frac{tf_{i,j}\,.\log\left(\frac{N-n_j}{n_j}\right)}{\sqrt{\sum_{vector}\left(tf_{i,j}\,.\log\left(\frac{N-n_j}{n_j}\right)\right)^2}}$$

where, N is the total number of documents and n_j is the number of documents which a term j is assigned. The formula in equation (3) had adopted, for instance, by Shi and Li (2013) with minor modification to consider the document length on the impact of weight normalized to the interval [0,1].

Radwan *et al*. (2006), computed the document weighting from the formula suggested by (Salton and Buckley, 1990) as follows:

$$w_{i,j} = \frac{0.5+0.5.\frac{tf_{i,j}}{\max tf}\,.\log\left(\frac{N}{n_j}\right)}{\sqrt{\left(0.5+0.5.\frac{tf_{i,j}}{\max tf}\right)^2.\left(\log\left(\frac{N}{n_j}\right)\right)^2}}$$

where, $w_{i,j}$ is the weight assigned to the term t_j in document D_i, $tf_{i,j}$ is the number of times that term t_j appears in document D_i, n_j is the number of documents indexed by the term t_j and finally, N is the total number of documents in the corpus.

Other researchers such as Lee *et al*. (2011), uses the Okapi rule (Salton and Buckley, 1988) for term weight calculation as follows:

$$w_{i,j} = \frac{tf_{i,j}}{tf_{i,j}+0.5+1.5*\frac{dl}{avgdl}}*\log\left(\frac{N}{n_j}\right)$$

where, *dl* is the length of the document and *avgdl* is the average length of documents.

Liu *et al*. (2011), took the size of each document into account and the parameter weight was defined as:

$$w_{i,j} = \frac{TF_{i,j}*\log\left(\frac{N}{n_j}+0.01\right)*\frac{\sum_{k=1}^{N} size(k)}{N}}{size(i)}$$

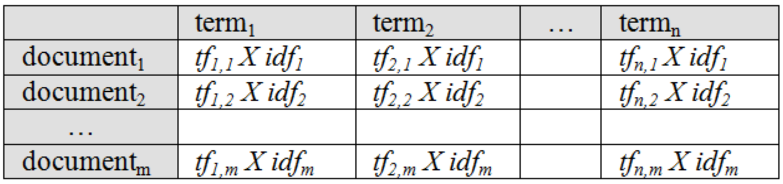
where, *size(i)* is the size of documents and the $\sum_{k=1}^{N} size(k)/N$ shows the average size of all documents in the data set.

We indicate to some of the principle term weighting schema here. More detailed discussion about global, local and normalized term weighting could be found, for instance, in (Fodor, 2002; Manning *et al*., 2008).

Dimensional Reduction

In general, the process of reducing the number of variables is done by utilizing two techniques: *Feature extraction* and *feature selection* (Fodor, 2002). Feature extraction, linear or non linear techniques, transforms the data in the high dimensional space into a space of lesser dimensions. Quite large number of documents with diverse terms will lead to large and sparse document-term matrix. Such large matrix leads to the problem of high and inefficient computation and increases the difficulty in detecting the relationships among terms (synonymy). To overcome these problems, linear feature extraction techniques could be applied during the preprocessing phase, such as Latent Semantic Indexing (LSI), Locality Preserving Indexing (LPI), Independent Component Analysis (ICA) or Random Projection (RP) (Han and Kamber, 2011; Palsonkennedy and Gopal, 2012; Tang *et al*., 2005; Thangamani and Thangaraj, 2010).

On the other hand, the feature selection techniques, supervised or unsupervised, attempt to acquire a subset of the original data. Since document clustering is an unsupervised process, the supervised techniques such as the Information Gain (IG) and X^2 statistics (CHI) could not be used with text clustering. Such techniques could be used with text classifications rather than clustering due to the presence of class label. Nevertheless, other unsupervised feature selection methods had been used with text clustering such as Document Frequency (DF), Term Contribution (TC) or Term Variance (TV) among other statistical techniques (Luying *et al*., 2005; Tang *et al*., 2005). Moreover, there are recently evolutionary algorithm based optimization methods for term or keyword selection, such as for instance the technique in (Shamsinejadbabki and Saraee, 2012).



	term$_1$	term$_2$	…	term$_n$
document$_1$	$tf_{1,1}$ X idf_1	$tf_{2,1}$ X idf_1		$tf_{n,1}$ X idf_1
document$_2$	$tf_{1,2}$ X idf_2	$tf_{2,2}$ X idf_2		$tf_{n,2}$ X idf_2
…				
document$_m$	$tf_{1,m}$ X idf_m	$tf_{2,m}$ X idf_m		$tf_{n,m}$ X idf_m

Fig. 2. A typical document-term matrix

Clustering Algorithm

Two commonly used categories of algorithms in document clustering: *Partitional* and *hierarchical* clustering. The most commonly used partitional clustering algorithms are k-means and its variations (Pavan *et al*., 2010; Steinbach *et al*., 2000; Velmurugan and Santhanam, 2010). These flat clustering algorithms group the documents into k predefined number of partitions based on the closest distance to the k centroids. While the family of hierarchical algorithms (divisive or agglomerative) construct a hierarchy by iteratively merging (or splitting in case of divisive) the most similar pair of partitions. Some researches used a hybrid of both approaches (Cutting *et al*., 1992). Others used different text based approaches such as the suffix tree based clustering algorithms (Wang *et al*., 2008; Zamir and Etzioni, 1999; Zeng *et al*., 2004).

There are certainly other conventional categories of clustering algorithms such as the density based, grid based and model based clustering, among others (Han and Kamber, 2011). However, to the best of the authors' knowledge, there were no attempts to cluster the documents using these categories of clustering algorithms, except a recent project headed by Prof. Han at university of Illinois to cluster the documents using the SCAN density based algorithm (Li, 2012). Finally, we have to say that documents had been clustered with other non-conventional algorithms such as the evolutionary-based algorithms. In this review, we shall discuss the most recent of these evolutionary-based algorithms.

Cluster Validation

The procedure of evaluating the quality of a clustering algorithm is known as *cluster validation.* Two mainly categories of cluster validity measures used in clustering, namely: Internal (unsupervised) and external (supervised) validity indices. Generally, a cluster validity index serves two purposes. Firstly, it can be used to determine the number of clusters and secondly, it finds out the corresponding best partition (Das *et al*., 2009). For that reason, these measures can be utilized as the fitness function(s) for the evolutionary algorithms. The internal validity indices, such as the Bayesian Information Criterion (BIC), Calinski-Harabasz index (CH), Dunn index and Davies-Bouldin index (DB) can handle the information presented in the data set itself (Mary and Kumar, 2012). While, the external validity indices, such as Entropy, Purity, Normalized Mutual Information and F-measures, can utilize external knowledge alongside the data set such as the given category labels by reviewer in advance.

On validity indices, Zhao and Karypis performed a comparison of selected validity measures applied specifically to document clustering (Zhao and Karypis, 2004). Halkidi *et al.* (2001) surveyed the widely known clustering algorithms in a comparative way and presented a review of clustering validity measures and approaches available. Rendon *et al.* (2011) made a recent comparison between the internal and external validity indices.

Early Studies on Document Clustering

In order to make this review as integral and accurate as possible and to pave the way to future possible hybrid algorithms utilizing from certain existing characteristics, we shall briefly highlight on some major surveys and/or reviews on document clustering. There must be a careful distinction not only among the algorithms used for clustering, but also between the data types that fit each algorithm, in which it is applied to two-dimensional data or multi-dimensional data as in the case of the text documents. Hence, this section is divided into two subsections. The first subsection is devoted to the studies that dealt with the conventional algorithms for document clustering (the dash-dotted line in Fig. 3). Meanwhile, the second subsection is devoted to the studies that dealt with the evolutionary algorithms for clustering the two dimensional data (the dotted line in Fig. 3). It should be noted that the evolutionary algorithm based clustering algorithms for 2D data might be useful for the document data.

Nevertheless, our main focus is on the Evolutionary Algorithm-based methods brought to bear specifically for document clustering (the dashed line in Fig. 3). Accordingly, we'll dedicate section 4 to list, categorize and criticize the latest studies on this issue.

Major Surveys on Conventional Document Clustering Algorithms

By conventional approaches we are specifically pointing out to two categories of clustering, namely the *partitional* and *hierarchical* algorithms. These two families of algorithms are the most commonly used algorithms for clustering the text documents.

The variations of the *k-means* algorithms are the most popular partitional clustering algorithms due to its ease of implementation and low time complexity. However, these algorithms have some drawbacks such as sensitivity to selection of initial centroids, sensitivity to outliers and the requirement to pre-specify the number of clusters. Whereas, the *hierarchical* algorithms provides more accurate results than those obtained from k-means algorithms. Nevertheless, the partitional algorithms also have some drawbacks such as high time complexity, producing the same result in all runs and the inability to reassign the initially wrong assigned points to clusters.

