Cluster optimisation in information retrieval using self-exploration-based PSO

Article · February 2016

DOI: 10.1504/IJIEI.2016.074513

3 authors, including:

Prakasha Shivanna
RNS Institute of Technology
11 PUBLICATIONS **23** CITATIONS

SEE PROFILE

Gt Raju
RNS Institute of Technology
61 PUBLICATIONS **146** CITATIONS

SEE PROFILE

Cluster optimisation in information retrieval using self-exploration-based PSO

S. Prakasha and G.T. Raju

Department of ISE,
R N S Institute of Technology,
Rajarajeshwarinagar Post, Channasandra,
Bangalore – 560098, India
Email: sprakashjpg@yahoo.co.in
Email: gtraju1990@yahoo.com

Manoj Kumar Singh*

Manuro Tech Research Pvt. Ltd.,
#20, 2nd Cross, Jyothi Nagar,
Chikkabettahalli,Vidyaranyapura Post,
Bangalore – 560097, India
Email: mksingh@manuroresearch.com
*Corresponding author

Abstract: Self-exploration capability is an important and necessary factor in all social communities where individual assumes to have their own intelligence. Macro social influencing factors are responsible for decision nature taken by an individual, whereas self-exploration process can be considered as a refinement of that decision by use of the cognitive capability to explore a number of surrounding possibilities. The mathematical model corresponding to the individual self-exploration process can be expressed with the help of the chaotic search method. In this paper, chaotic search-based self-exploration has integrated with social influenced-based particle swarm optimisation (PSO) to represent better computational model so that the complex optimisation problem could solve more efficiently. Two different levels of self-exploration called intrinsic cascade self-exploration and extrinsic cascade self-exploration have applied in association with PSO. This paper has applied the proposed concept to cluster documents data in the area of information retrieval and to achieve the global solutions for high dimensional numerical optimisation problems.

Keywords: cluster; information retrieval; particle swarm optimisation; PSO; self-exploration; chaotic search.

Reference to this paper should be made as follows: Prakasha, S., Raju, G.T. and Singh, M.K. (2016) 'Cluster optimisation in information retrieval using self-exploration-based PSO', *Int. J. Intelligent Engineering Informatics*, Vol. 4, No. 1, pp.91–115.

Biographical notes: S. Prakasha is currently working as Assistant Professor in the Department of ISE, R.N.S.R. Institute of Technology, Bangalore, India. His research interest includes data mining, and soft computing.

G.T. Raju is a Professor and the Head in the Department of CSE, R.N.S. Institute of Technology, Bangalore. His research interest includes data mining, and soft computing.

Manoj Kumar Singh is currently holding the post of Director in Manuro Tech. Research Pvt. Ltd., Bangalore, India. He has R&D background in advanced intelligent computing and solution development in various fields of engineering and technology. His field of technological research includes nano and quantum computing, soft computing, advanced machine learning, etc.

1 Introduction

In a number of applications, multidimensional data vectors required to separate by feature-based similarity like into a number of clusters or bins. There are huge applications of clustering algorithms to solve practical problems from different area like in data mining, data analysis; image processing and mathematical programming, etc. are few of them. In machine learning, scalability issue is an important and with the help of clustering concept, it is possible to cluster the training data priory and during the training period to reduce the computational complexity and to improve the generalisation performance. At present day by day, available quantity of data is increasing in a rapid manner; hence, it gets difficult to retrieve the relevant information efficiently. Therefore, there is need of methods to support the users for organising, exploring and searching features in the collections of textual data. In the area of information retrieval, exploration and utilisation of huge data are a challenge for researchers. Efficient document clustering can be very helpful in effectively navigating, summarising, and organising the information from the big dataset. With this approach, precision and recall can be improved very much in information retrieval and clustering can be great means to find the nearest neighbours of a document. The problem of document clustering can be defined in a simple manner as defining disjoint membership to documents into different clusters in such that documents under a cluster share some more similarity with each others in comparing with documents which are members of different clusters. Broadly, two different approach called agglomerative hierarchical and partitioning methods are generally applied under the clustering. Generally, algorithms which are based on the partitional clustering deliver better performance in terms of quality and computational efficiency for the larger text data base. Optimisation of global criterion a function, which carry different aspects of clusters characteristics and function of intra and inter cluster distances, is applied in partitional clustering algorithm.

Numerous works of literature related to cluster and its application in the area of information retrieval have motivated our research work; K-means clustering is very simple and fast efficient. This is most popular one and it is developed by MacQueen (1967). The easiness of K-means clustering algorithm made this algorithm used in several fields. The K-means algorithm is effective in producing clusters for many practical applications, but the computational complexity of the original K-means algorithm is very high, especially for large datasets. The K-means clustering algorithm is a partitioning clustering method that separates the data into K groups. One drawback in the K-means algorithm is that of a prior fixation of number of clusters (Yuan et al., 2004; Al-Shboul and Myaeng, 2009; Zhang and Xia, 2009; Yedla et al., 2010). A way of clustering using a

biologically inspired genetic algorithm was developed by Kamble (2010), which clusters data in dynamic form, the database is assumed to be clustered initially, and every new element is added as without the need of changing existing clustered database. Other ways of improving for K-means clustering algorithm offers improved simulation results which offers but also offers clustering result is more accurate and effective has been presented by Zhang and Fang (2013). Niknam et al. (2008) have presented an efficient hybrid evolutionary optimisation algorithm based on combining ant colony optimisation (ACO) and simulated annealing (SA), called ACO-SA, for cluster analysis. In this algorithm, the SA algorithm as a local searcher for each colony is considered. To evaluate the performance of the hybrid algorithm, it is compared with other stochastic algorithms viz. the original ACO, SA and k-means algorithms on several well-known real life datasets. Niknam and Amiri (2010) have proposed hybrid a evolutionary algorithm to solve nonlinear partitional clustering problem. It is the combination of fuzzy adaptive particle swarm optimisation (FAPSO), ACO and k-means algorithms, called FAPSO-ACO-K, which can find a better cluster partition. Details about particle swarm optimisation (PSO) have given by Kennedy and Eberhart (1995) and Clerc and Kennedy (2002). Chaos theory was discovered by Lorenz (1993). Since then, chaos theory has been applied for many research areas, such as mathematics, physics, engineering, biology, economics, and philosophy, etc. Bsoul et al. (2013) have applied document clustering to detect crime patterns. Evolutionary approach based on genetic algorithm for text document clustering has applied by Akter and Chung (2013). ASBO-based document clustering has been presented by Prakasha et al. (2013). Problem related with similar image retrieval from large data base has been presented by Liang et al. (2014), for fast similar image retrieval, especially for large-scale datasets with millions of images. SA-based clustering search algorithm for the rank aggregation problem has presented by Lorena et al. (2014). In case-based reasoning (CBR) two factors, weights determination and attributes reduction were taken together by Han et al. (2014), to improve the clustering quality. Probabilistic latent semantic analysis (PLSA) has applied by Zhao et al. (2014), to solve the scalability problem of data, by adapting the traditional EM algorithm into MapReduce. Chen et al. (2015) have proposed a disk-based metric access method, the space-filling curve and pivot-based tree (SPB-tree), to support a wide range of data types and similarity metrics. Habibi and Popescu-Belis (2015) have considered the keywords extraction from conversations to provide the relevant documents to participants. A fuzzy linguistic topological space along with a fuzzy clustering algorithm has presented (Chiang et al., 2015) to discover the contextual meaning in the web documents. Non-negative matrix factorisation-based cluster concept has applied by Ling et al. (2015) for development of system for extracting medication names and symptom names from clinical notes in healthcare domain.

2 Clustering in information retrieval

There is a fundamental assumption generally made when the cluster definition applied in the area of information retrieval and it can be defined as: all documents which share similar characteristics as according to the nature of information needs, either conceptual or terminologies or both can be considered under the same cluster. The assumption makes less sure that if a document appeared with respect to search requests than other

documents belong to the same cluster may also be useful. There is a relationship between documents by sharing a number of relevant terms and will cause to joint the same cluster. The purpose of information retrieval algorithm is to develop such kind of clusters that should have coherent characteristics in their member documents but also have some good quality of diversity so that maximum information can be extracted with a search request.

