# A DISCRETE PARTICLE SWARM OPTIMIZER FOR CLUSTERING SHORT-TEXT CORPORA

Leticia C. Cagnina and Marcelo L. Errecalde and Diego A. Ingaramo
*LIDIC (Research Group)-Universidad Nacional de San Luis*
*Ej. de los Andes 950. (D5700HHW) San Luis, Argentina.*
lcagnina,merreca,daingara@unsl.edu.ar

Paolo Rosso*
*Natural Language Engineering Lab.,DSIC, Universidad Politécnica de Valencia*
*Camino de Vera s/n 46022, Valencia, España*
prosso@dsic.upv.es

**Abstract** Work on "short-text clustering" is relevant, particularly if we consider the current/future mode for people to use 'small-language', e.g. blogs, text-messaging, snippets, etc. Potential applications in different areas of natural language processing may include re-ranking of snippets in information retrieval, and automatic clustering of scientific texts available on the Web. Despite its relevance, this kind of problems has not received too much attention by the computational linguistic community due to the high challenge that this problem implies. In this work, we propose the CLUDIPSO algorithm, a novel approach for clustering short-text collections based on a discrete Particle Swarm Optimizer. Our approach explicitly considers clustering as an optimization problem where a given arbitrary objective function must be optimized. We used two unsupervised measures of cluster validity with this purpose: the *Expected Density Measure* and the *Global Silhouette* coefficient. These measures have shown interesting results in recent works on short-text clustering. The results indicate that our approach is a highly competitive alternative to solve this kind of problems.

**Keywords:** Particle Swarm Optimization, Short-text Clustering, Clustering as Optimization

*The four authors acknowledge partial support by the MCyT TIN2006-15265-C06-04 project, the ANPCyT and the Universidad Nacional de San Luis.

## 1. Introduction

In clustering tasks the main goal is that the objects within a group be similar (or related) to one another and different from (or unrelated to) the objects in other groups [22]. When clustering tasks involve documents, different aspects can negatively affect the similarity estimation between documents and, in consequence, document clustering is usually harder than other problems addressed in cluster analysis research.

In those cases where clustering techniques are applied to collections containing *very short* documents, additional difficulties are introduced due to the low frequencies of the document terms. Research work on "short-text clustering" (that is, clustering of short-length documents) is relevant, particularly if we consider the current/future mode for people to use 'small-language', e.g. blogs, text-messaging, snippets, etc. Potential applications in different areas of natural language processing may include re-ranking of snippets in information retrieval, and automatic clustering of scientific texts available on the Web.

Clustering of short-text collections is one of the most difficult tasks in natural language processing and, in this work, we propose a new discrete Particle Swarm Optimizer algorithm for this kind of problems. Our approach explicitly considers clustering as an optimization problem where a given arbitrary objective function must be optimized. We used two unsupervised measures of cluster validity with this purpose, which have shown interesting results in recent works on short-text clustering.

The remainder of the paper is organized as follows. Section 2 presents some considerations about the particularities that arise when considering clustering as an optimization problem; here, we also describe the cluster validity measures that were used as objective function to be optimized. Section 3 describes in detail our proposed approach. In Section 4 some general features of the corpora used in the experiments are presented. The experimental setup and the analysis of the results obtained from our empirical study is provided in Section 5. Finally, some general conclusions are drawn and possible future work is discussed.

## 2. Clustering as Optimization

Document clustering consists in the assignment of documents to unknown categories. This task is more difficult than supervised text categorization because the information about categories and correctly categorized documents is not provided in advance. An important consequence of this lack of information is that in realistic document clustering problems, results cannot usually be evaluated with typical *external* measures like $F$-Measure or the Entropy, because the correct catego-

rizations specified by a human editor are not available. Therefore, the quality of the resulting groups is evaluated with respect to *structural* properties expressed in different *Internal Clustering Validity Measures* (ICVMs). Classical ICVMs used as cluster validity measures include the *Dunn* and *Davies-Bouldin* indexes, the *Global Silhouette* (GS) coefficient and new graph-based measures like the *Expected Density Measure* (EDM) (denoted $\bar{\rho}$) and the $\lambda$-Measure [20] (see [13] and [20] for more detailed descriptions of these ICVMs).