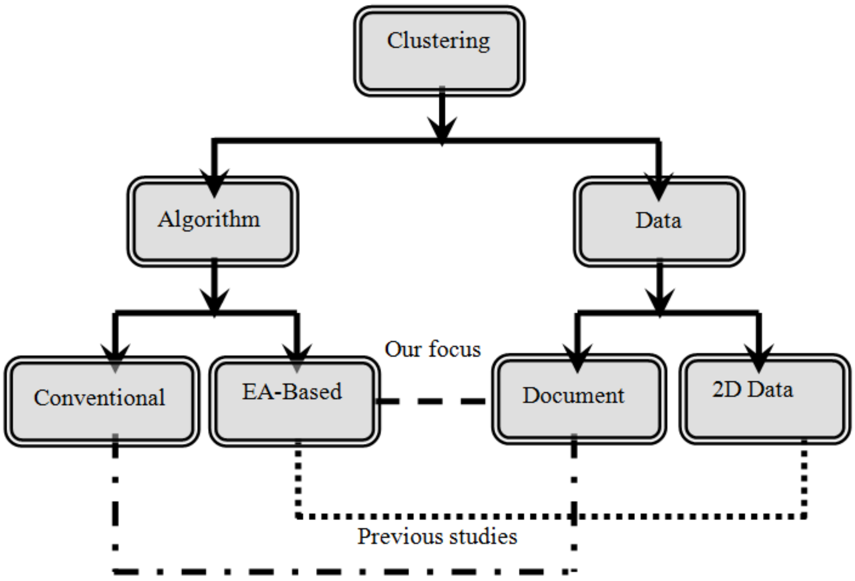


Fig. 3. The gap that we fill with our study

Peter Willett wrote one of the early critical reviews on document clustering (Willett, 1988). He discussed hierarchical agglomerative clustering methods that can be implemented on databases of nontrivial size. He also described the validation of document hierarchies; theoretically by the theory of random graph and empirically by characteristics of document collection that are to be clustered. The analysis was focused on the extensively used *single linkage* hierarchical method, with a description to other group of hierarchic agglomerative clustering methods like the *complete linkage*, *group average* and *Ward* methods.

After the pioneer hybrid strategy of combining the hierarchical and partitional clustering into one cluster-based browsing system done by (Cutting *et al.*, 1992; Steinbach *et al.*, 2000) did an excellent experimental study and comparison between the two main conventional approaches on document domain. For the hierarchical, they adopted three different schemas: Intra-Cluster Similarity Technique (IST), Centroid Similarity Technique (CST) and Unweighted Pair Group Method using Arithmetic Averages (UPGMA). Whereas, for partitional clustering, they adopted two schemas: K-means and its variation *bisecting k-means*. They came up with a contrary, yet interesting conclusion about applying the conventional clustering algorithms on the document data set. They showed the superiority of bisecting k-means over UPGMA, the best hierarchic schema they adopted on documents. In addition, they provided the explanation for this superiority. Their explanation was based on the analysis of the specific clustering algorithm used and the nature of the document data.

One more similar analysis is done by in (Amala Bai and Manimegalai, 2010). Among the different versions of conventional algorithms, they conducted their analysis via two schemas of partitional algorithms: Euclidian k-means (K-means) and Spherical k-means (SK-means) and one schema for hierarchical algorithms: Unsupervised Principle Direction Division Partitioning (PDDP). They assured the results of Karypis lab group (Steinbach *et al.*, 2000) on the ability of partitional algorithms to acquire better results than the hierarchical algorithms in certain initials clusters. Some of their assumptions raised the quality of the results, such as, assuming equal number of documents in all classes and stripping out the stop-word removal in the preprocessing phase.

Liping (2005) surveyed the text clustering from a different point of view. The survey shaded more lights on particular challenging problems in text clustering such as big volume, high dimensionality and complex semantics. The survey reviewed the suggested solutions for those problems and how they applied on some existing and well-known web systems, such as Unstructured Information Management Architecture (UIMA), the KArlsruhe ONtology and Semantic Web tool suite (KAON) and A General Architecture for Text Engineering (GATE).

A well-structured paper by Patel and Zaveri (2011) reviewed the web page clustering techniques. The paper presented the conventional algorithms with a swift overview to the optimization-based algorithm such as the Genetic Algorithms (GA). The document representation techniques and cluster evaluation measures had also been described briefly.

Fasheng and Lu (2011) demonstrated the common clustering algorithms. Namely, the hierarchical, partitional, density-based and self-organizing map algorithms. The paper analyzed the mentioned clustering algorithms and summarized the characteristics of each algorithm.

More Recently, Aggarwal and Zhai (2012) included a chapter to survey the text clustering algorithm. In addition to the frequent conventional distance based algorithms, some other new categories of algorithms have been introduced. These categories stated the feature selection based, word and phrase based and probabilistic text clustering algorithms. However, it didn't indicate any class of optimization-based nor evolutionary-based algorithms.

Major Surveys on EA-Based Data Clustering Algorithms

Evolutionary Algorithms (EAs) are population based metaheuristic optimization algorithms which use mechanisms inspired by the biological evolution, such as mutation, crossover, natural selection and survival of the fittest in order to refine a set of candidate solutions iteratively in a cycle (Weise, 2011). The EAs are mainly divided into four categories: Genetic Algorithms (GA), Genetic Programming (GP), Evolutionary Programming (EP) and Evolutionary Strategy (ES). Each of these constitutes a different approach. However, they are all inspired by the same principles shown in Fig. 4.

One of the early studies on data clustering was the notable and lengthy paper of Jain *et al*. (1999) that reviewed various deterministic and stochastic approaches to data clustering. The paper discussed statistical, fuzzy, neural network, knowledge-based and evolutionary approaches to data clustering. However, regarding the evolutionary-based approach, an indication had been given to only two early empirical studies on small data set; that is, fewer than 200 patterns. Nevertheless, the study assured a number of particular properties of evolutionary clustering among other reviewed algorithms. These properties are:

- The capability of searching more than one solution at a single run-time by virtue of the inherited population-based feature
- The ability to speed up performance due to the parallelism feature
- The uniqueness in EA-based algorithms in finding optimal solutions even when the criterion function is discontinuous
- And most importantly, the capability of the EAs of being the unique “globalized search technique” among other reviewed clustering algorithms

Jain *et al*. (1999) paper also shaded some lights on the domain of document clustering.

Sheikh *et al*. (2008) wrote a survey on the state-of-the-art GA-based data clustering techniques and their application to different problem domains. They stressed that GAs are the best known evolutionary techniques. The researchers commented shortly on merely two papers related to the document clustering domain.

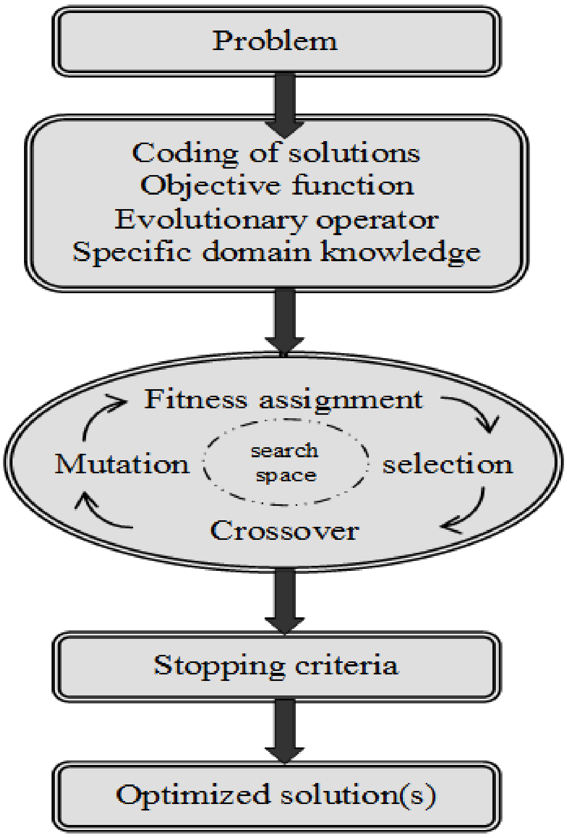


Fig. 4. Problem solution using EAs

In addition to the hard EA-based clustering, which had been covered in most of the previous surveys, the overlapped EA-based clustering had also been covered in (Hruschka *et al*., 2009). Besides that, this survey had included discussion on advance topics such as ensemble-based and multi objective evolutionary clustering. Moreover, it discussed a number of applications of EA clustering. Specifically application on image processing, bioinformatics, finance and Radial Basis Function (RBF) neural network design. Nevertheless, the domain of document clustering had barely been listed.

Text Document Clustering with Evolutionary Algorithms

We emphasize that none of the above surveys/studies addressed the detailed issue of document clustering from the EAs point of view. Accordingly, one of the main objectives of this paper is to cover the scope of EA-based clustering algorithms on the text document domain.

After careful analysis and detailed review of the recent researches in this filed, it appealed to us three main disciplines in dealing with the document clustering from the evolutionary algorithms point of view. Hence, the next subsections are organized accordingly, as shown in Fig. 5.

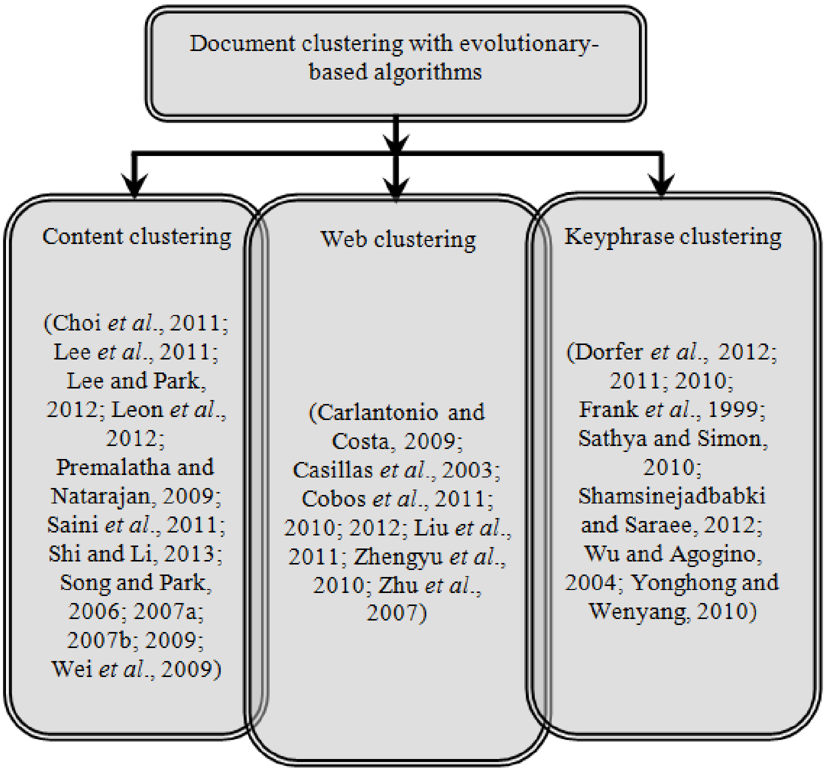

Document clustering with evolutionary-based algorithms
Content clustering
(Choi et al., 2011; Lee et al., 2011; Lee and Park, 2012; Leon et al., 2012; Premalatha and Natarajan, 2009; Saini et al., 2011; Shi and Li, 2013; Song and Park, 2006; 2007a; 2007b; 2009; Wei et al., 2009)
Web clustering
(Carlantonio and Costa, 2009; Casillas et al., 2003; Cobos et al., 2011; 2010; 2012; Liu et al., 2011; Zhengyu et al., 2010; Zhu et al., 2007)
Keyphrase clustering
(Dorfer et al., 2012; 2011; 2010; Frank et al., 1999; Sathya and Simon, 2010; Shamsinejadbabki and Saraee, 2012; Wu and Agogino, 2004; Yonghong and Wenyang, 2010)

