International Journal of Engineering & Technology, 7 (2.20) (2018) 396-400

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

An Efficient Modified K-Means and Artificial Bee Colony Algorithm for Mining Search Result from Web Database

V. Sabitha[1]*, Dr.S.K. Srivatsa[2]

[1]Research Scholar, Sathyabama University, Chennai.
[2]Senior Professor, Anna University, Chennai.
**Corresponding author E-mail:sabi0494@gmail.com*

Abstract

Nowadays, data growth is directly proportional to time and it is a challenge to store the data as well as retrieve the data in an organized fashion. The main goals of a web data clustering algorithm are to produce appropriate clusters for the end user, to assign the available data to the most relevant cluster, to respond the end user instantly. In this paper, we propose a new algorithm namely 'An Efficient Modified K-Means and Artificial Bee Colony Algorithm' to cluster web data. The proposed algorithm is the combination of K-means and Artificial Bee Colony (ABC) algorithm. The reasons for incorporating k-means algorithm are its simplicity and efficiency [9]. Initially, ABC algorithm is employed to achieve clustering [10] and this is followed by the application of k-means algorithm. The initial cluster centre is fixed by ABC algorithm. On experimental analysis, it is proved that the performance of An Efficient Modified K-Means and Artificial Bee Colony Algorithm is better than the other comparative algorithms, in terms of precision and recall. An annotation wrapper for the search site is automatically constructed and can be used to annotate new result pages from the same web database. From the annotated search result, frequently used websites are identified by using apriori Algorithm which involve pattern mining. The advantage of this new technique is fast operation on dataset containing items and provides facilities to avoid unnecessary scans to the database.

Keywords: *Web data clusters, modified k-means clustering, artificial bee colony algorithm, annotation generation, similarity measures.*

1. Introduction

Today's world revolves around data and the massive storehouse of data is the internet. Data growth is directly proportional to time and the stored data must be managed properly. Every day the internet deals with several petabyte (PB) of data. Thus, to distillate the useful information from the voluminous database is the major concern. Most of the information is in the form of digital data. Hence a mechanism is needed to systematize the data, such that the end users are able to retrieve the relevant data in a reasonable period of time. Data clustering is an effective mechanism that is meant for clustering related data together. This clustering approach paves way for finding the relevant data in minimal period of time.

The literal meaning of clustering is grouping; thus data clustering is systematizing the data into several classes based on the degree of relevance. Each class is denoted as a cluster and the entities within a cluster are closely related to each other. On the other hand, the entities of two different clusters will appear different [1-3]. Some of the main applicable areas of data clustering are found in data mining [4] and Content based Information Retrieval (CBIR) [5-8].

An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. The key element of this paradigm is the novel structure of the information processing system. It is composed of a large number of highly interconnected processing elements (neurons) working in unison to solve specific problems. An ANN is configured for a specific application, such as pattern recognition or data classification, through a learning process [7].

The main goals of a web data clustering algorithm are to produce appropriate clusters for the end user, to assign the available data to the most relevant cluster, to respond the end user instantly. In this paper, we propose a new algorithm namely 'An Efficient Modified K-Means and Artificial Bee Colony Algorithm' to cluster web data. The proposed algorithm is the combination of K-means and Artificial Bee Colony (ABC) algorithm. The reasons for incorporating k-means algorithm are its simplicity and efficiency [9]. Initially, ABC algorithm is employed to achieve clustering [10] and this is followed by the application of k-means algorithm. The initial cluster centre is fixed by ABC algorithm. On experimental analysis, it is proved that the performance of An Efficient Modified K-Means and Artificial Bee Colony Algorithm is better than the other comparative algorithms, in terms of precision and recall.