These unsupervised measures of cluster validity -or any arbitrary criterion function that gives a reasonable estimation of the quality of the obtained groups- can be used as an objective function whose optimization drives the entire clustering process. In this approach, adopted by diverse algorithms (e.g. K-means [14], Cobweb [10], Autoclass [3] and CLUTO [25]) the criterion function is explicit and can be easily stated. As observed in [25], this class of algorithms can be thought of as consisting of two key components: 1) the criterion function that the clustering solution optimizes, and 2) the actual algorithm that achieves this optimization.

For the first issue we selected the GS coefficient and the EDM $\bar{\rho}$ [20, 22], two ICVMs that have shown an adequate *correlation* degree with the categorization criteria of a human editor in recent works on clustering of short-text corpora [8, 13].

The GS measure is obtained computing the average cluster silhouette of all found clusters. The cluster silhouette of a cluster $C$ is the average silhouette coefficient of all objects belonging to $C$. The silhouette coefficient for the object $i$ is obtained as follows: $s(i) = \frac{b(i)-a(i)}{max(a(i),b(i))}$ with $-1 \leq s(i) \leq 1$. The $a(i)$ value denotes the average dissimilarity of the object $i$ to the remaining objects in its own cluster, and $b(i)$ is the average dissimilarity of object $i$ to all objects in the nearest cluster.

The EDM $\bar{\rho}$ of a clustering $\mathcal{C}$ is: $\overline{\rho}(\mathcal{C}) = \sum_{i=1}^{k} \frac{|V_i|}{|V|} \cdot \frac{w(G_i)}{|V_i|^{\theta}}$. $\mathcal{C} = \{C_1, .., C_k\}$ is the clustering of a weighted graph $G = \langle V, E, w \rangle$ and $G_i = \langle V_i, E_i, w_i \rangle$ is the induced subgraph of $G$ with respect to cluster $C_i$. The density $\theta$ of the graph from the equation $|E| = |V|^{\theta}$ where $w(G) = |V| + \sum_{e \in E} w(e)$, is computed as: $w(G) = |V|^{\theta} \Leftrightarrow \theta = \frac{\ln(w(G))}{\ln(|V|)}$.

An important issue to be considered is that those ICVMs can be used for driving or for evaluating the clustering algorithms but the real effectiveness of these algorithms only can be evaluated with external measures that incorporate the categorization criteria of the users. A very popular external measure used at this end is the $F$-measure.

In the context of clustering, $F$-Measure is an external validity measure that combines both, *precision* and *recall*. It may be formally defined as

follows. Let $D$ represents the set of documents, $\mathcal{C} = \{C_1, ..., C_k\}$ be a clustering of $D$ and $\mathcal{C}^* = \{C_1^*, \ldots, C_l^*\}$ designates the human reference classification of $D$. The *recall* of a cluster $j$ with respect to a class $i$, $rec(i,j)$ is defined as $|C_j \cap C_i^*|/|C_i^*|$. The *precision* of a cluster $j$ with respect to a class $i$, $prec(i,j)$ is defined as $|C_j \cap C_i^*|/|C_j|$. Thus, the $F$-measure of the cluster $j$ with respect to a class $i$ is $F_{i,j} = \frac{2 \cdot prec(i,j) \cdot rec(i,j)}{prec(i,j)+rec(i,j)}$ and the overall $F$-measure is defined as: $F = \sum_{i=1}^{l} \frac{|C_i^*|}{|D|} \cdot \max_{j=1,..,k}\{F_{i,j}\}$ A clustering result with an $F$-measure value equals to 1 corresponds to a "ideal" clustering, i.e., a grouping that exactly matches the clustering specified by a human expert.