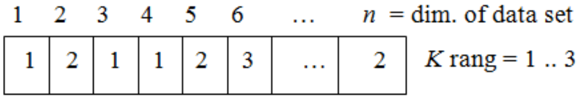
Fig. 5. Main researches' disciplines in document clustering with EAs

To make the discussion as clear as possible, we adopted some labeling scheme to refer to different research groups/disciplines. The first research group is going to be referred to as *content clustering*, since they dealt with the clustering of the entire textual contents of specific set of documents. The second group will be called *web document clustering*, as they examined other “web” features added to the clustering of the web/hypertext pages and lastly, the third group will be referred to by *keyword/keyphrase clustering*, as this group investigated the identification of groups of keywords/terms that best describe a specific set of documents. All of these researches are substantially discussed the document clustering problem from the evolutionary algorithms perspective. Each research is going to be discussed in depth in the later subsections showing its operation, characteristics and demonstrating the best of its results and weakness if any. Finally, since the objective function is the most distinguished portion of evolutionary algorithm, a summary of all fitness functions adopted for all disciplines is going be discuss in the next section (section 5).

Content Clustering

Wei *et al*. (2009) put forward a new dynamic method based on GA for document clustering. The method established on a new formula for describing the similarities of Chinese text documents. The formula took into account the partial similarity (up to 4 letters) of the keywords instead of full matching. The algorithm used floating point encoding and floating point crossover and mutation operators. The selection operator was a combination of choiceness and sorting. The sum of mean deviation of inter-class distance was used as the fitness function. The proposed algorithm didn't use elitism to allow the better chromosomes to carry on to the next generation. Finally the algorithm assumed that the number of categories *k* is given as an input parameter. The performance of the suggested GA methods showed better clustering results than k-means algorithm in term of the average of fitness function. The results obtained from 600 document chosen from CSSCI Chinese data set.

To show the potential power of the mutation operators and for a faster convergence Premalatha and Natarajan (2009) proposed clustering of documents based on GA with dynamic mutation operators and adaptive mutation rates. The idea is simply suggested *N* mutation operators with equal mutation ratio. After specific generations, the mutation operator that produces better average fitness values might increases its control ratio. Other parameters and operators of GA remained the same as in the standard GA. The fitness function is derived from the cosine similarity. The number of clusters *k* was fixed to 3 only. The representation, as we believe and, shown in Fig. 6 is suitable for small data set since each chromosome represent the entire set of documents. The method is assumed theoretically better than simple GA.



1	2	3	4	5	6	…	n = dim. of data set
1	2	1	1	2	3	…	2

Fig. 6. Chromosome representation in (Premalatha and Natarajan, 2009)

Saini *et al*. (2011) proposed a weighed fitness function that combined the semantic similarity measure along with other two standard similarity measures, namely Jaccard and cosine similarity. The algorithm used real encoding schema, standard crossover and mutation operators, roulette wheel selection operator, population size of 15 chromosomes and it didn't use elitism. The 1414 document handled in the implementation was taken from the cisi data set. Matlab software was the tool for implementing the algorithm, alongside with Matlab toolbox Text to Matrix Generator (TMG). The algorithm proposed single measure to combine weights from the Jaccard, Cosine and similarity measures. Thereafter, the algorithm used the genetic algorithm to optimize these weights. This study indicated that no significant improvement has been seen in average fitness value of overall generation.

Leon *et al*. (2012) proposed a niching based GA, which they claimed that it is robust to noise and able to determine the number of clusters automatically. The algorithm finds and maintains dense area or clusters in the solution space using GA and niching techniques. Each chromosome represents a candidate cluster (center and scale). The center evolved using GA while the scale or cluster size is updated using hill climbing procedure. The algorithm used sparse real and sparse binary encoding with specialized genetic operator suitable for this sparse representation. The fitness function was based on cluster center and cluster scale. The algorithm didn't use elitism. Two well-known data sets had been used. Namely, the 20-newsgroup and the TREC-7 with 2000 and 7454 text documents respectively. The algorithm claimed to achieve different degree of exploitation and exploration in searching for the optimal cluster prototypes. Moreover, the results indicated that, the proposed clustering process clusters the data in ways that sometimes go beyond the predefined document classes, by either splitting a class into several clusters or by forming a cluster that is distributed among several clusters.

A patented document Clustering algorithm using GA Model (CGAM) was invented by Shi and Li (2013). It is a GA based k-means that also took into consideration the impact of the outliers and part of the speech. Concerning the representation, the Algorithm constructed two VSMs. The first VSM composed by the named titles, nouns and verbs, while the second VSM composed by the remaining part of the speech words. The final VSM is a weighted combination from these two VSMs. The Selection operator was the roulette wheel which based on the probability of chromosome over the sum of all probabilities in population. The crossover and mutation operators were based on the floating point encoding schema. The fitness function was based on the cosine similarity measure between each sample and each center. On the contrary to other previous reviewed algorithms, elitism was used in this algorithm. The data set based on both Chinese text corpus and Reuters 21578. It should be noted that some of the algorithm's parameters had been selected in an empirical basis such as the number of iterations, number of elites' chromosomes and more importantly the number of clusters *k*. The results showed that CGAM achieved better than other GA based k-means algorithms and has been applied in Chinese national program of business intelligent system. The entire implemented system claimed to fit the practical needs of automatic text clustering, text categorization and topic detection against huge document sets.

Finally, a research group in the Korean Chonbuk National University reported a series of studies on document clustering with evolutionary algorithms. Few of these studies were on the semantic properties, whereas the most were on other similarity measures. Hence, we will focus on the latter studies in this review. In all studies, all of the data sets were adopted from the Reuter-21578 data collection with varying data set sizes between 100 and 1000 documents at maximum in one study. While most of their studies used 200 documents from the Reuter data collection. Moreover, a single fitness function applied mainly in all of the studies, namely the inverse of Davies-Bouldin Index (DBI) which was used to determine the number of clusters.

Initially, (Song and Park, 2006) focused on the representation by adopting a Modification to the Variable length Genetic Algorithm (MVGA). An indexing technique applied to encode the chromosome in order to indicate the location of each gene. Consequently, more effective genetic operators were introduced. MVGA designed to automatically adjust the influence between the diversity of the population and selective pressure during generations. The results which compared with the conventional Variable length Genetic Algorithm (VGA) showed that MVGA converged slightly faster than VGA with the first data set. Also, it showed that MVGA evolved much faster and more accurate than VGA with the second data set used.

The Subsequent researches concentrated on the concept of dimensional reduction. Song and Park (2007b) focused on GA with dimension reduction based on Singular Value Decomposition (SVD). While in the later two studies the focus was on another type of dimension reduction, namely the Latent Semantic Indexing (LSI) (Song and Park, 2007a; 2009). Template Numerical Toolkit (TNT) used for computing the SVD. TNT took more computation time than Matlab, but this toolkit provided higher quality and more reliable decomposition results. The results showed that the performance of the dimensionally-reduced VSM with GA is significantly superior to that of conventional GA in VSM. The proposed algorithms could retain high F-measure even with very high rates in term reduction.

A double layered GA (DLGA), with the graphical structure shown in Fig. 7, had been proposed to tackle the problem of Premature Convergence Phenomenon (PCP) in (Choi *et al*., 2011). PCP is the problem of converging to a local optimum rather than global optima in the solution space. The implemented system showed that DLGA is stronger against PCP compared to conventional genetic clustering algorithm. In addition, it showed that the document clustering using genetic algorithms performs better than the traditional clustering algorithms (K-means, Group Average).

In addition to the single objective function used in their previous researches, namely the DB index, they adopted another objective function based on Calinski and Harabasz's (CH) validity Index in (Lee *et al*., 2011; Lee and Park, 2012). Their results showed that the performance of these two multi objective algorithms is higher than those of traditional document clustering and general genetic-based algorithms, but the computational time for the multi objective algorithms have increased.

Web Document Clustering

Most of the web pages on the internet basically consist of a structured hypertext files. Hypertext representation inherits all the essential steps of the plan text representation and preprocessing. However, it takes advantage of the extra information in HTML files such as the metadata, title and the visual features (bold, italic, underline, emphasize, strong, headline) and more. Accordingly, further efforts will be needed in the preprocessing phase and new challenges will be added to employ these extra information to crop efficient algorithms.

