IJREAT International Journal of Research in Engineering & Advanced Technology, Volume 1, Issue 3, June-July, 2013
ISSN: 2320 - 8791
www.ijreat.org

An Enhanced Document Clustering Approach using Optimization Algorithm

Perumal P[1], Nedunchezhian R[2], Gomathi C[3]

[1] Associate Professor, Department of CSE, Sri Ramakrishna Engineering College, Coimbatore, Tamilnadu, India, 641 022

[2] Professor and Head, Department of IT, Sri Ramakrishna Engineering College, Coimbatore, Tamilnadu, India, 641 022

[3] PG Scholar, Department of CSE, Sri Ramakrishna Engineering College, Coimbatore, Tamilnadu, India, 641 022

Abstract

Fast and high quality document clustering is a crucial task in organizing information, search engine results, enhancing web crawling, and information retrieval or filtering. Recent studies have shown that the most commonly used partition-based clustering algorithm, the *K*-means algorithm, is more suitable for large datasets. In the existing system, WordNet enabled W-k means clustering algorithm significantly improves standard k-means generating useful and high quality cluster tags but not time efficiency. In this system, a novel document clustering algorithm based on the Harmony Search (HS) optimization method is proposed. By modelling clustering as an optimization problem, we first propose a pure HS based clustering algorithm that finds near-optimal clusters within a reasonable time. Then, harmony clustering is integrated with the K-means algorithm to achieve better clustering. Contrary to the localized searching property of K-means algorithm, the proposed algorithms perform a globalized search in the entire solution space. Additionally, the proposed system improves K-means by making it less dependent on the initial parameters such as randomly chosen initial cluster centres, therefore, making it more stable. Experimental results reveal that the HS integrated with K-means algorithm converges to the best known optimum faster than other methods and the quality of clusters are comparable.

Keywords— **Harmony search, K-means, Optimization, etc.,**

1. Introduction

Around the globe, news articles flood the Web every day from an extreme amount of major or minor news portals. It is completely impossible for an individual to be able to keep track of an event, or a series of related events, from an unbiased and truly informative point of view. While the amount of online information sources is exponentially increasing, so does the available online news content. For organizing this enormous amount of data, one of the most common approaches is the use of clustering techniques.

Clustering is an automatic learning technique aimed at grouping a set of objects into subsets or clusters. The goal is to create clusters that are coherent internally, but substantially different from each other. In plain words, objects in the same cluster should be as similar as possible, whereas objects in one cluster should be as dissimilar as possible from objects in the other clusters. A simple definition of clustering could be “the process of organizing objects into groups whose members are similar in some way”. Examples of document clustering include web document clustering for search users. The challenges which has given the motivation to use clustering of the news articles are (i) The number of available articles was large, (ii) A large number of articles were added each day, (iii) Articles corresponding to same news were added from different sources and (iv) The recommendations had to be generated and updated in real time.

The challenges which the clustering techniques normally should overcome are efficiency, ambiguity and synonymy. Efficiency, generated clusters have to be well connected from a notional point of view, despite the diversity in content and size that the original documents might have. For example, it is frequent for some news articles to belong to the same notional cluster, even though they do not share common words. The vice-versa is also possible: news articles sharing common words, while being completely unrelated to each other. Furthermore, having IR systems simply generate clusters of documents is not enough. The reason is that it is virtually impossible for humans to conceptualize information by merely browsing through hundreds of documents belonging to the same cluster. However, assigning meaningful labels to the generated clusters can help users conveniently recognize the content of each generated set and thus easily analyse the results.