With respect to the algorithm used for optimizing the EDM $\bar{\rho}$ and GS measures, we will describe our approach in the next section.

## 3. Our proposed approach: CLUDIPSO

Different *Particle Swarm Optimization* (PSO) approaches have been previously proposed in the literature to solve the clustering problem in general. However, few adaptations have been presented for document clustering. A PSO-based clustering algorithm that outperforms the $K$-means algorithm in image classification tasks is proposed in [16]. Van der Merwe and Engelbrecht presented an hybridization of the PSO and K-means algorithms for clustering general datasets. Basically, the result obtained by a $K$-means algorithm is used as a single particle in the initial swarm of the PSO algorithm [23]. In [24], Xiao presents an hybrid adaptation, based on the synergism of a PSO algorithm and a Self Organizing Map for clustering gene expression data. Cui proposes in [5] an hybrid method based on the combination of a PSO and a $K$-means algorithm in document clustering tasks. Firstly, a global search process is carried out by the PSO algorithm. Then, the best result obtained by the PSO algorithm is used by the K-means algorithm for determining the initial centroids.

Discrete PSO implementations were suggested in the research community for different combinatorial optimization problems [11, 4]. However, as far as we know, no approaches have been used for clustering short-text corpora.

Our proposal for this problem, named CLUDIPSO (CLUstering with a DIscrete PSO), is based on a PSO algorithm that operates on a population of particles. In CLUDIPSO, each valid clustering is represented as a particle. The particles are $n$-dimensional integer vectors, where $n =$ number of documents in the collection. The best position found so far for the swarm (*gbest*) and the best position reached by each particle (*pbest*) are recorded. The particles evolve at each iteration using two updat-

ing formulas, one for velocity (Equation (2)) and another for position. Since the task was modeled with a discrete approach, a new formula was developed for updating the positions (shown in Equation (1)). This modification was introduced to accelerate the convergence velocity of the algorithm (principal incoming of discrete PSO models).

$$par_{id} = pb_{id} \tag{1}$$

$$v_{id} = w(v_{id} + \gamma_1(pb_{id} - par_{id}) + \gamma_2(pg_d - par_{id})) \tag{2}$$

where $par_{id}$ is the value of the particle $i$ at the dimension $d$, $v_{id}$ is the velocity of particle $i$ at the dimension $d$, $w$ is the inertia factor [6] whose goal is to balance global exploration and local exploitation, $\gamma_1$ is the personal learning factor, and $\gamma_2$ the social learning factor, both multiplied by 2 different random numbers within the range [0..1]. $pb_{id}$ is the best position reached by the particle $i$ and $pg_d$ is the best position reached by any particle in the swarm.

It is important to note that in our approach the process of updating particles is not as direct as in the continuous case. In CLUDIPSO, the updating process is not carried out on all dimensions at each iteration. In order to determine which dimensions of a particle will be updated we do the following steps: 1) all dimensions of the velocity vector are normalized in the [0..1] range, according to the process proposed by Hu et al. [12] for a discrete PSO version; 2) a random number $r \in [0..1]$ is calculated; 3) all the dimensions (in the velocity vector) higher than $r$ are selected in the position vector, and updated using the Equation (1).

To help avoiding convergence to a local optimum, we used a dynamic mutation operator [2] which is applied to each individual with a *pm*-probability. This value is calculated considering the total number of iterations in the algorithm (*cycles*) and the current cycle number as the Equation (3) indicates:

$$pm = max\_pm - \frac{max\_pm - min\_pm}{max\_cycle} * current\_cycle \tag{3}$$

where $max\_pm$ and $min\_pm$ are the maximum and minimum values that $pm$ can take, $max\_cycle$ is the total number of cycles that the algorithm will iterate, and $current\_cycle$ is the current cycle in the iterative process. The mutation operation is applied if the particle is the same that its own *pbest*, as was suggest by [12]. The mutation operator swaps two random dimensions of the particle.