One of the pioneer researches in web document clustering with genetic algorithms presented by (Casillas *et al*., 2003). In this study, the algorithm was evaluated with a document set that were the output of a query in a search engine. That is a kind of clustered-based browsing. The assumptions were to provide a clustering for the search result without a prior knowledge of number of clusters *k* and to apply the clustering on small number of documents. Single objective function was used to estimate the number of clusters based on Calinski and Harabasz's (CH) rule. This function is approximately a kind of ratio of Between-Group Sum of Squared Distances (BGSS) to Within-Group Sum of Squared Distances (BGSS). Four data sets from a Spanish newspaper had been used containing 10, 12, 31 and 100 documents respectively. Unlike other followed researches, the representation was depended on the calculation of the Minimum Spanning Tree (MST). The experiments showed that at average the GA-based method got better results in a less time compared with CH-based method.

A lengthy and well-explained paper by (Carlantonio and Costa, 2009) developed a system called SAGH (Genetic Analytical System of Grouping Hypertexts) for clustering analysis of web documents based on genetic algorithms. The system was composed of seven modules. The first five modules were for preprocessing the hypertext, the sixth performed the cluster analysis and the seventh presented the results. SAGH used fixed size chromosome representation as shown in Fig 8. Selection was based on the classical roulette wheel selection. The crossover and mutation operators were oriented to groups.

The fitness function formulated on average silhouette width. The implemented system, which also applied elitism, didn't request any input parameters. The performance of SAGH system declared to be reasonably good. It recorded that for visualizing 400 documents it took 2 min and it took 30 sec for the 100 documents.

Zhengyu *et al*. (2010) enhanced their own work on web page document clustering presented in (Zhu *et al*., 2007). A Dynamic Genetic Algorithm (DGA) was designed then developed with Delphi language to overcome the shortages of their previous Hybrid Clustering Algorithm (HCA). The DGA improved the auto method of finding the number clusters *k*. It also improved the genetic operators, the fitness function and the encoding schema as well. DGA overcame the sensitivity in assigning the first page (*d'*) in its cluster, which might lead to incorrect number of clusters as shown in Fig. 9. The data set was 3300 downloaded web pages, arranged in 11 classes with 300 pages in each class. The genetic operators was nicely examined and modified to fit the problem. Specifically, the crossover adopted with changeable executive probability to achieve balance between selection pressure and convergence rate. While, the mutation adopted the Dynamic Splitting and Merging (DSAM) procedure to keep the number of cluster *k* fixed. i.e., when split was done for a large diameter cluster, another merge was done to two clusters with minimum centroids distance. Finally, a new third operator was introduced, called Local Adjustment (LA) operator, to overcome the weakness of genetic in local search compared with its ability in global search. The fitness functions in DGA made use of both concentrations (distances within each cluster) along with dispersion (distances among clusters) which was not taken into account in the HCA method. The enhanced encoding schema claimed to prevent falling in local optimization due to the variety between the fathers and child genes which wasn't in the previous schema.

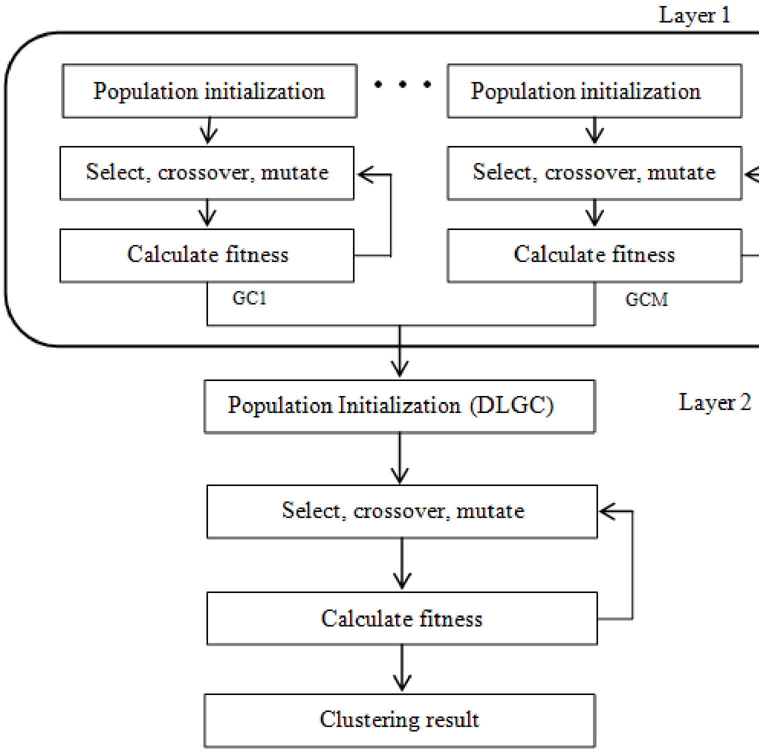

Layer 1
Population initialization
• • •
Population initialization
Select, crossover, mutate
Select, crossover, mutate
Calculate fitness
Calculate fitness
GC1
GCM
Population Initialization (DLGC)
Layer 2
Select, crossover, mutate
Calculate fitness
Clustering result

Fig. 7. The structure of the Double-Layered GA (DLGA) as proposed in (Choi *et al*., 2011)

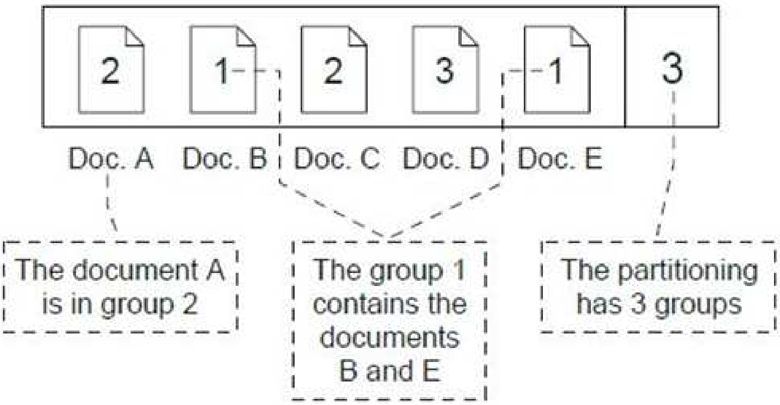
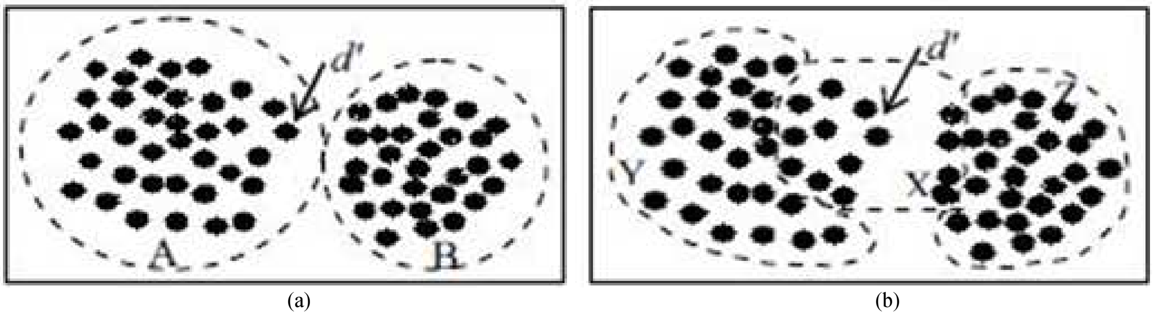

2
1
2
3
1
3
Doc. A
Doc. B
Doc. C
Doc. D
Doc. E
The document A is in group 2
The group 1 contains the documents B and E
The partitioning has 3 groups

Fig. 8. Fixed chromosome size representation by (Carlantonio and Costa, 2009)



(a)

(b)

Fig. 9. (a) A demo of two clusters (b) A demo of error clustering

Cobos *et al*. (2011; 2010; 2012) conducted a number of researches on web document clustering based on Evolutionary Computation (EC) algorithms and other optimization-based algorithms. The latest research (Cobos *et al*., 2012) was an approach for clustering the web document using genetic programming evolutionary algorithm. The novelty of this research was in obtaining the modified Bayesian Information Criteria (BIC) fitness function using the Genetic Programming (GP) in a reverse engineering view. The Representation was based on tree of expressions. As the genetic operators, rank selection, one- and two-point crossover and three kinds of mutation had been used. It is interesting to note that this new BIC fitness function presented better results than traditional BIC over 50 dataset based on DOMZ and 44 datasets based on ABIENT using a specific evolutionary algorithm.

Lastly, Liu *et al*. (2011) revealed a hypertext document clustering algorithm utilizing from additional information that may have more contribution for clustering, such as the Visual Features (VF). Precisely, it took into account the effects of text size, font and other appearance characteristics included in body, abstract, subtitle, keyword and title of the document. Hence, the weight of each term ($w_{i,j}$) was the ratio of weighted sum of each visual feature. The data set was taken from a Chinese corpus and the document similarity was presented by the cosine similarity. It is worth noting that the proposed VF-clustering algorithm made use of crossover and mutation thoughts of GA to improve the k-means algorithm. The analysis showed that the clustering result of the visual features was better than any single visual feature in representing documents. Although the VF-clustering algorithm adjusted the number of clusters *k* automatically using thoughts from GA, but it had introduced at least five unknown parameters for each weight of the visual feature used.

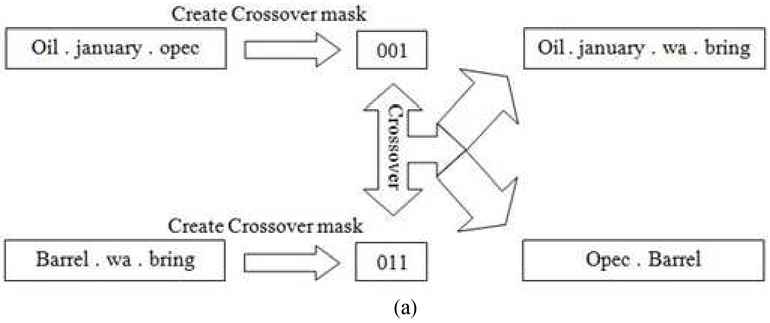
Keyword/keyphrase Clustering

The *keyword* is a significant or descriptive word within a document. The *keyphrase* is a phrase of two or more keywords to capture the main topic within a document. Early systems worked well in generating keywords/keyphrases for individual document, such as Keyphrase Extraction Algorithm (KEA) (Frank *et al*., 1999). The recent researches focused on finding keywords/keyphrases from the whole corpus for other clustering reasons. Such as: Clustering the keywords to improve the retrieval, reformulating the user queries through clustered terms (query expansion), or clustering the documents based on keywords selection/reduction. This subsection will review the recent work on this area.

