KNOWLEDGE ENGINEERING: PRINCIPLES AND TECHNIQUES
Proceedings of the International Conference on Knowledge Engineering,
Principles and Techniques, KEPT2007
Cluj-Napoca (Romania), June 6–8, 2007, pp. 25–32

A COMPARISON OF QUALITY CRITERIA FOR UNSUPERVISED CLUSTERING OF DOCUMENTS BASED ON DIFFERENTIAL EVOLUTION

D. ZAHARIE[(1)], F. ZAMFIRACHE[(2)], V. NEGRU[(3)], D. POP[(4)], AND H. POPA[(5)]

Abstract. This paper presents an analysis of different quality criteria for unsupervised clustering of documents. The comparative analysis is based on a differential evolution algorithm which allows the estimation both of the number of clusters and of their representatives. The proposed approach is tested on a classical data set in document clustering. The results illustrate the particularities of different clustering criteria and the ability of the proposed approach to identify both the number of clusters and their representatives.

1. Introduction

Clustering is one of the first steps in organizing large sets of electronic documents and it plays an important role in topic extraction and in guiding the documents browsing. In a general sense, clustering means identifying natural groups in data such that data in a group, called cluster, are sufficiently similar while data belonging to different groups are sufficiently dissimilar. Clustering is, usually, an unsupervised process based only on the data to be analyzed. However, most partitional algorithms (e.g. k-Means) need the knowledge of the expected number of clusters. When even this number is unknown the process is called unsupervised clustering.

The main issues in designing a document clustering system are: (i) finding an adequate encoding of the documents; (ii) finding appropriate similarity measures between documents and appropriate quality measures of the clustering result; (iii) choosing a clustering technique.

The ideal encoding of documents is highly task-dependent since certain features should be taken into account when dealing with documents containing only text and other kind of features when processing web-documents [8]. In the case of text clustering a common representation is that based on the vector-space model

2000 *Mathematics Subject Classification.* 62H30, 68T20.
Key words and phrases. document clustering, clustering criteria, differential evolution.

using the term weighting based on the term frequency combined with the inverse document frequency.

Because of the unsupervised character, evaluating the quality of a clustering result is a difficult task. Unfortunately there does not exist one unique quality criterion but at least two of them should be combined. A thorough analysis of the influence of different quality criteria and of their combinations on document clustering is presented in [9]. However this analysis is based on the hypothesis that the number of clusters is known and the quality criteria are used only to compare partitions containing the same number of clusters. One of the aims of this work is to analyze the effectiveness of some combinations of quality criteria when they are used in fully unsupervised document clustering.

From the traditional clustering techniques the partitional ones are, by far, the most used in document clustering. This is due not only to the fact that the partitional techniques are less computational expensive than the hierarchical ones but also to the fact that, as reported in [10], they lead to comparable or even better clustering performance. Partitional approaches can be interpreted as optimization problems having as aim to maximize the quality criteria. The involved optimization problem is a complex one, making simple searching strategies to be easily trapped in local minima. In the last decade a lot of evolutionary and other nature-inspired approaches in clustering have been proposed [4, 5]. Recently some evolutionary related approaches have been applied also to document clustering. In [3] is presented a Particle Swarm Optimization approach while in [1] is presented a document clustering method based on the Differential Evolution (DE) algorithm. Both approaches proved to behave better than k-Means but they are based on the knowledge of the number of clusters. In this work we propose an extension of the approach in [1] characterized by the fact that both the number of clusters and their representatives are evolved.

2. Preprocessing and representation of documents

The set of documents to be clustered should be first preprocessed in order to find an appropriate representation of each document. If the documents are interpreted as plain text they could be considered bag of words. Since not all words appearing in a document are relevant, a first step would be to just eliminate the words which are very common in the language. This is usually done by using a so-called stop list specific to the document language. There currently exist stop lists for different languages. In our experiments we used such a classical stop list for English.

Another common processing is that of stemming, i.e. reducing derived words to their root form. One of the most used algorithms, which we also used in our work, is that of Porter [6]. Even if stemming is a frequently used procedure it is not always beneficial for the clustering process, as is illustrated in [7].

