CINEFORUM
ISSN: 0009-7039
Vol. 64. No. 3S, 2024

A Comparison of Particle Swarm Optimization and its Normalized Variant

Nainika Kaushik[1]; Dr Manjot Kaur Bhatia[2]

[1] Department of Engineering and Technology, PhD Scholar, Jagannath University, Jaipur, India. Email: nainikakaushik13@gmail.com

[2]Department of Master of Computer Application, Professor, Jagan Institute of Management studies, Sector-5, Rohini, Delhi, India

Corresponding Author: Nainika Kaushik

Abstract: Particle Swarm Optimization is a metaheuristic optimization algorithm inspired by the social behaviour of bird flocks and fish schools, which has been widely used to solve complex optimization problems in various fields, including engineering, science, and technology. Over the years, numerous variants of the PSO algorithm have been proposed to address specific limitations or improve the performance of the original algorithm. This paper focuses on a different approach for efficiently mining and searching web content to provide users with meaningful results. We plan to apply support vector machine techniques and a variant of the Particle Swarm Optimization algorithm for web content search and retrieval, aiming to deliver efficient and high-quality outcomes. Here, we compare the results of the standard PSO method with the variant PSO technique and analyse the quality of the results in terms of accuracy. The normalized particle swarm optimization algorithm demonstrates superior performance in terms of accuracy compared to the standard particle swarm optimization approach.

Keywords: particle swarm optimization, Stemming, search engines, stop word removal, Support Vector machine

1.Introduction

Search engines have become the primary tool for users to access useful information on the internet. However, the search results provided by even the most widely used search engines are often unsatisfactory. This is because while users input appropriate keywords, the majority of the returned pages are irrelevant. Developing effective web search mechanisms requires addressing two key challenges: how to extract web pages that are

relevant to the user's interests, and how to rank these potentially related pages according to their relevance. Evaluating the effectiveness of a web search approach necessitates measures of semantic similarity. Traditionally, users have provided manual assessments of relevance or semantic similarity, but this process is very laborious and expensive.

The study of semantic similarity between words has been a crucial aspect of information retrieval and natural language processing. Semantic similarity is a concept whereby a set of terms within term lists are assigned a metric based on the degree of likeness in their meaning. Measuring the semantic similarity between words is a vital component in various web-based tasks, including relation extraction, community mining, document clustering, and automatic meta-data extraction. Furthermore, the immense amount of data involved in processing user queries necessitates that search engines employ caching techniques to provide users with a list of relevant and authoritative results promptly.

Modern search engines often utilize caching techniques to enhance query response times and optimize the use of available storage. One approach is to cache posting lists of frequently queried terms, while an alternative method involves storing entire result pages or individual documents. The latter approach presents the advantage of enabling the same document to serve multiple queries, thereby using cache space efficiently. However, this method also carries the drawback of longer query response times, as the result pages must be constructed using the cached documents. Search-result caches can be static, subject to periodic updates based on historical data, or dynamic, storing results according to the query sequence. Additionally, a combination of both static and dynamic caching techniques may be employed. The vast information available on the World Wide Web poses a challenge in identifying relevant content, which web mining aims to address. Web mining involves the application of machine learning and data mining techniques to automatically extract meaningful patterns and relationships from large collections of web data. This field can be subdivided into three main areas: web content mining, web structure mining, and web usage mining, which collectively work to extract knowledge from web-based data.

2. Methods

The proposed system employs a web crawler to gather data from the web. Subsequently, text pre-processing is applied to the database to remove irrelevant data, and the term

frequency of each term is calculated. The relevancy of the data is then evaluated using a support vector machine, and the results are optimized utilizing particle swarm optimization variant. The system's workflow is illustrated in Figure 1.