Wu and Agogino (2004) established one of the pioneer researches of evolutionary algorithm on keyphrases. They had used the NSGA-II algorithm with two objectives. The first objective was the number of phrases selected and the second objective was the measure of dispersion of the phrase over the textual units in the document. Their results indicated that the algorithm can extract a good keyphrase set just by processing a set of documents in a certain domain without the need of any domain-specific knowledge or prior training. To assess the quality of the extracted phrases, a human evaluation procedure by total of six evaluators was carried out. It reported that over 80% of the keyphrases were accepted from the chosen data set. The data set was 34 papers taken from American Society of Mechanical Engineering-Design Theory and Methodology (ASME-DTM) conference. As a measure of performance and on a 1.8 GHz workstation, the algorithm took 5 h to converge.

Shamsinejadbabki and Saraee (2012) presented a GA-based method for keyword selection for document clustering. A new Modified Term Variance (MTV) measuring method was proposed to evaluate the grouping of terms. Binary representation was used for the presence or absence of a specific term in the phrase. The selection operator employed the standard roulette wheel selection. The crossover and mutation were also standards as shown in Fig. 10. The fitness function was based on the proposed MTV without using elitism in the algorithm. As a performance metric, the MTV-method showed better average accuracy and F1-measure comparing with the traditional Term Variance (TM) and Document Frequency (DF) methods over data set taken from Reuter-21578 corpus collection. It is also worth mentioning that there were some unknown parameters introduced by this algorithm for the GA operators and for the genetic encoding schema.

Sathya and Simon (2010) implemented a genetic-based algorithm to find out the combination of terms extracted from online documents. First a crawler was used to extract the terms from the documents then GA was used to generate the combination of terms. Thereafter, the results obtained from the GA were applied to IR system as a kind of query expansion. The fitness function was a ratio of the number of times the keywords appeared in the whole document over the total number of documents in the data set. Floating point representation was used to encode the chromosomes. Basic GA operators were applied. Namely, the selection operator was tournament selection and the crossover operator was the single point crossover. As a final result, the proposed system with the query expansion feature claimed to be more efficient than the traditional systems in terms of precession and recall metrics. The results had been evaluated over a data set consisting of 1000 documents chosen within a specific domain.



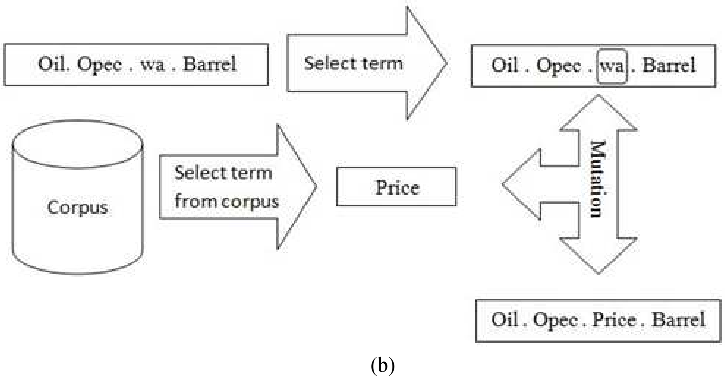


Fig. 10. (a) Crossover operation (b) mutation operation The GA combination operators in (Shamsinejadbabki and Saraee, 2012)

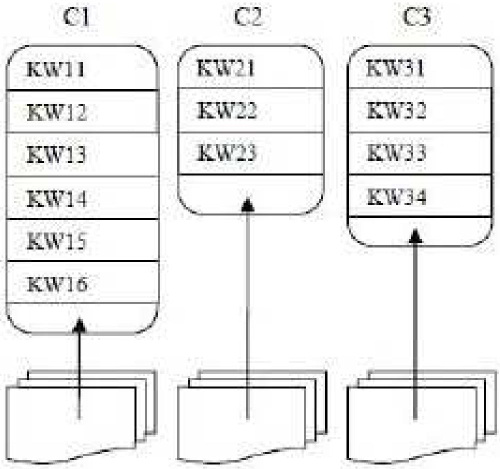


Fig. 11. Keyword cluster defined by one solution candidate

Yonghong and Wenyang (2010) introduced a genetic algorithm method for text clustering based on terms selection, or more precisely, terms reduction. The main characteristics of the proposed algorithm are: Binary bit-string representation, roulette wheel selection, standard crossover and mutation, no elitism used and the fitness function was based on the cosine similarity. It is worth to say that no data set was mentioned in the paper and the method had been proven mathematically. The research provided analysis and theorem proof that the algorithm can provide higher performance in computational complexity, clustering effects and high dimensional data clustering.

Dorfer *et al*. (2010) initially proposed a simple evolutionary strategy algorithm for keyword clustering. Next, in Dorfer *et al*. (2011) analyzed the performance of four different kinds of evolutionary algorithms for keyword clustering. Lastly, in Hooper and Paice (2005c) they presented a population diversity analysis in keyword cluster optimization using four different types of evolutionary algorithm. Namely Genetic Algorithm (GA), genetic algorithm with strict Off Spring Selection (OSGA), Evolution Strategy (ES) and the multi-objective elitist Non-Dominated Sorting Genetic Algorithm (NSGA-II). A keyword clustering solution is defined as a list of lists of keywords as shown in Fig. 11. The system conducted with the Heuristic Lab Software. The data set was taken from the TREC-9 conference 2000, which contained 36,890 publication information entries.

The base of Dorfer *et al*. (2012) researches was the developed fitness function which consists of six weighted parameters. Hence, these parameters needed a lot of weightening factors and parameter tuning to obtain meaningful results. The Final comparison results, with a specific parameter tuning for each algorithm, showed that the ES generates highly similar solutions then other EAs, whereas the OSGA maintains the diversity until the end of the runs.

The Objective Functions used in Document Clustering

The objective function (or fitness function) is the measure that evaluates the optimality of the generated evolutionary algorithm's solutions in the search space. In clustering domain, the fitness function refers to the adequacy of the partitioning. Accordingly, it needs to be formulated carefully, taken into consideration that the clustering is an unsupervised process. Different objective functions generate different solutions even form the same evolutionary algorithm. Presuming also that the fitness could either be a minimization or a maximization optimization function. Moreover, the algorithm could be formulated with one objective function or with multi objective functions. To sum up, “choosing optimization criterion is one of the fundamental dilemmas in clustering” (Das *et al*., 2009).

Broadly speaking, there are several measures appeared in the lectures to define the *proximity* (similarity or difference) between two documents or among set of documents. Examples of the similarity measures are Dice, Jaccard, Overlap and Cosine similarity measures. Examples of distance measures are the Minkowski, Mahalanobis, Euclidean and Manhattan distance measures. Beside proximity, there are measures to judge the correctness of the clustering such as the *internal* and *external validity indexes*, as mentioned earlier in section 2.6. Moreover, there are the inter-cluster measures that gauge the *separation* among clusters (such as single linkage, complete linkage, average linkage, centroids or ward methods) and the intra-clustering measures that gauge the *cohesion* within the components of a cluster (such as maximum, radius or average methods). What is interesting to know that all of the above categories of measures had been used in a way or another as an objective function to the evolutionary-based algorithms for document clustering.

The first column of Table 1-3 summarize the objective functions for the reviewed researches. The parameters of each function are explained briefly in the second column. The classification of optimality and the class of the employed measure are listed in the following columns.

Based on the observation for the functions and as presented in Table 1 and 2, we found out that the content and web document researches applied most of the measures, namely the inter and/or intra clustering, the proximity and the validity index measures. Additionally, most of these researches dealt with the problem as a maximization problem, except in (Wei *et al*., 2009) and (Cobos *et al*., 2011; 2010; 2012) because the intra-clustering and BIC are minimization in its nature. While in (Lee *et al*., 2011) and (Choi *et al*., 2011; Lee *et al*., 2011; Lee and Park, 2012; Song and Park, 2006; 2007a; 2007b), the researchers adopted the inverse of the DB index to convert the problem into a maximization problem.