To alleviate the limitations of K-means algorithm, here involves the use of optimization methods that optimize a pre-defined clustering objective function. Specifically, optimization based methods define a global objective function over the quality of clustering algorithm and traverse the search space trying to optimize its value. Any general purpose optimization method can serve as the basis of this approach such as Genetic Algorithms (GAs) , Ant Colony Optimization and Particle Swarm Optimization, which have been used for web page and image clustering. Since stochastic optimization approaches are good at avoiding convergence to a locally optimal solution, these approaches could be used to find a global near-optimal solution. However the stochastic approaches take a long time to converge to a globally optimal partition. Harmony Search (HS) is a new meta-heuristic optimization method imitating the music improvisation process where musicians improvise the pitches of their instruments searching for a perfect state of harmony. HS has been very successful in a wide variety of optimization problems, presenting several advantages over traditional optimization techniques such as: (a) HS algorithm imposes fewer mathematical requirements and does not require initial value settings for decision variables, (b) as the HS algorithm uses stochastic random searches, derivative information is also unnecessary, and (c) the HS algorithm generates a new vector, after considering all of the existing vectors, whereas methods such as GA only consider the two parent vectors. These three features increase the flexibility of the HS algorithm.

2. Related Work

Argyris Kalogeratos and Aristidis Likas (2011) [1] proposed the idea of synthetic cluster prototype that has been computed by first selecting a subset of cluster objects and then computing the representative of these objects and finally selecting important features thus introduced the MedoidKNN synthetic prototype that favors the representation of the dominant class in a cluster.

Ming-Chao Chiang, Chun-Wei Tsai and Chu-Sing Yang (2011) [14] proposed a pattern reduction algorithm for reducing the computation time of K-means and also K-means-based clustering algorithms, which work by compressing and removing at each iteration patterns that are unlikely to change their membership. This can also be applied to many other iterative clustering algorithms such as kernel-based and population-based clustering algorithms.

Shanfeng Zhu, et.al, (2009) [19] proposed a probabilistic model called field independent clustering model (FICM) to explicitly handle each field in a document separately and also incorporated the distinct word distributions of each field to integrate the discriminative abilities as well as to select the most suitable component probabilistic model for each field.

Jim Z.C. Lai and Tsung-Jen Huang (2011) [9] proposed a method to resolve the problem of a non-optimal solution for Double linked algorithm (DLA) while keeping the corresponding advantage of low computational complexity.

Greg Hamerly, Erez Perelman and Brad Calder (2006) [8] provided a detailed comparison of using *k*-means and multinomial clustering for SimPoint and also they have showed that *k*-means performs better than the recently proposed multinomial clustering approach. As a further process, they have proposed two improvements to the prior multinomial clustering approach in the areas of feature reduction and picking of simulation points which allow multinomial clustering to perform as well as *k*-means. Finally, they have proposed a new metric, cluster-purity for determining how effective the multinomial algorithm is at characterizing a program. As a result, there is a small improvement in accuracy with a small reduction in the average number of simulation points.

Yeming Hu, Evangelos Milios and James Blustein (2010) [23] proposed an automated framework Document Clustering using Iterative Class-Based Feature Selection (DCIFS) which iteratively updated the feature set for clustering and improved the clustering performance over baselines in most cases. In order to eliminate noisy features selected in the automated framework, Interactive Document Clustering Using Iterative Class-Based Feature Selection (IDCIFS) has been proposed in which users are invited to confirm whether a feature is good or not. Experiments have been showed that IDCIFS improves clustering performance over pure iterative DCIFS. Also their framework with user interaction reduced the effect of noisy features as feature reweighting gives higher weights to the user accepted features.

Julian Sedding and Dimitar Kazakov (2004) [10] proposed a method to cluster documents by meaning which is relevant to language understanding. Also, naive, syntax-based disambiguation has been attempted by assigning each word a part-of-speech tag and by enriching the ‘bag-of-words’ data representation often used for document clustering with synonyms and hypernyms from WordNet. As a result, quality increases with the number of clusters.

Congnan Luo, Yanjun Li and Soon M. Chung (2009) [5] proposed a method to use the neighbours and link along with the cosine function in different aspects of the K-means and bisecting K-means algorithms for clustering documents. This system has used a new method of selecting the initial centroids and then a new similarity measure to determine the closest cluster centroid and finally, a new heuristic function for the bisecting K-means algorithm to select a cluster to split.

