Journal of Theoretical and Applied Information Technology
31st May 2024. Vol.102. No. 10

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

HEREDITARY APPROACH FOR FINEST DISSECTION SYSTEM FOR WEB DATA

[1]M V GANESWARA RAO,[2]LAKSHMI MANASA B, [3] BALAJI TATA, [4]DR. KARUNA ARAVA
[5]ANE ASHOK BABU, [6]PRAVEEN TUMULURU, [7] N.JAYA

[1]Dept. of ECE, Shri Vishnu Engineering College for Women, Bhimavaram, AP.
[2]Dept. of ECE, Sri Venkateshwara College of Engineering, Bangalore India.
[35]Dept. of ECE, PVP Siddhartha Institute of Technology, Vijayawada, A.P, India
[4]Dept. of CSE, University College of Engineering Kakinada, JNTUK, Kakinada, A.P,
[6]Dept. of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, A.P,
[7]Dept. of EIE, Faculty of Engineering and Technology, Annamalai University, Chidambaram, Tamil Nadu

Email: ganesh.mgr@gmail.com , manasa.7006@gmail.com , balajitata@pvpsiddhartha.ac.in, karunagouthana@jntucek.ac.in, , ashokbabu@pvpsiddhartha.ac.in , praveenluru@gmail.com, jayanavaneethan@rediffmail.com

ABSTRACT

The rise of electronic data has spawned an ocean of untapped information, laying the groundwork for web mining. Simultaneously, the surge in computer technology has flooded databases with vast amounts of data, propelling the realms of Web Science and Big Data Analytics into the forefront. Web Science, an engineering process, delves into large datasets, seeking patterns amidst the chaos. Yet, extracting intrinsic structures from this vast expanse poses a formidable challenge, hindering efforts to organize them into coherent groups. Existing clustering algorithms often fall short of meeting the diverse needs of web applications. This spurred our team to pioneer an innovative algorithm, poised to offer greater applicability and resilience in this dynamic landscape.

The driving force behind this research endeavor is the creation of a Machine Learning framework aimed at extracting technological insights from web data sources. The authors advocate for an Optimal Segmentation System employing a Machine Learning approach with dual objectives: firstly, to preprocess unstructured and semi-structured web documents and establish an efficient data representation structure to facilitate the application of both supervised and unsupervised techniques. Subsequently, the system prioritizes segmenting the preprocessed web data by hybridizing Genetic Approach with clustering techniques, mirroring biological evaluation processes with self-learning capabilities. Extensive experimentation has been conducted to validate the performance results of the proposed framework across various orders of magnitude, confirming its efficacy as claimed.

Keywords: *Hereditary, Machine Learning, Finest, Dissection*

1. INTRODUCTION

The proliferation of data sources and their accessibility on the World Wide Web (www) has ushered in new avenues for academic and social engagement. This is underscored by the proliferation of diverse web mining and analytical techniques. The endeavor to extract hitherto unknown patterns from vast volumes of web data is encapsulated as knowledge retrieval from web data, as illustrated in Figure 1.

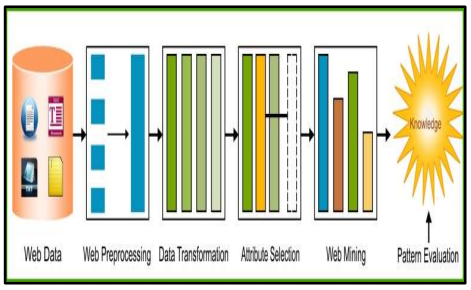

Fig. 1. Process Of Web Mining

Web data, often gathered from informal settings like weblogs, emails, web pages, and chat rooms, presents a rich source for exploration. To harness

its potential, storing web data in a structured format is essential. While various methods exist, many concentrate on extracting the textual structure and its nuances.

Typically, web mining approaches operate under the assumption that a web document can be represented by a sequence of words. To gauge the significance of each word, a vector representation is commonly employed, involving the following phases.
To efficiently process web documents for analysis, the following steps are typically involved:

1. **Tokenization**: Split the web file into individual words using space as the delimiter.

2. **Filtering**: Remove words that lack specific meaning, such as common stop words (e.g., "the", "and", "is").

3. **Stemming**: Apply the Porter Stemmer algorithm to reduce words to their common root form. This helps to consolidate variations of the same word (e.g., "running" and "ran" both stem to "run").

4. **Data Structure Construction**: Create a data structure optimized for analyzing large collections of web documents without relying on explicit semantic information. This structure should facilitate efficient storage and retrieval of information, enabling various analytical tasks without the need for semantic understanding. Examples of such structures include inverted index for information retrieval or term-document matrices for document similarity analysis. These structures allow for rapid querying and manipulation of web data without the overhead of semantic processing.
Web data transformation involves extracting features from web documents. Unlike open-ended data mining, web mining requires a structured approach due to the unorganized nature of the data. Attribute selection algorithms are employed to identify important features for a given task. Without relevant features, these algorithms may not be practical for unsupervised learning.