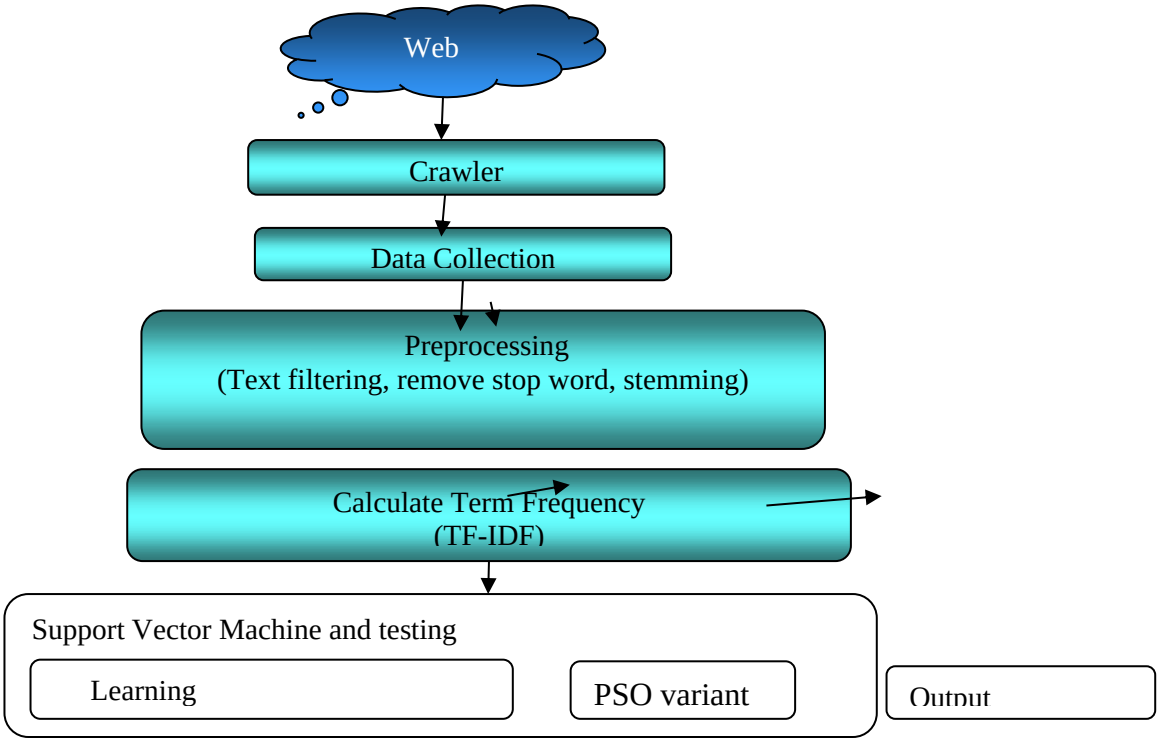


Figure 1 Workflow of the System.

Methodology

1 SUPPORT VECTOR MACHINE(SVM)

Support Vector Machines are a class of supervised learning algorithms widely used for classification and regression tasks. SVM operates by finding a hyperplane that best separates the data into distinct classes. The goal is to identify the hyperplane that maximizes the margin between the positive and negative examples.

To determine the most relevant web pages, we can use a Support Vector Machine to classify the pages as relevant or irrelevant. This approach has been explored in previous research and has shown promising results in improving the accuracy of web search engines.((Kaushik et al., 2020) The use of SVMs in document ranking tasks has been investigated, with researchers finding that while SVMs are not necessarily better than other machine learning methods, they can be effectively applied to the problem of

learning how to rank documents in ad hoc retrieval (Manning et al., 2008).SVM approaches have the advantage of being able to handle high dimensional feature spaces, making them well-suited for text classification tasks (Naidu et al., 2014).Support vector machine classification has also been employed in other web mining tasks, such as extracting relevant content from Indian web pages (Prakash & Raman, 2018). Support vector machine techniques can be applied to the problem of identifying relevant web pages for a given query. SVM-based approaches have demonstrated their effectiveness in improving the accuracy and performance of web search systems.

2 PARTICLE SWARM OPTIMIZATION(PSO)

A brief introduction to Particle Swarm Optimization

Particle Swarm Optimization is a population-based optimization algorithm inspired by the social behavior of bird flocking or fish schooling. In PSO, a set of candidate solutions, called particles, move through the search space, guided by their own best-known position and the overall best-known position. At each iteration, the velocity and position of each particle are updated based on the particle's own best position and the global best position. Researchers have developed various modified versions of the basic PSO algorithm, known as nPSO (nonlinear particle swarm optimization) (Xu et al., 2021)(Sengupta et al., 2018). These nPSO algorithms often incorporate additional mechanisms to improve the algorithm's exploration and exploitation capabilities, such as incorporating inertia weights, constriction factors, or using different update equations. The nPSO approach has been applied to a variety of optimization problems, including web page ranking and document retrieval tasks. Compared to the standard PSO, the nPSO variants have demonstrated improved performance in terms of convergence speed and solution quality. The nPSO (modified Particle Swarm Optimization) approach has been applied to a wide range of optimization problems, including web page ranking and document retrieval tasks. Compared to the standard PSO algorithm(Kaushik & Bhatia,2020), the nPSO variants have demonstrated improved performance in terms of convergence speed and solution quality. These nPSO algorithms often incorporate additional mechanisms, such as inertia weights or constriction factors, to enhance the algorithm's exploration and exploitation capabilities, leading to better optimization results across various applications.