After these steps, each document will be a multi-set of terms (stemmed words) which can have a large number of elements. In order to reduce the dimensionality

corresponding to a document the terms having a low-frequency over the entire set of documents could be eliminated. In our approach we avoided to apply this in order to not eliminate terms which could play an important role in discriminating the clusters. In order to obtain a numerical finger-print of each document a score is assigned to each term belonging to the document. One of the most used approaches is that of using the term frequency in the document and the frequency of documents containing that term. Let us consider a set of N documents, D_1, D_2, ..., D_N and let t be a term. The weight of term t with respect to document D_i is defined as follows:

$$w(t, D_i) = f(t, D_i) \log(N/F(t)) \tag{1}$$

where $f(t, D_i)$ denotes the relative frequency of term t in the document D_i and $F(t)$ denotes the number of documents which contain the term t. It is easy to see that $w(t, D_i) \in [0, \log(N)]$. The minimal value corresponds to terms belonging to all documents and large values are obtained for terms which appear very frequently but only in one document. In order to limit the size of a document description, only the weights corresponding to terms in the document were stored. Each document can be thus described by the set of weights corresponding to the terms it contains: $\{w(t, D_i); t \in D_i\}$. Based on this representation the classical cosine similarity measure between two documents, D_i and D_j, can be defined as follows:

$$s(D_i, D_j) = \frac{\sum_{t \in D_i \wedge t \in D_j} w(t, D_i) w(t, D_j)}{\|D_i\| \|D_j\|} \tag{2}$$

where $\|D\| = \sqrt{\sum_{t \in D} w(t, D)^2}$ is in fact the Euclidean norm of the vector of weights corresponding to terms in D. Even if in the method implementation each document is described by the weights of terms belonging to the document in the following we shall formally consider that each document corresponds to a vector having the length equal with the number of terms in the entire set of documents. In this way all operations on vectors (summation and multiplication) are also valid on documents.

3. Choosing quality criteria for unsupervised clustering

The aim of partitional clustering is to find a partition $(C_1, \ldots, C_k)$ of the set $D = \{D_1, \ldots, D_N\}$ of documents such that $D = \cup_{i=1}^{k} C_i$ and $C_i \cap C_j = \emptyset$ for all $i \neq j$. A cluster C_r can be described by the set of all documents belonging to it or by a representative R_r which could be an element of D or another element from the vector space corresponding to the documents encoding ($[0, \log(N)]^\tau$, τ being the number of all terms in the set of documents). Based on their representatives, the clusters can be constructed by assigning each document to the nearest representative. A particular case of representatives is represented by the clusters centers, $R_r = (\sum_{D \in C_r} D)/n_r$, n_r being the number of elements of C_r.

The aim of the clustering process is to find that partition which maximizes the similarity between the elements of the same cluster and minimizes the similarity between elements belonging to different clusters. Thus partitional clustering can be formulated as an optimization problem involving one or multiple optimization criteria. Typical quality criteria of a partition are compactness, connectedness and separability.

Compactness is a measure of the concentration of data inside a cluster. It should be maximized and can be expressed as either the averaged similarity between all pairs of elements in the cluster or the total similarity between the elements in the cluster and its center. The most used is the second variant, characterized through

$$\mu_1(C_1,\ldots,C_k)=\sum_{r=1}^{k}\sum_{D\in C_r}s(D,R_r)=\sum_{r=1}^{k}\|S_r\| \tag{3}$$

where S_r is the sum of all documents in C_r and $R_r = S_r/n_r$ is the center of C_r. When the cluster representatives are not necessarily their centers then the last equality is no necessarily true.

Connectedness evaluates the degree to which similar documents (neighboring data) have been placed in the same cluster. The corresponding measure is [4]:

$$\mu_2(C_1,\ldots,C_k)=\frac{1}{N}\sum_{i=1}^{N}\frac{1}{L}\sum_{j=1}^{L}\gamma(D_i,D_{\nu(i,j)}) \tag{4}$$

where $D_{\nu(i,j)}$ denotes the j-th nearest neighbor of document D_i,

$$\gamma(D_i,D_{\nu(i,j)})=\begin{cases}1/j & \text{if } D_i \text{ and } D_{\nu(i,j)} \text{ are placed in the same cluster}\\ 0 & \text{otherwise}\end{cases} \tag{5}$$

and L is the number of considered nearest neighbors. While compactness favors spherical clusters, connectedness allows the generation of arbitrary shaped clusters.