K. Venkata Ratnam, H. Devaraju and Y. Ramesh Kumar (2012) [21] proposed a novel mechanism for multi view point with different similarity measure so that more informative assessment of similarity can be achieved. As a result, theoretical analysis and empirical examples have been showed that MVS is potentially more suitable than the popular cosine similarity for text documents. Also it has provided efficient results than single view point Clustering mechanisms.

3. Proposed Work

Recent studies have shown that the most commonly used partitioning based clustering algorithm, the K-means algorithm, is more suitable for large datasets. However, the K-means algorithm may generate a local optimal clustering. Although K-means algorithm is straightforward, easy to implement and works fast in most situations, it suffers from some major drawbacks that make it unsuitable for many applications. The first disadvantage is that the number of clusters K must be specified in advance [13]. In addition, since the summary statistic that is maintained for each cluster by K-means algorithm is simply the mean of samples assigned to that cluster, the individual members of the cluster can have a high variance and hence the mean may not be a good representative for the cluster members. Further, as the number of clusters grows into the thousands, K-means clustering becomes untenable, approaching O (m2) comparisons where m is the number of documents.

However, for relatively few clusters and a reduced set of pre-selected features, K-means performs well. Another major drawback of the K-means algorithm is its sensitivity to initialization. Lastly, the K-means algorithm converges to local optima, potentially leading to clusters that are not globally optimal. To alleviate the limitations of traditional partition based clustering methods discussed above, particularly the K-means algorithm, different techniques have been introduced in recent years. One of these techniques involves the use of optimization methods that optimize a pre-defined clustering objective function. Specifically, optimization based methods define a global objective function over the quality of clustering algorithm and traverse the search space trying to optimize its value.

4. Architecture

A. Keyword Extraction

At its input stage, this system crawls and fetches news articles from major or minor news portals from around the world. This is an offline procedure and once articles as well as metadata information are fetched, they are stored in the centralized database from where they are picked up by the following procedures. A key procedure of the system as a whole, which is probably as least as significant as the clustering algorithm that follows it, is preprocessing of text on the fetched article's content, that results into the extraction of the keywords each article consists of. Keyword extraction handles the cleaning of articles, the extraction of the nouns [3], the stemming as well as the stop word removal process. Following, it applies several heuristics to come up with a weighing scheme that appropriately weighs the keywords of each article based on information about the rest of the documents in our database.

Next comes the pruning of words that appear with low frequency throughout the corpus and are unlikely to appear in more than a small number of articles. Keyword extraction in essence generates the term-frequency vector [12] for each article that is used by the information retrieval techniques that follow treating it as a 'bag of words'.

Text summarization is the categorization of the articles on a predetermined set of classes and also contains some additional steps deployed in order to extract useful information from the data [5]. The intention of text summarization is to express the content of a document in a condensed form that meets the needs of the user. Far more information that can realistically be digested is available on the World-Wide Web and in other electronic forms. News information, biographical information, minutes of meetings missed isn't possible to read everything one would want to read and so some form of information condensation is needed. Following the retrieval techniques, information is transmitted back to the end user.

B. Enriching synsets

The Word Net is a lexical reference system, which organizes different rhetorical relations into hierarchies. The main relation among words in WordNet is synonymy, as between the words significant and important or apple and fruit. Synonyms are the words that specify the same concept and are interchangeable in many contexts and also grouped into unordered sets (synsets). Each of WordNet's 117 000 synsets is linked to other synsets by means of a small number of "conceptual relations." Moreover, a synset has a brief definition and additionally one or more example sentences demonstrating the use of the synset members. Word forms with several different meanings are represented in as many distinct synsets.