The nPSO approach has been applied to a wide range of optimization problems, including web page ranking and document retrieval tasks. Compared to the standard PSO algorithm, the nPSO variants have demonstrated improved performance in terms of convergence speed and solution quality. These nPSO algorithms often incorporate additional mechanisms, such as inertia weights or constriction factors, to enhance the algorithm's exploration and exploitation capabilities, leading to better optimization results across various applications. The incorporation of these advanced techniques in the nPSO algorithms has enabled them to outperform the standard PSO in many real-world optimization scenarios, making them a more effective choice for tasks such as web search and information retrieval.

The PSO Algorithm

The Particle Swarm Optimization algorithm involves several key steps. First, the algorithm initializes with a group of random particles and then searches for optimal solutions by updating generations. In each iteration, each particle is updated by following two "best" values. The first is the best solution the particle has achieved thus far, referred to as the personal best or "pbest". The second "best" value tracked by the particle swarm optimizer is the globally best value obtained so far by any particle in the population, known as the global best or "gbest".

Following the identification of the personal best and global best solutions, the particle swarm optimization algorithm proceeds to update the particle's velocity and position using the corresponding mathematical equations(1,2). This iterative process continues until a certain stopping criterion is met, such as a predefined number of iterations or when a satisfactory level of performance is achieved.

$$v[\,] = v[\,] + c1 * rand(\,) * (pbest[\,] - present[\,]) + c2 * rand(\,) * (gbest[\quad] - present[\quad]) \quad (1)$$

$$present[\,] = present[\,] + v[\,] \quad (2)$$

The particle's velocity is denoted as v, and the current particle's position is represented as present. The term 'rand' refers to a randomly generated number within the range of 0 and 1. The parameters c1 and c2 are known as learning factors. In most cases, these are assigned values of c1 = 2 and c2 = 2.

A new parameter "n" is incorporated into the algorithm. This parameter typically decreases linearly from 0.99 to 0.1 over the course of the iterations, simulating the progress of the particles toward the target. However, due to the randomness involved, the value of "n" is compared to a randomly generated value. Since "n" starts at a maximum value, it is more likely to be greater than the random value, especially in the early stages of the iteration. This allows the particles to move at their normal rates during the initial phases of the optimization process. If "n" is less than the random value, then the values of "o" and "s" are updated accordingly, which in turn updates the particle's velocity "v" and position "present".

The "o" parameter in the algorithm is a randomly generated number, which is set to 0.3 in this study. As the iteration progresses, the value of the "n" parameter, which typically decreases linearly, is more likely to be smaller than the randomly generated value. This, in turn, leads to updates in the values of "o" and "s", which subsequently alter the particle's velocity "v" and position "present". Additionally, the "d" parameter, which is set between 0 and 1 and taken as 0.95 in this research, serves to reduce the particle's velocity to a certain extent and limit the changes in its position. These calculations are performed to gradually decrease the particle's speed during the optimization process.

Advantages of the PSO algorithm :

The Particle Swarm Optimization algorithm offers several advantages, including its simplicity in implementation, its ability to handle non-linear and non-differentiable objective functions, and its suitability for parallel computing. Furthermore, PSO has demonstrated its effectiveness in solving a wide range of optimization problems, from function optimization to real-world applications such as job scheduling and power system optimization.

The performance of the standard PSO algorithm, however, can be affected by premature convergence, which can lead to suboptimal solutions. To address this issue, various modifications and extensions to the original PSO algorithm have been proposed, including the introduction of a normalized PSO (nPSO) variant.