Separability measures how the various clusters are different from each other. Unlike the previous measures this one should be minimized. A measure, corresponding to the case when the representatives are the clusters centers is:

$$\mu_3(C_1,\ldots,C_k)=\sum_{r=1}^{k}n_r s(R_r,R)=\sum_{r=1}^{k}n_r s(S_r,S) \tag{6}$$

where S is the sum of all documents in the set and $R = S/N$. In [1] is used a different separability measure, based on the maximal similarity between the clusters representatives (which usually are not the clusters centers but vectors generated by the evolutionary algorithm):

$$\mu_4(C_1,\ldots,C_k)=\sum_{r=1}^{k}\max_{q=\overline{1,k},q\neq r}s(R_q,R_r) \tag{7}$$

In order to obtain a quality clustering criteria, the above measures should be combined. Different combinations were proposed in the literature. For instance in [9], besides other measures which are not presented here, was analyzed μ_1/μ_3, which proved to be the best. In [1] was used $1/(\mu_4/\mu_1 + \epsilon)$ with $\epsilon > 0$ a small correction value.

When the number of clusters is not predefined and the quality criteria are used to compare partitions having different numbers of clusters we have to take into account the natural bias of different measures to favor a small or a large number of clusters. Let us suppose that all document vectors are normalized ($\|D_i\| = 1, i = \overline{1, N}$). If the representatives are the centers of the clusters then the compactness measure μ_1 satisfies

$$\mu_1(C_1, \ldots, C_k) = \sum_{r=1}^{k} \| \sum_{i=1}^{n_r} D_{l(i)} \| \leq \sum_{r=1}^{k} \sum_{i=1}^{n_r} \|D_{l(i)}\| = N \tag{8}$$

where $l(r, i)$ denotes the index of the ith document in the cluster C_r. Since the trivial clustering which correspond to the case when each document is in its own cluster is characterized by a value of μ_1 equal to N it follows that by using only the μ_1 criterion and letting k to vary, the maximum will be attained for the maximal value of k. A similar behavior was remarked in the case when the representatives are not necessarily the centers but in this case this fact cannot be theoretically proven so easy.

On the other hand, it is easy to see that the maximal value of connectedness is obtained if all documents are assigned to one cluster, thus when trying to maximize μ_2 a small number of clusters is favored. In the case of the separability measure μ_3 (which should be minimized) the following relations hold:

$$\mu_3(C_1, \ldots, C_k) = \sum_{r=1}^{k} n_r \frac{S_r^T \cdot S}{\|S_r\|\|S\|} \geq \frac{1}{\|S\|} \sum_{r=1}^{k} n_r \frac{S_r^T \cdot S}{n_r} = \frac{1}{\|S\|} \sum_{i=1}^{N} D_i^T \sum_{j=1}^{N} D_j \tag{9}$$

since $\|S_r\| \leq \sum_{i=1}^{n_r} \|D_{l(r,i)}\| = n_r$. The last term in eq. 9 is the value of μ_3 corresponding to the case of N clusters, thus it follows that by minimizing μ_3 one maximizes k. On the other hand, by minimizing the separability measure μ_4, the partitions having a small number of clusters are favored.

If the number of clusters should be estimated, the optimization criterion should involve two measures which are characterized through opposed dependence on the number of clusters (otherwise trivial partitions having either 1 cluster or N clusters are obtained). In Table 1 are summarized all possible combinations of the above four measures by marking which ones favor the increase of the number of clusters and which favor their decrease. Combinations where both measures favor the same modification on the number of clusters are not appropriate when this number should be estimated. Thus the criteria μ_1/μ_3 which proved to have a good behavior in the case of a fixed number of clusters is no more appropriate in the case

of a variable number, since by favoring high values of k it leads to an overestimation of the number of clusters. Combining μ_2 with μ_4 one obtains a criterion which favor the small number of clusters and could lead to an underestimation of k. On the other hand, the criteria $\mu_1\mu_2$ used in [4], $1/(\mu_4/\mu_1+\epsilon)$ used in [1] and those obtained by combining μ_2 with μ_3 (μ_2/μ_3) or μ_3 with μ_4 ($1/(\mu_3\mu_4)$) ensures the compromise between small and large values of k.