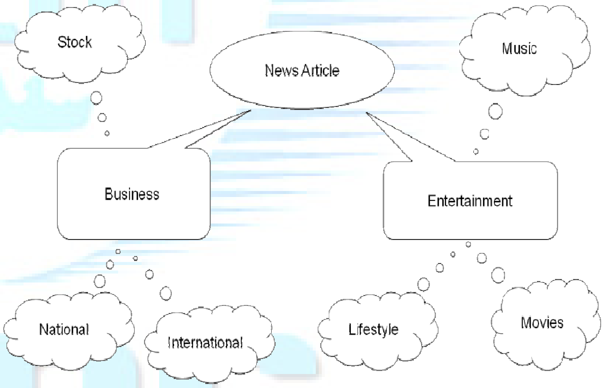


Fig. 1 Hypernym example

Initially, for each given keyword of the article, we generate its graphs of hypernyms leading to the root hypernym. Following, we combine each individual hypernym graph to an aggregated one. There are practically two parameters that need to be taken into consideration for each hypernym of the aggregate tree-like structure in order to determine its importance: the depth and the frequency of appearance.

For example, Fig. 1 depicts the hypernym example for six terms. If the hypernym is higher in the graph, then it is more global. If the hypernym is lower in the graph, then there is less chance to occur many graph paths, i.e. its frequency of appearance is low. We can also see that each term might have multiple graph paths that lead from the term itself to the root, i.e. 'entity' node.

Most importantly, given any verb, noun, adverb and adjective, WordNet can provide results regarding hypernyms, hyponyms, meronyms or holonyms. Using these graph-like structures, we can search the Word Net database for all the hypernyms of a given set of words, and then weigh them appropriately.

C. Harmony search algorithm

Harmony Search (HS) is a new meta-heuristic [22] optimization method imitating the music improvisation process where musicians improvise their instruments pitches searching for a perfect state of harmony. The superior performance of the HS algorithm has been demonstrated through its application to different problems. Here, we provide a brief introduction to the main algorithm and the interested reader can refer to for theoretical analysis of the exploratory power of the HS algorithm. The overall procedure of the algorithm is described [16] in Fig. 2.

1) Initialize the problem and algorithm parameters

In this step, the optimization problem is specified as follows:

Minimize f(x) [12] subject to:

$$\begin{aligned} & g_i(x) \geq 0 \qquad i = 1,2,\ldots.m \\ & h_j(x) = 0\ j = 1,2,\ldots.p \qquad (1) \\ & LB_k \leq x_k \leq UB_k\ k = 1,2,\ldots n \end{aligned}$$

Where f(x) is the objective function, m is the number of inequality constraints and p is the number of equality constraints and n is the number of decision variables. The lower and upper bounds for each decision variable k are LBk and UBk respectively [15]. The HS parameters are declared in the equation (4.1). These are the harmony memory size (HMS), or the number of solution vectors in the harmony memory, the probability of memory considering (HMCR), [7] the probabilityof pitch adjusting (PAR), and the number of improvisations (NI), or stopping criterion. The harmony memory (HM) is a memory location where all the solution vectors (sets of decision variables) are stored. This HM is similar to the genetic pool in the GA. The HMCR, which varies between 0 and 1, is the rate of choosing one value from the historical values stored in the HM, while 1-HMCR is the rate of randomly selecting one value from the possible range of values.

2) Initialize the harmony memory

In this step, the HM matrix represented in equation (4.2) is filled with as many randomly generated solution vectors as the HMS allows:

$$HM = \left[\begin{array}{ccccc|c} x_1^1 & x_2^1 & \cdots & x_{n-1}^1 & x_n^1 & f(x_1) \\ x_1^1 & x_2^1 & \cdots & x_{n-1}^1 & x_n^1 & f(x_2) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_1^{HMS-1} & x_2^{HMS-1} & \cdots & x_{n-1}^{HMS-1} & x_n^{HMS-1} & f(x_{HMS-1}) \\ x_1^{HMS} & x_2^{HMS} & \cdots & x_{n-1}^{HMS} & x_n^{HMS} & f(x_{HMS}) \end{array}\right] \qquad (2)$$

The initial harmony memory is generated from a uniform distribution in the ranges [LB_i,UB_i], where 1≤ i ≤ n. This is done in the equation (4.3).