3. Experiments

The user interface initially presents three tabs located (Figure 2) in the top left corner of the screen:

- Admin
- Search Engine
- Exit

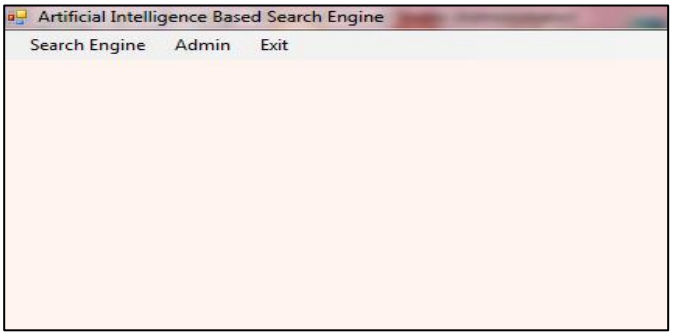


Figure 2: UI Of the System

The updated user interface comprises the following tabs(figure 3):

- Crawl Data
- View Data
- Add Stop Words
- Exit

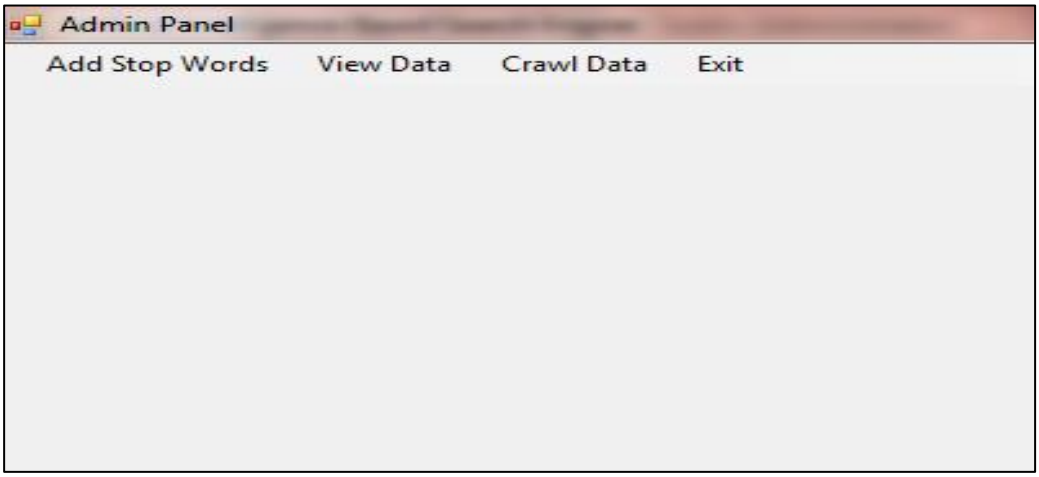
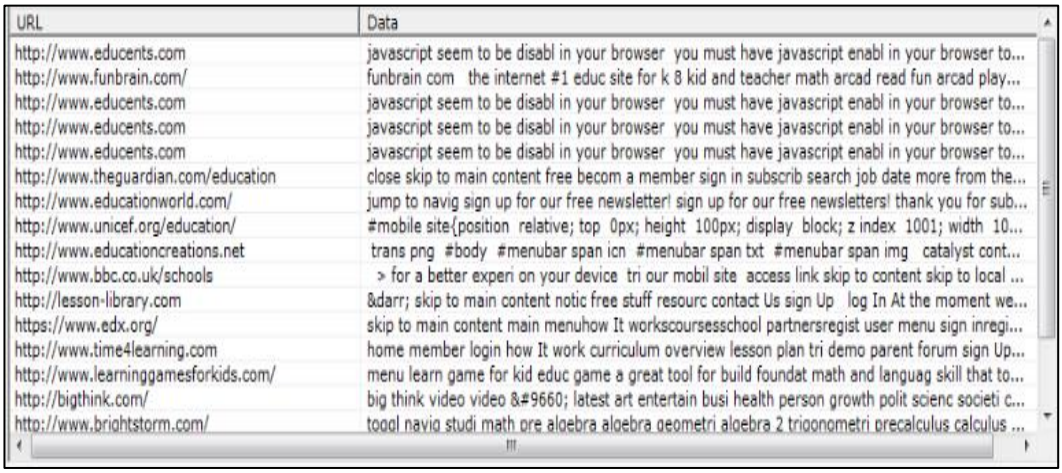


Figure 3: Admin Panel UI

If the user selects the "Add Stop Words" option, a pop-up window will appear, prompting the user to input a stop word. The system currently utilizes a predefined list of commonly used stop words. However, if a researcher wishes to exclude a specific word that is not included in the built-in stop word list, they can easily add their custom stop word by entering it and clicking the "Save" button.

The "View Data" tab allows users to access a list of links and their associated data stored in the database(Figure 4). When this tab is selected, the system will display the collection of URLs along with the corresponding information connected to each web address.