Figure 1 Cycle of cluster optimisation in information retrieval (see online version for colours)

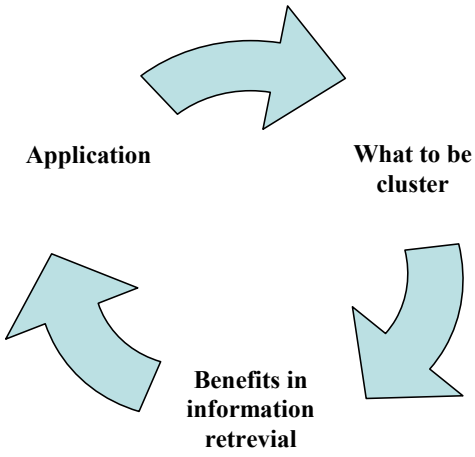


Table 1 Applications of clustering in information retrieval

Application	*What is to be cluster*	*Benefit in information retrieval*
Search result clustering	Search results	More effective information presentation to user
Scatter-gather	(Subsets of) collection	Alternative user interface: 'search without typing'
Collection clustering	Collection	Effective information presentation for exploratory browsing
Language modelling	Collection	Increased precision and/or recall
Cluster-based retrieval	Collection	Higher efficiency: faster search

3 Vector space model for document representation

Representation is a primary and fundamental requirement for a number of information-based applications like in data management, information filtering, information retrieval, indexing, classification, and clustering tasks. The vector space model (VSM) is very powerful method to represent the documentation for various applications. Generally, in VSM, a document is represented by a vector whose dimensions are represented by terms (or phrases). There is a quantitative incremental entry in a dimension place if the corresponding term occurs in the document.

To define cluster in clustering algorithms, the dataset which has to clustered is represented by a set of vectors $X = \{x_1, x_2, \ldots, x_n\}$, where the each vector x_i is a feature vector correspond to a feature in the dataset. Proper selection of feature which could

represent the object of interest in feature vector is very important. In VSM model, a term weight vector (D) is calculated from the content of a document such that each component of term weight vector quantitatively represent the relatively significant (by occurrence frequency) of the term with respect all other documents in available in dataset. Mathematically, it can formulate as estimation of vector $D = \{w_1, w_2, \ldots\ldots\ldots\ldots, w_n\}$, where w_i (i = 1, 2, ……., n) is the term weight of the term t_i in one document. The frequently applied weighting scheme combines the term frequency with inverse document frequency (TF-IDF). With this approach weight of term i in document 'j' is given below:

$$w_{ji} = tf_{ji} \times idf_{ji} = tf_{ji} \times \log\left({}^{n}\!/\!_{df_{ji}} \right) \tag{1}$$

where tf_{ji} is the number of occurrences of term i in the document j; df_{ji} indicates the term frequency in the collections of documents; and 'n' is the total number of documents in the collection. The formulation approach of defining TF-IDF weighting scheme has some technical significant of clustering assumption. It will increase the weight of those terms which have a higher occurrence in a smaller set of documents in comparison to those which have frequently occurred over the larger set of documents.

3.1 Similarity measure

To shape the quality of cluster, choice of metric function as a proximity definition, is a significant one to interpret as clusters in the multidimensional space where documents exist. Requirement to achieve a proper separation while keeping generality too, Euclidian distance-based metric also known as the l^2 norm of the difference vector is a fine choice so that neither too small nor too large clusters could develop. This metric is the direct distance between two objects in a linear space defined below:

$$d_2(p, q) = \sqrt{\sum_{k=1}^{n} \left(p_k - q_k \right)^2} \tag{2}$$

where p, q are either points or vectors.

4 Particle swarm optimisation

PSO is very much similar to evolutionary computation in certain aspects, like it is a technique in that, a population which represents the potential solutions to the problem under consideration is utilised to explore the solution space in hope to meet the better one. However, there are some remarkable differences also like rather than generational progress as in EC, there is single population achieve the progress because of social influences with time. In PSO, each individual solution of the population has an adaptive parameter called velocity (position change), which has an inspirational change because of leader and cognition factor. With this change solution moves in the search space. Moreover, there is a memory for each individual to remember the best position in the search space has visited in the past. Thus, solution movement is a vector sum of its best previously visited position and towards the best individual of topological consideration.

Two different possibilities are considered most, in one case there is a global neighbourhood and another one is a local neighbourhood. In the global variant, each solution gets additive change being proportional to the difference of the present position with its best previous position and best particle in the whole population respectively. While in local variant, each solution gets additive change being proportional to the difference of the present position with its best previous position and best particle in the restricted neighbourhood respectively. In this work, we have considered the global variant to get convergence faster.

Mathematical formulation of PSO is simple and computationally efficient. Suppose the size of the search space is D dimensional, then for the i^{th} solution, there are three vectors, current position, velocity and best previously visited position respectively each one of dimension size D as defined below:

- Position vector

$$X_i = \left[x_{i1}, x_{i2}, ..., x_{iD}\right].$$

- Velocity vector

$$V_i = \left[v_{i1}, v_{i2}, ..., v_{iD}\right].$$

- Memory vector

$$P_i = \left[p_{i1}, p_{i2}, ..., p_{iD}\right].$$

$$V_{id}^{(n+1)} = \chi\left[wV_{id}^{n} + C_1 r_1\left(P_{id}^{n} - X_{id}^{n}\right) + C_2 r_2\left(P_{gd}^{n} - X_{id}^{n}\right)\right] \tag{3}$$

$$X_{id}^{(n+1)} = X_{id}^{n} + V_{id}^{(n+1)} \tag{4}$$

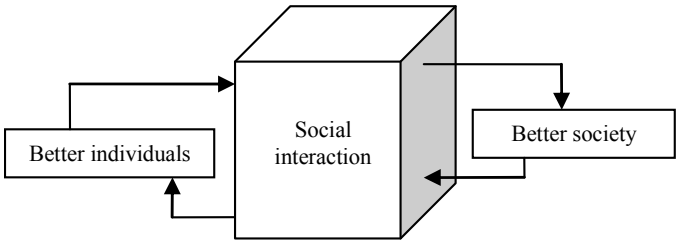
Mathematically, velocity and position change can be defined by equation (3) and equation (4) for $(n + 1)^{th}$ iteration. Index 'g' represents the best particle in the population, 'm' is the best seen by that particular solution Where 'w' is called inertia weight; C_1, C_2 are two positive constants, C_1 is called cognitive parameter. C_2 is called social parameter. χ is a constriction factor.

5 Self-exploration requirement and its advantage

If we observe the inherent development nature of any social structure, there is closed-loop-based driving rule exist between quality of members and formed a society by them through social interaction mechanism, as shown in Figure 2.

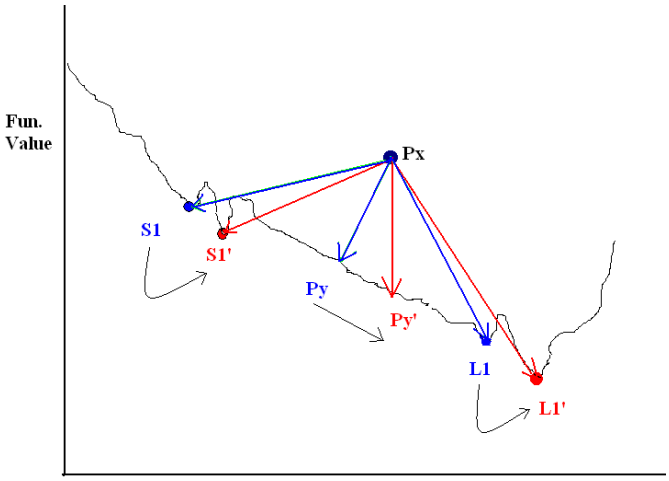
Better individuals help to form the better society and developed social culture makes individual further better, in return society get improves and this cyclic process keeps continuing. It is very important that individual should have high levels of cognition status, which makes individual not to just drive by social factor, but also have their own cognitive intelligence to improve his fitness.