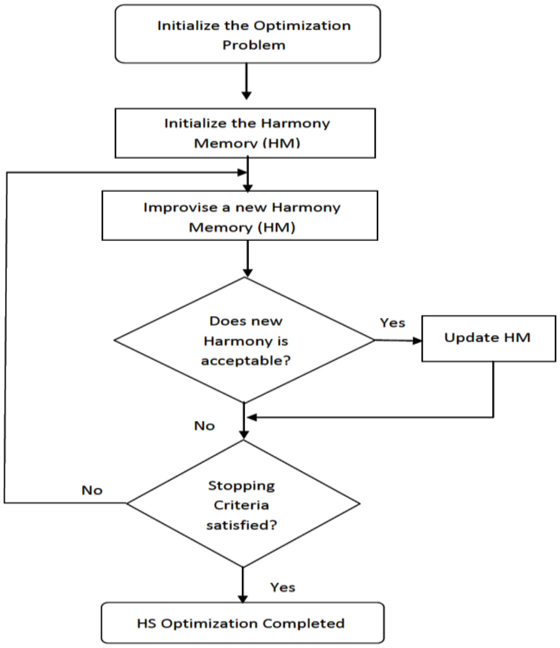
$$x_i^j = LB_i + r * (UB_i - LB_i),\ j = 1,2,\ldots..HMS \qquad (3)$$

Where r ~ U(0,1) and U is a uniform random number generator and x is the set of each decision variable.

3) Improvise a new harmony

Generating a new harmony is called improvisation. A New Harmony vector x' = (x'_1, x'_2,……, x'_n) is generated based on three rules: memory consideration, pitch adjustment, and random selection. In the memory consideration, the value for a decision variable is randomly chosen from the historical values stored in the HM with the probability of HMCR.

Every component obtained by the memory consideration is examined to determine whether it should be pitch-adjusted. This operation uses the PAR parameter, which is the probability of pitch adjustment. Variables which are not selected for memory consideration will be randomly chosen


Initialize the Optimization Problem
Initialize the Harmony Memory (HM)
Improvise a new Harmony Memory (HM)
Does new Harmony is acceptable?
Yes
Update HM
No
No
Stopping Criteria satisfied?
Yes
HS Optimization Completed

from the entire possible range with a probability equal to 1-HMCR.

Fig. 2 Flowchart of HS algorithm

4) Update harmony memory

If the New Harmony vector, x' = (x'$_1$, x'$_2$,……, x'$_n$) has better fitness value than the worst harmony in the HM, the new harmony is included in the HM and the existing worst harmony is excluded from it.

5) Check stopping criterion

The HS is terminated when the stopping criterion (e.g., maximum number of improvisations) has been met. Otherwise, Steps 3 and 4 are repeated. We note that in recent years, some researchers have improved the original HS algorithm. In a paper [12] there is a proposal for an improved variant of HS by using varying parameters. The intuition behind these algorithms is as follows. Although the HMCR and PAR parameters of HS help the method in searching for globally and locally improved solutions, respectively, however PAR and bw parameters have a profound effect on the performance of the HS. Thus, fine tuning these two parameters is very important. Of the two parameters, bw is more difficult to tune because it can take any value from (0, ∞). To address these shortcomings of HS, a new variant of HS, called the Improved Harmony Search (IHS), is proposed. IHS dynamically updates PAR according to the equation (4.4) as,

$$\mathrm{PAR(t)} = PAR_{min} + \frac{(PAR_{max} - PAR_{min})}{NI} \times \mathrm{t} \qquad (4)$$

where PAR(t) is the pitch adjusting rate for generation t, PAR$_{min}$ is the minimum adjusting rate, PAR$_{max}$ is the maximum adjusting rate, t is the generation number and NI is the maximum number of generations. In addition, bw is dynamically updated as represented in the equation (4.5) as:

$$\mathrm{bw(t)} = bw_{max}\,exp\left(\frac{\ln\left(\frac{bw_{min}}{bw_{max}}\right)}{NI} \times \mathrm{t}\right) \qquad (5)$$

where bw(t) is the bandwidth for generation t, bw$_{min}$ is the minimum bandwidth and bw$_{max}$ is the maximum bandwidth. In order to overcome the parameter setting problem of HS which is very tedious and could be another daunting optimization problem, [6] a variant of HS which eliminates tedious and difficult parameter assigning efforts is proposed.

