Journal of Theoretical and Applied Information Technology
15th June 2022. Vol.100. No 11

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

MACHINE LEARNING-BASED OPTIMAL SEGMENTATION SYSTEM FOR WEB DATA USING GENETIC APPROACH

N. SILPA[1], Dr. V. V. R. MAHESWARA RAO[2]

[1]Research Scholar, Department of Computer Science and Engineering,
Centurion University of Technology and Management, Odissa, India

[2]Professor, Department of Computer Science and Engineering,
Shri Vishnu Engineering College for Women, Bhimavaram, India

E-mail: [1]nrusimhadri.silpa@gmail.com, [2]mahesh_vvr@yahoo.com

ABSTRACT

The rapid emergence of computer technology has led to the storage of vast amounts of information in databases. The increasing popularity of electronic data has also created vast amounts of unlabeled information. The potential of extracting valuable knowledge from such digital data has created the basis for the researchers towards important research areas such as Web engineering, Data Science, Big Data Analytics etc. Web engineering is a process utilized for exploring patterns in large databases. However, finding intrinsic structures in large amounts of data becomes a distinctive challenge to organize them into meaningful groups. Many of the existing clustering algorithms are not appropriately suitable for all kinds of web applications.

This prompted the present researchers to develop a machine algorithm that is more applicable and robust in real-time to get technological intelligence especially from web data sources. The authors propose an Optimal Segmentation System using a Machine Learning approach (MLOSS) with twin objectives. Initially, MLOSS performs the pre-processing step on the unstructured and semi-structured web documents to prepare efficient data representation structure for applying either supervised and unsupervised techniques. Later as a part of second objective, the proposed system emphasizes on the segment of the preprocessed web data using clustering techniques with an hybridization of the Genetic Approach, that mimic the biological evaluation process having the self-learning proficiency. To validate the performance results of the proposed framework in numerous orders of magnitude, much experimentation is done and, the results have proven this as claimed.

Keywords: *Machine Learning, Web Mining, Genetic Algorithms, Clustering, Un-Supervised Techniques*

1. INTRODUCTION

The increasing number of data sources and the ease of accessing them on www have created new opportunities for academic and social activities. This is evidenced by the emergence of various web mining and analytical techniques. The practice of extracting previously unknown patterns from massive amounts of web data is known as knowledge retrieval from web data as presented in Figure 1.

The web data is usually collected in informal settings like weblogs, emails, web pages, chat rooms etc. To extract the most out of the web documents, it demands to store the web data in a structured data structure. Although there are many ways to do this, most of them are focused on extracting the text structure and its idiosyncrasies. Most web mining approaches rely on the idea that a given web document can be represented by a series of words. To be able to determine the importance of a given word, a vector representation is typically used, with the following phases.

- ✓ Tokenize the web file using space as the delimiter
- ✓ Clear the words that do not have any specific meaning
- ✓ Use the porter stemmer algorithm to get the common root word for the words with the same root word.
- ✓ Construct a data structure that enables very efficient analysis of huge web document collections without using any explicit semantic information.

Journal of Theoretical and Applied Information Technology
15th June 2022. Vol.100. No 11

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

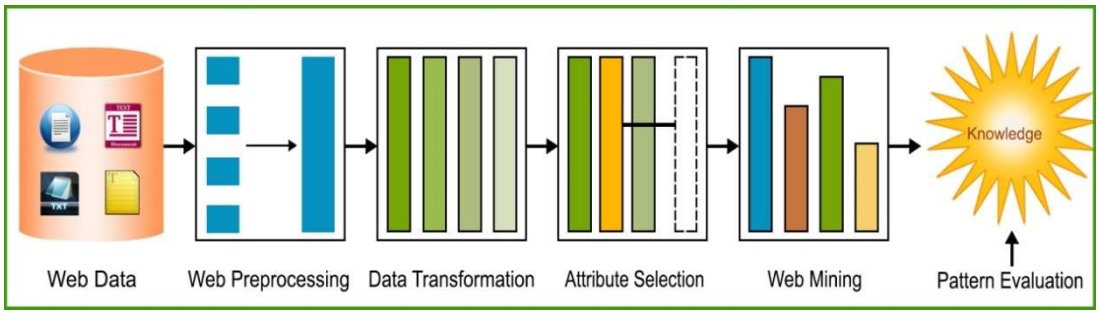

Figure 1: Process of Web Mining

Web data transformation is a process that generates features from web documents. Unlike open-ended data mining, web mining is not structured so it requires an organized procedure to perform this task.

Attribute selection algorithms are ad hoc methods that are used to identify the features that are important to a given cardinality. If the features are not available, the algorithms may be impractical for unsupervised learning.

Web data engineering is a procedure that identifies the intrinsic findings within and across a large amount of web data. It can be performed in various forms such as knowledge mining from web data, fixed domain web mining, supervised web mining, unsupervised web mining etc.