URL	Data
http://www.educents.com	javascript seem to be disabl in your browser you must have javascript enabl in your browser to...
http://www.funbrain.com/	funbrain com the internet #1 educ site for k 8 kid and teacher math arcad read fun arcad play...
http://www.educents.com	javascript seem to be disabl in your browser you must have javascript enabl in your browser to...
http://www.educents.com	javascript seem to be disabl in your browser you must have javascript enabl in your browser to...
http://www.educents.com	javascript seem to be disabl in your browser you must have javascript enabl in your browser to...
http://www.theguardian.com/education	close skip to main content free becom a member sign in subscrib search job date more from the...
http://www.educationworld.com/	jump to navig sign up for our free newsletter! sign up for our free newsletters! thank you for sub...
http://www.unicef.org/education/	#mobile site{position relative; top 0px; height 100px; display block; z index 1001; width 10...
http://www.educationcreations.net	trans png #body #menubar span icn #menubar span txt #menubar span img catalyst cont...
http://www.bbc.co.uk/schools	> for a better experi on your device tri our mobil site access link skip to content skip to local ...
http://lesson-library.com	↓ skip to main content notic free stuff resourc contact Us sign Up log In At the moment we...
https://www.edx.org/	skip to main content main menuhow It workscoursesschool partnersregist user menu sign inregi...
http://www.time4learning.com	home member login how It work curriculum overview lesson plan tri demo parent forum sign Up...
http://www.learninggamesforkids.com/	menu learn game for kid educ game a great tool for build foundat math and languag skill that to...
http://bigthink.com/	big think video video ▼ latest art entertain busi health person growth polit scienc societi c...
http://www.brightstorm.com/	toggl navig studi math pre algebra algebra geometri algebra 2 trigonometri precalculus calculus ...

Figure 4 : View Data

To add a new URL to the database, the user can navigate to the "Crawl Data" tab and click on it, which will display a new user interface(Figure 5). The user can then copy the desired web link from their browser and paste it into the provided space within this interface.

Figure 5: Crawl Data

Upon clicking the "Go!" button(Figure 5), the full data associated with the URL will be displayed. Next, the user can proceed to the crucial preprocessing stage. Once the user has accessed the data, they can initiate the preprocessing process by clicking the "Preprocessing" button(Figure 6). The preprocessing will include:

- Stemming (Using Stemming Porter Algorithm)
- Removal Of Stop Words

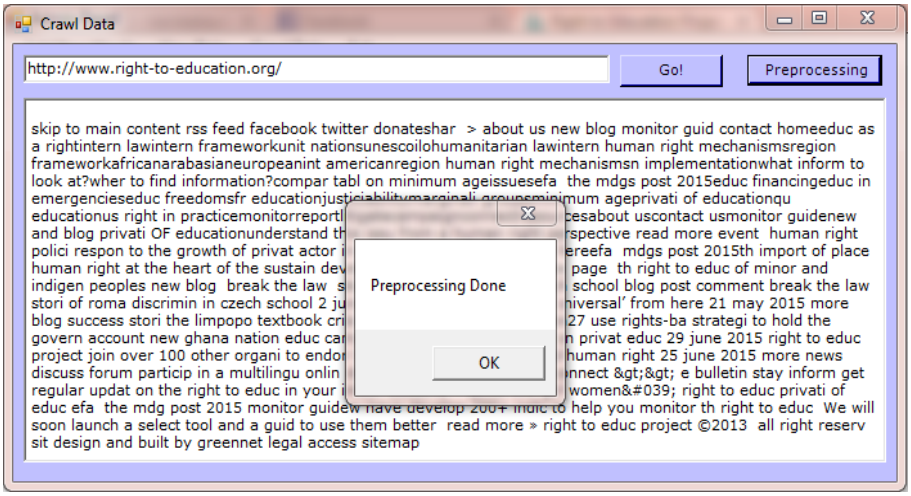


Figure 6: Preprocessing Data

Upon completion of the preprocessing stage, a notification will appear, informing the user that the preprocessing has been successfully executed. The user can then proceed to click the "OK" button to acknowledge the notification. Subsequently, the user can move forward to the next steps in the workflow.

Next, the user can select the "TF-IDF" option to compute the term frequency-inverse document frequency for each word(Figure 7).Upon clicking this button, the system will prompt the user to save the preprocessed data. Following this, the user can submit the preprocessed data to initiate the calculation of the word frequencies.

