American Journal of Intelligent Systems. 2011; 1(1): 10-15
DOI: 10.5923/j.ajis.20110101.02

Intelligent Extended Clustering Genetic Algorithm for Information Retrieval Using BPEL

N. El-Bathy[1,*], C. Gloster[1], I. Kateeb[1], G. Stein[2]

[1]Department of Electronics, Computer, and Information Technology, North Carolina A&T State University, Greensboro, NC, USA
[2]Department of Computer Science, Lawrence Technological University, Southfield, MI, USA

Abstract In this paper, the problem of clustering intelligent web using K-means algorithm has been analyzed and the need for a new data clustering algorithm such as Genetic Algorithm (GA) is justified. We propose an Intelligent Extended Clustering Genetic Algorithm (IECGA) using Business Process Execution Language (BPEL) to be an optimal solution for data clustering. It improves the efficiency and performance for retrieving a proper information results that satisfy user's needs. The proposed IECGA uses several mutation operators simultaneously to produce next generation. This series of random mutation process depend on chromosome best fitness in the population and rely on high relevancy as well. The mutation operation will guarantee the success of IECGA for data clustering since it expands the search. So the highly effective mutation operators the greater effects on the genetic process. Finally, IECGA for data clustering gives the user needed documents based on similarity between query matching and relevant document mechanism. The results obtained from the web intelligent search engine are optimal.

Keywords BPEL, Clustering Genetic Algorithm, K-Means, Intelligent Agent, Information Retrieval

1. Introduction

The search for information relies on the ability to programmatically adapt over time to find new methodologies necessary to break data into meaningful clusters[1,8], and[9]. With data constantly changing, it is desired to develop an algorithm capable of clustering in a way that is relevant to the data that is being clustered[10]. In order to tackle this problem, the algorithm must have the ability to try numerous ways of clustering a particular data set. The algorithm must evolve the sets of best fit for a predetermined number of times until an optimal clustering of data is achieved[2].

In an attempt to allow for this capability, the use of an extended genetic algorithm has been proposed. It provides a way of clustering data that is relevant to the type of data being clustered, with the ability to adapt over time to changes in subjects of topics of desired data. By developing the proposed algorithm, data can evolve into information in a way that produces robust flexibility[10].

The remaining material of this paper is structured as follow: In Section 2, the clustering process is introduced. In Section 3, Intelligent Extended Clustering Genetic Algorithm (IECGA) is structured. In Section 4, the results are provided. Finally, the conclusion and future work are given in Section 5 and section 6.

* Corresponding author:
nielbath@ncat.edu (N. El-Bathy)
Published online at http://journal.sapub.org/ajis

2. Clustering Process

Intelligent clustering and learning establishes groups of users exhibiting similar browsing patterns and provides useful knowledge[2]. While K-Means clustering method is useful and efficient, it lacks the ability to intelligently evolve over time to user browsing patterns and collected data applications. While other algorithms have made numerous advancements, there is still a lack of the ability to evolve[3].

2.1. Extended Genetic Algorithm Purpose

Nowadays, various applications of genetic algorithms are in the early stages of being used to actively and efficiently cluster data based on relevancy. Such applications aim to produce information that has adapted over time based on user requests.

The purpose of our clustering algorithm is to divide set of N documents into K clusters, where the sum of distances D between clusters' documents is the least possible. This means that when clustering algorithm has been completed, the set will be divided into K proper subsets with no documents in more than one such subset of the documents. Each subset has the closest grouping of documents possible with K clusters.

Figure 1 illustrates the clustering process. The x and y axis represent word weights, and the "documents" being clustered only have 2 words. They can be displayed as a 2 dimensional such that it would be put into 3 clusters by clustering algorithm.

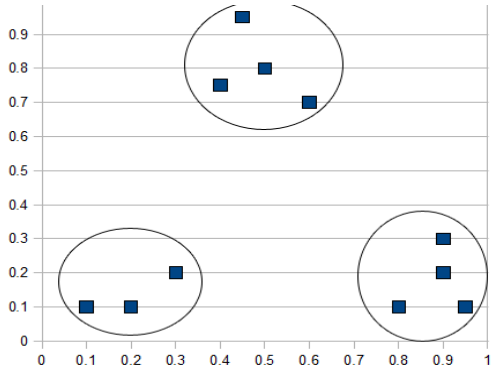

Figure 1. Clustering Process.