Supervised web mining is used to classify web documents. Before the model can be used, pre-defined classes are used to classify the new web documents. The goal is to train the algorithm to automatically classify incoming news stories according to their topic. The training documents are labeled with a class. The model is then used to assign the correct class to the new web document.

In terms of real-time applications, unsupervised web mining is more advantageous than most supervised methods. The unsupervised method aims to find a set of documents that are both similar and dissimilar in terms of their contents. The result is typically a set of clusters. Usually, the quality of the cluster is better if the documents in the cluster are more similar than those in the opposite cluster.

Due to its unsupervised web mining nature, it is not sufficient for most web applications. It is therefore recommended that web science scientists use an optimized method such as genetic algorithms, for solving non-linear and complex domain problems.

In the present work, the authors propose a Machine Learning based Optimal Segmentation System using a genetic algorithm. The system uses a novel approach that works with the inherent characteristics of genetic algorithms.

The present research paper is prepared as follows. Section 1.2 provides the important associated work and inspired motivations of the proposed work. Section 1.3 presents the major research contributions of the authors. In 1.4, the investigation results along with analysis are showcased. Finally, section 1.5 gives a summary regarding the present paper and then pays a focus on future work.

2. LITERATURE REVIEW AND MOTIVATIONS

2.1 Literature Review

Over a two-decade period, the present authors conducted a detailed literature review to investigate the major contributions and prominent research paths in the field of web mining. The researchers [1, 2, 3, 4, 5, 6, 9, 10, 22, 23] have been studying web mining techniques and technologies to develop new solutions for real-world problems. This section provides a thorough review of the literature on various developments in web mining as well as clustering techniques, as shown in Figure 2, in order to provide context for the proposed work.

There have been many attempts [16, 17, 19, 33, 35] at information retrieval. In 1975, the first attempt at automatic indexation was made [36]. This work demonstrates that the concept of searching for information has gained increasing attention as a result of the rise of the WWW, which directs current research toward web mining.

The majority of the authors [20, 21, 24] acknowledge that the task of segmenting web documents is a prominent path of web mining and is extremely difficult in the rapidly growing online and digital era. As a result, the current work selected Web document categorization to segment the documents into meaningful parts to derive actionable intelligence for web scientists.

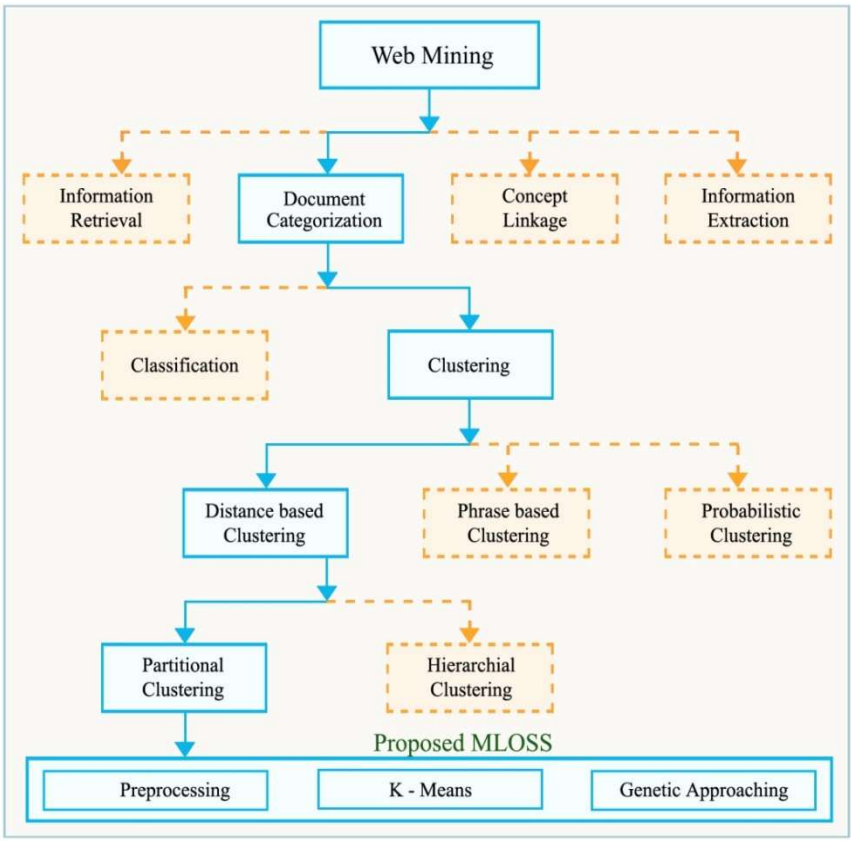

Figure 2: Related Work of Web Mining

Especially, web Document Categorization is an encouraging domain for web researchers. Numerous papers have been published in this field. The web document categorization method can be used in two different ways: first, it can be done by automatically grouping similar documents (Unsupervised Learning) [7, 11, 32, 34] or by assigning keywords to specific documents (Supervised Learning) [19, 29, 33]. Among these, web document categorization using unsupervised techniques is a current research area, and this is the motivation for the current research work.