2. Related Works

The related work has been analyzed with the help of different research papers as defined below:

S. Kalyani, et al., (2011) Safety assessment is a major concern in the planning and operation of a power system studies. Conventional method of evaluating the security Played by computer simulation implies long and generates large result. Secure/ insecure under given operating condition and contingency this article presents a K-means approach for ranking the states of the power system. This article demonstrates how the traditional K-means clustering algorithm can be modified to be used with

advantage as a classification algorithm. The proposed algorithm combining particle swarm optimization (PSO) with the traditional k-means algorithm to meet the requirements of a classifier. PSO K-means clustering technique proposed base is implemented in IEEE 30 Buses, 57, 300 standard test security systems and bus 118 bus to static and transient security assessment. The simulation results of the algorithm are compared with the K-means clustering without supervision proposed, which use different methods for initializing the cluster center.

Dervis Karaboga et al., (2011) Artificial Bee Colony Algorithm (ABC), which is one of the most recently introduced optimization algorithms, simulates intelligent foraging behavior of a swarm of bees. Clustering Analysis, used in many disciplines and applications, is an important tool and a descriptive task trying to identify homogeneous groups of objects based on the values of their attributes. In this work, ABC is used for data aggregation on reference problems and the performance of the ABC algorithm is compared with particle Swarm Optimization (PSO) algorithm and nine other literature classification techniques. Thirteen typical test data sets from the UCI Machine Learning Repository are used to demonstrate the results of these techniques. The simulation results indicate that the ABC algorithm can be effectively used for multivariate data clustering.

R.J. Kuoet al., (2012) although cluster analysis algorithms are constantly improving, most clustering algorithms has yet to define the number of clusters. Thus, this study proposes a dynamic clustering approach based on the novel particle swarm optimization (PSO) and Genetic Algorithm (GA) (DCPG) algorithm. The proposed algorithm DCPG Data can be automatically ammunition by examining the data without a number of pre-specified clusters. The results of calculation of four reference data sets indicate that the algorithm has better DCPG validity and stability of the dynamic clustering approach based on binary PSO (DCPSO) and dynamic clustering approach based on GA (DCGA) algorithms. In addition, the algorithm is applied to DCPG group BOMs (BOM) for Advantech Company in Taiwan. Grouping the results can be used to classify products that share the same materials in clusters.

S. Rana et al., (2013) the data grouping is the most popular data analysis method in data mining. It is the method that the parties of the data object to significant groups. It has been applied in many fields such as image processing, pattern recognition and learning of the machine where the data sets are many shapes and sizes. The most popular and other conventional K-means algorithms suffer from a discount from their initial choice in the selection of local optima and center of gravity. This article presents a new improved algorithm named Adaptive Boundary small particles Swam Optimization (BR-APSO) algorithm with the limit restriction policy. The BR-APSO proposed algorithm is tested on new data sets, and its results are compared with those of PSO, NM-PSO, PSO-K and K-means algorithms. It has been found that the proposed algorithm is robust, generates more accurate results and its convergence speed is also faster compared to other algorithms.

3. Research Methodology

Clustering is the development of distinguishing usual federations or clusters in multidimensional statistics established on some similarity processes [6]. The proposed efficient modified k-means and artificial bee colony algorithm consist of four phases are: Data modeling, Similarity measures selection, clustering model and annotation generation.