Figure 2 Individuals and society interaction



In PSO model, both, cognisant part (its own previous best value, derived from memory and experience) and social part (derived from the path set by society or group) are available. But each individual does not apply their own intelligence in fitness improvement, instead all time progress is completely depending upon passive experience of their own and the leader at present. This will cause of diversity loss in population and in result there may be premature convergence to a suboptimal solution. If an individual does not apply their intelligence for self-exploration with each state of progress, it may lose the better solution, if it is available nearby. Hence, the fundamental requirement of making society better by the better individuals does not fulfil. In short, in PSO model there is a missing part of natural society and that is – 'contribution of individual intelligence', which is used to think, explore and exploit to make individual better.

Figure 3 Self-exploration advantage in association with PSO for a particular iteration over landscape (see online version for colours)



In Figure 3, the status for a member in solution population for function minimisation problem has shown over a part of the landscape for a particular iteration. Benefit of self-exploration has shown in association with driving influence factors in PSO by self best and the leader of the population. Assume that member Px has to update its position and at present it has self best position at S1 and leader position at L1. In result, the resultant driving force will shift the position of the Px to Py. If there is self-exploration available, self best position has been shifted from s1 to s1' which has higher fitness in the vicinity and similar way leader position has been shifted to L1' from L1. These shifts

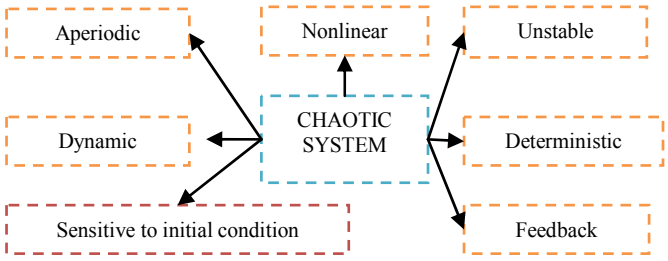
have delivered the more optimal influence and resultant influence forced to shift Px to take the new position Py', which has better fitness comparatively, hence, there is faster optimal convergence. With the same process either it is possible to prevent the solution to trap in local minima or at least it will apply best efforts to come out from there which may be cause of success with time. In summary, the effect of self-exploration makes individual and leader better, in result optimal and faster convergence.

6 Chaotic search

In the area of mathematics, engineering and sciences, modelling the problem in terms maximisation or minimisation of highly nonlinear functions is a common approach to define the problem is the important steps towards obtaining the optimal solution but at the same handling such complexity is a challenging task. Heuristic algorithms have given high hopes in this area, and in result, it became a growing field of research. This difficulty associated with these problems can be explained by the complexity of the objective function, available constraints, and availability of multiple local minima and the limitations of many optimisation methodologies like trapping in local minima by gradient decent-based optimisation method.

With available characteristics in chaotic process, chaos theory can be considered to use in the development of novel techniques for global optimisation. By use of chaotic sequences instead of quasi random number generators seems to be a powerful strategy for improving many traditional heuristic algorithms and can help in escape of local minima points. A chaotic system is having a number of important characteristics as shown in Figure 4.

Figure 4 Characteristics of chaotic system (see online version for colours)



There are two function parts under chaos-based optimisation method:

1 chaotic samples generation and its mapping in design point

2 evaluation with respect to objective function, (if betterment observed preserve it) and explore further with new chaotic sample.

It is possible to define the larger change and a small change in consecutive change with each generated new chaotic sample. Hence, there are two stages under which search takes place, first explore and estimate the closer region of optimal solution and next apply fine or local search to obtain the optimum one. In the first stage for global search candidate, a points Xc are moved in the search range [L, U] by means of the vector of chaotic sample α generated by chaotic map Z in each iteration as shown in the global search algorithm.

The output of the global search stage is taken as starting search input for local search algorithm and more detail has given below. During the chaotic local search; the parameter, step size (λ) is very important in defining the convergence characteristic. Around the present solution what will be the smallest ergodic range of adjustment will also decide by parameter λ. The step size is providing the control facility in generating the new trial solution from current best solution. A small of λ help to explore the local region while larger value has effect in global search.

Algorithm 1 Global search algorithm

```
Initialise α
    For (m1 = 1, …….., M1)
        {
        Let Xc = L + α(U–L)
            If (m1==1) then (Xn=Xc);
        Let Δf = f(Xc) – f(Xn)
            If (Δf < 0) then (Xn=Xc)
        α = Z(α)
        }
```

Algorithm 2 Local search algorithm

```
Initialise λ and β
        For (m2 = 1, …….., M2)
    {
            If r < 0.5
                {
    (Where r ∈ uniformly generated random number in range [0 1])
                Let Xc = Xn + λβ |U – Xn|
            else
                Let Xc = Xn – λβ |Xn – L|
                }
        Let Δf = f(Xc) – f(Xn)
            If (Δf < 0) then (Xn=Xc)
        β = Z(β)
}
```

Lozi map has considered in this paper to define the chaotic mapping as it defined below:

$$C(t+1) = 1 - P.\left|C(t)\right| + y(t) \tag{5}$$

$$y(t) = Q.C(t) \tag{6}$$

where 't' is the iteration number. The parameters used in this work are P = 1.7 and Q = 0.5, these values show the sensitivity with respect to initial condition. To limit the

value of chaotic output, the values of y are normalised in the range [0, 1] by transformation as it is given below:

$$Z(t) = \frac{y(k) - \lambda_{\min}}{\lambda_{\max} - \lambda_{\min}} \tag{7}$$

where generated chaotic outcomes y achieve the range value in [–0.6418, 0.6716] hence [$\lambda_{\max}$, $\lambda_{\min}$] taken as [0.6418, –0.6716].

7 Hybridisation of PSO with chaotic search

Fundamentally, PSO can consider as a very abstract model of social influence to improve the quality of society individual. But if consider from a practical point of view, society is a very complex organisation and various factors drive the state of society from one time to another time and or from one state to another. Many times these driving factors are at macro level like influence where factors can be observed, whereas many times they appeared on micro level where clear observation is not possible but always play an important role in defining the success of society. Self-exploration is one of the examples of micro level operator, which is defined as the utilisation of individual cognitive intelligence to find the other near better possibility with respect to the status it achieved by analysing a number of possibilities existed in the surrounding. Characteristics of chaos support the behaviour of self-exploration and chaotic optimisation is a means to achieve that.

We have applied three different approaches to hybridise the chaotic search process (COP) with PSO, as their operational structures have shown in Figure 5 to Figure 7. In first approach called HPSOGLC, in each iteration, both global and local chaotic search have applied sequentially for each member of the population created by PSO to generate the chaotic population as shown in Figure 5. This chaotic new population is considered as input for PSO to generate new solutions and the process will continue until terminating criteria does not satisfy. In second case called HPSOLC, the process creates a new chaotic population without global search to explore the effect of local search as shown in Figure 6. In the third case called PSOCC, as shown in Figure 7, PSO is cascaded with COP in terms of final output of PSO system is the input of the COP system which have a local search facility.

8 Experimental analysis

Proposed methods for all the three cases have applied with PSO in various test bench numerical optimisation and to evaluate the performances of clustering in document clustering, compactness of clusters have measured based on the total intra cluster distance applied to find the fitness of the cluster results. The clustering results obtained in all the three different cases are comparing together as well as with PSO and COP performances also. The MATLAB (7.10) based software environment has applied to develop the all algorithms. All experiments have repeated for 10 independent trials. Parameter value for PSO in all cases of simulation defined as: $C_1 = C_2 = 0.5$, $\chi = 0.75$ and the inertia weight

value decreases from 1.2 towards 0 with iterations. The parameters of Lozi map are taken as $P = 1.7$; $Q = 0.5$; with initial condition defined randomly in the range of [0 0.5].

7.1 Numeric optimisation

Five different benchmark test functions (F1 to F5), as shown in Table 2 with their optimal value in Table 3, have taken as a problem of minimisation to define the experiments with two different dimension size of problem equal to 10 and 30. F1 and F2 are unimodal while F3, F4 and F5 are multimodal problems. Maximum number of allowed iterations in COP are equal to 2,000 in which half of that has applied for global search and remaining iterations has applied for local search. There were 1,000 iterations for PSO and further 1000 iterations have applied for PSOCC hybridisation. In case of HPSOLC and HPSOGLC, associated PSO has given iterations equal to 500 along with 100 iterations in local search for HPSOLC while HPSOGLC contains 50 iterations in global search and 50 iterations in local search for each member of the population under solution iteration by chaotic optimisation. Size of population in PSO in all cases has taken as 50. In all test cases, ten independent trials have given to analyse the performance in terms of best and worst solution achieved along with mean and standard deviation available in observed solutions.