Table 1. The objective functions used in the content clustering researches of section 4.1

Objective function	The function's parameters	Type of optimality	Type of measure	Reference (s)
$E=\sum_{j=1}^{K}\sum_{x_i\in C_j}\left(x_i-x_j^*\right)^2/n_j$	x_j*= the cluster center of c_j. n_j = no. of documents in c_j.	Min.	Intra clustering	(Wei *et al*., 2009)
$F=\dfrac{\sum_{j=1}^{p_c}\dfrac{\dfrac{m_{ij}\bullet O_j}{\lVert m_{ij}\rVert\bullet\lVert O_j\rVert}}{P_i}}{N_c}$	m_{ij} = jth document belongs to cluster i. O_j = centroids of the ith cluster. = no. of documents belong to cluster C_i. N_c = no. of clusters.	Max.	Variation of cosine similarity	(Premalatha and Natarajan, 2009)
$SBCSM=W_1*Semantic\_Similarity+W_2*Co\sin e\_Similarity+W_3*Jaccard\_Similarity$	W_1,W_2,W_3 = weights in range [0,1].	Max.	Weighted similarity	(Saini *et al*., 2011)
$\mathrm{f_i}=\dfrac{\sum_{j=1}^{N}\mathrm{W_{ij}}}{\sigma_i^2}$ where $\sigma_i^2=\dfrac{\sum_{j=1}^{N}W_{i,j}d_{,ij}^2}{\sum_{j=1}^{N}W_{ij}},W_{ij}=\exp\left(\dfrac{d_{i,j}^2}{2\sigma_i^2}\right)$	f_i =fitness value for the ith candidate center σ_i^2 = measure of dispersion of the candidate center. $W_{i,j}$ = is the measure of how typical is x_j belongs to cluster i. $d_{i,j}^2$ = Jaccard or Cosine distance of point x_j from center i.	Max.	Variation of inter and intra clustering	(Leon *et al*., 2012)

$f=\dfrac{1}{1+\sum_{j=1}^{k}\sum_{x_i\in C_j}(1-\cos(x_i,C_j))}$	k = no. of clusters.
	$Cos(x_i,C_j)$ = is the cosine similarity between sample x_i and center C_j.
		Max.	Cosine similarity	(Choi *et al*., 2011;
				Song and Park, 2006;
				2007a; 2007b; 2009)

$F=\dfrac{1}{DB}$	DB = Davies-Bouldin index.	Max.	Internal cluster validity index	(Shi and Li, 2013)
$F=\max(F_{DB}\wedge F_{CH})$	F_{DB} = Davies-Bouldin index. F_{CH} = Calinski and Harabasz index.	Max.	Internal cluster validity indexes	Lee *et al*., 2011; Lee and Park, 2012)

Table 2. The objective functions used in the web document clustering researches of section 4.2

Objective function	The function's parameters	Type of optimality	Type of measure	Reference (s)
$VRC = \frac{between\ cluster\ sum(seperation) / k - 1}{within\ cluster\ sum(cohesion) / n - k}$	k = no. of clusters.	Max.	Internal cluster	(Casillas *et al*., 2003)
	Variance Ratio Criterion (VRC) of Calinski and Harabasz's.		validity index	
$OF(chromosome) = \sum_{i=1}^{n} \frac{S(i)}{n}$	n = no. of documents.	Max.	Silhouette Coefficient	(Carlantonio and Costa, 2009)
$S(i) = Silhouette(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$	n = no. of documents.			
$S(i) = Silhouette(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$	*a(i)* = the arithmetic mean of the			
	distances of the document *i* to each document of group *i* . *b(i)* = the distance of object *i* to the next neighboring cluster.			
$F = \frac{\sum_{i=1}^{c} \sum_{x_j \in A_i} d\left(x_j, \bar{v}_i\right)}{\sum_{i=1}^{c} \sum_{j=1, i \neq j}^{c} d\left(\bar{v}_i, \bar{v}_j\right)}$	c = the no. of clusters (chromosome).	Max.	Inter and intra clustering	(Zhengyu *et al*., 2010;
	A_i = all the clusters of population. *d(x,y)* = cosine similarity. V_i (i=1,2, … c) are the centers.			Zhu *et al*., 2007)
$BIC = n * \ln\left(\frac{SSE}{n}\right) + k * \ln(n)$	n = no. of documents.	Min.	Bayesian	(Cobos *et al*., 2011;
$SSE = \sum_{j=1}^{k} \sum_{i=1}^{n} \left(p_{ij} * \left(1 - SimCos\left(x_i, c_j\right)\right)^2\right)$	k = no. of clusters.		Information criterion	Cobos *et al*., 2010)
$p_{ij=} \begin{cases} 1\, if\ x_i \in c_j \\ 0\ otherwise \end{cases}$	*SSE* = Sum Square Error.			
$BIC = n * \ln\left(\frac{SSE}{n * ADBC}\right) + k * \ln(n)$	n = no. of documents.	Min.	Bayesian information	(Cobos *et al*., 2012)
$SSE = \sum_{j=1}^{k} \sum_{i=1}^{n} \left(p_{ij} * \left(1 - SimCos\left(x_i, c_j\right)\right)^2\right)$	k = no. of clusters.		Criterion	
$ADBC = \frac{2}{n * (n-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \left(1 - SimCos\left(c_i, c_j\right)\right)$	*SSE* = Sum Square Error.			
$p_{ij=} \begin{cases} 1\, if\ x_i \in c_j \\ 0\ otherwise \end{cases}$	*ADBC* = Average distance between centers.			
It is not a pure GA, but rather an improvement to the k-means algorithm using two of the GA operators, specifically: 1- Mutation = to change the cluster centers. i.e., the value of the centers 2- Crossover = to split/merge the clusters. i.e., changing the number *k* in k-means algorithm				(Liu *et al*., 2011)

These setting and observation are useful especially when it comes to the issue of implementing more than one conflicting objective function in the multi objective evolutionary algorithms.

On the contrary, the keyword/key phrase clustering showed diversity in formulating or choosing the objective function. Except for the first of the two functions presented in which is a kind of separation measure, all of rest of these clustering algorithms used either generated or statistical measures to define the objective function. Column 4 in Table 3 illustrates the category of each objective function as summarized from the reviewed research. Note also that, most of the objective functions are tend to be maximization except in the two objective functions of (Wu and Agogino, 2004) and the weighted function of parameters in (Dorfer *et al*., 2012) and in (Dorfer *et al*., 2011; 2010) respectively.

It is also important to know that each implemented algorithm has its own characteristics. These characteristics were previously highlighted in the previous sections. The emphasis, however, was on the objective function which is the milestone of the evolutionary algorithms as it evaluates solutions fitness. The ultimate aim is to make these objective functions comparable and to be developed more easily in later studies.

Conclusion and Future Directions

Document Clustering is the research issue of increasingly many studies. After each research stage, researchers combined and classified these studies in reviews or survey papers. A number of these previous reviews dealt with the specific nature of the text document clustering problem and the corresponding conventional solutions for it. The rest of the reviews explicitly discussed the evolutionary algorithm for clustering the generated two dimensional data, whilst the document clustering is high dimensional problem in its nature.

Table 3. The objective functions used in the keyphrase clustering researches of section 4.3

Objective function	The function's parameters	Type of optimality	Type of measure	Reference (s)
$M_c = {}^{N}\!/_{E}$ $where \quad E = D\left[1-\left(1-\frac{1}{D}\right)^{T}\right]$ 1st objective: M$_c$= measure of dispersion. 2nd objective: no. of phrases selected.	N = number of textual Units that actually contains the phrase D = no. of textual unit in the repository. T = total occurrence of the phrase.	Min.	Inter Clustering and the frequency of the phrases	(Wu and Agogino, 2004)
$fitness(ch_i) = mtv(ch_i, th) * \ln(ch_i + 1)$ $mtv = \sum_{i=1}^{N}[f_{i,j,th} - f_{i,th}]^{2}$	*mtv* = modified term variance. *th* = no. of terms should be in document. $f_{i,j,th}$ = frequency of term *i* in document *j*. $f_{i,th}$ = frequency of term *i* in corpus.	Max.	Statistical variance	(Shamsinejadbabki and Saraee, 2012)
$F = 1 - \frac{n}{N}$	*n* = occurrence of term in document. *N* = total no. of documents.	Max.	A simple ratio of frequencies	(Sathya and Simon, 2010)
$F = \frac{1}{N}\sum_{i=1}^{N} Cos(C, D_i)$ $where\ c = \frac{1}{n}\sum_{k=1}^{n} d_{ij}$	*N* = no. of documents. d_{ij} = Weight of term *i* in document (*tfidf*).	Max.	Text set density	(Yonghong and Wenyang, 2010)

$F = w_1A + w_2B + w_3C +$
$w_4D + w_5E + w_6G$
	W_i = weight parameters.
	A=distinct documents/total documents.
	B=no. of doc.s assigned to keyword clusters
	C=mean average cluster confidence.
	D=mean average document confidence.
	E=the σ of no. of doc.s assigned.
	G=no of generated clusters (k).
		Min.	Weighted function
			of parameters
				(Dorfer *et al.*, 2012;
				Dorfer *et al.*, 2011;
				Dorfer *et al.*, 2010)

In this review, we firstly summarized some significant of those review studies. Additionally and as a main target scope, we had reviewed several research papers that dealt specifically with the clustering of documents from the evolutionary algorithm point of view. Besides that, details for the general model for document clustering have been described. Different term weighting schemas, stemming algorithms, cluster validity indices and a list of dimensional reduction techniques suitable for document clustering have been shown. A number of sources to the data sets had been provided. Finally, various objective functions from range of research papers have been carefully grouped, classified and illustrated.

When dealing with document clustering from evolutionary algorithm point of view, three groups of researches had been explored. The first group of research focused merely on the textual contents of the documents without any additional information. Whereas, the second group of researches focused on the web text document and made use of the metadata, visual and other features associated with these documents. All of those two types of researches benefited from standard measures to define its fitness function, such as the cosine similarity or the measure of separation between clusters and so on. The third researches' group, the keyword/keyphrase clustering, took a different turn in employment its version of evolutionary algorithms in document clustering. In these algorithms most of the fitness functions were derived from the statistical concepts of frequency for keyword, keyphrase, terms or document in the dataset. Besides the chosen or derived objective function, it should be noted that each implemented algorithm has its own added characteristics such as: Introducing an efficient encoding schema, modifying or adding new evolutionary operators, minimizing or even canceling the unknown input parameters for the algorithm, implementing hybrid algorithm based on another existing method, or enhancing the algorithm performance.

Because the notation of “good cluster” cannot be precisely defined, there were many algorithm developed for clustering including the evolutionary algorithm. A number of issues still open and needs further research. For instance, most of the research assumed hard clustering when partitioning the document data. Hence, there is a need to investigate the performance of the algorithms with the overlapped or fuzzy clustering. Likewise, the majority of EA-based algorithms carried out with single objective function. For that reason, more efforts are required to consider the emerging multi objective EA-algorithms. In addition, the group-oriented EA operators rather than the "bitwise" operators need more attention. Outside the scope of the algorithm design, the effect of applying the optional dimension reduction process should also taken into consideration along with the keyphrase feature selection methods. The authors are currently working in these directions. Finally, there is a need to incorporate and assess these document clustering algorithms into applications such as query expansion and cluster-based browsing.