D. Clustering process

In this step, the resulting centroid from HS algorithm is given to K-means algorithm, where instead of assuming initial centroid, this is taken as best initial centroid. The proposed system is described in the Fig. 3.

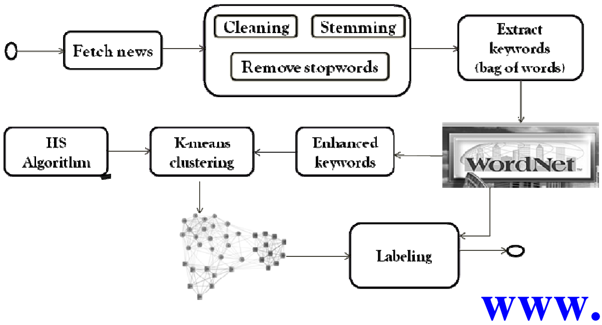


Fig .3 Architecture of the proposed system

K-means is an algorithm to classify or to group objects based on attributes/features into K (positive integer) number of group. The grouping is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid. Thus, the purpose of K-mean clustering is to classify the data.

If the number of data is bigger than the number of cluster, for each data, we calculate the distance to the centroid and get the minimum distance. This data is said belong to the cluster that has minimum distance from this data. For calculating the distance between cluster centroid to each object, we are using the following distance function,

Euclidian distance where the distance between two data points is defined [4] in the equation (4.6) as:

$$\mathrm{d(a,b)} = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \qquad (6)$$

where a and b are data points and n is the dimensionality of data.Then we assign all the data to this new centroid. This process is repeated until no data is moving to another cluster anymore. Mathematically this loop can be proved convergent.

E. Labeling process

The generated clusters are finally forwarded for labeling, taking also advantage of the Word Net database. The labeling process outputs suggested tags for the given cluster. Cluster assignments and labels [4] are the output of the proposed approach.

Thus, pre-processing is done as the first step with the process of stemming and stopwords removal. After this, the keywords are enriched with the external database where hypernyms are generated. These are given as source to the harmony search algorithm, where best initial centroid is chosen. This centroid is given to K-means clustering algorithm where a number of clusters are formed. At last the clusters are given corresponding labels.

5. Results and Discussion

The dataset used here is 20 news-group dataset where a large number of documents are available for usage and they are analyzed offline. By comparing the results of the existing system, intra cluster similarity gets increased when using Harmony Search algorithm before clustering in this system.

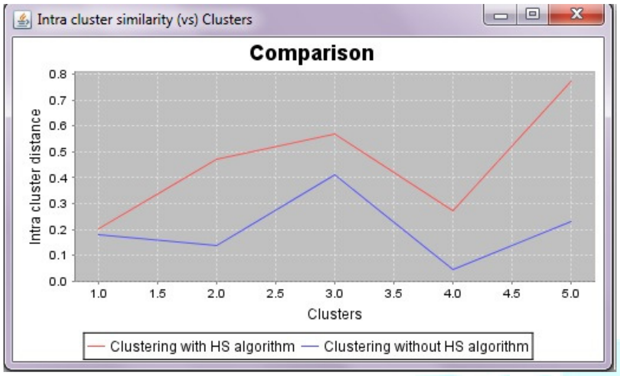


Fig. 4Intra cluster similarity among clusters

As represented in the Fig. 4, the value is for clustering using HS integrated with K-means algorithm gives 0.78 and the clustering without HS algorithm gives 0.23 as intra cluster similarity value.