Web Document Clustering is an unsupervised method that can be used to organize web documents into segments or objects. The traditional approach for web document clustering involves the recognition of the attributes of a data set. For both quantitative and categorical data, the problem has been studied [26, 28, 29, 35]. The authors choose Distance-based Clustering over Phrase-based Clustering and Probabilistic Clustering for web document clustering.

The Distance-based methods are intended for measuring the similarity of web documents between them. The finding of content similarity is a primary issue in web mining. Towards that, content similarity functions can be used in the combination of clustering algorithms such as hierarchical clustering, partitioned clustering etc. The proposed work concentrates on partitioned clustering to segment the web documents efficiently.

There have been various attempts at partitioned based K-Means algorithm by many authors [1, 7, 20] to overcome the problems and to address many complex issues. The research review [24, 27, 28, 31, 33, 53] is evident that most of the previous works have investigated the clustering problem in the light of choice of replacement centroid that has the highest SSE, detection of outliers, reducing the SSE with post-processing, incremental updating of centroid etc. With these findings, the current researchers continue their investigation into the benefits of partition-based clustering techniques.

After that, the present authors extended their literature survey to the prominent contributions in the area of web document pre-processing and the suitability of genetic to clustering techniques.

In [8, 13, 14, 15] researchers have paid attention to various preprocessing techniques to improve the efficiency of web document categorization. The main techniques that they used are stop word removal, stem-based preprocessing, and TF-IDF. Due to the complexity of natural language processing, the need for stemming techniques has been highlighted by very few authors [13, 14, 25].

The researchers [12, 18, 25, 30, 31] have explained the biological evaluation procedures related to computer evaluation in the light of genetic algorithms. Furthermore, the authors [3, 18, 27, 30] presented the web document clustering based on a genetic algorithm. Some more web researchers [1, 2, 18] have noticed that K-Means is a widespread technique for segmenting web documents, resulting in local optimization. To overcome that, very few researchers utilized the genetic approach, but require more iteration for finding improved clusters. They expressed that there is a necessity for a more proficient algorithm to give better clusters in less iteration.

2.2 Motivations

Due to the increasing volume of web data, it is becoming very challenging for web researchers to extract knowledge. In addition, with the increasing number of documents that can be accessed through computer networks, the need for extracting web content from them has become a basic motivation for the proposed Machine Learning framework. Although web mining techniques are good for handling low-dimensional data, they are not suitable for massive amounts of data. Due to this, many of the techniques are not optimized for high-dimensional data.

Motivated by the above considerations, here the researchers aim to enhance the clustering process by employing the Genetic algorithm. Towards this, the objectives of the proposed work are accordingly framed as the following:

- ✓ Design and implementation of effective pre-processing techniques - A web database is composed of large volumes of documents that are gathered from various sources. Due to the nature of the data, it tends to be cluttered and inconsistent. This is one of the main reasons to implement more professional pre-processing techniques.
- ✓ Construction of effective mining-ready data structure - To arrange the pre-processed web data in a good format, the data structure must be designed and modelled. The structure must determine the appropriate set of keywords for each web document in the corpus.
- ✓ Design and development of optimal clustering algorithm - Due to the complexity of web databases, the extraction of various types of hidden and previously unknown knowledge is typically done in various ways. This process involves designing and developing an optimal clustering algorithm in the nutshell of the Genetic Approach.

3. PROPOSED WORK

The quality and efficiency of web data segmentation are often evaluated through the use of clustering techniques. This study introduced the Machine Learning based Optimal Segmentation System (MLOSS) by concentrating on all the stages of web-based clustering techniques with an advantage of genetic framework to get optimal solutiion as presented in Figure 3.

3.1. Web Data Pre-Processing

The preprocessing of web documents is a crucial component of the Web mining process. This discipline involves extracting structured representations from web documents. In the literature, it has been observed that the time and effort spent in preprocessing are usually neglected or less attractive compared to the other tasks. So, the authors in this paper aim to implement complete web preprocessing techniques before segmenting the web documents.

3.1.1. **Web Document Collection:** The initial phase of web data preparation of the proposed MLOSS is the collection of web documents. The gathering of web documents is highly dependent on the aim of web mining over the World Wide Web.

3.1.2. **Tokenization:** The next step of MLOSS splits the web content into pieces and removes all punctuation marks. The tokens are then divided by replacing a space between the words. The set of obtained tokens of all web documents is considered as the dictionary of MLOSS.