2.2. Extended Genetic Algorithm Mechanism

In our clustering algorithm, each document is stored both as a set of weights and a set of words that the weights correspond to. The set of weights is the ratio of each word's occurrences to the sum of all words in the document's occurrences. To simplify some of the computations involved, each document's set of words contains every word that appears in any of the other documents, but with a weight of zero if it does not actually occur within that document. Euclidean distance is utilized in computing the similarity to quantify the distance between the documents in each cluster. The average of the distances between all documents in each cluster to each other, as if they were points in an n-dimensional space is used as our "quality" for each cluster. In an n-dimensional space, n is the number of words in each document.

The following math is used to find D, the average distance between the documents in the i^{th} cluster of set C of clusters.

$$P = C_i \times C_i \quad (1)$$

$$d(A,B) = \sqrt{(A_1 - B_1)^2 + (A_2 - B_2)^2 + \ldots + (A_{|A|} - B_{|B|})^2} \quad (2)$$

$$D = (\frac{1}{|P|}) \sum_{i=1}^{|P|} d(P_{i_0}, p_{i_2}) \quad (3)$$

The variable P is used to hold the Cartesian product of the set of documents in the cluster with itself, creating a set of pairs of documents. Each pair in P contains two documents from within the cluster, and to find the average distance between any two documents in the cluster, each pair's distance will need to be computed. The function d is the Euclidean distance between two sets. D, the average distance between the documents in C_i, is calculated by finding the sum of all distances of P's elements and finding the quotient of that and the cardinality of P.

3. The Intelligent Extended Genetic Algorithm (IECGA)

The structure of genetic algorithm is extended to hold multiple populations in the population space [12]. The Intelligent Extended Clustering Genetic Algorithm (IECGA) is designed using artificial intelligence methodologies, not geometric approaches, to the clustering problem[4] and[10].

Our proposed method uses a genetic algorithm to find an ideal clustering solution instead of a more mathematical method such as the k-means algorithm. This key difference allows for more adaptive behavior within our clustering method. Also, web services can induce very large amounts of data.

As it is important to manipulate data accurately and efficiently, Business Process Execution Language (BPEL) approach has been proposed. It implements dynamic service capabilities with genetic algorithms to apply reasoning and flexible service workflows[8,9], and[10].

This paper builds a utility-based intelligent agent that implements a faster genetic algorithm with greater efficiency than the original algorithm. The genetic algorithm support a flexible service composition mechanism while having the ability to improve efficiency over time, all while reusing previously tested efficiency. While systems can be made bigger, modern paradigm breakthroughs are evolving to make systems smarter. The orchestrations of genetic algorithms provide flexible service workflows that can quickly adapt to changes. BPEL Composes, or orchestrates, the services into business flows.

Web service is a technology that enables programs to communicate through Hypertext Transfer Protocol (HTTP) on the Internet[6]. Service standards are effective platforms for publishing services. These standards are Web Services Description Language (WSDL), Extensible Markup Language (XML), and Simple Object Access Protocol (SOAP). WSDL provides a model for describing services. XML adds an intelligent level to distribute information on the Internet. SOAP exchanges structured information in the implementation of the service[7].

The orchestration of web services is supported by Business Process Execution Language (BPEL)[5]. In the study solution of this research, the process is simply designed, deployed, monitored, and administered within a framework provided by Oracle BPEL Process Manager. BPEL enables linking two or more services as one piece of a process[5].

The study solution of this research orchestrates the web services collectively within a BPEL process.

Chromosomes are encoded to represent a genetic algorithm and to be parsed into tree structures, which prevents syntax crossovers and allows for mutation stages. Once proper genetic algorithms are put in place, the desired service item from the web part can be requested. Upon this initial request, the first generation of information retrieval is randomly generated, which can lead to a slight decrease of efficiency. What makes up for this initial sacrifice in performance is that as the workflow processes information, the algorithm creates a new generation of logic and the results are assessed based on goodness of fit to results. As new logic workflows are developed, they can be selected and mutated to produce better results. As this process continues, eventually the service matchmaking with user requirements

can be provided in such a way to enable increased efficiencies over time. Upon delivery of the user request, the generation cycle is terminated.

3.1. Data Preparation

In order to cluster a set of documents, the documents get assigned a number. The numbers are grouped together creating the clusters. Once the clusters are created, relevancy of the corresponding document to the other document can be determined, and an overall cluster score can be calculated. A set of documents is obtained; each unique document is given a number that will act as a document abstraction through clustering process.

3.2. The Algorithm

The proposed algorithm does not have a normalization step as it does not use centroids to define the clusters like the traditional clustering method. Figure 2 describes Intelligent Extended Clustering Genetic Algorithm (IECGA).