Another observation is that as the number of articles increases, the Clustering Index (CI) [20] difference of HS-k means compared to k-means gets wider. This is because of the fact that while this experimentation data set grows larger; the probability of hypernyms occurring also increases. Therefore, this clustering approach has a better chance of selecting clusters with improved connectivity while at the same time keeping different clusters well separated from each other.

In order to determine the efficiency of the each clustering method, the evaluative criteria of CI is used. Intuitively, since the most efficient clusters are the ones containing articles close to each other within the cluster, while sharing a low similarity with articles belonging to different clusters, CI focuses on increasing the first measure (intra-cluster similarity) while decreasing the second (inter-cluster similarity). The clustering index is represented in the equation (5.1) as

$$CI = \frac{\bar{\sigma}^2}{\bar{\sigma} + \bar{\delta}} \qquad (7)$$

Where $\bar{\sigma}$ is the average intra-cluster similarity and $\bar{\delta}$ is the average inter-cluster similarity.

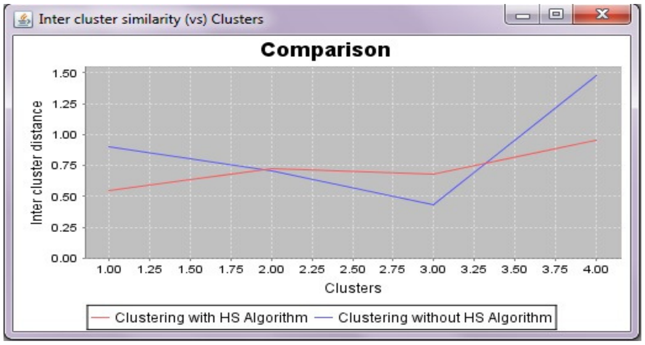


Fig. 5 Inter cluster similarity among clusters

The results presented in Fig. 5, gives the inter cluster similarity measure for the clusters in the case of WordNet enriched executions of the k-means algorithm. It is clearly depicted that clustering using HS integrated with K-means algorithm gives significantly improved clustering results when applied in the data set, regardless of the number of articles or the category they belong to.

6. Conclusion

When compared to the existing work, enhancing k-means clustering using NLP, the proposed work eliminates the random initialization problem found. Also, experimental results reveal that the proposed algorithms can find better clusters and the quality of clusters is comparable. A better improvement over the standard k-means algorithm in terms of high intra-cluster similarity and low inter-cluster similarity is found. Furthermore, the resulting labels are with high precision the correct ones as compared with their category tagging counterparts. The advantage of this system over existing system is that the influence of the improperly chosen initial cluster centers will be diminished by enabling the algorithm to explore the entire decision space over a number of iterations and simultaneously increasing its fine-tuning capability around the final decision. Therefore, it will be more stabilized and less dependent on the initial parameters such as randomly chosen initial cluster centers, while it is more likely to find the global solution rather than a local one.

References

[1] Argyris Kalogeratos and Aristidis Likas (2011), ‘Document clustering using synthetic cluster prototypes’, Elsevier science, Data & Knowledge Engineering, Vol.70, pp 284–306.

[2] Blum, C., and Andrea R. (2003), ‘Metaheuristics in Combinatorial Optimization: Overview and Conceptual Comparison’, ACM Computing Surveys, Vol. 35, pp 268–308.

[3] Bouras.C, Tsogkas.V (2008), ‘Improving text summarization using noun retrieval techniques’, Lecture Notes in Computer Science, Knowledge-Based Intelligent Information and Engineering Systems, Vol. 5178, pp 593–600.

[4] Bouras. C, Tsogkas. V (2012), ‘A clustering technique for news articles using WordNet’, Elsevier Science, Knowledge Based Systems, Vol. 36, pp 115-128.

[5] Congnan Luo, Yanjun Li and Soon M. Chung (2009), ‘Text document clustering based on neighbors’, Elsevier

science, Data & Knowledge Engineering, Vol.68 pp 1271–1288.