Web data engineering aims to uncover intrinsic insights within large amounts of web data. It encompasses various forms such as knowledge mining from web data, fixed domain web mining, supervised web mining, and unsupervised web mining. Each form serves to extract valuable information from web sources through systematic analysis and processing.

Supervised web mining involves classifying web documents by training algorithms with pre-defined classes. The model learns to automatically categorize new web documents based on their content, typically using labeled training data. This approach is useful for tasks like classifying news stories by topic.

In real-time applications, however, unsupervised web mining often holds more advantages over supervised methods. Unsupervised methods aim to group documents based on similarities and differences in their content, resulting in clusters. The quality of these clusters improves when documents within a cluster are more similar to each other than to documents in other clusters.

Despite its advantages, unsupervised web mining alone may not suffice for many web applications. To address this limitation, web science researchers may opt for optimized methods like genetic algorithms, especially for tackling nonlinear and complex domain problems. These algorithms can enhance the effectiveness of unsupervised methods in extracting valuable insights from web data.
In this study, the authors introduce a Machine Learning-based Optimal Segmentation System employing a genetic algorithm, offering a novel approach tailored to the innate characteristics of genetic algorithms.
The research paper is structured as follows:
- Section 1.2 outlines the important related work and the inspirations behind the proposed approach.
- Section 1.3 highlights the major contributions of the authors to the field.
- Section 1.4 presents the findings of the investigation along with a detailed analysis.
- Finally, in section 1.5, a concise summary of the paper is provided, followed by a discussion on future research directions.

2. LITERATURE REVIEW

In the past two decades, a plethora of researchers [1, 2, 3, 4, 5, 6, 9, 10, 22, 23] have delved into web mining techniques and technologies, striving to devise novel solutions for real-world challenges. This section provides an extensive literature survey, encapsulating the myriad developments in web mining, as depicted in Fig 2.

Numerous endeavors [16, 17, 19, 33, 35] have been undertaken in the realm of information retrieval. The pioneering attempt at automatic indexation in 1975 [36] underscores the growing significance of information search, catalyzed by the emergence of the World Wide Web.

Researchers have predominantly focused on diverse approaches for extracting web data, aiming to unearth specific knowledge from vast datasets. Of particular interest is web document categorization, a domain that has garnered significant attention from web researchers. A multitude of papers have contributed to this field, with web document categorization methods typically employed in two ways: automatic grouping of similar documents (Clustering or Unsupervised Learning) [7, 11, 32, 34], or assigning keywords to specific documents (Classification or Supervised Learning) [19, 29, 33]. Among these approaches, web document categorization using unsupervised techniques stands as a prominent and well-established research area.

Web Document Clustering, an unsupervised method, facilitates the organization of web documents into segments or clusters. Traditionally, this involves identifying the attributes of a dataset. This problem has been extensively studied [26, 28, 29, 35] for both quantitative and categorical data. Among the various approaches, the authors opt to focus on Distance-based Clustering.

Distance-based Clustering methods are designed to gauge the similarity between web documents. Determining content similarity is pivotal in web mining. To achieve this, content similarity functions are employed in conjunction with clustering algorithms such as hierarchical clustering and partitioned clustering. In the proposed work, the emphasis is placed on partitioned clustering to efficiently segment web documents.

a. Motivations

- The burgeoning volume of web data presents a formidable challenge for web researchers seeking to extract knowledge. Moreover, as the accessibility of documents via computer networks continues to expand, there is a growing imperative to extract web content efficiently. This serves as a fundamental motivation for the proposed Machine Learning framework.
- While web mining techniques excel in handling low-dimensional data, they often struggle to cope with massive datasets. As a result, many existing techniques are not optimized for high-dimensional data.

  Motivated by these challenges, the researchers aim to enhance the clustering process by leveraging the Genetic algorithm. To achieve this, the objectives of the proposed work are formulated as follows:

- To address the challenges posed by large volumes of cluttered and inconsistent web data, the following steps are proposed:

  1. **Effective Pre-processing Techniques: ** Develop and implement professional pre-processing techniques to clean and standardize the web data. This involves tasks such as removing noise, handling missing values, and normalizing data formats to ensure consistency.

  2. **Construction of Mining-Ready Data Structure: ** Design and implement a structured data format to organize the pre-processed web data effectively. This data structure should facilitate efficient storage and retrieval of information, including the identification of relevant keywords for each document in the dataset.

  3. **Optimal Clustering Algorithm Design: ** Given the complexity of web databases, extracting hidden and previously unknown knowledge requires a sophisticated approach. Develop an optimal clustering algorithm leveraging Genetic Approach to cluster the web data effectively. This algorithm should be capable of identifying meaningful patterns and relationships within the dataset, enabling insightful analysis and decision-making.