Figure 5 Internal hybridisation of PSO with COP contains global as well as local search (HPSOGLC) (see online version for colours)

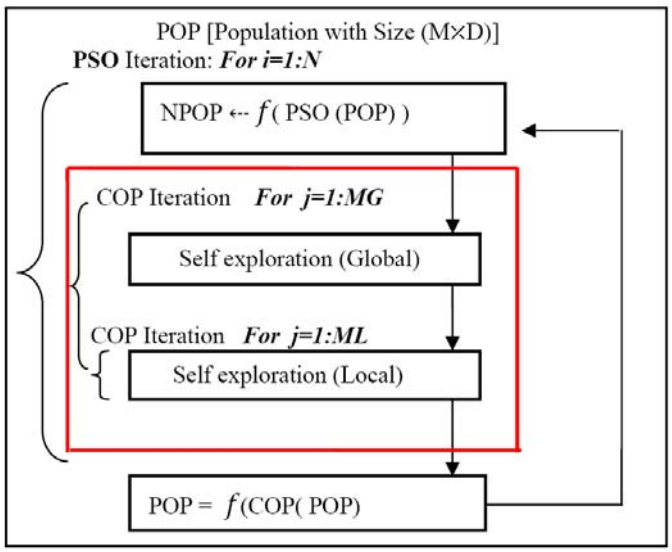


Figure 6 Internal hybridisation of PSO with COP contains local search (HPSOLC)

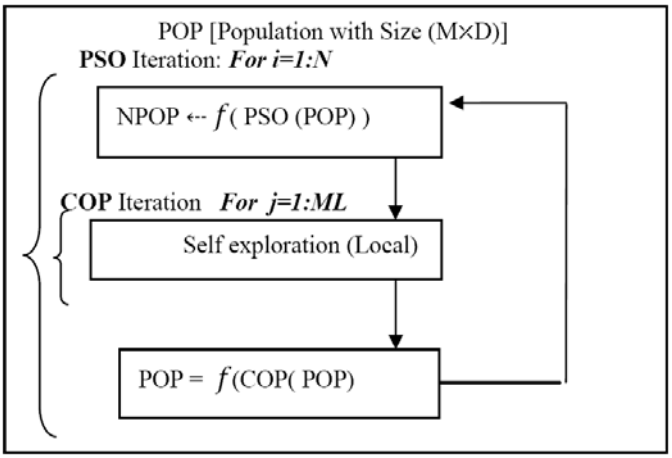
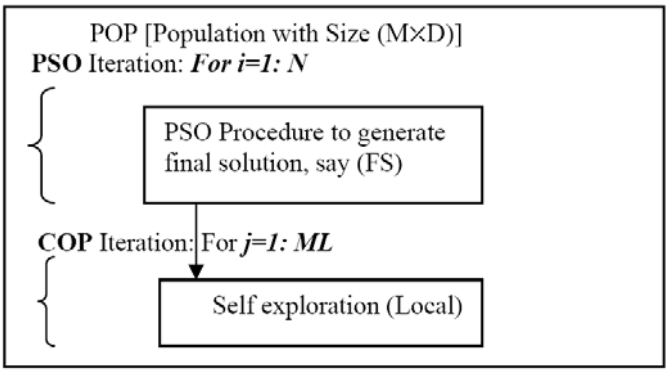


Figure 7 Cascade hybridisation of PSO with COP contains local search (PSOCC)



With respect to the spherical model problem of dimension size 10 and 30, obtained performances have shown in Table 4 and in Table 5. It can observe from Table 4 that PSO has delivered the most appreciable result in compare to others for dimension size 10 where as HPSOGLC has shown finest performance for dimension size 30. Convergence characteristics of all the methods have shown in Figure 8 and in Figure 9. It can observe very clearly that HPSOGLC has the fastest rate of convergence in comparison to others. In second test function generalised Rosenbrock performances have shown in Table 6 and in Table 7. For dimension size 10 best performances have given by HPSOGLC but overall mean performance in terms of minimum value has shown by HPSOLC and this situation became reverse for dimension size 30. Convergence characteristics for both dimension sizes have shown in Figure 10 and Figure 11 and it can be concluded that fastest convergence appeared with HPSOGLC. There is small improvement shown by PSOCC in comparing to PSO. In generalise Rastrigin function, performance for both dimensions have shown in Table 8 and in Table 9 and it can observe that PSOGLC has given the best results among all. Convergence characteristics have shown in Figure 12 and in Figure 13 and can observe there is faster and smoother convergence in performance of HPSOLC. In Ackley function, performances have shown in Table 10 and

in Table 11 and in compare to others HPSOLC and HPSOGLC performances are superior and same even this similarity can also be observed in convergence characteristics as shown in Figure 14 and in Figure 15. For generalise Griewank function performances for both dimensions have shown in Table 12 and in Table 13. For dimension equal to 10, PSO has shown slightly better result in comparing to self-exploration-based method, but as dimension increases to 30, limitation of handling higher dimension appeared where as HPSOLC and HPSOGLC have shown the robustness in performances. Convergence characteristics have shown in Figure 16 and in Figure 17 for all methods. HPSOGLC have shown the fastest rate in comparison to others. Performances of COP in all test cases are suboptimal shows the limitation of self-exploration alone in terms of solution development. Performance of cascade system is nearly same as PSO performance hence PSOCC convergence characteristics have not shown in the plots.

Table 2 Five test function for numerical optimisation

Test function
Sphere model $F_1(x) = \sum_{i=1}^{N} x_i^2$
Generalised Rosenbrock $F_2(x) = \sum_{i=1}^{N-1} 100\left(x_{i+1}^2 - x_i\right)^2 + \left(x_i - 1\right)^2$
Generalise Rastrigin $F_2(x) = \sum_{i=1}^{N-1} 100\left(x_{i+1}^2 - x_i\right)^2 + \left(x_i - 1\right)^2$
Ackley's function $F_2(x) = \sum_{i=1}^{N-1} 100\left(x_{i+1}^2 - x_i\right)^2 + \left(x_i - 1\right)^2$
Generalise Griewank function $F_2(x) = \sum_{i=1}^{N-1} 100\left(x_{i+1}^2 - x_i\right)^2 + \left(x_i - 1\right)^2$

Table 3 Search range for different function and their minimum value

Function	*Range*	*fmin*
F1	[–100 100]	0
F2	[–30 30]	0
F3	[–5.12 5.12]	0
F4	[–32 32]	0
F5	[–600 600]	0

Table 4 Performances for spherical model test function with dimension size equal to 10

Perf	*COP*	*PSO*	*PSOCC*	*HPSOLC*	*HPSOGLC*
Best	5.4e-03	4.9e-19	4.9e-19	1.5e-12	1.e-010
Worst	3.5e+03	3.1e-07	3.0e-07	3.1e-07	2.6e-05
Mean	6.7e+02	4.4e-08	4.3e-08	7.2e-08	2.6e-06
Std. dev.	1.4e+03	9.5e-08	9.5e-08	1.1e-07	8.0e-06

Figure 8 Convergence characteristics with respect to spherical model with problem dimension size 10 (see online version for colours)

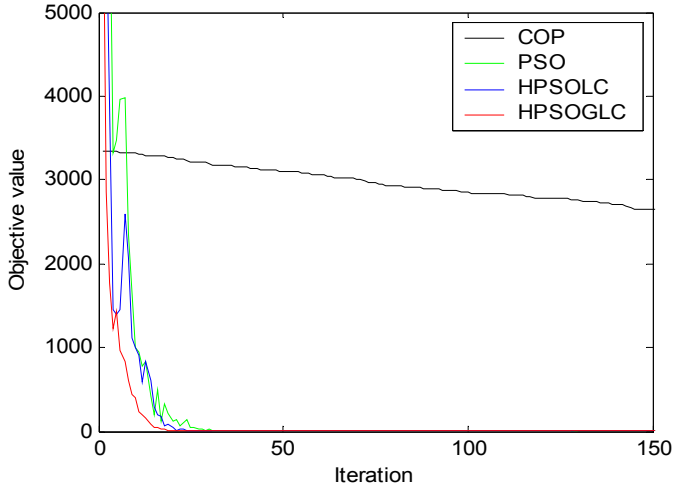


Table 5 Performances for spherical model test function with dimension size equal to 30