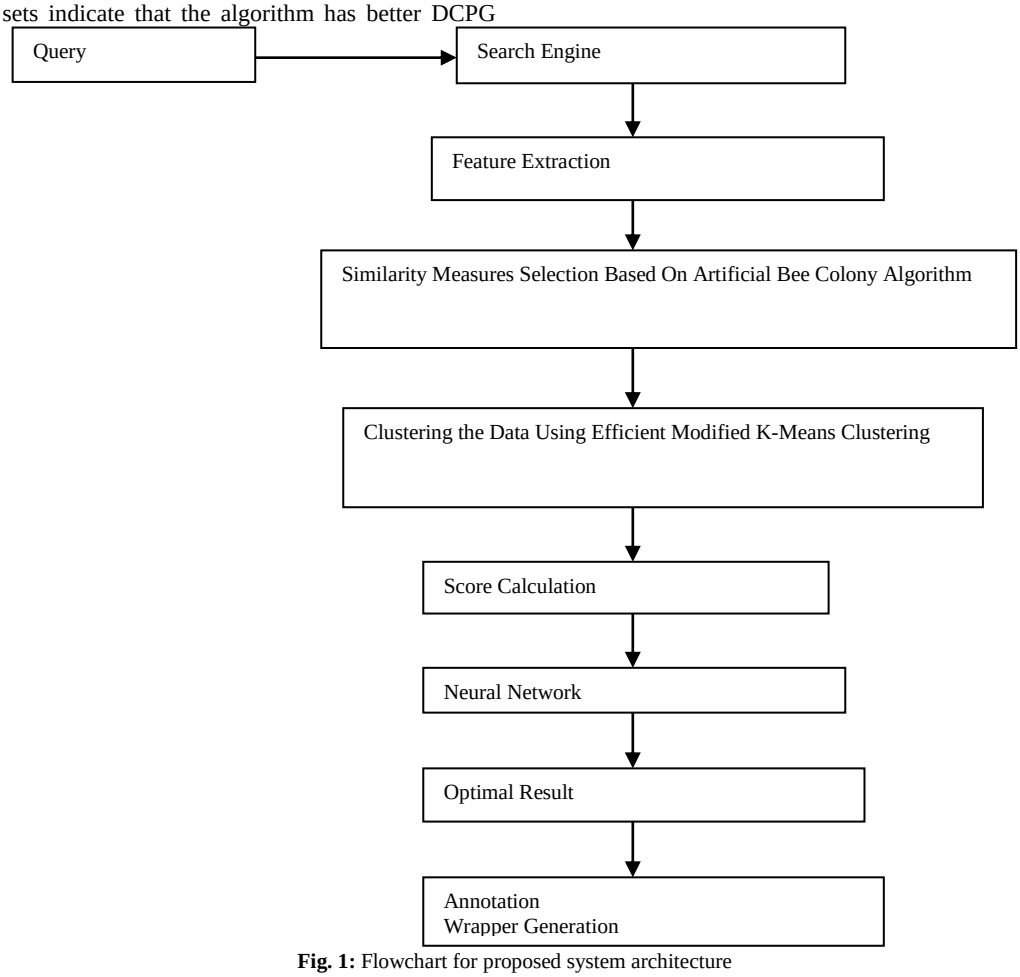

Fig. 1: Flowchart for proposed system architecture

Data Extraction

A web search engine is a software system that is designed to search for information on the World Wide Web. The search results are generally presented in a line of results often referred to as search results record (SRR). Web database has multiple search result record. Each SRR refer to an entity. SRR from web database contain multiple data units (or instances). Each SRR refer to an entity. Data units are different from text node.

Text node is surrounded by pair of HTML tags. Data units are texts that semantically represent the single concept of an entity. These data units are not used for application such as deep web data collection and internet comparison shopping. Here the annotation is done on the basis of data units. The data units are annotated by assigning meaningful labels to them.

Annotation problem has become significant problem due to the rapid growth of the deep web and the need to query multiple web mining, it is imperative that is data units are correctly labeled so they can be appropriately organized and stored for subsequent machine processing. Note that the search sites that have web service interfaces, it may be easier to annotate their SRRs because the semantic meanings of their data units more clearly describe in WSDL. However that very few search sites have web services interfaces. Therefore it is still necessary to extract and annotate data from legacy HTML pages.

In this system we first extract the SRR page from the given web database. Then the data units are identified and aligned such that the aligned data units are belong to the same attributes or concepts. We then design different basic annotator to annotate data units of each aligned group. These different basic annotator results are combined to determine appropriate label for each data unit groups. Finally the annotator wrapper is generated for the corresponding WDBs which are used to annotate new SRRs retrieved for different queries. Result page from web database has multiple records (SRR). Each SRR contain multiple data units each of which describes one aspects of real world entity.