3. PROPOSED WORK

In this study, the authors introduce the Optimal Segmentation System, a novel approach aimed at enhancing the quality and efficiency of web data segmentation. This system represents a comprehensive framework designed to evaluate and improve the segmentation of web data using clustering techniques.

The Optimal Segmentation System incorporates advanced algorithms and methodologies to achieve superior segmentation results. By leveraging clustering techniques, the system aims to organize web data into meaningful clusters, thereby facilitating the extraction of valuable insights and patterns.

Through rigorous evaluation and experimentation, the authors demonstrate the effectiveness of the Optimal Segmentation System in improving the quality and efficiency of web data segmentation. The system offers a valuable tool for researchers and practitioners seeking to analyze and extract

knowledge from large volumes of web data.

Overall, the introduction of the Optimal Segmentation System represents a significant contribution to the field of web data analysis, providing a robust and scalable solution for segmentation tasks.

3.1Web Data Pre-Processing

The preprocessing of web documents plays a pivotal role in the web mining process, involving the extraction of structured representations from the raw web content. Despite its significance, preprocessing is often overlooked or given less attention compared to other tasks in the literature. Therefore, the authors of this paper aim to implement comprehensive web preprocessing techniques before segmenting the web documents.

1. **Web Document Collection**: The initial phase of the MLOSS (Machine Learning-based Optimal Segmentation System) involves collecting web documents. The selection of web documents depends on the specific objectives of the web mining task.

2. **Tokenization**: The next step in MLOSS is tokenization, where the web content is split into individual tokens and all punctuation marks are removed. The tokens are then separated by spaces, forming the dictionary of MLOSS.
3. **Stop Word Removal**: Stop words are common terms that provide structural elements to a language rather than conveying significant content. However, they can introduce noise and confusion in the web mining process. Therefore, the removal of stop words helps streamline the analysis and minimize the dimensionality of the proposed Vector Space Model (VSM). Fig X illustrates a few examples of stop words commonly found in English.

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	

Fig 2 Few English Stop Words

Stemming is a crucial step in reducing the size of the Vector-Space-Model (VSM) employed in web document preprocessing. Its goal is to convert words to their common base form, thereby minimizing or avoiding derivational forms. This process involves disregarding inflectional and derivational variations of words. The MLOSS adopts a statistical N-Gram stemmer for this purpose, which estimates the proportion of N-Grams commonly found in words. This method is language-independent and utilizes string manipulation techniques to transform inflated words into their stemmed counterparts. N-Grams are sequential characters extracted from text, and documents containing N-Grams are analyzed to identify their root forms and associated words. The statistical analysis technique used for identification is known as Inverse Frequency Document.
After the cleaning process, the web documents still contain a large set of features. To address this, MLOSS employs attribute selection to minimize the size of the feature set. This step aids in selecting the most relevant attributes for the specific problem domain, thus streamlining subsequent analysis tasks.

Term Contribution(TC): The Term Contribution (TC) is computed based on the terms that contribute to the similarity of all the web documents collected. In essence, TC quantifies the importance of each term in relation to the overall similarity between documents. This calculation involves assessing the frequency of each term across the entire document collection and evaluating its significance in

determining document similarity. Terms that appear frequently across multiple documents are deemed more influential and contribute more to the overall similarity measure. By computing TC, analysts can identify key terms that play a significant role in capturing the essence of the document collection and informing subsequent analysis and decision-making processes.

3.2Web Data Encoding and Structure Representation

Document encoding is the process of converting web data into a format suitable for web mining. This is typically achieved by assigning a binary value (1 or 0) to each term in the web document, indicating whether the term is present or absent.

The term weight \(w(d, t) \) represents the importance of a term \(t \) in a given web document \(d \). It is calculated by dividing the frequency of the term in the document (word importance) by the total number of words in the document. This normalization ensures that the term weight reflects the relative importance of the term within the document.

Once term weights are computed for all terms in the document collection, they are translated into a vector space model (VSM). In this model, each document is represented as a vector in a high-dimensional space, with each dimension corresponding to a unique term. The value of each dimension (term) in the vector represents the weight of the term in the document.

In a simple dimensional vector space model, the terms that appear in the web document are listed, and the value of each term corresponds to the number of times it appears in the document. This representation provides a straightforward way to visualize the contents of the document and identify the most frequent terms.