1. Create initial random population P of N individuals
2. i ← 0
3. If i is equal to the number of desired generations, return the best individual of the most recent generatic
4. P_{i+1} ← empty set
5. B ← the most fit individual of the previous generatic
6. Add B to P_{i+1}
7. Insert into P_{i+1} mutate(B)
8. Repeat 7 until P_{i+1} has N individuals
9. Evaluate the fitness for all individuals from P
10. i ← i + 1
11. Goto 4.

Figure 2. IECGA Algorithm.

3.3. Fitness

The fitness of an individual is computed based on the "distances" between the words or other tracked items appearing within a document. The items are compared by their weights, meaning the ratio of their appearances to the total sum of words in the document. These weights are then treated as if they were coordinated for the document's point on an n-dimensional grid, where n is the number of different words appearing within the set of documents being clustered by IECGA algorithm.

In IECGA algorithm, an individual with a lower fitness value actually represents a solution of greater quality than one with a greater fitness value. This is because the quality of the clustering solution is the closeness of the items being clustered. Only the most individual fit is passed on to the next generation. The fitness for a chromosome is found through repetition of the math used for finding the similarity of the documents in a cluster. For each chromosome in the generation, the fitness is computed by finding the average of the similarities for each cluster. By using this method, the fitness is also the average distance between any two documents in any one cluster in the solution.

3.4. Mutation

Mutation is a way that changes the population to produce the best solution[11][13]. The IECGA clustering process involves a series of mutations that will evolve over time taking only the mutations with a high relevancy, and mutating those further. The IECGA algorithm used one type of mutation. This type is known as a one-point mutation. Either a single document's position is moved through the chromosome, switching its place in the clusters with another document, or the point at which a cluster is organized is moved. Through the repeated use of these two types of mutations, the solution can create a generation consisting of a multitude of clustering possibilities.

To further increase the genetic diversity present in each generation of the IECGA, the algorithm includes a step where a new individual is added to the population. This individual is randomly generated with each generation iterated, to create additional diversity, even without the crossover step's inclusion in the algorithm.

3.5. Crossover

The proposed algorithm removed crossover step although it is a key part of numerous genetic algorithms. The reason is that crossover decreases the efficiency of our algorithm. It would build new chromosomes out of sections from two different chromosomes, creating new generations with greater diversity. The lesser number of generations required comes with a cost in the form of a drop in efficiency.

Currently, our genetic algorithm stores each chromosome as a sequence of characters representing the documents. The order of the characters in our chromosomes is of great importance and no repeats are allowed. Traditional genetic algorithms use a series of bits which represent in turn a series of operations and values. Using crossovers in the source code of our genetic algorithm negatively affects the efficiency of the algorithm more than it would lower the amount of generations required.

With just our current generation loop utilizing only varying degrees of mutations, we are likely creating the same chromosomes which would result from crossovers. The proposed genetic algorithm is simply a way to go through a vast number of possible solutions with greater speed and efficiency than other strategies. With or without crossovers, our genetic algorithm should arrive at the same value.

4. Experimental Results

The IECGA algorithm is tested on set of sample data. The data is based on 50 generations/iterations of the IECGA or

K-means respectively, using the same random sample set of 15 documents with 600 words each. Figure 3 serves as decent evidence that the solutions from our Extended Genetic Algorithm are generally closer clustered than those generated by k-means, even if k-means can find a solution faster. Figure 3 defines GA 1 and GA 2 as the two graphed trials of the genetic algorithm.

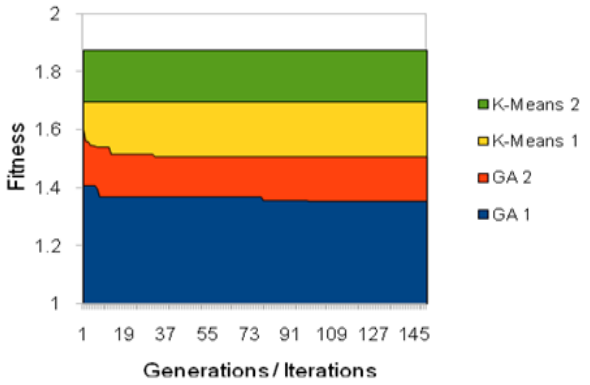

Figure 3. IECGA and K-Means Comparison.