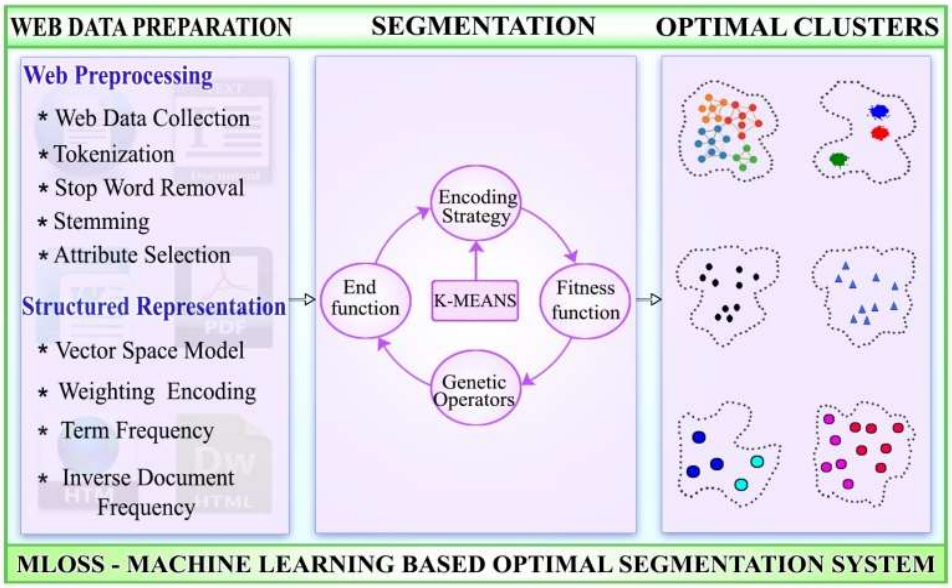

Figure 3: Architecture of Proposed MLOSS

3.1.3. **Stop Word Removal:** Stop words are terms that provide structure to a language instead of content. Yet, they may lead to confusion too in the entire web mining process. Moreover, the discard of stop words minimizes the dimensionality of the proposed VSM. Few stop words in English are presented in Figure 4.

a	further	myself	to
about	had	no	too
above	hadn't	nor	under
after	has	not	until
again	hasn't	of	up
against	have	off	very
all	haven't	on	was
am	having	once	wasn't
an	he	other	we
and	he'd	ought	we'd
any	he'll	our	we'll
are	he's	ours	we're
aren't	himself	ourselves	we've
as	his	out	were
at	how	over	weren't
be	how's	own	what
because	i	same	what's
before	i'd	shan't	when
being	i'll	she	where's
below	i'm	she'd	which
between	i've	she'll	while
both	if	she's	who
but	in	should	who's
by	into	shouldn't	whom
can't	is	so	why
cannot	isn't	some	why's
could	it	than	with
couldn't	it's	that	won't
did	its	that's	would
didn't	itself	the	wouldn't
do	her	their	you
does	here	theirs	you'd
doesn't	here's	them	you had
doing	hers	themselves	you'll
down	herself	then	you will
during	him	there	you're
each	more	there's	you've
few	most	these	your
for	mustn't	they	yours'
from	my	they'd	

Figure 4: Few English Stop Words

3.1.4. **Stemming:** The MLOSS resolves the issue of reducing the size of the Vector-Space-Model (VSM) by employing stemming. The goal of stemming is to minimize or avoid derivational forms of each word to common-base form. This procedure involves avoiding the use of inflectional forms and derivational associated forms of a word. The MLOSS uses the statistical N-Gram stemmer to identify common words. It does so by estimating the proportion of N-Grams that are in general.

The method is very interesting and it is language independent. It makes use of the string method to convert word inflation into its stemming word. N-Grams are parallel characters that are extracted from a given text. A document or data that contains N-Grams is then analyzed for its root form and associated words. The statistical analysis method used to identify them is known as the Inverse Frequency Document.

3.1.5. **Attribute Selection:** Even after the completion of the cleaning process of the web documents, a large set of features still exists. To minimize the size of the set, the MLOSS steps towards attribute selection. This step helps in selecting the best attributes for the problem domain.

a. ***Term Contribution(TC):*** The TC is computed based on which term contribute to the similarity of all the web documents collected.

$$TC(t_k) = \sum_{i,j \cap i \neq j} f(t_k, D_i) * f(t_k, D_j) \quad (1)$$

Where $f(t_k,D_i)$ is the Term-Frequency-Identifier of k^{th} term of i^{th} web document.

b. ***Term Variance:*** It is calculated in a web document by averaging the scores of terms that exhibit high document frequency.

$$v(t_i) = \sum_{j=1}^{N} [f_{ij} - \bar{f}_i]^2 \quad (2)$$

Where f_{ij} is the frequency of i^{th} term of j^{th} web document, (f_i) is mean frequency.

c. ***Term Variance Quality:*** The total variance is used to calculate the quality of a phrase in this metric.

$$q(t_i) = \sum_{j=1}^{n} {f_{ij}}^2 - \frac{1}{n}\left[\sum_{j=1}^{n} f_{ij}\right]^2 \quad (3)$$

Here f_{ij} is the frequency of i^{th} term of j^{th} web document.