Perf	*COP*	*PSO*	*PSOCC*	*HPSOLC*	*HPSOGLC*
Best	8.8e-02	2.8e+02	7.4e-02	2.5e+01	5.3e-03
Worst	2.6e+04	1.4e+03	1.3e+03	9.0e+02	1.6e-02
Mean	1.1e+04	8.7e+03	3.7e+02	1.4e+02	1.0e-02
Std. dev.	1.2e+04	4.2e+02	5.1e+02	2.7e+02	3.9e-03

Figure 9 Convergence characteristics with respect to spherical model test function with problem dimension size 30 (see online version for colours)

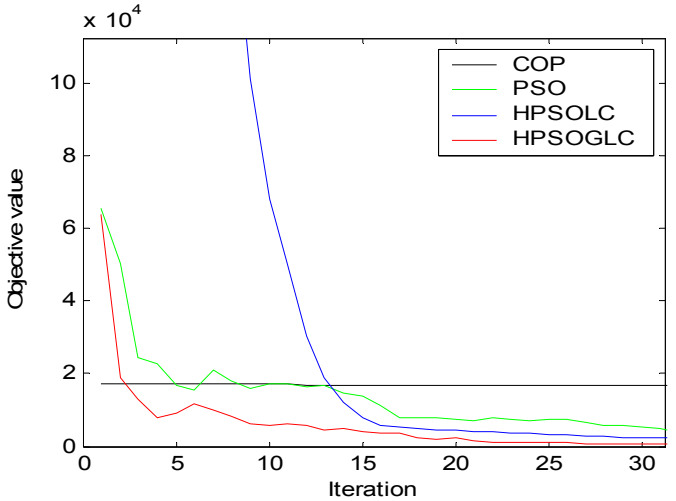


Table 6 Performances for generalised Rosenbrock test function with dimension size equal to 10

Perf	*COP*	*PSO*	*PSOCC*	*HPSOLC*	*HPSOGLC*
Best	5.2e+00	5.6e+00	4.6e+00	3.9e-001	3.2e-02
Worst	2.0e+06	5.6e+02	5.6e+02	4.6e+00	2.1e+01
Mean	6.0e+05	1.0e+02	9.9e+01	1.4e+00	3.0e+00
Std. dev.	8.2e+05	1.7e+02	1.7e+02	1.2e+00	6.5e+00

Figure 10 Convergence characteristics with respect to generalise Rosenbrock with problem dimension size 10 (see online version for colours)

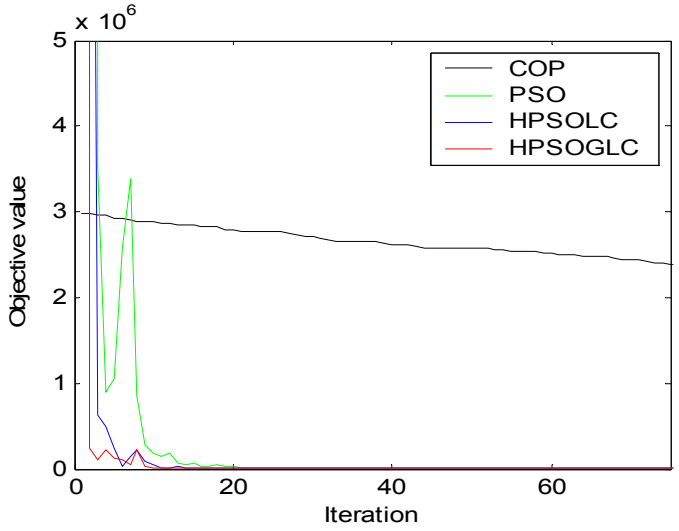


Table 7 Performances for generalise Rosenbrock test function with dimension size equal to 30

Perf	*COP*	*PSO*	*PSOCC*	*HPSOLC*	*HPSOGLC*
Best	2.8e+01	1.0e+04	2.8e+01	1.9e+01	2.3e+01
Worst	4.8e+07	2.7e+05	2.6e+05	1.5e+02	1.4e+02
Mean	1.8e+07	1.0e+05	6.4e+04	6.2e+01	5.1e+01
Std. dev.	1.8e+07	7.7e+04	9.4e+04	5.1e+01	4.2e+01

Table 8 Performances for Generalise Rastrigin test function with dimension size equal to 10

Perf	*COP*	*PSO*	*PSOCC*	*HPSOLC*	*HPSOGLC*
Best	9.0e+00	1.6e+01	1.6e+01	3.9e-06	3.0e-05
Worst	1.4e+02	3.3e+01	3.4e+01	1.7e-04	1.6e-04
Mean	4.3e+01	2.5e+01	2.5e+01	5.4e-05	8.7e-05
Std. dev.	3.8e+01	6.8e+00	6.8e+00	6.3e-05	4.3e-05

Figure 11 Convergence characteristics with respect to generalise Rosenbrock with problem dimension size 30 (see online version for colours)

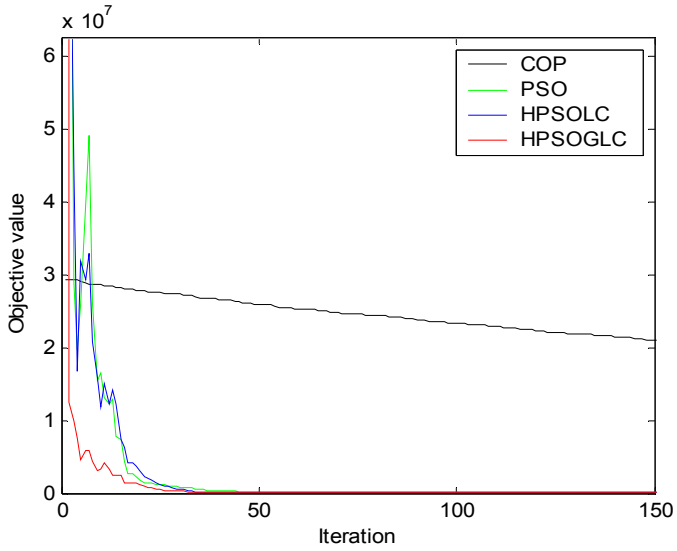


Figure 12 Convergence characteristics with respect to generalise Rastrigin with problem dimension size 10 (see online version for colours)

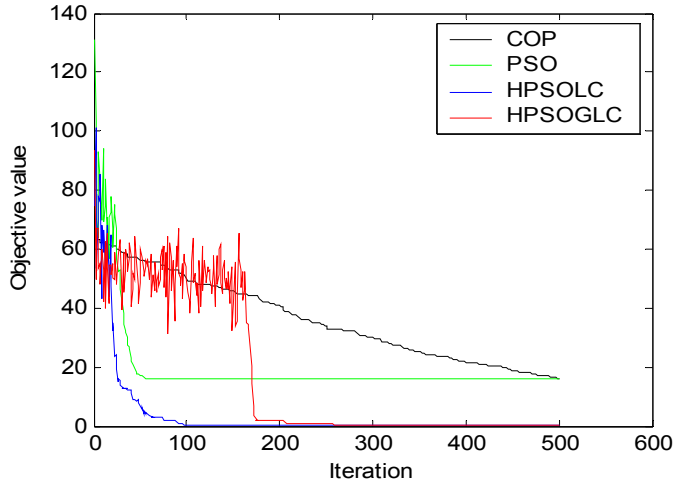


Table 9 Performances for generalise Rastrigin test function with dimension size equal to 30

Perf	*COP*	*PSO*	*PSOCC*	*HPSOLC*	*HPSOGLC*
Best	5.1e+01	8.8e+01	8.3e+01	1.6e-03	2.3e-03
Worst	3.9e+02	1.6e+02	1.5e+02	1.0e-02	1.6e-02
Mean	2.2e+02	1.2e+02	1.2e+02	4.7e-03	6.0e-03
Std. dev.	1.4e+02	2.3e+01	2.4e+01	3.1e-03	4.2e-03