Acknowledgement

We would like to acknowledge Faculty of Computer Science and Information Technology, Universiti Putra Malaysia for supporting our work.

Funding Information

This article has been funded by Fundamental Research Grant Scheme (FRGS) under Ministry of Higher Education, Malaysia.

Author's Contributions

Sarmad Makki: He had analyzed comprehensively the works that are related. He had contributed on preparation, development and modification of this manuscript.

Razali Yaakob: He had contributed on drafting, analyzing, reviewing, final correction and approval of this manuscript.

Norwati Mustapha: She had contributed on analyzing, reviewing it critically for significant intellectual content, final approval of this manuscript.

Hamidah Ibrahim: She had contributed on analyzing, reviewing it critically for significant intellectual content, final approval of this manuscript.

Ethics

There is no ethical issues will be arised since everybody had been agreed based on their contribution.

References

Aggarwal, C.C. and C. Zhai, 2012. A survey of text clustering algorithms. Mining Text Data, 1: 77-128. DOI: 10.1007/978-1-4614-3223-4_4

Amala Bai, V.M. and D. Manimegalai, 2010. An analysis of document clustering algorithms. Proceedings of the IEEE International Conference on Communication Control and Computing Technologies, Oct. 7-9, IEEE Xplore Press, Ramanathapuram, pp: 402-406. DOI: 10.1109/ICCCCT.2010.5670585

Benbrahim, H. and M. Bramer, 2009. Text and hypertext categorization. Artificial Intellig. Int. Perspective, 5640: 11-38. DOI: 10.1007/978-3-642-03226-4_2

Carlantonio, L. and R.M. Costa, 2009. Exploring a Genetic Algorithm for Hypertext Documents Clustering. In: Intelligent Text Categorization and Clustering, Nedjah, N., L. Macedo, J. Mourelle, F. Kacprzyk and G. França *et al*. (Eds.), Springer, Berlin Heidelberg, pp: 95-117.

Casillas, A., M.T.G. Lena and R. Martيnez, 2003. Document Clustering Into an Unknown Number of Clusters Using a Genetic Algorithm. In: Text, Speech and Dialogue, Matouڑek, V. and P. Mautner (Eds.), Springer, Berlin Heidelberg, pp: 43-49.

Choi, L.C., J.S. Lee and S.C. Park, 2011. Double Layered Genetic Algorithm for Document Clustering. In: Software Engineering, Business Continuity and Education, T.H. Kim, H. Adeli, H.K. Kim, H.J. Kang and K. Kim *et al*. (Eds.), Springer, Berlin Heidelberg, pp: 212-218.

Cobos, C., 2011. DMOZ Data Sets, for clustering and categorization.

Cobos, C., M. Mendoza and E. Leń, 2011. A hyper-heuristic approach to design and tuning heuristic methods for web document clustering. Proceedings of the IEEE Congress on Evolutionary Computation (CEC), Jun. 5-8, IEEE Xplore Press, New Orleans, pp: 1350-1358. DOI: 10.1109/CEC.2011.5949773

Cobos, C., C. Montealegre, M.F. Mejيa, M. Mendoza and E. Leń, 2010. Web document clustering based on a new niching memetic algorithm, term-document matrix and Bayesian information criterion. Proceedings of the IEEE Congress on Evolutionary Computation (CEC), Jul. 18-23, IEEE Xplore Press, Barcelona, pp: 1-8. DOI: 10.1109/CEC.2010.5586016

Cobos, C., L. Muȍz, M. Mendoza, E. Leń and E. Herrera-Viedma, 2012. Fitness Function Obtained from a Genetic Programming Approach for Web Document Clustering Using Evolutionary Algorithms. In: Advances in Artificial Intelligence-Iberamia Pavń, J., N.D. Duque-Méndez and R.F. Fernلndez (Eds.), Springer, Colombia, pp: 179-188.

Cutting, D.R., D.R. Karger, J.O. Pedersen and J.W. Tukey, 1992. Scatter/gather: A cluster-based approach to browsing large document collections. Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Jun. 21-24, New York, pp: 318-329. DOI: 10.1145/133160.133214

Das, S., A. Abraham and A. Konar, 2009. Metaheuristic Pattern Clustering-An Overview. In: Metaheuristic Clustering, Das, S., A. Abraham and A. Konar (Eds.), Springer, Berlin Heidelberg, pp: 1-62.

Dorfer, V., S. Winkler, T. Kern, G. Petz and P. Faschang, 2012. Analysis of Single-Objective and Multi-Objective Evolutionary Algorithms in Keyword Cluster Optimization. In: Computer Aided Systems Theory-Eurocast, Moreno-Dچaz, R., F. Pichler and A. Quesada-Arencibia (Eds.), Springer, Berlin Heidelberg, pp: 408-415.

Dorfer, V., S.M. Winkler, T. Kern, S.A. Blank and G. Petz *et al*., 2011. On the performance of evolutionary algorithms in biomedical keyword clustering. Proceedings of the 13th Annual Conference Companion on Genetic and Evolutionary Computation, Jul. 12-16, New York, pp: 511-518. DOI: 10.1145/2001858.2002041

Dorfer, V., S.M. Winkler, T. Kern, G. Petz and P. Faschang, 2010. Optimization of keyword grouping in biomedical information retrieval using evolutionary algorithms. Proceeding of the 22th European Modeling and Simulation Symposium, (MSS' 10), Morocco, pp: 25-30.

Fasheng, L. and X. Lu, 2011. Survey on text clustering algorithm. Proceedings of the 2nd International Conference on Software Engineering and Service Science, Jul. 15-17, IEEE Xplor Press, Beijing, pp: 901-904. DOI: 10.1109/ICSESS.2011.5982485

Fodor, I.K., 2002. A survey of dimension reduction techniques: Center for applied scientific computing. Lawrence Livermore National Laboratory.

Frakes, W.B. and R. Baeza-Yates, 1992. Information Retrieval: Data Structures and Algorithms. 1st Edn., Prentice Hall, Englewood Cliffs, ISBN-10: 0134638379, pp: 504.

Frank, E., G.W. Paynter, I.H. Witten, C. Gutwin and C.G. Nevill-Manning, 1999. Domain-specific keyphrase extraction. Proceedings of the 16th International Joint Conference on Artificial Intelligence, (CAI' 99), Stockholm, Sweden.

Halkidi, M., Y. Batistakis and M. Vazirgiannis, 2001. On clustering validation techniques. J. Intellig. Inform. Syst., 17: 107-145. DOI: 10.1023/A:1012801612483

Han, J. and M. Kamber, 2011. Data Mining: Concepts and Techniques: Concepts and Techniques. 3rd Edn., Elsevier, Burlington, ISBN-10: 0123814804, pp: 744.

Harman, D., 1991. How effective is suffixing? J. Am. Society Inform. Sci., 42: 7-15.

Hooper, R. and C. Paice, 2005a. Algorithm implementations. Lancaster University.

Hooper, R. and C. Paice, 2005b. Other stemmers. Lancaster University.

Hooper, R. and C. Paice, 2005c. The Paice/Husk stemming algorithm. Lancaster University.

Hruschka, E.R., R.J.G.B. Campello, A.A. Freitas and A.P.L.F. De Carvalho, 2009. A survey of evolutionary algorithms for clustering. IEEE Trans. Syst. Man Cybernet. Part C: Appli. Rev., 39: 133-155. DOI: 10.1109/TSMCC.2008.2007252

IndiraPriya, P. and D.K. Ghosh, 2013. A survey on different clustering algorithms in data mining technique. Int. J. Modern Eng. Res., 3: 267-274.

Jain, A.K., M.N. Murty and P.J. Flynn, 1999. Data clustering: A review. ACM Comput. Surveys, 31: 264-323. DOI: 10.1145/331499.331504

Lang, K., 2008. The 20 Newsgroups data set.

Lee, J.S., L.C. Choi and S.C. Park, 2011. Multi-Objective Genetic Algorithms, NSGA-II and SPEA2, for Document Clustering. In: Software Engineering Business Continuity and Education, Kim, T.H.H. Adeli, H.K. Kim, H.J. Kang and K. Kim *et al*. (Eds.), Springer Berlin Heidelberg, pp: 219-227.

Lee, J.S. and S.C. Park, 2012. Document clustering using multi-objective genetic algorithms on MATLAB distributed computing. Proceedings of the International Conference on Information Science and Applications, May 23-25, IEEE Xplore Press, Suwon, pp: 1-6. DOI: 10.1109/ICISA.2012.6220980

Leon, E., J. Gomez and O. Nasraoui, 2012. A genetic niching algorithm with self-adaptating operator rates for document clustering. Proceedings of the 8th Latin American Web Congress, Oct. 25-27, IEEE Xplore Press, Cartagena de Indias, pp: 79-86. DOI: 10.1109/LA-WEB.2012.22

Lewis, D.D., 2004. Reuters-21578 Data Set.

Li, X., 2012. TDSCAN: A density based algorithm for text documents clustering. Department of Computer Science. University of Illinois at Urbana-Champaign.

Liping, J., 2005. Survey of text clustering. Department of Mathematics, The University of Hong Kong.

Liu, L., J. Kang, J. Yu and Z. Wang, 2011. Document clustering method based on visual features. Proceedings of the International Conference on and 4th International Conference on Cyber, Physical and Social Computing Internet of Things, Oct. 19-22, IEEE Xplore Press, Dalian, pp: 458-462. DOI: 10.1109/iThings/CPSCom.2011.69

Lovins, J.B., 1968. Development of a stemming algorithm. Mechanical Trans. Comput. Linguist., 11: 23-31.