Table 1 Loss Sample Dimentional Vsm

	News	shopping	Blogs	Games	Finance
WebDoc1	5	0	5	8	0
WebDoc2	0	7	0	0	0
WebDoc3	0	12	0	5	5
WebDoc4	7	5	4	2	0
WebDoc5	8	1	9	8	3
WebDoc6	8	4	6	7	7
WebDoc7	35	12	22	14	7
WebDoc8	5	8	9	9	3
WebDoc9	0	9	20	3	2

3.3Web Data Segmentation using K-Means

In the K-Means clustering algorithm, the formatted web data collected by the vector space model is combined and utilized to segment the web documents into groups or clusters. Each cluster is represented by a centroid point, which serves as a reference point for the documents assigned to that cluster.

Initially, each web document is assigned to the nearest centroid point based on its similarity to the centroid. Then, the centroid point of each cluster is recalculated as the mean of all the documents assigned to that cluster. This process continues iteratively until the centroid points no longer change significantly, indicating convergence.

The iteration involves adjusting the centroid points based on the contents similarity of the web documents allocated to each cluster. By iteratively refining the centroid points, K-Means effectively partitions the web documents into cohesive clusters based on their similarity in the vector space model.

Procedure:

The K-Means clustering algorithm typically follows these steps:

1. **Initialization**: Choose K web documents as the initial centroid points.
2. **Iteration**:
- Assign each web document to its nearest centroid point, forming K segments.
- Calculate the centroid point of each segment by averaging the positions of all web documents assigned to that segment.
- Repeat these steps until the centroid points remain constant (convergence).

These iterative steps ensure that each web document is assigned to the nearest centroid point and that the centroid points are updated to reflect the central tendency of the documents in each segment. This process continues until the centroids

stabilize, indicating that the clustering has converged.

The Euclidean distance (DE) between two web documents \(w_{\text{a}} \) and \(w_{\text{b}} \), denoted by their word vectors \(\vec{t}_{\text{a}} \) and \(\vec{t}_{\text{b}} \), is calculated as follows:

\[DE(w_{\text{a}}, w_{\text{b}}) = \sqrt{\sum_{i=1}^{n} (t_{\text{a}_i} - t_{\text{b}_i})^2} \]

where \(t_{\text{a}_i} \) and \(t_{\text{b}_i} \) are the components of the word vectors \(\vec{t}_{\text{a}} \) and \(\vec{t}_{\text{b}} \) respectively, and \(n \) is the dimensionality of the word vectors.

The quality of segmentation is evaluated using the Sum of Squared Error (SSE) approach. This involves calculating the squared error for each web document, which is the Euclidean distance from the nearest centroid, and then summing up these squared errors across all documents. The formal definition of SSE is as follows:

\[SSE = \sum_{i=1}^{K} \sum_{w \in C_i} DE(w, c_i)^2 \]

where \(K \) is the number of clusters, \(C_i \) is the set of web documents assigned to cluster \(i \), and \(c_i \) is the centroid of cluster \(i \). \(DE(w, c_i) \) represents the Euclidean distance between web document \(w \) and the centroid \(c_i \).

When comparing two sets of clusters generated by two distinct K-Means runs, the set with the lower SSE is chosen, indicating that the centroids of this clustering better represent the web documents in their respective clusters.

3.4 Web Data Segmentation using Genetic Approach

It sounds like you're describing a hybrid approach that combines genetic algorithms with the K-Means clustering technique for optimization. This approach aims to improve the performance of K-Means by leveraging genetic algorithms' ability to search for optimal solutions.

In this hybrid technique, the initial phase involves generating K clusters using the K-Means algorithm. Then, genetic algorithms are applied to optimize these clusters further by creating a new population and improving it through biological-inspired operators.

The encoding strategy plays a crucial role in this process. It involves representing the clusters as binary chromosomes, where each chromosome corresponds to a cluster and indicates the presence or absence of web documents within that cluster. For example, if a web document belongs to a particular cluster, its corresponding bit in the chromosome is set to 1; otherwise, it's set to 0.
By using this encoding technique, the genetic algorithm can manipulate and evolve the clusters to find a more optimal solution, ultimately improving the performance of the K-Means algorithm in determining cluster centroids and assignments. This hybrid approach combines the strengths of both algorithms to overcome the limitations of K-Means' random initialization and improve its effectiveness in finding optimal cluster solutions.

Fig 3. Binary Chromosome Of Web Document

Fitness Function: Exactly, the fitness function serves as the guiding principle for the genetic algorithm in evaluating the quality of potential solutions. In the context of clustering, particularly with K-Means, the fitness function typically aims to optimize the cohesion within clusters while maximizing the separation between them.

Cohesion refers to how closely related the data points within a cluster are to each other, typically measured by minimizing the distance between points within the same cluster. Separation, on the other hand, relates to the dissimilarity between clusters, often measured by maximizing the distance between cluster centroids or minimizing the overlap between clusters.