Figure 13 Convergence characteristics with respect to generalise Rastrigin with problem dimension size 30 (see online version for colours)

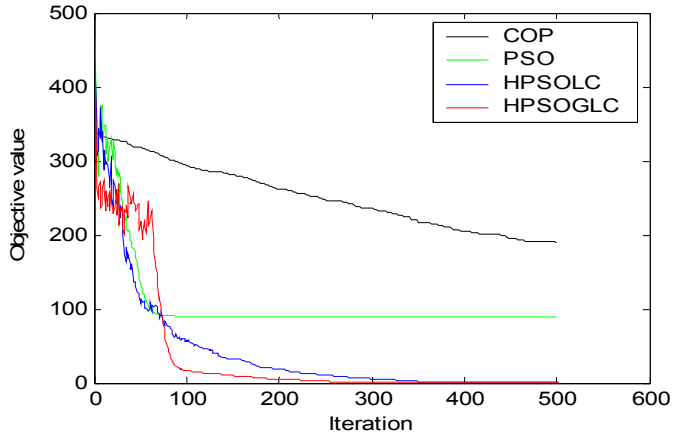


Table 10 Performances for Ackley's function test function with dimension size equal to 10

Perf	*COP*	*PSO*	*PSOCC*	*HPSOLC*	*HPSOGLC*
Best	1.4e+01	3.1e-07	3.1e-07	5.8e-06	1.4e-05
Worst	1.5e+01	1.6e+00	1.6e+00	6.4e-04	6.8e-04
Mean	1.5e+01	2.8e-01	2.8e-01	1.2e-04	1.1e-04
Std. dev.	5.1e-01	6.0e-01	6.0e-01	2.0e-04	2.1e-04

Figure 14 Convergence characteristics with respect to Ackley's function with problem dimension size 10 (see online version for colours)

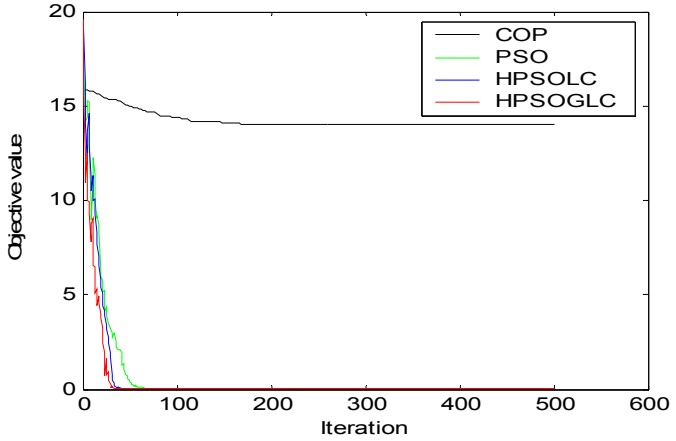


Table 11 Performances for Ackley's function test function with dimension size equal to 30

Perf	*COP*	*PSO*	*PSOCC*	*HPSOLC*	*HPSOGLC*
Best	1.5e+01	5.5e+00	5.5e+00	2.4e-02	2.5e-02
Worst	1.9e+01	9.8e+00	9.7e+00	4.0e-02	4.0e-02
Mean	1.8e+01	8.7e+00	8.4e+00	3.2e-02	3.2e-02
Std. dev.	8.7e-01	1.4e+00	1.4e+00	5.0e-03	4.7e-03

Figure 15 Convergence characteristics with respect to Ackley's function with problem dimension size 30 (see online version for colours)

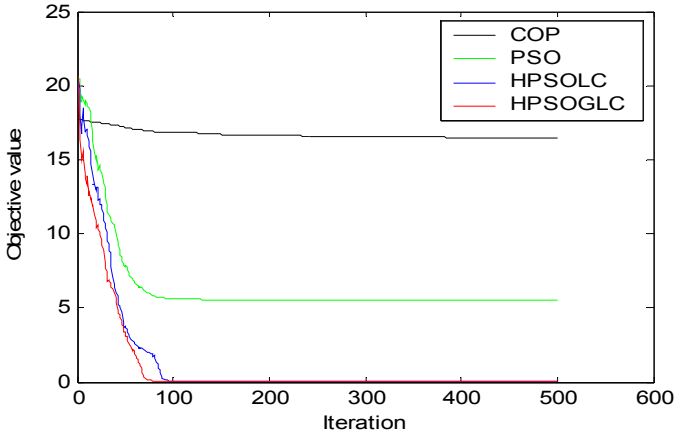


Figure 16 Convergence characteristics with respect to generalise Griewank function with problem dimension size 10 (see online version for colours)

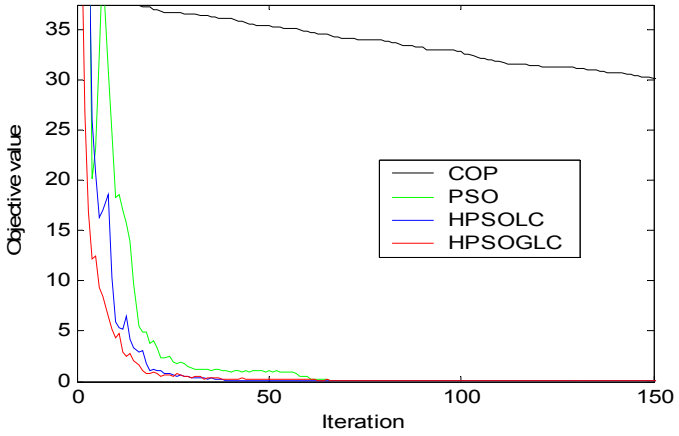


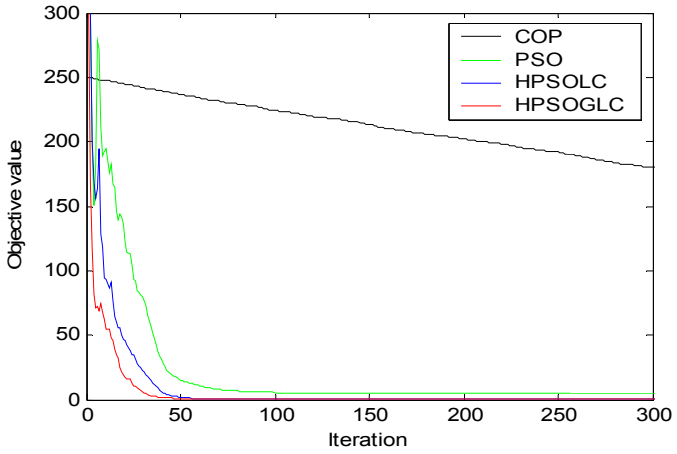
Table 12 Performances for Generalise Griewank function test function with dimension size equal to 10

Perform	*COP*	*PSO*	*PSOCC*	*HPSOLC*	*HPSOGLC*
Best	3.4e+00	1.2e-02	1.2e-02	4.2e-02	5.7e-02
Worst	3.2e+01	1.2e-01	1.2e-01	1.6e-01	2.1e-01
Mean	1.4e+01	7.1e-02	7.1e-02	9.4e-02	1.1e-01
Std. dev.	7.9e+00	2.9e-02	2.9e-02	3.9e-02	5.2e-02

Table 13 Performances for Generalise Griewank function test function with dimension size equal to 30

Perf	*COP*	*PSO*	*PSOCC*	*HPSOLC*	*HPSOGLC*
Best	1.9e-01	5.2e+00	1.4e-01	1.5e-02	2.0e-02
Worst	2.1e+02	1.3e+01	1.2e+01	8.3e-02	7.9e-002
Mean	7.9e+01	8.0e+00	4.6e+00	3.4e-02	4.3e-02
Std. dev.	1.0e+02	2.8e+00	5.2e+00	1.9e-02	1.8e-02

Figure 17 Convergence characteristics with respect to generalise Griewank function with problem dimension size 30 (see online version for colours)