Luying, L., K. Jianchu, Y. Jing and W. Zhongliang, 2005. A comparative study on unsupervised feature selection methods for text clustering. Proceedings of the IEEE International Conference on Natural Language Processing and Knowledge Engineering. 30 Oct.-1 Nov., IEEE Xplore Press, pp: 597-601. DOI: 10.1109/NLPKE.2005.1598807

Manning, C.D., P. Raghavan and H. Schütze, 2008. Introduction to Information Retrieval. 1st Edn., Cambridge University Press, New York, ISBN-10: 0521865719, pp: 482.

Mary, S.A.L. and K.R.S. Kumar, 2012. A density based dynamic data clustering algorithm based on incremental dataset. J. Comput. Sci., 8: 656-664. DOI: 10.3844/jcssp.2012.656.664

NIST, 2000. TREC Data Set.

Paice, C.D., 1990. Another stemmer. SIGIR Forum, 24: 56-61. DOI: 10.1145/101306.101310

Palsonkennedy, R. and T.V. Gopal, 2012. Matching LSI for scalable information retrieval. J. Comput. Sci., 8: 2083-2097. DOI: 10.3844/jcssp.2012.2083.2097

Patel, D. and M. Zaveri, 2011. A Review on Web Pages Clustering Techniques. In: Trends in Network and Communications, Wyld, D., M. Wozniak, N. Chaki, N. Meghanathan and D. Nagamalai (Eds.), Springer Berlin Heidelberg, pp: 700-710.

Pavan, K.K., A.A. Rao, A.V.D. Rao and G.R. Sridhar, 2010. Single pass seed selection algorithm for k-means. J. Comput. Sci., 6: 60-66. DOI: 10.3844/jcssp.2010.60.66

Piatetsky-Shapiro, G., 1993. Kdnuggets Data Repositories.

Porter, M., 2006. The porter stemming algorithm.

Porter, M.F., 1980. An algorithm for suffix stripping. Program: Electronic Library Inform. Syst., 14: 130-137.

Premalatha, K. and A.M. Natarajan, 2009. Genetic algorithm for document clustering with simultaneous and ranked mutation. Modern Applied Sci., 3: 75-82. DOI: 10.5539/mas.v3n2p75

Radwan, A.A.A., B.A. AbdelLatef, A.A. Ali and O.A. Sadek, 2006. Using genetic algorithm to improve information retrieval systems. Enformatika, 17: 6-12.

Rendon, E., I.M. Abundez, C. Gutierrez, S.D. Zagal and A. Arizmendi *et al.*, 2011. A comparison of internal and external cluster validation indexes. Proceedings of the 5th WSEAS International Conference on Computer Engineering and Applications, (CEA' 11), Puerto Morelos, Mexico, pp: 158-163.

Saini, M., D. Sharma and P.K. Gupta, 2011. Enhancing information retrieval efficiency using semantic-based-combined-similarity-measure. Proceedings of the International Conference on Image Information Processing, Nov. 3-5, IEEE Xplore Press, Himachal Pradesh, pp: 1-4. DOI: 10.1109/ICIIP.2011.6108982

Salton, G. and C. Buckley, 1988. Term-weighting approaches in automatic text retrieval. Inform. Process. Manage., 24: 513-523. DOI: 10.1016/0306-4573(88)90021-0

Salton, G. and C. Buckley, 1990. Improving retrieval performance by relevance feedback. J. Am. Society Inform. Sci., 41: 288-297. DOI: 10.1002/(sici)1097-4571(199006)41:4<288::aid-asi8>3.0.co;2-h

Salton, G., A. Wong and C.S. Yang, 1975. A vector space model for automatic indexing. Commun. ACM, 18: 613-620. DOI: 10.1145/361219.361220

Sathiyakumari, K., V. Preamsudha, G. Manimekalai and M.P. Scholar, 2011. A survey on various approaches in document clustering. Int. J. Comput. Technol., 2: 1534-1539.

Sathya, A.S. and B.P. Simon, 2010. A document retrieval system with combination terms using genetic algorithm. Int. J. Comput. Electrical Eng., 2: 1-6.

Shamsinejadbabki, P. and M. Saraee, 2012. A new unsupervised feature selection method for text clustering based on genetic algorithms. J. Intellig. Inf. Syst., 38: 669-684. DOI: 10.1007/s10844-011-0172-5

Sheikh, R.H., M.M. Raghuwanshi and A.N. Jaiswal, 2008. Genetic Algorithm Based Clustering: A Survey. Proceedings of the 1st International Conference on Emerging Trends in Engineering and Technology, Jul. 16-18, IEEE Xplore Press, Nagpur, Maharashtra, pp: 314-319. DOI: 10.1109/ICETET.2008.48

Shi, K. and L. Li, 2013. High performance genetic algorithm based text clustering using parts of speech and outlier elimination. Applied Intellig., 38: 511-519. DOI: 10.1007/s10489-012-0382-8

Song, W. and S.C. Park, 2006. Genetic algorithm-based text clustering technique: Automatic evolution of clusters with high efficiency. Proceedings of the 7th International Conference on Web-Age Information Management Workshops, (IMW' 06), pp: 17-17.

Song, W. and S.C. Park, 2007a. Analysis of web clustering based on genetic algorithm with latent semantic indexing technology. Proceedings of the 6th International Conference on Advanced Language Processing and Web Information Technology, Aug. 22-24, IEEE Xplore Press, Luoyang, Henan, China, pp: 21-26. DOI: 10.1109/ALPIT.2007.77

Song, W. and S.C. Park, 2007b. An efficient method of genetic algorithm for text clustering based on singular value decomposition. Proceedings of the 7th International Conference on Computer and Information Technology, Oct. 16-19, IEEE Xplore Press, Aizu-Wakamatsu, Fukushima, pp: 53-58.

Song, W. and S.C. Park, 2009. Genetic algorithm for text clustering based on latent semantic indexing. Comput. Math. Applic., 57: 1901-1907. DOI: 10.1016/j.camwa.2008.10.010

Steinbach, M., G. Karypis and V. Kumar, 2000. A comparison of document clustering techniques. University of Minnesota.

Tang, B., M. Shepherd, M. Heywood and X. Luo, 2005. Comparing Dimension Reduction Techniques for Document Clustering. In: Advances in Artificial Intelligence, Kégl, B. and G. Lapalme (Eds.), Springer Berlin Heidelberg, pp: 292-296.

Thangamani, M. and P. Thangaraj, 2010. Integrated clustering and feature selection scheme for text documents. J. Comput. Sci., 6: 536-541.

van-Rijsbergen, C.J., 1979. Information Retrieval. 2nd Edn., London: Butterworths.

Velmurugan, T. and T. Santhanam, 2010. Computational complexity between k-means and k-medoids clustering algorithms for normal and uniform distributions of data points. J. Comput. Sci., 6: 363-368.

Wang, J., Y. Mo, B. Huang, J. Wen and L. He, 2008. Web Search Results Clustering Based on a Novel Suffix Tree Structure. In: Autonomic and Trusted Computing, Rong, C., M. Jaatun, F. Sandnes, L. Yang and J. Ma (Eds.), Springer Berlin Heidelberg, pp: 540-554.

Wei, J.X., H. Liu, Y.H. Sun and X.N. Su, 2009. Application of genetic algorithm in document clustering. Proceedings of the International Conference on Information Technology and Computer Science, Jul. 25-26, IEEE Xplore Press, Kiev, pp: 145-148. DOI: 10.1109/ITCS.2009.269

Weise, T., 2011. Global optimization algorithms-theory and application.

Willett, P., 1988. Recent trends in hierarchic document clustering: A critical review. Inf. Process. Manag., 24: 577-597. DOI: 10.1016/0306-4573(88)90027-1

Wu, J.L. and A.M. Agogino, 2004. Automating keyphrase extraction with multi-objective genetic algorithms. Proceedings of the 37th Annual Hawaii International Conference on System Sciences, Jan. 5-8, IEEE Xplore Press.
DOI: 10.1109/HICSS.2004.1265278

Xiao, Y., 2010. A Survey of Document Clustering Techniques and Comparison of LDA and moVMF. North Carolina State University.

Yonghong, Y. and B. Wenyang, 2010. Text clustering based on term weights automatic partition. Proceedings of the 2nd International Conference on Computer and Automation Engineering, Feb. 26-28, IEEE Xplore Press, Singapore, pp: 373-377.
DOI: 10.1109/ICCAE.2010.5451390

Zamir, O. and O. Etzioni, 1999. Grouper: A dynamic clustering interface to web search results. Comput. Netw., 31: 1361-1374.
DOI: 10.1016/S1389-1286(99)00054-7

Zeng, H.J., Q.C. He, Z. Chen, W.Y. Ma and J. Ma, 2004. Learning to cluster web search results. Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Jul. 25-29, Sheffield, United Kingdom, pp: 210-217.
DOI: 10.1145/1008992.1009030

Zhao, Y. and G. Karypis, 2004. Empirical and theoretical comparisons of selected criterion functions for document clustering. Machine Learn., 55: 311-331.
DOI: 10.1023/B:MACH.0000027785.44527.d6

Zhengyu, Z., H. Ping, Y. Chunlei and L. Lipei, 2010. A dynamic genetic algorithm for clustering web pages. Proceedings of the 2nd International Conference on Software Engineering and Data Mining, Jun. 23-25, IEEE Xplore Press, Chengdu, pp: 506-511.

Zhu, Z., Y. Tian, J. Xu, X. Deng and X. Ren, 2007. An improved partitioning-based web documents clustering method combining GA with ISODATA. Proceedings of the 4th International Conference on Fuzzy Systems and Knowledge Discovery, Aug. 24-27, IEEE Xplore Press, Haikou, pp: 208-213.
DOI: 10.1109/FSKD.2007.165