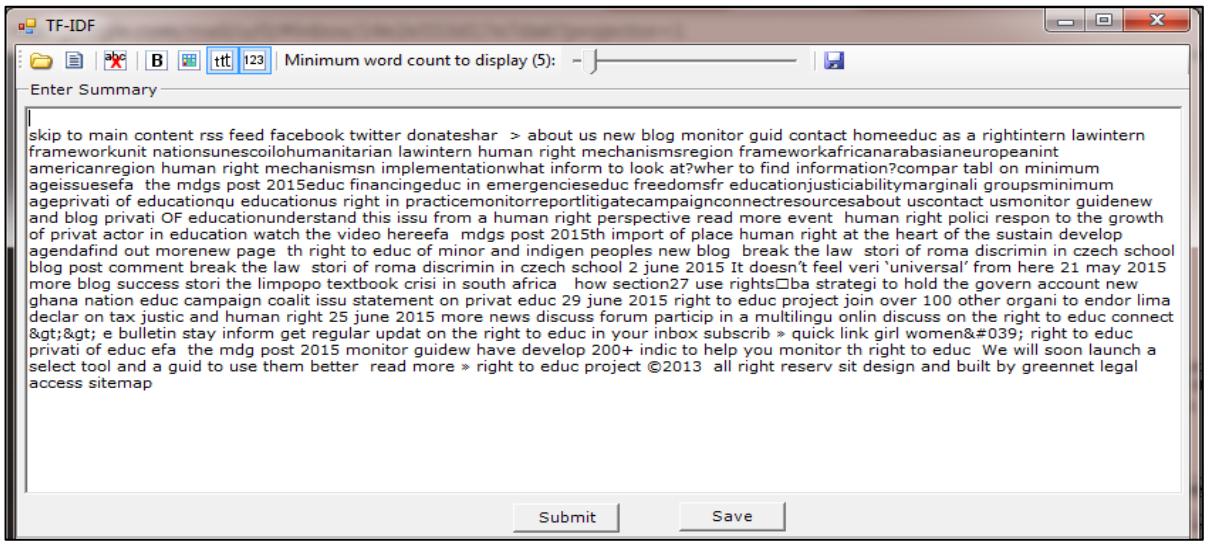


Figure 7: Submit For TF-IDF Calculation

Upon clicking the "Submit" button, the system will display a status window indicating the process of reading the number of words in the document. Once the preprocessing of

the data is complete, a visual representation in the form of an image will be presented, depicting the frequency distribution of the words within the preprocessed data (Figure 8.).

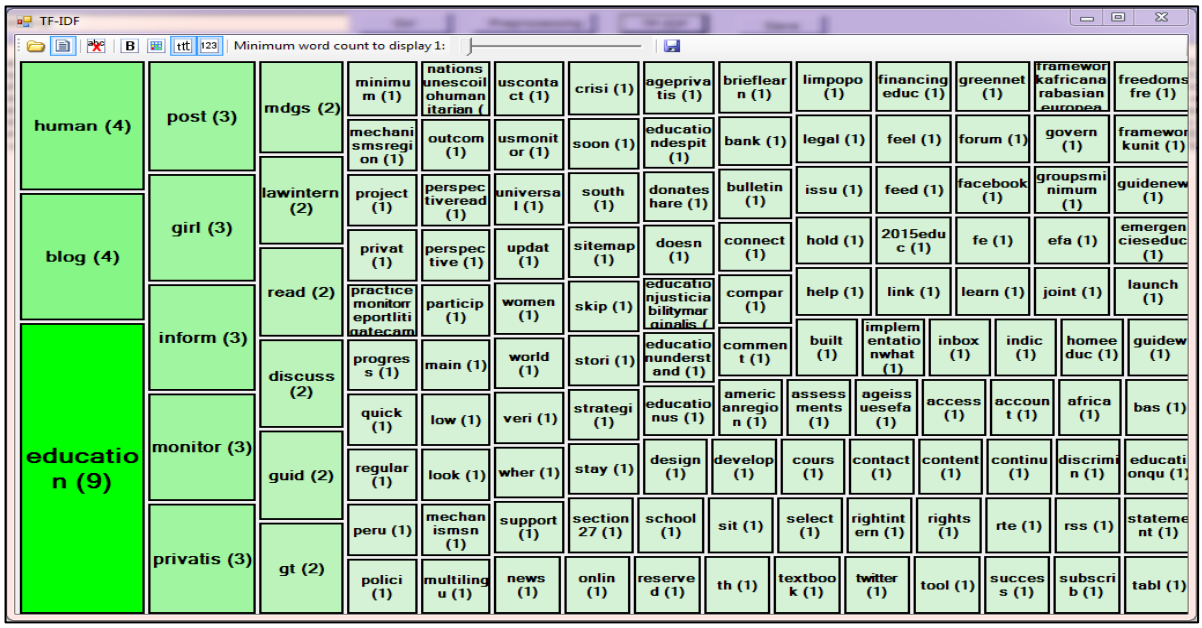


Figure 8 : TF-IDF Image

The final step in the process is to select the "Save" option, which will automatically store the URL and its associated data in the database (Figure 9).