An example for search result is shown below with both original HTML page and the HTML source.

Fig. 2: Search result from Amazon.com

Artificial Bee Colony Algorithm

Artificial Bee colony algorithm is used for similarity measures of data has to be selected from the data modeling. The most basic ABC algorithm consists of three phases. They are initialization, employed, onlooker and scout bees phase. Each phase is replayed until the maximum count of iterations is reached. In the initial phase, the count of solutions and the control parameters are fixed. The employed bees phase deals with the search of new high quality food sources in the nearby locality of old food source. The new food source is then evaluated for its fitness, which is then followed by the comparison of the old and the new food source by means of greedy selection. The so collected knowledge about the food source is distributed among the onlooker bees present in the beehive.

In the next phase, the onlooker bees follow a probabilistic approach to select the food sources with respect to the information provided by the employed bees. This is followed by the calculation of the fitness function of the food source, which is located nearby the selected food source. Finally, the old and the new food sources are compared by the greedy selection.

In the final phase, the employed bees turn to scout bees, when their solutions cannot be enhanced within a predefined count of iterations. The solutions so found by the bees are dropped out. At this point, the scout bees search for new food source again. By this functionality, the poor solutions are dropped out. These three phases continue its process until the stopping point is reached [11, 12].

Efficient Modified K-Means Clustering

There have been a lot of research works done in the past to improve the K-Mean clustering technique. All they wanted to improve the clustering result and fix the limitations of previously proposed method. We get influenced by those thesis and research works and decided to make some modification that can more efficient and faster. We worked on the problem to find initial cluster(R). Also we worked to find initial centroid. In last part of our algorithm, we try to minimize calculation by finding the feasible points of working. When we combined all of these points, we found an effective modified k-means algorithm

In the first part, we found the number of cluster. Then assign data points to initial cluster. Then we go for find an effective and useful initial centroid. In these two parts; there have some loops and its calculation but these make a better clustering rather than other clustering methods. For make algorithm more intelligent, some procedure must add. Intelligence of algorithm will help to determine the number of cluster and which points are the initial centroids.

In last portion of modified algorithm, we worked only those feasible data points which have chance to change current cluster and move to new cluster. We also make a short list of points for calculate. It minimizes calculation, that's why it was capable to save time on behalf of original standard k-means algorithm.

In step I, we calculate the equation and get a concept about the number of cluster. We take a concept about the cluster number, but this calculation is not the final decision.

From step II, we assign a value in x for maintain cluster number.

By step III, IV and V, algorithm assigns a new cluster and assigns data points to this cluster.

From step VI, algorithm finalizes the initial cluster's member data points, and gets decision to start a new cluster.

Step VII; find the initial centroids for clusters.

By help of step VII, step VIII finalized a stable cluster and centroid.

In step IX, X, XI and XII, there have tricks to find feasible data points those have chance to change current cluster. We worked only those interval's points. Normally other points don't move clusters. For this step, algorithm saves a lot of time. It minimizes a lot of calculation.

To get decision, step XIII is used. In this step; algorithm take decision about the algorithm continue or all clustering is finished.

ABC colony algorithm is an efficient population based optimization algorithm and it imitates the behavior of real bees. The k-means algorithm is efficient and fast, however the problem is on finding initial cluster point. This work proposes to locate the initial cluster point with the help of bees and these clusters are refined by the k-means algorithm.

Annotation Wrapper Generation

In this phase, uses six basic annotators; such as table annotator(TA), query-based annotator(QA), schema value annotator(SA),frequency-based annotator (FA), in-text prefix/suffix annotator(IA), common knowledge annotator(CA) are used to label the data unit group. In table annotator the aligned data units are arranged in the table format and the column name is used to label the group. In schema value annotator uses the schema value such as publisher author and title for labeling.