Finally, these filters rank all of the terms, thus only the highest-ranked terms are used as selected features for the next stage of MLOSS.

3.2. Web Data Encoding and Structure Representation

Document encoding is a process that involves encoding web data in a format that is suitable for web mining. It is done by setting the element encoding value to one or zero if the word is used in the web document.

The term weight w(d,t) is calculated by taking the word importance of the given web document and dividing it by the number of words. The document weight w(d,t) is then translated into a vector space model.

A simple dimensional vector space model shows the terms that appear in the web document. The no. of times a word appears in the web document determines its value.

Table 1: MLOSS Sample Dimensional VSM

	News	Shopping	Blog	Games	Movies
WebDoc1	4	0	4	7	0
WebDoc2	0	6	0	0	0
WebDoc3	0	11	0	4	4
WebDoc4	6	4	3	2	0
WebDoc5	9	1	8	7	2
WebDoc6	7	3	5	6	7
WebDoc7	33	13	21	13	7
WebDoc8	4	9	9	9	2
WebDoc9	0	8	19	2	1

3.3. Web Data Segmentation using K-Means

The K-Means combines the formatted web data collected by the vector space model with the formatted web data. It segments the web documents into groups.

In the VSM, a web document is allotted to a centroid point. The centroid point of the segment is further modified based on web documents allocated for it based on the similarity of its content until the centroid point remain constant.

Procedure:

Step1. Find primary centroid point by choosing K web documents
Step2. Iterate
Step3. From K-Segments allocating each web document to its nearest centroid point
Step4. Re-calculate the point-of-centroid of each segment
Step5. Till centroid points remain constant

The allocation and relocation of web documents to clusters in the above technique are done using Euclidian distance DE. Given two web documents wd_a and wd_b, denoted with their word vectors $\vec{t_a}$ & $\vec{t_b}$, the D_E between the two web words is given as follows

$$D_E(\vec{t_a}, \vec{t_b}) = \left(\sum_{t=1}^{m} |w_{t,a} - w_{t,b}|^2\right)^{\frac{1}{2}} \quad (4)$$

Here, term-set TS = $\{t_{S1}, t_{S2}, \ldots, t_{Sm}\}$. Equally stated before, in the VSM TF-IDF is used as term weight, that is $w_{t,a} = TF - IDF(wd_a, t)$.

The statistical mean is used in the computation and re-computation of the centroid. In this situation, the quality of segmentation is identified by using the Sum of Squared Error approach (SSE). Calculate each web document's error or its Euclidean distance from the nearest centroid, first, and then the overall sum of the squared errors. When faced with two sets of clusters generated by two distinct K-Means runs, choose the one with the lower squared error, as this indicates that the centroids of this clustering are a better representation of the web documents in their cluster.

Journal of Theoretical and Applied Information Technology
15th June 2022. Vol.100. No 11

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The formal definition of the SSE is as follows:

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist(C_i, x)^2 \quad (5)$$

As a result, the usual k-mean technique ensured that only locally optimal solutions were found. Furthermore, weak initial centroids and random initialization of the K-Means technique result in low clustering accuracy.

3.4. Web Data Segmentation using Genetic Approach

The K-Means algorithm is commonly used for determining a cluster. However, due to its complexity, the algorithm tends to take many decisions at random initialization and initial centroids. The authors propose a hybrid technique that combines the use of a genetic approach with K-Means.

Optimization techniques find the optimal solution in certain circumstances. It involves finding the optimal cluster size. The initial phase of this process is to generate K clusters that were processed by the K-Means. Further, the biological-inspired operators of MLOSS create a new population and improve it.

The encoding strategy involves finding the initial population from the K-Means clustering algorithm's survival patterns. After the fitness function has evaluated the survival patterns, the next step is the generation of the next biological population.

3.4.1. **MLOSS Encoding:** The encoding technique of MLOSS is a step in the genetic process that helps identify the potential solution to a problem and make the algorithm capable of processing it. It is performed by selecting a suitable population from the generated clusters. For example, web document cluster {wd1, wd4, wd5, wd8} is coded as a binary chromosome with a length of 10 as described in Figure. 5. The existence of a web document is encoded with 1 and the other is encoded with 0.

Figure 5: Binary Chromosome of Web Document

3.4.2. **MLOSS Fitness Function:** The fitness function's characteristics are key to the success of a genetic algorithm. The genetic algorithm finds the optimal fit between the various outputs generated by the fitness function. The cohesion or separation of a cluster is determined by the sum of the nearby centroids of its members.

$$\text{ClusterCohesion}(C_i) = \sum_{X \in C_i} \text{Proximity}(x, c_i) \quad (6)$$

$$\text{ClusterSeparation}(C_i, C_j) = \text{promity}(c_i, c_j) \quad (7)$$

Where c_i is the centroid of a cluster C_i and c is the overall centroid. The similarity is calculated with a possible proximity measure that displays a comparable function's value.