[6] Geem Zong Woo, Joong Hoon Kim and Loganathan G. V. (2001), 'A New Heuristic Optimization Algorithm: Harmony Search', SIMULATION 76:2, pp 60-68.

1)

[7] Geem Zong Woo, Sim K-B (2010), 'Parameter-setting-free harmony search algorithm', Elsevier science, Applied Mathematics and Computation, Vol. 217 (8), pp 3881–3889.

[8] Greg Hamerly, Erez Perelman and Brad Calder (2006), 'Comparing Multinomial and K-means clustering for SimPoint', Proc. IEEE International Symposium on Performance Analysis of Systems and Software, pp 131-142.

[9] Jim Lai.Z.C and Tsung-Jen Huang (2011), 'An agglomerative clustering algorithm using a dynamic k-nearest-neighbor list', Elsevier science, Information Sciences, Vol. 181, pp 1722–1734.

[10] Julian Sedding and Dimitar Kazakov (2004), 'WordNet-based Text Document Clustering', Proceedings of the 3rd Workshop on ROobust Methods in Analysis of Natural Language Data, pp 104-113.

[11] Lee Kang Seok, Geem Zong Woo (2005), 'A new meta-heuristic algorithm for continuous engineering optimization: harmony search theory and practice', Elsevier Science, Computer Methods in Applied Mechanics and Engineering, Vol. 194(36-38), pp 3902–3933.

[12] Mahdavi M., Fesanghary M., and Damangir E. (2007), 'An improved harmony search algorithm for solving optimization problems', Elsevier science, Applied Mathematics and Computation, Vol. 188, pp 1567–1579.

[13] Mehrdad Mahdavi and Hassan Abolhassani (2009), 'Harmony *K*-means algorithm for document clustering', Springer Science, Data Mining and Knowledge Discovery, Vol. 18, pp 370-391.

[14] Ming-Chao Chiang, Chun-Wei Tsai and Chu-Sing Yang (2011), 'A time-efficient pattern reduction algorithm for k-means clustering', Elsevier science, Information Sciences Vol.181 pp 716–731.

[15] Rana Forsati, et.al, (2013), 'Efficient stochastic algorithms for document clustering', Elsevier Science, Information Sciences, Vol. 220, pp 269–291.

[16] Sachin A. Patil and D. A. Patel (2013), 'Improved Harmony Search Algorithm and Its Applications in Mechanical Engineering', Proc. International Journal of Engineering Science and Innovative Technology (IJESIT) Volume 2(1).

[17] Salton.G, Wong.A, Yang.C.S (1975), 'A vector space model for automatic indexing', Communications of the ACM 18 (11), pp 613–620.

[18] Sayantani Ghosh, Mr. Sudipta Roy, and Samir Bandyopadhyay. K (2012), 'A tutorial review on Text Mining Algorithms', International Journal of Advanced Research in Computer and Communication Engineering Vol. 1, Issue 4.

[19] Shanfeng Zhu, et.al, (2009), 'Field independent probabilistic model for clustering multi-field documents', Elsevier science, Information Processing and Management, Vol. 4425, pp 331–342.

[20] Taeho Jo and Malrey Lee (2007), 'The Evaluation Measure of Text Clustering for the Variable Number of Clusters', Proc. International symposium on Neural Networks, pp 871-879.

[21] Venkata Ratnam. K, Devaraju. H and Ramesh Kumar. Y (2012), 'A Novel K-variant Algorithm for Document Clustering', International Journal of Engineering Science and Advanced Technology, Vol. 2, pp 1018-1022.

[22] Xin-She Yang (2009), 'Harmony Search as a Metaheuristic Algorithm', Springer Berlin Heidelberg, Studies in Computational Intelligence Vol.191, pp 1-14.

[23] Yeming Hu, Evangelos Milios and James Blustein (2010), 'Interactive Document Clustering Using Iterative Class-Based Feature Selection', Proc. ACM Symposium on Applied Computing, pp 1143-1150.