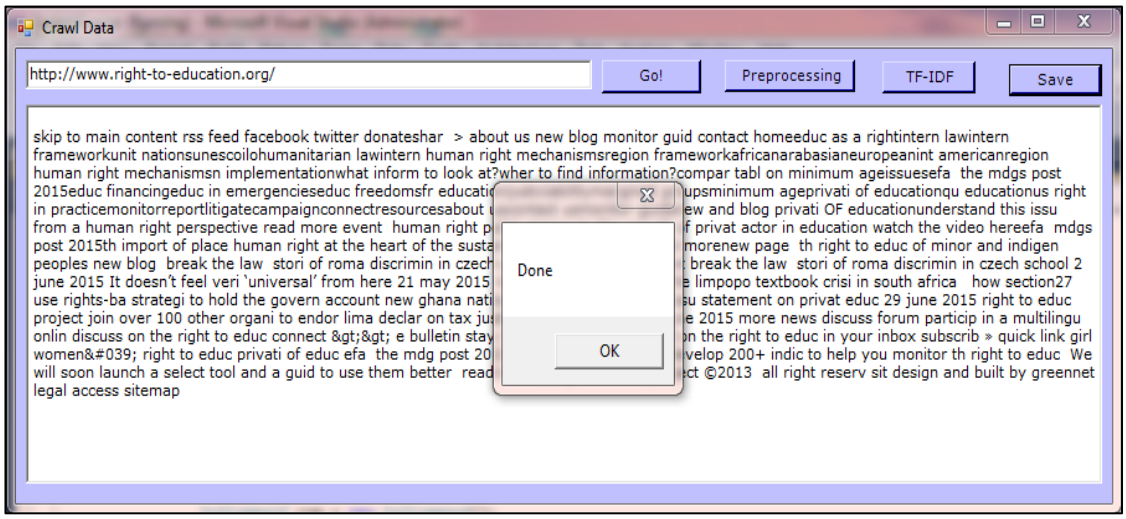


Figure 9: Save The Preprocessed Data

Users can verify the saved URL by navigating to the "View Data" tab, where the stored web address and associated information will be displayed in the database(Figure 10).

URL	Data
https://www.gov.uk/government/organisations/d...	skip to main content gov uk us
http://www.topuniversities.com/courses/enginee...	skip to main content topuniver
http://www.eng.cam.ac.uk/	skip to main content studi at ca
https://www.engineering.cornell.edu/	skip to main content search th
https://engineering.stanford.edu/	skip to main content sunet log
http://vlib.org/Engineering	engineering en es fr zh the w
http://www.engin.umich.edu/college/	michigan engin skip to main co
http://www.research.ibm.com/	select a country region unit st
http://www.topuniversities.com/courses/enginee...	skip to main content topuniver
http://engineering.illinois.edu/	class of 2015! src ws engr il
http://research.microsoft.com/en-us/	microsoft translat | research a
https://research.google.com/	skip to content home research
https://www.cs.washington.edu/	jump to navig univers of washi
http://www.right-to-education.org/	skip to main content rss feed f
http://www.right-to-education.org/	skip to main content rss feed f

Figure 10 : URL Check In Database

4. RESULT AND CONCLUSION

The user interface of the system is presented(Figure 11.), and the user can select the "Search Engine" tab.

Figure 11 : UI For Testing

Web search engines have widely adopted the keyword-driven search model. Upon user input, a new interface will appear, enabling the search of relevant keywords and the retrieval of corresponding links stored in the system's database(Figure 12).