## 4. Data Sets

The complexity of clustering problems with short-text corpora demands a meticulous analysis of the features of each collection used in

the experiments. For this reason, we will focus on specific characteristics of the collections such as document lengths and its closeness with respect to the topics considered in these documents. We attempt with this decision to avoid introducing other factors that can make incomparable the results.

We select for the experimental work the CICling-2002 collection, perhaps the only short-text collection that has been considered in a significative number of research works on short-text clustering [15, 1, 17, 13, 8]. CICling-2002 corpus is considered a high complexity collection since its documents are narrow domain scientific abstracts (short-length documents with an high vocabulary overlapping). Our choice of this collection is not casual. In the majority of works that have used CICling-2002, this corpus has shown a higher difficulty degree than the other collections considered. Therefore, if a good performance on this collection is achieved, we can be confident that good results will be also obtained with other easier corpus.

In order to verify this last assertion we also used the Micro4News corpus, a collection recently proposed in [8]. Micro4News is a collection significatively easier than CICling-2002 with respect to the length of documents and vocabulary overlapping. However, other features such as the number of groups and number of documents per group were maintained the same for both collections in order to obtain comparable results.

Space limitations prohibit a more detailed explanation of these corpora, but the interested reader can obtain more information in [7].

## 5. Parameter Settings and Analysis of Results

The documents of CICling-2002 and Micro4News used in the experiments were represented using the popular *Vector Space Model* and the "SMART codifications" [19] associated. In this case, we used the cosine similarity and the codification *ntc* that refers to the scheme where the weight for the $i$-th component of the vector for the document $d$ is computed as $tf_{d,i} \times \log(\frac{N}{df_i})$ and then cosine normalization is applied. Here, $N$ denotes the number of documents in the collection, $tf_{d,i}$ is the term frequency of the $i$-th term in the document $d$ and $df_i$ refers to the document frequency of $i$-th term over the collection.

We performed 50 independent runs per problem, with 10,000 iterations (*cycles*) per run. CLUDIPSO used the following parameters: swarm size = 50 particles, dimensions at each particle = number of documents ($N$), $pm\_min = 0.4$, $pm\_max = 0.9$, inertia factor $w = 0.9$, personal and social learning factors for $\gamma_1$ and $\gamma_2$ were set to 1.0. The parameter settings were empirically derived after numerous experiments.

Our results were compared with the results obtained with other three clustering algorithms: $K$-means, MajorClust [21] and DBSCAN [9]. $K$-means is one of the most popular clustering algorithms and, MajorClust and DBSCAN are representative of the density-based approach to the clustering problem. Basically, these two last algorithms attempt to separate the set of objects (documents) into subsets of similar densities. Our motivation for choosing these two density-based algorithms was to compare the performance of our algorithm with other approaches that also attempt to maximize the density of the resulting groups. Furthermore, MajorClust has shown in recent works to be one of the most successful algorithms for document clustering in general and short-text clustering problems in particular. A significative difference between the algorithms considered is whether the algorithm requires information about the number correct of groups ($k$) or not. This information has to be provided to $K$-means and CLUDIPSO but MajorClust and DBSCAN determine the cluster number $k$ automatically.

### 5.1 CICLing2002

We focus our analysis of the results obtained by the different algorithms considering the EDM $\bar{\rho}$ (that we will refer as $\bar{\rho}$ from now on) and $F$-measure values (Table 1) and the GS and $F$-measure values (Table 2). The $F$-measure values are listed in order to show the correlation between these and the metric values. In Table 1 we can observe that CLUDIPSO and MajorClust obtain the highest values of $\bar{\rho}_{avg}$ and $\bar{\rho}_{min}$ for this collection. However, CLUDIPSO is outperformed by both density-based algorithms (DBSCAN and MajorClust) if we consider the results of $\bar{\rho}_{max}$. In order to understand this last result, it is important to consider that both density-based algorithms can generate clusterings with different number of groups. Furthermore, in previous works we have observed that higher values of $\bar{\rho}$ can usually be obtained when the result has a smaller number of groups. CLUDIPSO and $k$-means only can generate clusterings with a fixed number of groups and, therefore, it is impossible for these algorithms to reach these $\bar{\rho}$ values. In that sense, the GS measure is not affected by the number of clusters obtained and thus, it can be more informative to consider the GS values shown in Table 2. In this table, we can observe that CLUDIPSO clearly outperforms the GS values of all the remaining algorithms. On other hand, if we now consider the $F$-measure values obtained by CLUDIPSO when it used $\bar{\rho}$ (Table 1) and GS (Table 2) as objective functions, in both cases this algorithm obtained excellent results with respect to the $F$-measure, outperforming the remaining algorithms considered. We can also appre-