Figure 4 presents sample set of 15 documents as a demonstration of clustering. The document set has been simplified to only have 2 different words in each document. The values on the X and Y axes are the word weights of those two words in the documents. Figure 4a shows the documents arranged on 2-dimensional grid without any clustering information applied. Figure 4b and Figure 4c differ in that the documents have been colored and circled to designate the different clusters within the set of documents. Figure 4b has been clustered using the k-means algorithm, while with Figure 4c our genetic algorithm is used to find a clustering solution.

The results are listed in Table 1 were collected over 15 test runs of both clustering methods on the same data set. The table shows the statistics collected from IECGA and K-Means algorithms to demonstrate their relative performance capabilities. The values given are the fitness of the final clustering solution generated by each run, which means that the lower fitness are from better solutions, while higher fitness values are worse solutions. As each method uses a random starting point, there is room for variation in solutions.

Table 1. IECGA and K-Means Performance

	IECGA	K-Means
Maximum	1.66384	1.86476
Average	1.56938	1.67881
Minimum	1.35574	1.40269

From this data, we can observe that on average, our IECGA algorithm excels k-means clustering algorithm. The test runs did not find as good a solution with k-means as the best solution from the IECGA algorithm, and even the worst solution from the IECGA algorithm is of better fitness than the average solution from k-means.

While the data collected does not represent all possible input cases, and cannot claim to represent all of them, it shows a trend of the IECGA algorithm exceeding the performance shown the clustering process we had used previously.

5. Conclusions

The main objective of this paper is to address the applicability of potential Intelligent Extended Clustering Genetic Algorithm (IECGA) to solve the efficiency and limitation problems in data clustering. The proposed algorithm solution has markedly increased the success of document clustering and relevancy between query-matching and relevant documents as shown. The Intelligent Extended Clustering Genetic Algorithm (IECGA) had a marginally high efficiency and performance than simple K-Means due to the concurrent mutation processes with a high relevancy. The experimental results showed that the proposed algorithm outperforms K-Means algorithm.

The Intelligent Data Clustering is a challenging research problem that arises in many applications. This Extended Genetic Algorithm can be used for many different applications requiring data mining, information retrieval, computational biology, text categorization and image annotation.

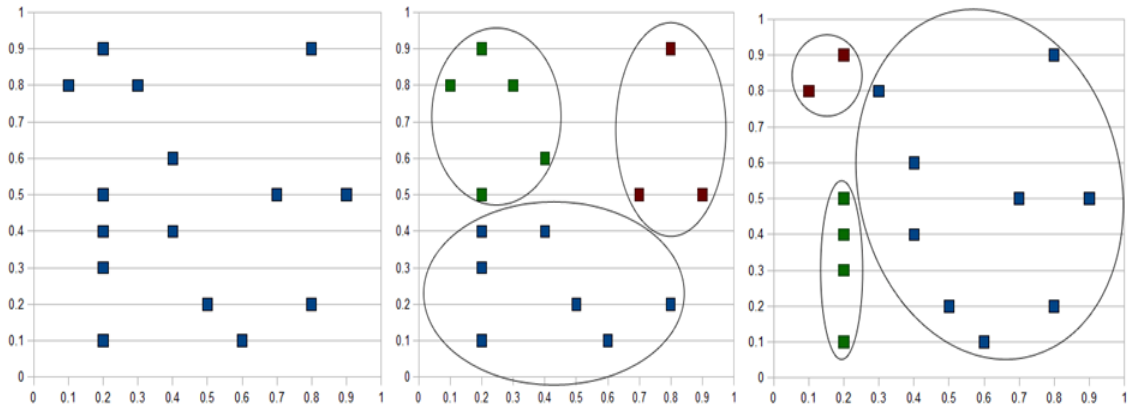

Figure 4. a) Documents without clustering (left), b) K-means Clustering Results (middle), c) IECGA Results (right).

The Intelligent Extended Clustering Genetic Algorithm (IECGA) enhances an organization's ability to collect information faster at lower cost and to make accurate decisions. Implementing IECGA provides acceptable benefits in terms of agility and integrity.

The orchestrations of genetic algorithms by implementing Business Process Execution Language (BPEL) allow flexible service workflows to be immediately adjusted to modifications and make systems smarter.

6. Future Work

Our future work will concentrate on the implementation of the IECGA to large data sets, generalization of the proposed approach to general graph structures, and investigation of the possibility of integrating multiple sources of data for improving the clustering quality.

A comparison between Intelligent Extended Clustering Genetic Algorithm (IECGA) and DNA algorithms is an essential future work. The goal is measuring the quality and the efficiency of IECGA and DNA algorithms in retrieving information and clustering data. DNA algorithms have been solved expensive problems in different areas including machine learning, finite automata, and relational data modeling.