In the case of your hybrid approach, where genetic algorithms are combined with K-Means, the fitness function could indeed be designed to evaluate the cohesion and separation of clusters. This evaluation might involve calculating the sum of distances between data points and their respective centroids within each cluster, as well as the distances between centroids of different clusters.
By maximizing cohesion and minimizing separation through the fitness function, the genetic algorithm can effectively guide the optimization process towards finding an optimal clustering solution. This ensures that the resulting clusters are compact and well-separated, leading to more meaningful and useful cluster assignments.

The selection process you described is commonly known as "roulette wheel selection" in genetic algorithms. It's a method used to select individuals (or clusters, in this case) from the population for mating, with a probability proportional to their fitness.

Here's how it works:

1. **Fitness Evaluation**: First, the fitness function is used to evaluate the fitness of each cluster in the population. This fitness function could be based on the cohesion and separation characteristics you mentioned earlier.

2. **Probability Calculation**: Once the fitness values are obtained, they are used to calculate the probability of selection for each cluster. Higher fitness values result in higher probabilities of selection.

3. **Roulette Wheel**: Imagine a roulette wheel where each cluster occupies a segment whose size is proportional to its selection probability. The wheel is spun, and a marker selects a segment. The cluster corresponding to the selected segment is chosen for the mating pool.

4. **Repetition**: This process is repeated until the desired number of clusters is selected for mating.

Roulette wheel selection ensures that clusters with higher fitness values (better solutions) have a higher chance of being selected for reproduction, mimicking the process of natural selection in biological evolution. This helps in driving the genetic algorithm towards better solutions over successive generations.

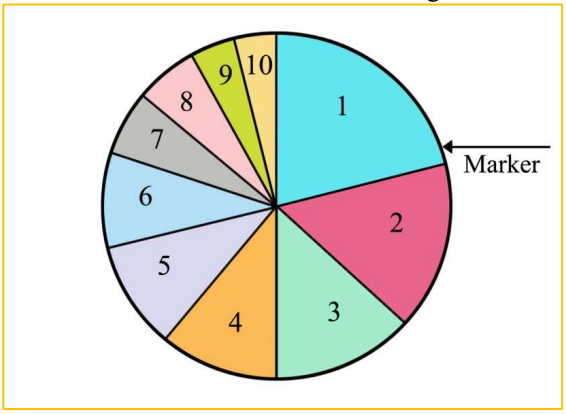

Fig 4 Roulette Wheel Parent Selection Process

The crossover function, also known as recombination, is a crucial genetic operator that facilitates the exchange of information between parent chromosomes to generate new offspring. In the context of your hybrid approach combining genetic algorithms with K-Means, the crossover function would operate on the binary chromosomes representing clusters

Here's how the crossover function typically works:1. **Selection of Parent Chromosomes**: First, two parent chromosomes are selected from the mating pool using the selection process you described earlier, such as roulette wheel selection.

2. **Crossover Point Generation**: Next, a crossover point is randomly chosen along the length of the parent chromosomes. This crossover point determines where the exchange of genetic material will occur between the parent chromosomes

3. **Exchange of Genetic Information**: The genetic information (bits) to the right of the crossover point in one parent chromosome is exchanged with the corresponding genetic information in the other parent chromosome. This exchange results in the creation of two child chromosomes.

4. **Creation of Child Chromosomes**: After the exchange of genetic information, two child chromosomes are formed by combining the genetic material from the parent chromosomes.

5. **Repeat**: This process is repeated for each pair of parent chromosomes until the desired number of child chromosomes is generated.

The crossover function enables exploration of the solution space by combining beneficial characteristics from different parent chromosomes, potentially leading to offspring with improved fitness compared to their parents. By iteratively applying crossover and other genetic operators, the genetic algorithm can evolve a population of clusters towards more optimal solutions for the clustering problem.

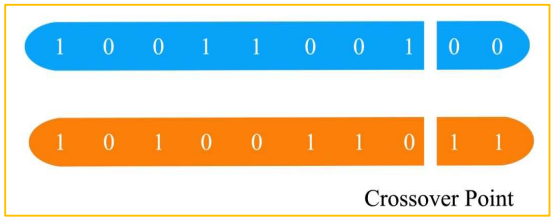

Fig 5 Example Of Crossover Point

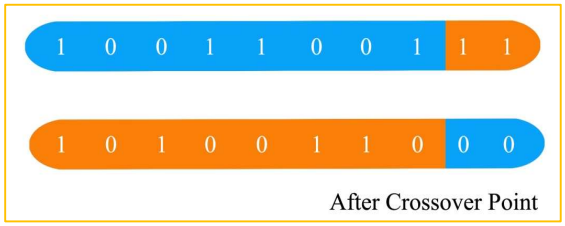

Fig 6 Result Of Crossover Function

Mutation Function: Mutation is another essential genetic operator that introduces diversity into the population by randomly altering individual chromosomes. In your case, where binary chromosomes represent clusters, mutation serves to explore the solution space by flipping bits within the chromosomes.