If m and n are two web documents vectors, then

$$\cos(m, n) = \frac{m.n}{\|m\|\|n\|} \quad (8)$$

Where, $m.n = \sum_{k=1}^{x} n_k m_k$ and $\|m\|$ the length of vector m, $\|m\| = \sqrt{m.m}$

3.4.3. **MLOSS Operator:** Based on biological principles, the genetic operators - selection, mutation, and crossover are implemented on the clusters to produce new generations.

a. ***MLOSS Selection Process:*** The selection process involves selecting the genes from the mating pool. The selection is performed by spinning a roulette wheel with a marker and choosing the number of slots that corresponds to the fitness function of the cluster as presented in Figure. 6.

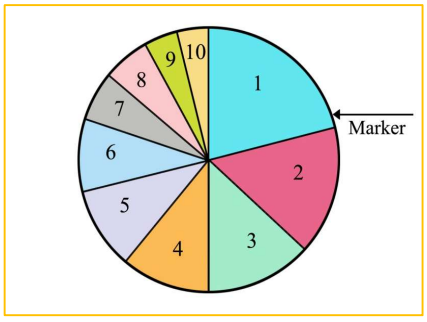

Figure 6: MLOSS Roulette Wheel Parent Selection Process

b. ***MLOSS Crossover Function:*** For each parent's chromosomes, the crossover point is generated. The information exchanged between the two parent chromosomes is then used to create two child chromosomes as presented in figure 7 & 8.

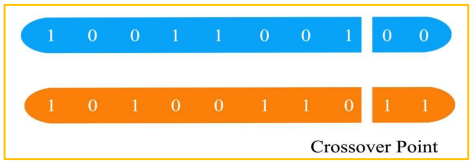

Figure 7: Example of Crossover Point

Journal of Theoretical and Applied Information Technology
15th June 2022. Vol.100. No 11

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

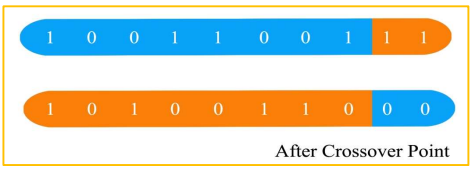

Figure 8: Result of Crossover Function

c. ***MLOSS Mutation Function:*** Fixed probability function is used to mutate binary chromosomes using bit position values. The bit position is then changed by flipping its value as shown in Figure 9.

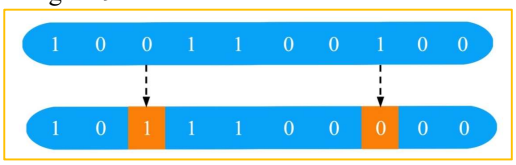

Figure 9: Process of MLOSS Mutation

d. ***MLOSS Genetic Algorithm:*** The MLOSS genetic algorithm begins with the initial population preparation step. The fitness function then evaluates the survival capacity of the generated population. The crossover operators and the mutation operators are then selected and executed in parallel to produce the new population.

Algorithm:

1. Begin
2. Create initial population using MLOSS Encoding
3. Estimate population by MLOSS Fitness Function
4. Produce new population by Looping
   a. MLOSS Selection Process
   b. MLOSS Single Point Crossover
   c. MLOSS Bit-by-Bit Mutation
5. Till End Condition is Reached
6. Arrive Optimal Web Document Clusters

Finally, the Genetic algorithm of MLOSS with all its sub-processes reaches the optimal solution in segmenting web documents in less time.

4. EXPERIMENTAL ANALYSIS

The proposed MLOSS algorithm is the hybridization of K-Means and genetic algorithms. The number cluster generated by the algorithm is derived by taking the data from the vector space model.

A series of experiments have been performed to showcase the accuracy of MLOSS over the number of generations of proposed genetic algorithm in web data segmentation as shown in Figure. 10.

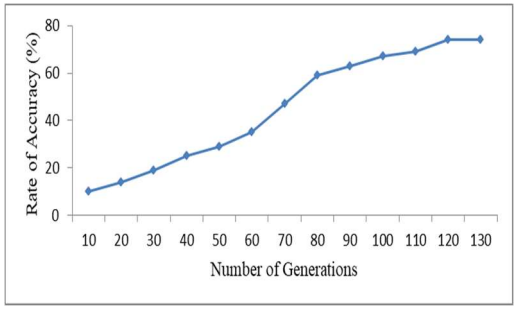

Figure 10: Accuracy Rate of MLOSS

The MLOSS performed various cross over probabilities for different iterations. It showed good performance at the average mean Cp is 0.5 as described in Figure. 11.

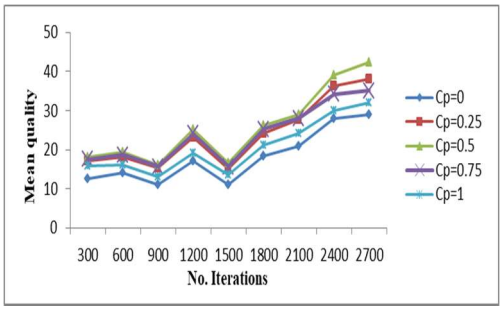

Figure 11: Mean Quality of MLOSS

The MLOSS takes a little more time over Standard K-Means, however, it outperforms in generating optimal segments than K-Means as shown in Figure.12.