*Table 1.* CICLing2002: $\bar{\rho}$ and $F$-measure values for the different algorithms.

| **Algorithm** | $\bar{\rho}_{avg}$ | $\bar{\rho}_{min}$ | $\bar{\rho}_{max}$ | $F_{avg}$ | $F_{min}$ | $F_{max}$ |
|---|---|---|---|---|---|---|
| K-Means | 0.87 | 0.84 | 0.91 | 0.46 | 0.35 | 0.57 |
| MajorClust | 0.92 | 0.91 | 0.94 | 0.43 | 0.37 | 0.58 |
| DBSCAN | 0.91 | 0.88 | 0.95 | 0.47 | 0.42 | 0.56 |
| CLUDIPSO | 0.92 | 0.91 | 0.93 | 0.63 | 0.42 | 0.74 |

*Table 2.* CICLing2002: GS and $F$-measure values for the different algorithms.

| **Algorithm** | $GS_{avg}$ | $GS_{min}$ | $GS_{max}$ | $F_{avg}$ | $F_{min}$ | $F_{max}$ |
|---|---|---|---|---|---|---|
| K-Means | 0.07 | -0.06 | 0.22 | 0.46 | 0.35 | 0.57 |
| MajorClust | 0.14 | -0,24 | 0.36 | 0.43 | 0.37 | 0.58 |
| DBSCAN | 0.08 | -0.11 | 0.21 | 0.47 | 0.42 | 0.56 |
| **CLUDIPSO** | **0.39** | **0.36** | **0.41** | **0.6** | **0.5** | **0.72** |

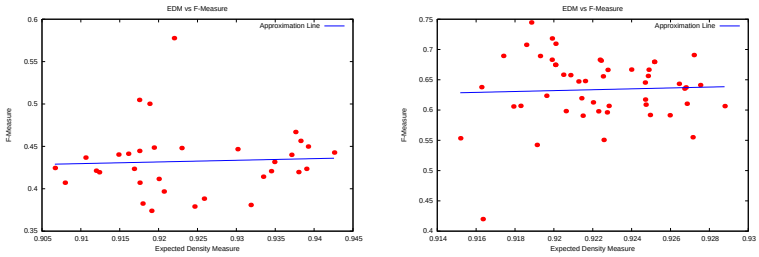


*Figure 1.* CICLing2002: $\bar{\rho}$ vs $F$-measure for MajorClust (left) and CLUDIPSO (right).

ciate this good performance of CLUDIPSO with respect to MajorClust in the Figure 1 where the $F$-measure values obtained by CLUDIPSO and MajorClust are compared. Here, we can observe that the majority of results produced by CLUDIPSO have $F$-measure values greater than 0.55. These results differ significatively of those obtained by MajorClust, which obtains a majority of $F$-measure values lower than 0.5.