Here's how the mutation process typically works:

1. **Selection of Chromosomes**: Randomly select chromosomes from the population for mutation. This selection can be done with a fixed probability for each chromosome.

2. **Mutation Operation**: For each selected chromosome, iterate through its bits. At each bit position, determine whether to flip the bit based on a fixed mutation probability.

3. **Bit Flipping**: If the decision is made to mutate the bit at a particular position, flip its value. For example, if the bit is 0, change it to 1, and vice versa.

4. **Repeat**: Repeat the mutation process for all selected chromosomes in the population.

By introducing random changes to the chromosomes, mutation helps prevent premature convergence to suboptimal solutions and promotes exploration of new regions in the solution space. However, since mutation is applied with a fixed probability, it typically occurs at a low rate to maintain the stability of the population and prevent excessive disruption of potentially good solutions.

Overall, the combination of selection, crossover, and mutation operators in genetic algorithms allows for the iterative improvement of solutions over generations, ultimately leading to the discovery of optimal or near-optimal solutions to the clustering problem.

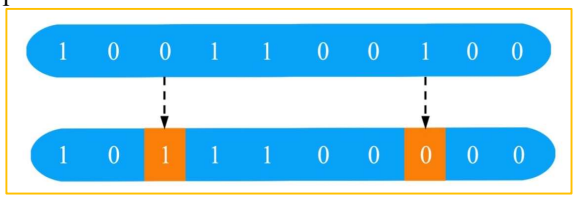

Fig.7 Process Of Mutation

It seems like you're describing two different studies or projects: one involving the use of a genetic algorithm for clustering web documents, and another involving the creation of a data logger system for environmental sensor data with IoT integration and cloud storage.

In the first study, a genetic algorithm is employed to segment web documents into clusters efficiently. The algorithm involves several steps, including

encoding, fitness evaluation, selection, crossover, and mutation, ultimately leading to the optimal segmentation of web documents into clusters. This approach aims to improve the efficiency of clustering while reducing the computational time required.

In the second study, a data logger system is developed to collect and store sensor data, typically integer values representing environmental parameters, into EEPROM memory. Additionally, with IoT integration, the collected data is transmitted to a web page for visualization. To make the system cost-effective, free cloud storage services are utilized to host the sensor data, enabling universal access without incurring rental fees for hosting a dedicated web page.

Both studies highlight the use of modern technologies such as genetic algorithms, IoT, and cloud computing to solve real-world problems efficiently and cost-effectively. Integrating oxygen level monitoring into a weather monitoring system represents a significant enhancement to the existing setup. By incorporating this additional environmental parameter, the proposed model can provide more comprehensive and accurate insights into atmospheric conditions.Here's how the proposed model differs from the existing setup:1. **Web-Enabled Monitoring**: Unlike the existing manual data collection and analysis process, the proposed model utilizes web-enabled tools for monitoring. This enables real-time data collection and remote access to weather and oxygen level information, enhancing the system's accessibility and usability.
2. **Integration of Oxygen Level Monitoring**: The existing weather monitoring system may focus on traditional meteorological parameters such as temperature, humidity, and precipitation. However, the proposed model expands the scope by incorporating oxygen level monitoring. This addition allows for a more holistic understanding of environmental conditions and potential impacts on human health and ecosystems.
3. **Prediction and Forecasting Enhancement**: While the existing setup may provide weather predictions and forecasts based on conventional parameters, the proposed model enhances forecasting capabilities by considering oxygen levels. Changes in oxygen levels can indicate variations in air quality, pollution levels, and overall environmental health, which can influence weather patterns and human well-being.
4. **Comprehensive Data Analysis**: With the inclusion of oxygen level data, the proposed model enables more comprehensive data analysis and predictive modeling. By analyzing correlations between weather parameters and oxygen levels, the system can provide insights into complex atmospheric processes and phenomena, improving the accuracy of forecasts and predictions.

Overall, the proposed model offers a more advanced and sophisticated approach to weather monitoring and forecasting by integrating oxygen level monitoring. This integration enhances the system's ability to provide valuable insights into environmental conditions and their potential impacts on various aspects of life and ecosystems.

4. EXPERIMENTAL ANALYSIS

Combining the K-Means algorithm with a genetic algorithm in a hybrid approach for web data segmentation is an innovative strategy. This hybridization leverages the strengths of both algorithms to potentially improve the accuracy and efficiency of clustering in the vector space model.

Here's how the proposed algorithm works:

1. **Data Preparation**: The algorithm begins by preparing the input data, typically represented in a vector space model. This model represents each web document as a vector in a high-dimensional space, where each dimension corresponds to a feature or term frequency.

2. **Determining Number of Clusters**: The number of clusters generated by the algorithm is derived from the vector space model. This could involve techniques such as analyzing the distribution of data points or using metrics like the silhouette score to determine the optimal number of clusters.