(μ_1,μ_2) [4]	(μ_1,μ_3) [9]	(μ_1,μ_4) [1]	(μ_2,μ_3)	(μ_2,μ_4)	(μ_3,μ_4)
$(\uparrow,\downarrow)$	$(\uparrow,\uparrow)$	$(\uparrow,\downarrow)$	$(\downarrow,\uparrow)$	$(\downarrow,\downarrow)$	$(\uparrow,\downarrow)$

TABLE 1. Influence of different combinations of criteria on the evolution of the number of clusters when the combined clustering criterion is maximized ($\uparrow$ - favor the increase, $\downarrow$ - favor the decrease)

4. Application of differential evolution for unsupervised document clustering

In [1] is illustrated the fact that Differential Evolution (DE) provides better results than classical Genetic Algorithms and Particle Swarm Optimization when applied to document clustering. This is why we chose to extend this approach in order to deal with the case of an unknown number of clusters. DE is a simple evolutionary approach based on a particular type of recombination which involves three randomly selected parents in order to obtain an offspring. The differences between our approach and that in [1] are related with the population elements encoding and with the recombination operator.

In the approach we developed (called extended DE - eDE) each element of a population corresponds to a partition and has the following components: $(k, R_1, \ldots, R_k)$ where $k \in \{k_{min}, \ldots, k_{max}\}$ is the number of clusters and R_r, $r = \overline{1,k}$ are representatives of the clusters (vectors of weights associated to terms in the documents set). At the start of the evolutionary process, the population elements are randomly initialized: k is randomly selected from its range, $\{k_{min}, \ldots, k_{max}\}$, and for each cluster the representative is randomly selected from the entire set of documents, D. At each iteration of the evolutionary process for each element e_i ($i = \overline{1,m}$) of the population the following operations are executed:
(i)Three other distinct elements e_{j_1}, e_{j_2} and e_{j_3} are randomly selected from the population.
(ii)A new trial element $e' = (k', R'_1, \ldots, R'_{k'})$ is constructed as follows: $k' = \lfloor |k^{(j_1)} + F \cdot (k^{(j_2)} - k^{(j_3)})| \rfloor$ with a probability p and remains $k^{(i)}$ with the probability $1-p$. For each r the representative R'_r is constructed as a linear combination of randomly selected representatives from those three elements: $R'_r = R^{(j_1)}_{r_1} + F \cdot (R^{(j_2)}_{r_2} - R^{(j_2)}_{r_2})$ with probability p and remains R_r with probability $1-p$.
(iii)The trial element is evaluated with respect to the chosen optimization criteria and if it is better than the original element, e_i, then it replaces e_i.

This iterative process continues until a given number of generations is reached. The parameters $p \in (0, 1]$ and $F \in (0, 2)$ influences the convergence properties of the DE algorithm but in this study we did not give a particular attention to these. They were fixed on $p = 0.5$ and $F = 0.75$ (the average of the values used in [1]). Based on the results reported in [1] we also hybridized the DE approach with k-Means: after a given number of DE generations (e.g. 50), k-Means is applied to all elements of the population.

Algorithm	Error ± stdev	Entropy ± stdev	No.clusters ± stdev	Success ratio
k-Means	0.25797± 0.00753	0.52077± 0.02606	3	10/10
eDE + $\mu_1\mu_2$	0.24129± 0.08577	0.32948± 0.15888	6.8± 1.469	0/10
eDE +μ_1/μ_3	0.29683± 0.01234	0.37365± 0.05982	10	0/10
eDE+μ_1/μ_4	0.29958± 0.072117	0.48675± 0.13626	5.8± 2.749	1/10
eDE+μ_2/μ_3	**0.18442± 0.04454**	**0.32838± 0.05580**	**4± 1.549**	**6/10**
eDE+μ_2/μ_4	0.37988± 0.10048	0.69751± 0.15007	2.6± 0.66332	4/10
eDE+1/($\mu_3\mu_4$)	0.37569± 0.06833	0.74449± 0.13602	5.2± 2.0396	3/10

TABLE 2. Clustering results for a small set of documents (210 documents belonging to 3 classes and containing 14284 terms)