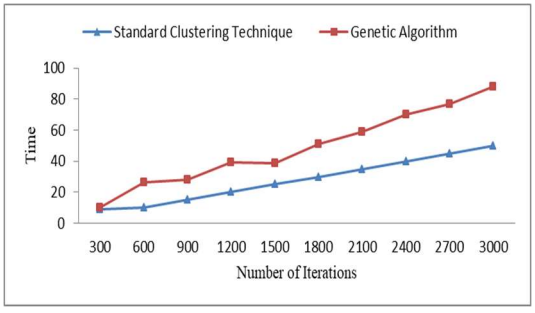

Figure 12: Processing Time of MLOSS

5. CONCLUSIONS

Web engineering is a multidisciplinary field of research that focuses on extracting valuable information from unstructured web documents. The aim of this work was to find optimal segments of web documents using unsupervised mining

techniques. The method was formulated by combining the K-Means and genetic algorithm. The proposed MLOSS framework enables organizations to segment their web documents into various data segments for actionable intelligence. The proposed work's results demonstrate that combining genetic algorithms with both supervised and unsupervised techniques provides a future research in the era of web engineering.

REFERENCES:

[1] Chuang Shan and Yugen Du, "A Web Service Clustering Method Based on Semantic Similarity and Multidimensional Scaling Analysis", Hindawi, Scientific Programming, Volume 2021, pp.01-12, 2021.

[2] C. Sun, L. Lv, G. Tian, Q. Wang, X. Zhang, and L. Guo, "Leverage label and word embedding for semantic sparse web service discovery," Mathematical Problems in Engineering, vol. 2020, pp. 1–8, 2020.

[3] I. Lizarralde, C. Mateos, A. Zunino, T. A. Majchrzak, and T. M. Gronli, "Discovering web services in social web service repositories using deep variational autoencoders," Information Processing & Management, vol. 57, no. 4, 2020.

[4] P. Ashok kumar and S. Don, "Link-Based Clustering Algorithm for Clustering Web Documents", Journal of Testing and Evaluation, DOI: 10.1520/JTE20180497, 2019.

[5] Muhammd Jawad Hamid Mughal, "Data Mining: Web Data Mining Techniques, Tools and Algorithms: An Overview", International Journal of Advanced Computer Science and Applications, Vol. 9, No. 6, pp.208-2015, 2018.

[6] Zuping Zhang, Jing Zhao and Xiping Yan, "A Web Page Clustering Method Based on Formal Concept Analysis", Information 2018, MDPI, 2018.

[7] Thamme Gowda1 and Chris Mattmann, "Clustering Web Pages Based on Structure and Style Similarity", 2016 IEEE 17th International Conference on Information Reuse and Integration, pp.175-180, 2016.

[8] Mitali Srivastava, Rakhi Garg, P. K. Mishra, "Analysis of Data Extraction and Data Cleaning in Web Usage Mining", ICARCSET 2015, ACM, pp.01-06, 2015.

[9] Chen-Hau Wang, Ching-Tsorng Tsai, Chai-Chen Fan, Shyan-Ming Yuan, "A Hadoop Based Weblog Analysis System", 7th International Conference on Ubi-Media Computing and Workshops, IEEE, pp.72-77, 2014.

[10] Xindong Wu, Xingquan Zhu, et.al., "Data Mining with Big Data", IEEE Transactions on Knowledge and Data Engineering, Vol. 26, No. 1, pp: 97-107, 2014.

[11] Duc Thang Nguyen, Lihui Chen, "Clustering with Multiviewpoint-Based Similarity Measure", IEEE Transactions on Knowledge and Data Engineering, VOL. 24, NO. 6, pp:987-1000, 2012.

[12] K. Santra, C. Josephine Christy, "Genetic Algorithm and Confusion Matrix for Document Clustering", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 2, pp: 322-328, 2012.

[13] Anjali Ganesh Jivani, "A Comparative Study of Stemming Algorithms", Int. J. Comp. Tech. Appl.(IJCTA), Vol 2 (6), pp: 1930-1938, 2011.

[14] Giridhar N S, Prema K.V, .V Subba Reddy, "A Prospective Study of Stemming Algorithms for Web Text Mining", Ganpat University Journal of Engineering & Technology, Vol.-1, Issue-1, pp: 28-34, 2011.

[15] Michal Munk, Martin Drlík, "Impact of Different Pre-Processing Tasks on Effective Identification of Users' Behavioral Patterns in Web-based Educational System", International Conference on Computational Science (ICCS 2011), Published by Elsevier Ltd, pp. 1640–1649, 2011.