3. **Hybridization**: The K-Means algorithm is applied to the input data to initially partition it into clusters. However, instead of relying solely on the random initialization of centroids, the genetic algorithm is employed to refine these initial clusters and optimize their quality.

4. **Genetic Algorithm Optimization**: The genetic algorithm operates on the clusters produced by K-Means, using genetic operators such as selection, crossover, and mutation to iteratively improve the clustering solution. The fitness function evaluates the quality of clusters based on

cohesion, separation, or other relevant metrics.

5. **Experimentation and Evaluation**: A series of experiments are conducted to assess the accuracy and effectiveness of the hybrid algorithm in web data segmentation. This involves evaluating clustering results against ground truth data or using metrics such as precision, recall, and F1 score to quantify performance.

Indeed, the figure referenced likely presents the results of experiments conducted to evaluate the accuracy and effectiveness of the proposed hybrid algorithm for web data segmentation. These results would demonstrate the algorithm's performance in generating meaningful clusters from web data, highlighting its potential advantages over traditional clustering methods.

The benefits of the hybrid approach, which combines K-Means with a genetic algorithm and derives the number of clusters from the vector space model, are multifold:

1. **Improved Accuracy**: By integrating genetic algorithms, which can optimize cluster quality based on predefined fitness functions, the hybrid approach may yield more accurate clustering results compared to conventional K-Means.

2. **Enhanced Efficiency**: The use of genetic algorithms can help overcome K-Means' sensitivity to initial centroid placement, potentially reducing the number of iterations needed to converge to optimal clusters and improving computational efficiency.

3. **Adaptability to Data Complexity**: Deriving the number of clusters from the vector space model allows the algorithm to adapt to the complexity and structure of the input data, ensuring that the clustering process is tailored to the specific characteristics of web data.

4. **Robustness to Noise and Outliers**: Genetic algorithms' ability to explore diverse solutions and mitigate the impact of noise and outliers in the data can enhance the robustness of the clustering process, leading to more reliable segmentation results.

Overall, the proposed hybrid algorithm represents a promising approach to web data segmentation, offering the potential to outperform traditional clustering methods in terms of both accuracy and efficiency. The results presented in the figure would provide empirical evidence supporting the effectiveness of the hybrid approach and its suitability for various web data analysis tasks..

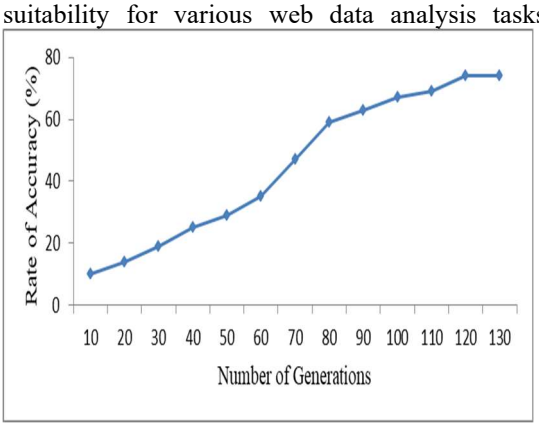

Fig 8 Accuracy Rate

The utilization of various crossover probabilities across different iterations is a method commonly employed in genetic algorithms to explore different search spaces and optimize solutions effectively. The performance evaluation, particularly focusing on the average mean Cp as described in Figure 11, likely provides valuable insights into the algorithm's behavior and effectiveness under different crossover probability settings.

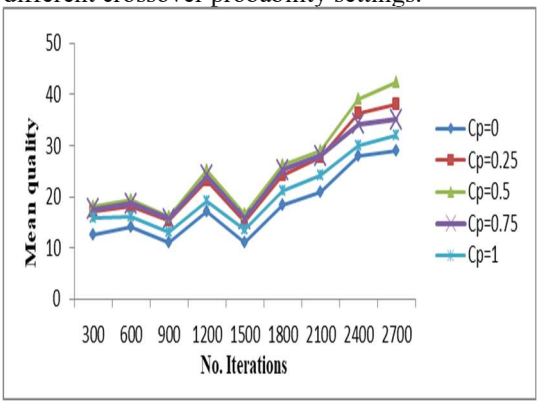

Fig 9 Mean Quality

The statement indicates that while the proposed hybrid algorithm may take slightly more time compared to the standard K-Means algorithm, it outperforms in generating optimal segments. This suggests that the trade-off in computational time is justified by the improved quality of the clustering results achieved by the hybrid approach.

Here's a breakdown of what this statement implies:
1. **Time Consideration**: Despite taking a bit longer to execute than standard K-Means, the hybrid algorithm is still considered acceptable in terms of computational time. The additional time required for the hybrid approach is likely due to the

incorporation of genetic algorithms, which involve additional iterations and computational overhead compared to the straightforward implementation of K-Means.