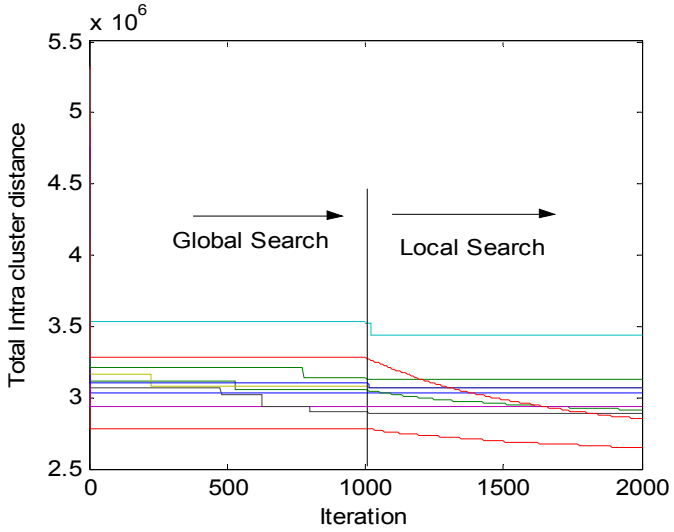


7.2 Clustering in information retrieval

To capture the wide practical cases a synthetic dataset contains three categories of document (total 300 documents), each category carry 100 documents and each document contains 30 keywords. Frequency of availability of keywords in all categories is uniformly distributed random number. The range of frequency appeared for the first ten in the first category is [10, 20], for next ten keywords in a range of [0, 1] and last ten is in the range of [0, 3]. The range of frequency appeared for the first ten in the second category is [0, 2], for next ten keywords in a range of [10, 20] and last ten is in the range of [0, 2]. The range of frequency appeared for the first ten in the third category is [0, 2], for next ten keywords in a range of [0, 3] and last ten is in a range of [10, 20].Weighting

scheme has applied over generated dataset as according to (1). Experiments have done with all developed methods to obtain the compact and better quality of clusters. For COP there are 1,000 iterations for global search and 1,000 iterations for local search and initial cluster centroids have defined randomly between minimum and maximum range of weighted document data. Convergence performances for all ten independent trials have shown in Figure 18. There are two parts in the plot, results up to 1,000 iterations have given by global search of COP while later part is the result of local search defined under COP. It can observe that there is not a very significant improvement in the intra cluster distance with local search facility.

Figure 18 Convergence characteristics with respect to COP in cluster problem (see online version for colours)



PSO and PSOCC-based cluster convergence characteristics have shown in Figure 19. In the first phase 500 iterations have given to PSO, which have a population size equal to 20 and next 500 iterations have applied with local search of COP. As it was observed in numerical optimisation, there is also no significant change appeared.

In HPSOLC and HPSOGLC, inherent PSO has given iteration of 100 with population size 20 and initial cluster centroid population have defined by a random sample selection method from dataset. Local search inherited with HPSOLC has given 100 iterations for each solution member where as for HPGLC there were 50 iterations for global search and 50 for local search. Convergence performance for HPSOLC and HPSOGLC have shown in Figure 20 and in Figure 21. It can observe that nature of convergence is smooth in both cases, but once HPSOLC could not converge towards optimality while HPSOGLC has converged properly in all trials.

The performance of all methods in terms of intra cluster distance (ICD) with respect to 10 independent trials have shown in Table 14. Table 15 contains average F-measure value (F-MV) and purity estimated from obtaining clusters with success rate in terms of convergence of defining clusters. It can be observed that HPSOGLC has delivered the minimum intra cluster distance in compare to others as well as best possible value of F-measure and purity along with a 100% success rate.

Figure 19 Convergence characteristics with respect to PSO and PSOCC in cluster problem (see online version for colours)

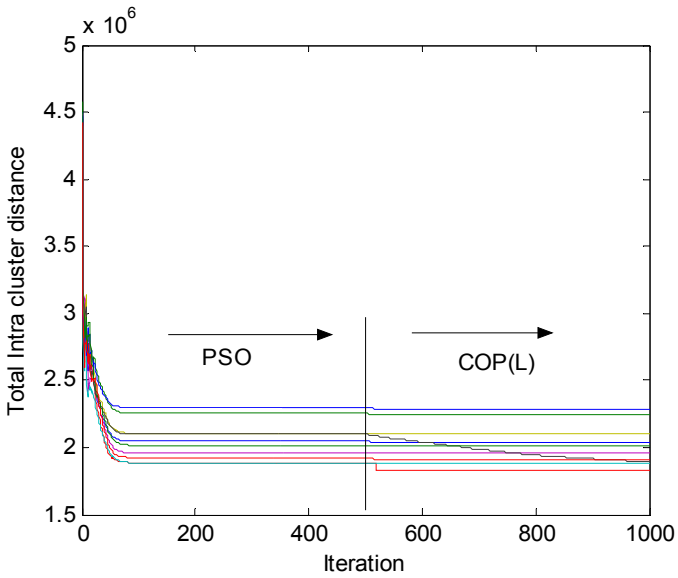
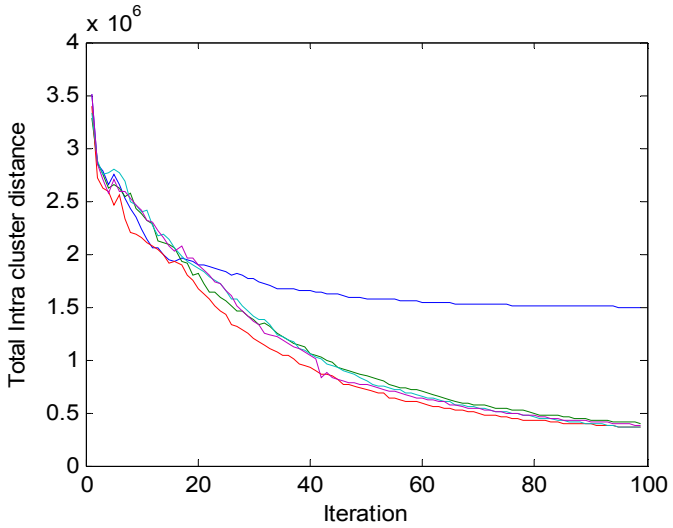


Figure 20 Convergence characteristics with respect to HPSOLC in cluster problem (see online version for colours)



For comparison purpose of convergence characteristics for best performance of each method have shown in Figure 22 and the relative intra cluster distance also has presented in Figure 23.

Figure 21 Convergence characteristics with respect to HPSOGLC in cluster problem (see online version for colours)

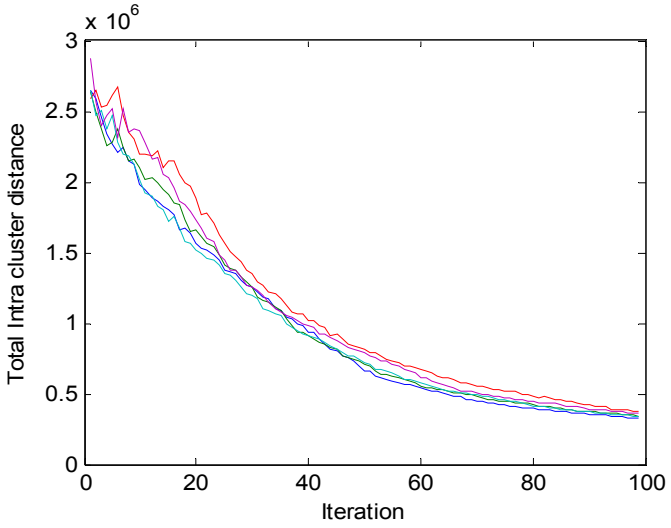


Table 14 Intra cluster distance (ICD) performance for clusters by different method in 10 independent trials

ICD	*COP*	*PSO*	*PSOCC*	*HPSOLC*	*HPSOGLC*
Best	2.6e+06	1.9e+06	1.8e+06	3.5e+05	3.2e+05
Worst	3.4e+06	2.3e+06	2.2e+06	1.5e+06	3.7e+05
Mean	3.0e+06	2.0e+06	2.0e+06	6.0e+05	3.5e+05
Std. dev.	2.1e+06	1.4e+05	1.5e+05	5.0e+06	2.0e+04

Figure 22 Convergence characteristics of best solution (see online version for colours)