In frequency based annotator frequently available data units is used to label. In text prefix suffix annotator the prefix or the suffix of the data units is used for labeling. Common knowledge annotator uses the basic knowledge for labeling. Each annotator can independently assign labels to data units based on certain features of the data units. Moreover different annotator may produce different label for the obtain group of data unit. Hence to select more suitable label for the group a probabilistic model is applied to combine the results from different annotators into a single label. It is highly flexible so that the existing basic annotators may be modified and new annotators may be added easily without affecting the operation of other annotators.

Annotation wrapper is a description of the annotation rules for all the attributes in the result page. After the annotation is completed wrapper is generated automatically for the annotated result group. The wrapper can be applied to efficiently annotating the SRRs retrieved from the same WDB with new queries.

Also the frequently used websites are identified from the annotated group. So that the mostly used websites are known and they can be displayed first in the result page. This makes the process more efficient. Also it avoids unnecessary scan of the database. The architecture diagram of the system is shown in the figure in which the search result is obtained from the web database. Before annotating the result it check the wrapper whether it is annotated earlier or not. If not annotation process takes place and then wrapper is generated and the frequent available websites are identified and then the result is displayed.

4. Performance Analysis

The proposed system performance is evaluated on the basis of two factors that is precision and Recall. The precision and recall is calculated for performance of alignment and performance of annotation. The precision for performance of alignment is as follows.

$$\text{Precision}=\frac{\textit{correctly aligned data units}}{\textit{aligned data units}} * 100$$

$$\text{Recall}=\frac{\textit{data units that are correctly aligned}}{\textit{manually aligned data units}} * 100$$

Table 1 represent the performance calculation for alignment in which the average value for precision and recall is about 98%. And for each domain it is more than 96%.

Table 1: Performance of Alignment

Domain	Alignment For Precision	Alignment For Recall
Book	98.4	97.3
Game	98.7	98.0
Music	99.0	99.1
Average	98.7	98.1

The performance of each alignment features as mentioned in alignment phase is given below in which over all alignment give the best result than the individual one. Here the tag path gives accurate result next to overall result. That means while calculating individually tag path give more accurate result than other features.

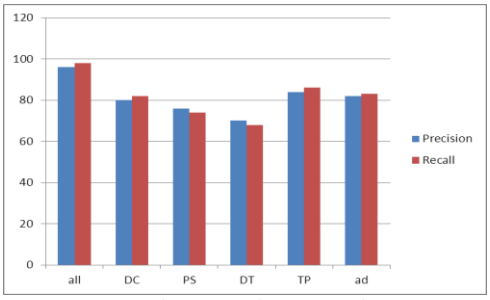

Fig. 3: Performance of alignment features

The basic formula used to calculate precision and recall for annotation is as follows

$$\text{Precision}=\frac{\textit{correctly annotated data units}}{\textit{data units annotated}} * 100$$

$$\text{Recall}=\frac{\textit{data units correctly annotated}}{\textit{manually annotated data units}} * 100$$

The table 2 shows the Performance of annotation face in which the average precision and recall is nearly 97%. And for each domain it results more than 95%.

Table 2: Performance of Annotation

Domain	Alignment For Precision	Alignment For Recall
Book	97.4	96.3
Game	97.7	97.0
Music	97.0	97.7
Average	97.3	97.0

The performance of the basic annotator are compared and shown in the fig 4. The evaluation shows the combination of all Annotators give the most accurate result than finding each one individually. Comparing others table annotator gives nearly an accurate result.

Fig. 4: Performance of basic annotator

5. Conclusion

This paper proposes a web data clustering algorithm namely An Efficient Modified K- Means and Artificial Bee Colony algorithm. The main purpose of clustering is to group relevant data together, such that the degree of relevance between the data in a cluster is more and the data between two clusters show lesser degree of relevance. ABC algorithm is used as the global search optimizer and k-means is employed as the local solution optimizer. This work employs four different datasets for checking the performance of An Efficient Modified K- Means and Artificial Bee Colony algorithm. The experimental results of An Efficient Modified K- Means and Artificial Bee Colony algorithm are satisfactory, when it is compared with the several algorithms. Using wrapper the annotation become efficient for even a new queries. Here we also use the frequent item set retrieval to know the result set which is more in annotator group. It is also used to list down the trusted sites in the data base.