2. **Optimal Segmentation**: The primary advantage of the hybrid algorithm is its ability to generate optimal segments, surpassing the performance of standard K-Means. This suggests that the hybrid approach effectively leverages the genetic algorithm's optimization capabilities to produce higher-quality clusters, possibly with better cohesion, separation, or overall clustering accuracy.
3. **Visualization in Fig**: The statement refers to the visualization provided in the figure, which likely depicts a comparison between the clustering results obtained by the hybrid algorithm and those of standard K-Means. The figure would likely showcase the superiority of the hybrid approach in terms of cluster quality or another relevant performance metric.
Overall, while the hybrid algorithm may require slightly more computational time than standard K-Means, its ability to generate optimal segments justifies this trade-off, making it a preferable choice for web data segmentation tasks where accuracy and clustering quality are paramount.

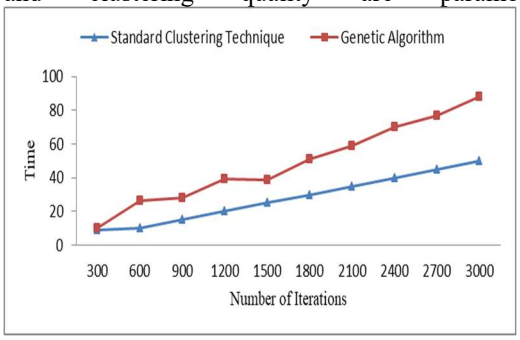

Fig 10 Processing Time

5.CONCLUSIONS:

The introduction and exploration of various aspects of web mining, as described in your summary, highlight its importance as a multidisciplinary field that aims to extract valuable insights from unstructured web documents. Here's a breakdown of the key points mentioned:

1. **Multidisciplinary Nature**: Web mining draws upon techniques from multiple disciplines, including statistics and machine learning, to analyze and extract useful information from the vast amount of unstructured data available on the web. This interdisciplinary approach allows researchers to leverage diverse methods and tools to address the challenges posed by web data analysis.

2. **Combination of Techniques**: The proposed method combines the K-Means clustering algorithm with genetic algorithms to segment web documents effectively. This hybrid approach harnesses the strengths of both techniques to optimize the segmentation process and generate actionable intelligence from web data. By integrating these techniques, organizations can gain valuable insights into their web content and make informed decisions based on segmented data.

3. **Addressing the Growing Volume of Web Data**: With the ever-increasing volume of web data, there is a growing need for more sophisticated web mining techniques. The proposed framework aims to meet this need by providing organizations with a comprehensive approach to segmenting web documents and extracting valuable information efficiently.

4. **Potential for Extension and Exploration**: The proposed framework can be extended to incorporate other mining algorithms and machine learning techniques, offering flexibility and adaptability to different research contexts and applications. By using a genetic approach, the framework can be further enhanced to accommodate a wide range of data mining tasks and challenges.

Overall, the paper contributes to the advancement of web mining research by introducing a comprehensive framework for segmenting web documents and exploring various aspects of the field. It underscores the importance of leveraging advanced techniques and interdisciplinary approaches to extract meaningful insights from the vast and complex world of web data.

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
[8] Information Reuse and Integration, pp.175-180, 2016.
[9] 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.
[10] 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.
[11] 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.
[12] 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.
[13] 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.
[14] Anjali Ganesh Jivani, "A Comparative Study of Stemming Algorithms", Int. J. Comp. Tech. Appl.(IJCTA), Vol 2 (6), pp: 1930-1938, 2011.
[15] 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.
[16] 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.
[17] 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]
[18] 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]
[19] 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.
[20] Brijendra Singh, Hemant Kumar Singh, "Web Data Mining Research: A Survey", 978-1-4244-5967-4/10, IEEE, pp: 1-10, 2010.
[21] 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]
[22] 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]
[23] 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.
[24] 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.
[25] 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]
[26] 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.

[27] 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.

[28] 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]

[29] 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.

[30] 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.

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

[32] 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.

[33] 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.

[34] 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.

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

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

[36]M. V. Subbarao, J. T. S. Sindhu, Y. C. A. Padmanabha Reddy, V. Ravuri, K. P. Vasavi and G. C. Ram, "Performance Analysis of Feature Selection Algorithms in the Classification of Dry Beans using KNN and Neural Networks," 2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), Erode, India, 2023, pp. 539-545, doi: 10.1109/ICSCDS56580.2023.10104809.

[37] M V Ganeswara Rao, P Ravi Kumar, T Balaji, " A High Performance Dual Stage Face Detection Algorithm Implementation using FPGA Chip and DSP Processor " , Journal of Information Systems and Telecommunication (JIST), 2022, pp 241-248, doi: 10.52547/jist.31803.10.40.241

.