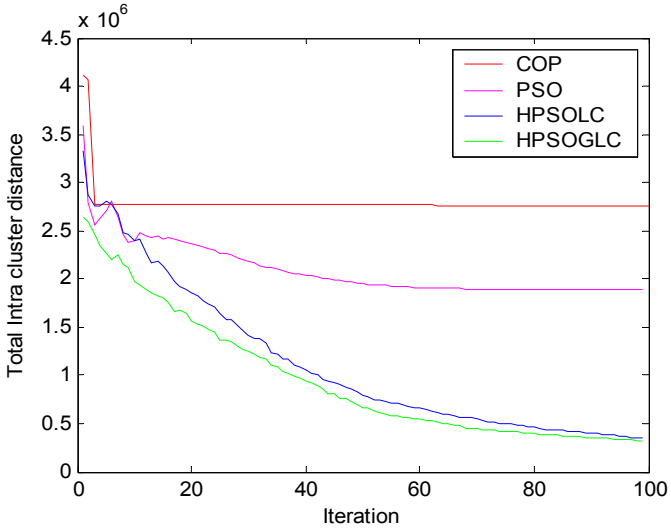


Figure 23 Normalised relative intra cluster distance (see online version for colours)

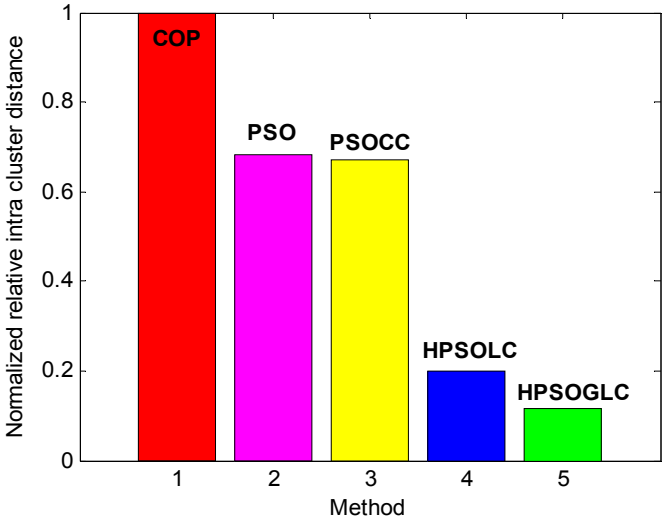


Table 15 Average performances of defined clusters

Perform	*COP*	*PSO*	*PSOCC*	*HPSOLC*	*HPSOGLC*
F-MV	0.85	1.0	1.0	1.0	1.0
Purity	0.81	1.0	1.0	1.0	1.0
Success rate (%)	40	90	90	90	100

8 Conclusions

A self-exploration scheme which enhances the performance of PSO very much has presented in this paper. Integration has defined at various structural levels to perform the self-exploration along with social influence. Integration of self-exploration which contains global as well as local search by chaotic method with PSO has given a better solution with a faster rate of convergence. Various possibilities of interaction have explored and it is observed that facility of solution exploration with a chaotic map at the time of progress phase has delivered a remarkable advantage compare to other stages. The proposed method can consider as a powerful solution method to those applications where optimisation is an important task and difficult to achieve the optimal solution.

Acknowledgements

This research work is completed in Manuro Tech Research Pvt. Ltd.; Bangalore, India. The authors expressed their thanks to associated members for their support.

References

Akter, R. and Chung, Y. (2013) 'An evolutionary approach for document clustering', *EECS 2013*, Elsevier, Vol. 4, pp.370–375.

Al-Shboul, B. and Myaeng, S-H. (2009) 'Initializing K-means using genetic algorithms', *World Academy of Science, Engineering and Technology*, Vol. 54, No. 30, pp.114–118.

Bsoul, Q., Salim, J. and Zakaria, L.Q. (2013) 'An intelligent document clustering approach to detect crime patterns', *ICEEI*, Elsevier, pp.1181–1187.

Chen, L., Gao, Y., Li, X., Jensen, C.S. and Chen, G. (2015) 'Efficient metric indexing for similarity search', *Data Engineering (ICDE)*, pp.591–602, DOI: 10.1109/ICDE.2015.7113317.

Chiang, I., Liu, C., Tsai, Y. and Kumar, A. (2015) 'Discovering latent semantics in web documents using fuzzy clustering', *IEEE Transactions on Fuzzy Systems*, Vol. 23, No. 6, pp.2122–2134, DOI: 10.1109/TFUZZ.2015.2403878.

Clerc, M. and Kennedy, J. (2002) 'The particle swarm optimization-explosion stability, and convergence in a multidimensional complex space', *IEEE Trans. Evol. Comput*., Vol. 6, No. 1, pp.58–73.

Habibi, M. and Popescu-Belis, A. (2015) 'Keyword extraction and clustering for document recommendation in conversations', *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 23, No. 4, pp.746–759, DOI: 10.1109/TASLP.2015.2405482.

Han, M., Cao, Z. and Li, Y. (2014) 'An improved case-based reasoning method based on fuzzy clustering and mutual information', *IEEE, Intelligent Control and Information Processing (ICICIP)*, pp.293–300, DOI: 10.1109/ICICIP.2014.7010266.

Kamble, A. (2010) 'Incremental clustering in data mining using genetic algorithm', *International Journal of Computer Theory and Engineering*, June, Vol. 2, No. 3, pp.1793–8201

Kennedy, J. and Eberhart, R. (1995) 'Particle swarm optimization', in *Proceedings of IEEE International Conference on Neural Networks*, Vol. 4, pp.1942–1948.

Liang, Y., Dong, L., Xie, S., Na, L.V. and Xu, Z. (2014) 'Compact feature based clustering for large-scale image retrieval', *Multimedia and Expo Workshops (ICMEW), IEEE International Conference*, pp.1–6, DOI: 10.1109/ICMEW.2014.6890597.

Ling, Y., Pan, X., Li, G. and Hu, X. (2015) 'Clinical documents clustering based on medication/symptom names using multi-view nonnegative matrix factorization', *IEEE Transactions on NanoBioscience*, Vol. 14, No. 5, pp.500–504, DOI: 10.1109/TNB.2015.2422612.

Lorena, L.H.N., Lorena, A.C., Lorena, L.A.N. and De Leon Carvalho, A.C.P. (2014) 'Clustering search applied to rank aggregation', *Intelligent Systems (BRACIS), 2014 Brazilian Conference*, pp.198–203, DOI: 10.1109/BRACIS.2014.44.

Lorenz, E.N. (1993) *The Essence of Chaos*, University of Washington Press, Seattle, WA.

MacQueen, J.B. (1967) 'Some methods for classification and analysis of multivariate observations', *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, Vol. 1, pp.281–297.

Niknam, T. and Amiri, B. (2010) 'An efficient hybrid approach based on PSO, ACO and k-means for cluster analysis', *Elsevier Applied Soft Computing*, Vol. 10, No. 1, pp.183–197.

Niknam, T., Firouzi, B.B. and Nayeripour, M. (2008) 'An efficient hybrid evolutionary algorithm for cluster analysis', *World Applied Sciences Journal*, Vol. 4, No. 2, pp.300–307.

Prakasha, S., Singh, M.K. and Raju, G.T. (2013) 'Clustering of text document based on ASBO', *Wulfenia Journal*, Vol. 20, No. 6, pp.152–165.

Yedla, M., Pathakota, S.R. and Srinivasa, T.M. (2010) 'Enhancing k-means clustering algorithm with improved initial center', *International Journal of Computer Science and Information Technologies (IJCSIT)*, Vol. 1, No. 2, pp.121–125.

Yuan, F., Meng, Z.H., Zhangz, H.X. and Dong, C.R. (2004) 'A new algorithm to get the initial centroids', *Proceedings of the 3rd International Conference on Machine Learning and Cybernetics*, pp.26–29.

Zhang, C. and Fang, Z. (2013) 'An improved k-means clustering algorithm', *Journal of Information & Computational Science*, Vol. 10, No. 1, pp.193–199.

Zhang, C. and Xia, S. (2009) 'K-means clustering algorithm with improved initial center', in *Second International Workshop on Knowledge Discovery and Data Mining (WKDD)*, pp.790–792.

Zhao, Y., Chen, Y., Liang, Z., Yuan, S. and Li, Y. (2014) 'Big data processing with probabilistic latent semantic analysis on MapReduce', *Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*, pp.162–166, DOI: 10.1109/CyberC.2014.37.