5. Results and further work

The experimental analysis is based on a classical dataset consisting of 3891 documents representing abstracts corresponding to three categories: CISI, CRANFIELD and MEDLINE (ftp://ftp.cs.cornell.edu/pub/smart). The total number of terms remained after preprocessing is 283720. In a direct vector space encoding this would mean to work with vectors having 283720 components. In order to analyze different clustering criteria we randomly selected a subset of the dataset consisting of 210 documents. In order to evaluate the quality of the obtained partition we used two measures based on the knowledge of the real assignment of data to classes: error ratio (ratio of documents pairs which either belong to the same class and have been assigned to different clusters or they belong to different classes and were assigned to the same cluster) and the classical entropy measure [9]. The results obtained for this set (for a population of 15 elements and 50 generations, for a number of clusters limited to $\{2, \ldots, 10\}$ and for 10 independent

runs) are presented in Table 2. These results confirm the remarks presented in the previous section and suggest that the best behavior is obtained by combining the connectedness and separability measures. In the case of the set of all 3891 documents the results obtained by the DE-based approach using the criterion μ_2/μ_3 are 0.01366±0.00109 (error ratio) and 0.05347±0.003511 (entropy) while k-Means led to 0.03753±0.00499 (error ratio) and 0.09374±0.01347 (entropy).

An extended experimental analysis based on other documents collections, including web documents will be further conducted. Another aspect to be analyzed is that of reducing the number of terms considered into the clustering process. The present analysis was intentionally based on the entire set of terms in order to have a reference result for future variants based on reduced feature vectors.

Acknowledgement. This work was supported by the project MindSoft (RO-CNCSIS 1385/05).

References

[1] A. Abraham, S. Das, A. Konar; Document Clustering Using Differential Evolution, Proceedings of CEC 2006 - IEEE Congress on Evolutionary Computation, pp. 1784- 1791, 2006.

[2] A. Casillas, M. T. Gonzalez de Lena, R. Martynez; Document Clustering into an unknown number of clusters using a Genetic Algorithm, Proc. of 6th International Conference on Text, Speech and Dialogue, LNCS 2807, pp.43-49, 2003.

[3] X. Cui, T. E. Potok, P. Palathingal, Document Clustering using Particle Swarm Optimization, Proc. of the 2005 IEEE Swarm Intelligence Symposium, June, 2005, Pasadena, California, USA, pp. 185-191, 2005.

[4] J. Handl, J. Knowles; Multiobjective clustering with automatic determination of the number of clusters. Technical Report TR-COMPSYSBIO-2004-02. UMIST, Manchester, 2004.

[5] T. Krink, S. Paterlini; Differential Evolution and Particle Swarm Optimization in Partitional Clustering, Computational Statistics and Data Analysis, Volume 50, Issue 5, pp. 1220-1247, 2006.

[6] M.F. Porter; An Algorithm for Suffix Stripping, Program, 14(3), pp. 130-137, 1980.

[7] M.P. Sinka, D.W.Corne; A Large Benchmark Dataset for Web Document Clustering, in Abraham, A., Ruiz-del-Solar, J., Koeppen, M. (eds.), Soft Computing Systems: Design, Management and Applications, Volume 87 of Frontiers in Artificial Intelligence and Applications, pp. 881-890, 2002.

[8] M.P. Sinka, D.W.Corne; Evolving Document Features for Web Document Clustering: A Feasability Study, Proceedings of CEC 2004 IEEE Congress of Evolutionary Computation, Portland, USA, 2004.

[9] Y. Zhao, G. Karypis; Criterion Functions for Document Clustering. Experiments and Analysis,Technical Report of Army HPC Research Center, Minneapolis, #01-40, 2002, (available online: http://citeseer.ist.psu.edu/zhao02criterion.html)

[10] Y. Zhao, G. Karypis; Hierarchical clustering algorithms for document datasets, Data Mining and Knowledge Discovery, Volume 10, Number 2, pp. 141-168, 2005.

[1,2,3,4,5] Department of Computer Science, West University of Timişoara, bv. V. Pârvan, no. 4, 300223, Timişoara

E-mail address: [1]dzaharie@info.uvt.ro, [2]zflavia@info.uvt.ro, [3]vnegru@info.uvt.ro, [4]danielpop@info.uvt.ro,[5]hpopa@info.uvt.ro