## 5.2 Micro4News

An obvious question that arises from the previous experiments is if the good performance of CLUDIPSO with respect to the other algorithms considered can also be expected with other more simple collections. Therefore, we also analyze the results obtained with Micro4News, a collection with documents significatively larger than CiCling-2002 that

refer to well differentiated topics. In this case, the results shown in the Tables 3 and 4 are similar to those obtained in the previous collection respect to the $\bar{\rho}$ and the GS measures. With respect to the $F$-measure, CLUDIPSO once more achieves very high values, and it reaches (when GS is used as objective function) the highest possible $F$-measure value ($F_{max} = 1$) that corresponds to the optimum clustering (respect to the criteria of a human expert). Based on these results we can conclude that CLUDIPSO, used as an optimization algorithm of different ICVMs like $\bar{\rho}$ or GS, gives very good $F$-measure values in collections with diverse complexity levels. This suggests that the mechanisms used in this algorithm for clustering of documents usually agree with the grouping criteria of a human expert and it deserves additional research work.

*Table 3.* Micro4News: $\bar{\rho}$ and $F$-measure values for the different algorithms.

| **Algorithm** | $\bar{\rho}_{avg}$ | $\bar{\rho}_{min}$ | $\bar{\rho}_{max}$ | $F_{avg}$ | $F_{min}$ | $F_{max}$ |
|---|---|---|---|---|---|---|
| K-Means | 0.99 | 0.89 | 1.07 | 0.69 | 0.46 | 0.96 |
| MajorClust | 1.08 | 1.05 | 1.1 | 0.9 | 0.76 | 0.96 |
| DBSCAN | 1.05 | 1.01 | 1.1 | 0.82 | 0.71 | 0.88 |
| CLUDIPSO | 1.07 | 1.06 | 1.07 | 0.93 | 0.87 | 0,96 |

*Table 4.* Micro4News: GS and $F$-measure values for the different algorithms.

| **Algorithm** | $GS_{avg}$ | $GS_{min}$ | $GS_{max}$ | $F_{avg}$ | $F_{min}$ | $F_{max}$ |
|---|---|---|---|---|---|---|
| K-Means | 0.39 | 0.05 | 0.74 | 0.69 | 0.46 | 0.96 |
| MajorClust | 0.69 | 0.64 | 0.74 | 0.9 | 0.76 | 0.96 |
| DBSCAN | 0.54 | 0.36 | 0.67 | 0.82 | 0.71 | 0.88 |
| **CLUDIPSO** | **0.72** | **0.69** | **0.74** | **0.93** | **0.85** | **1** |

## 6. Conclusions and Future Works

In this work we present two new ideas for clustering short-text corpora: 1) a novel discrete PSO-based algorithm adapted for this kind of problems (CLUDIPSO) and 2) the use of two interesting ICVMs ($\bar{\rho}$ and GS) as an explicit objective function to be optimized. The results obtained by CLUDIPSO indicate that our approach is a highly competitive alternative to solve problems of clustering short-text corpora.

At the present time, we are testing our approach with other short-text collections and we are also defining a new continuous PSO version that uses the $\bar{\rho}$ and GS ICVMs.

# References


[1] M. Alexandrov, A. Gelbukh and P. Rosso. An Approach to Clustering Abstracts. In *Proc. of the 10th Int. NLDB-05 Conference*, LNCS, Springer-Verlag, 3513:8–13. 2005.

[2] L. Cagnina, S. Esquivel and R. Gallard. Particle Swarm Optimization for Sequencing Problems: a Case Study. In *Congress on Evolutionary Computation*, pages 536–541, USA, 2004.

[3] P. Cheeseman and J. Stutz. Bayessian Classification (AutoClass): Theory and Results. In *U. Fayyad, G. Piatetsky-Shapiro, P. Smith and R. Uthurusamy (Eds.), Advances in Knowledge Discovery and Data Mining*, pages 153–180. 1996.

[4] M. Clerc. Discrete Particle Swarm Optimization Illustrated by the Traveling Salesman Problem. In *New optimization techniques in engineering*, pages 219–239. 2004.

[5] X. Cui, T. Potok and P. Palathingal. Document Clustering using Particle Swarm Optimization. In *Proc. of IEEE Swarm Intelligence Symposium (SIS-2005)*. 2005.