[16] N. El-Bathy, C. Gloster, I. Kateeb, G. Stein, "Intelligent Extended Clustering Genetic Algorithm for Information Retrieval Using BPEL", American Journal of Intelligent Systems, Vol 1(1): pp: 10-15, 2011. [Information retrieval]

[17] S.Vijayalakshmi, Dr.D.Manimegalai, "Query based Text Document Clustering using its Hypernymy Relation", International Journal of Computer Applications, Volume 23– No.1, pp:13-16, 2011. [Information retrieval]

[18] Atul Kamble, "Incremental Clustering in Data Mining using Genetic Algorithm", International Journal of Computer Theory and Engineering, Vol. 2, No. 3, June, pp:326-328, 2010.

[19] Brijendra Singh, Hemant Kumar Singh, "Web Data Mining Research: A Survey", 978-1-4244-5967-4/10, IEEE, pp: 1-10, 2010.

[20] Muhammad Rafi, M. Shahid Shaikh, Amir Farooq, "Document Clustering based on Topic Maps", International Journal of Computer Applications (0975 – 8887), Volume 12– No.1, December 2010. [document Introd]

[21] Ponmuthuramalingam P, T. Devi, "Effective Term Based Text Clustering Algorithms", International Journal on Computer Science and Engineering Vol. 02, No. 05, pp; 1665-1673, 2010. [document Clustering]

[22] V.V.R. Maheswara Rao, Dr. V. Valli Kumari, "A Novel Lattice Based Research Frame Work for Identifying Web User's Behavior with Web Usage Mining", Springer LNCS-CCIS, ISSN: 1865-0929, Vol. 101, Part 1, pp. 90-99, 2010.

[23] V.V.R.Maheswara Rao, Dr. V. Valli Kumari, Dr. KVSVN Raju "Study of Visitor Behavior by Web Usage Mining" Springer LNCS-CCIS, Vol. 70, pp. 181-187, 2010.

[24] Bader Aljaber , Nicola Stokes, James Bailey, Jian Pei, "Document clustering of scientific texts using citation contexts", Springer Science+Business Media, LLC , pp: 2009. [concept linkage]

[25] Pencheva T., Atanassov K., Shannon A., "Modelling of a Roulette Wheel Selection Operator in Genetic Algorithms Using Generalized Nets", BIOAUTOMATION, 13 (4), pp: 257-264, 2009.

[26] H. Chim and X. Deng, "Efficient Phrase-Based Document Similarity for Clustering," IEEE Trans. Knowledge and Data Eng., vol. 20, no. 9, pp. 1217-1229, Sept. 2008.

[27] Martin Krallinger, Alfonso Valencia, Lynette Hirschman, "Linking genes to literature: text mining, information extraction, and retrieval applications for biology", Genome Biology 2008. [Information retrieval, concept linkage]

[28] Olfa Nasraoui, Maha Soliman, Esin Saka, Antonio Badia and Richard Germain, "A Web Usage Mining Framework for Mining Evolving User Profiles in Dynamic Web Sites", IEEE Transactions on Knowledge And Data Engineering, Vol.20, Issue.2, pp.1-13, 2008.

[29] X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J. McLachlan, A. Ng, B. Liu, P.S. Yu, Z.-H. Zhou, M. Steinbach, D.J. Hand, and D. Steinberg, "Top 10 Algorithms in Data Mining," Knowledge Information Systems, vol. 14, no. 1, pp. 1-37, 2007.

[30] Elizabeth Leon, Olfa Nasraoui, and Jonatan Gomez, "ECSAGO: Evolutionary Clustering with Self Adaptive Genetic Operators", IEEE Congress on Evolutionary Computation, pp:1768-1775, 2006.

[31] S. M. Khalessizadeh, R. Zaefarian, S.H. Nasseri, and E. Ardil, "Genetic Mining: Using Genetic Algorithm for Topic based on Concept Distribution", International Journal of Engineering and Applied Sciences, pp:51-54, 2005.

[32] He, X.; Ding, C.H.; Zha, H., Simon, H.D. "Automatic topic identification using webpage clustering", 2001 IEEE International Conference on Data Mining, San Jose, CA, USA, 29 November–2, pp. 46–54, December 2001.

[33] Kwon, O.W.; Lee, J.H. Web page classification based on k-nearest neighbor approach. In Proceedings of the 5th international workshop on Information Retrieval with Asian Languages, Hong Kong, China, 30 Sep–1 Oct 2000; pp. 9–15, 2000.

[34] S. Guha, R. Rastogi, K. Shim. CURE: An Efficient Clustering Algorithm for Large Databases. ACM SIGMOD Conference, 1998.

[35] P. Anick, S. Vaithyanathan. Exploiting Clustering and Phrases for Context-Based Information Retrieval. ACM SIGIR Conference, 1997.

[36] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing", Communications of the ACM, Vol.18, Issue.11, pp.613–620, 1975.