References

[1] Das S & Konar A, "Automatic image pixel clustering with an improved differential evolution", *Applied Soft Computing*, Vol.9, (2009), pp.226–236.

[2] Das S, Abraham A & Konar A, "Automatic clustering using an improved differential evolution algorithm", *IEEE Transaction on Systems, Man, and Cybernetics–Part A: Systems and Humans*, Vol.38, (2008), pp.218–237.
[3] Das S, Abraham A & Konar A, "Automatic clustering with a multi-elitist particle swarm optimization algorithm", *Pattern Recognition Letters*, Vol.29, (2008), pp.688–699.
[4] Han J & Kamber M, *Data Mining: Concepts and Techniques, second ed.*, Morgan Kaufman, San Francisco, (2006).
[5] Baeza-Yates R & Ribeiro-Neto R, *Modern Information Retrieval, Addison Wesley*, ACM Press, New York, (1999).
[6] Hammouda KM & Kamel MS, "Efficient phrase-based document indexing for web document clustering", *IEEE Transactions on Knowledge and Data Engineering*, Vol.16, (2004), pp.1279–1296.
[7] Kalashnikov DV, Chen ZS, Mehrotra S & Nuray-Turan R, "Web people search via connection analysis", *IEEE Transactions on Knowledge and Data Engineering*, Vol.20 (2008), pp.1550–1565.
[8] Aggrawal CC & Reddy CK, *Data Clustering Algorithms and Applications*, CRC Press, (2014).
[9] MacQueen J, "Some methods for classification and analysis of multivariate observations", *Proc. 5th Berkeley Symp. Math. Stat. Probability*, (1967).
[10] Karaboga D, "An idea based on honey bee swarm for numerical optimization", *Technical Report-TR06, Erciyes University, Engineering Faculty, Computer Engineering Department*, (2005).
[11] Karaboga D, Gorkemli B, Ozturk C & Karaboga N, "A comprehensive survey: artificial bee colony (ABC) algorithm and applications", *Artificial Intelligence Review*, Vol.42, No.1,(2014), pp.21-57.
[12] Karaboga D & Ozturk C, "A novel clustering approach: Artificial Bee Colony (ABC) algorithm", *Applied soft computing*, Vol.11, No.1,(2011), pp.652-657.
[13] Lee S & Lee W, "Evaluation of time complexity based on max average distance for K-means clustering", *Int. J. Security Appl.*, (2012), pp.449–454.
[14] Manning C, Raghavan P & Schütze H, *Introduction to Information Retrieval*, Cambridge University Press, Cambridge, England, (2008).
[15] Reddy D & Jana PK, "Initialization for K-means clustering using voronoi diagram", *Procedia Technol.*, Vol.4, (2012), pp.395–400.
[16] Wu X, Kumar V, Ross Quinlan J, Ghosh J, Yang Q, Motoda H, McLachlan G, Ng A, Liu B, Yu P, Zhou ZH, Steinbach M, Hand D & Steinberg D, "Top 10 algorithms in data mining", *Knowl. Inf. Syst*., Vol.14, (2008), pp.1–37.
[17] He H, Meng W, Yu C & Wu Z, "Automatic Extraction of Dynamic Record Sections From Search Engine Result Pages", *VLDB J.*, Vol.13, No.3, (2012), pp.256-273.
[18] Liu W, Meng X & Meng W, "Vide: A Vision-Based Approach for Deep Web Data Extraction", *IEEE Trans. Knowledge and Data Eng*., Vol.22, No.3,(2010), pp.447-460.