An adapted DNA algorithm will be our aim to improve the processes of the intelligent information retrieval. Such adapted DNA algorithm will be part of search lifecycle architecture [14][15].

The new search lifecycle architecture can be designed, developed, and tested within Service-Oriented Architecture (SOA) software application. It can be created applying services composition methodology.

Our adapted DNA algorithm and new intelligent architecture can help modernize medical and health care industries. These industries experience numerous challenges. The modernization requires an innovative solution to determine diagnosis of diseases and the best treatment. This solution discovers related diseases to doctors' original diagnosis and quickly reassesses the situation if their diagnosis is incorrect. It also should eliminate unnecessary treatments and testing. It shortens time spent in hospitals for patients. The future work will focus on creating such a solution in the form of an intelligent information retrieval lifecycle architecture based adapted DNA algorithm using Service-Oriented Architecture (SOA).

Dedication

The primary author of this paper, Dr. Naser El-Bathy, has dedicated this research to the Egyptian writer, Ibrahim El-Bathy, who passed away in 1979. His methods in writing and dedication to the publishing and newspaper industries are what drove him to choose this concentration for this research. His interest in research and analysis sparked an interest and embedded the idea for this research in Dr. Naser El-Bathy. Ibrahim El-Bathy's achievements in the field of medicine, publishing and newspaper industries are reported in several Arabic newspapers and magazines during his life and after his death.

REFERENCES

[1] R. Akerkar and P. Lingras, Building an intelligent web – theory and practice. Sudbury, Massachusetts: Jones and Bartlett Publishers, 2008

[2] W. Li, X. Zhang, and X. Wei, "Semantic web-oriented intelligent information retrieval system," IEEE BMEI Proceedings of the International Conference on BioMedical Engineering and Informatics, Washington, DC, Vol. 01, 2008

[3] R. H. Sheikh, M. M. Raghuwanshi, and A. N. Jaiswal, "Genetic algorithm basedcClustering: a survey," First International Conference on Emerging Trends in Engineering and Technology, IEEE, Nagpur, Maharashtra, pp. 314 – 317, 2008

[4] B. Coppin, Artificial intelligence illuminated. Sudbury, Massachusetts: John and Bartlett Publishers, 2004

[5] M. P. Papazoglou and W. Heuvel, "Service oriented architectures: approaches, technologies and research issues," The VLDB Journal, Springer Berlin / Heidelberg, vol. 16, Number 3, pp. 389–415, 2007

[6] R. H. Abrahiem, "A new generation of middleware solution for a near-real-time data warehousing architecture," Electro/Information Technology IEEE International Conference, Chicago, IL, United States, pp. 192–197, May 2007

[7] R. G. Reynolds and J. M. Stefan, "Web services, web searches, and cultural algorithms," Systems, Man and Cybernetics, IEEE International Conference, vol.4, pp. 3982 – 3987, 2003

[8] N. El-Bathy, P. Chang, G. Azar, and R. Abrahiem , "An intelligent search of lifecycle architecture for modern publishing and newspaper industries using SOA," IEEE, Electro/Information Technology, Normal, IL, United States, pp. 1-7, 2010

[9] N. El-Bathy and G. Azar, "Intelligent information retrieval and web mining architecture applying service-oriented architecture," KG. Saarbrücken, Germany: LAP LAMBERT Academic Publishing AG & Co. 2010

[10] N. El-Bathy, G. Azar, M. El-Bathy, and G. Stein, "Intelligent lifecycle architecture of disease diagnosis based CGA using SOA," KG. Saarbrücken, Germany: LAP LAMBERT Academic Publishing AG & Co. 2011

[11] H. P. Pfeifer, "An exhaustive analysis of recombination and mutation variances for genetic algorithms," Protocol Labs, Munich, 2010

[12] U. Aickelin and K. Dowsland, " Exploiting problem structure in a genetic algorithm approach to a nurse rostering problem," Journal of Scheduling, UK, May 2008.

[13] F. Petzzella, G. Morganti, and G. Ciaschetti,"A genetic

algorithm for the flexible job-shop scheduling problem," Computers and Operations Research, Volume 35, Issue 10, pp. 3202-3212, October 2008.

[14] N. El-Bathy, C. Gloster, and I.Kateeb, " Intelligent clustering extended DNA algorithm using service-oriented architecture," Unpublished.

[15] N. El-Bathy, C. Gloster, and I.Kateeb, " Intelligent lifecycle architecture of information retrieval based DNA algorithm using service-oriented architecture," Unpublished.