[6] R. Eberhart and Y. Shi. A Modified Particle Swarm Optimizer. In *International Conference on Evolutionary Computation, IEEE Service Center*, Anchorage, AK, Piscataway, NJ, 1998.

[7] M. Errecalde and D. Ingaramo. *Short-text Corpora for Clustering Evaluation*, LIDIC, 2008. Url: http://www.dirinfo.unsl.edu.ar/∼ia/resources/shortexts.pdf

[8] M. Errecalde, D. Ingaramo and P. Rosso. Proximity Estimation and Hardness of Short-text Corpora. In *5th International Workshop on Text-based Information Retrieval (TIR-2008)*.

[9] M. Ester, H. Kriegel, J. Sander and X. Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proc. of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pages 226–231. 1996.

[10] D. Fisher. Knowledge Acquisition via Incremental Conceptual Clustering. *Machine Learning*, 2:139–172, 1987.

[11] A. Guner and M. Sevkli. A Discrete Particle Swarm Optimization Algorithm for Uncapacitated Facility Location Problem. In *JAEA*. 2008.

[12] X. Hu, R. Eberhart and Y. Shi. Swarm Intelligence for Permutation Optimization: a Case Study on n-queens Problem. In *Proc. of the IEEE Swarm Intelligence Symposium*, pages 243–246, 2003.

[13] D. Ingaramo, D. Pinto, P. Rosso and M. Errecalde. Evaluation of Internal Validity Measures in Short-Text Corpora. In *Proc. of the CICLing 2008 Conference*, pages 555–567. LNCS, vol. 4919, Springer-Verlag. 2008.

[14] J. MacQueen. Some Methods for Classification and Analysis of Multivariate Observations. In *Proc. 5th Symp. Math. Statis, Prob.*, pages 281–297. 1967.

[15] P. Makagonov, M. Alexandrov and A. Gelbukh. Clustering Abstracts Instead of Full Texts. In *Proc. of the TSD-2004 Conference*, 3206:129–135. 2004.

[16] M. Omran, A. Engelbrecht and A. Salman. Particle Swarm Optimization Method for Image Clustering. In *IJPRAI*, 19:297–321. 2005.

[17] D. Pinto, J. Benedí and P. Rosso. Clustering Narrow-domain Short Texts by Using the Kullback-Leibler Distance. In *Proc. of the CICLing 2007 Conference*, LNCS, Springer-Verlag, 4394:611–622. 2007.

[18] D. Pinto and P. Rosso. On the Relative Hardness of Clustering Corpora. In *Proc. of the Text, Speech and Dialogue 2007 Conference - TSD07, LNAI*. 4629:155–161. Springer-Verlag. 2007.

[19] G. Salton. The Smart Retrieval System: Experiments in Automatic Document Processing, Prentice Hall. 1971.

[20] B. Stein, S. Meyer zu Eissen and F. Wisbrock. On Cluster Validity and the Information Need of Users. In *Proc. of the 3rd IASTED*, pages 216–221. 2003.

[21] B. Stein and O. Niggemann. On the Nature of Structure and its Identification. In *Proc. of the 25th International Workshop on Graph Theoretic Concepts in Computer Science - WG99*, LNCS, Springer-Verlag, 1665:122–134. 1999.

[22] P. Tan, M Steinbach and V. Kumar. *Introduction to Data Mining*. 2006.

[23] D. Van and A. Engelbrecht. Data Clustering using Particle Swarm Optimization. In *Congress on Evolutionary Computation, Special Session on Design Optimisation with Evolutionary Computation(CEC'03)*, pages 215–220. 2003.

[24] X. Xiao, E. Dow, R. Eberhart, Z. Ben Miled and R. Oppelt. Gene Clustering Using Self-Organizing Maps and Particle Swarm Optimization. In *IPDPS '03: Proc. of the 17th International Symposium on Parallel and Distributed Processing*. 2003.

[25] Y. Zhao and G. Karypis. Empirical and Theoretical Comparison of Selected Criterion Functions for Document Clustering. *Machine Learning*, 55:311–331, 2004.