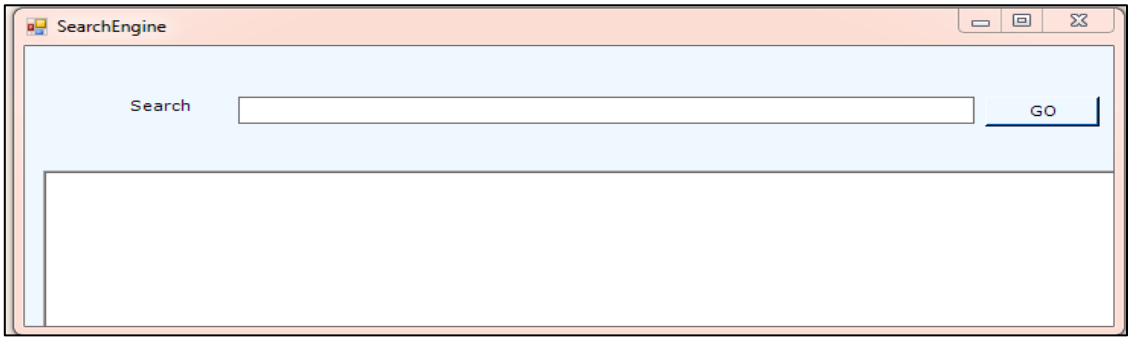


Figure 12 : Keyword-Based Search

In order to examine the functionality, let us conduct a test by searching for a relevant keyword, such as "education", and then clicking the search button (Figure 13.).

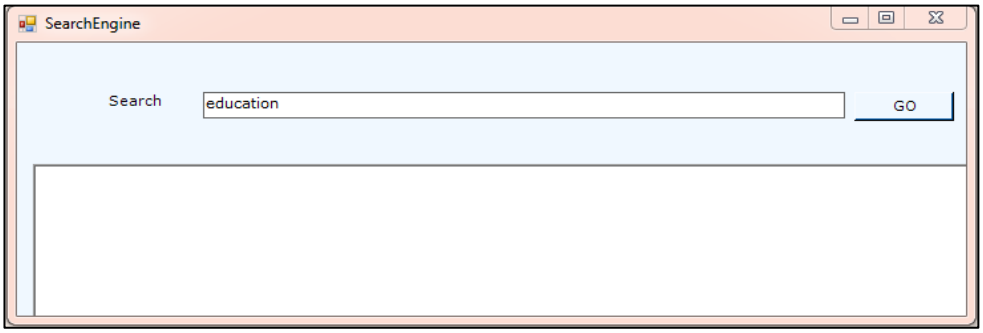


Figure 13 : Example 1

The system retrieves a number of URLs relevant to the search query "education" and ranks them according to their priority (Figure 14.). This prioritization is achieved through the implementation of a Support Vector Machine and a normalized variant of the Particle Swarm Optimization algorithm.

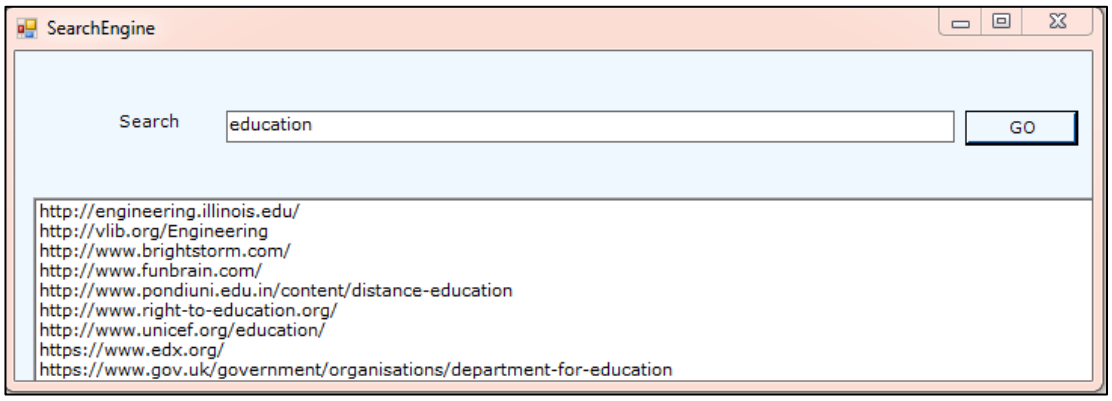


Figure 14 : Output

The proposed system's output was evaluated against the performance of the standard particle swarm optimization algorithm to assess its accuracy.

After completing a search, the prevailing question in every searcher's mind is whether they have found the most relevant information or if they are missing crucial items. Additionally, searchers hope to avoid retrieving an excessive amount of irrelevant content. While obtaining comprehensive results while avoiding extraneous information is challenging, if not unfeasible, it is possible to evaluate the performance of a search

strategy using two primary metrics: precision and recall. Precision represents the ratio of relevant records retrieved to the total number of records retrieved, both relevant and irrelevant. Recall denotes the proportion of relevant records retrieved out of the total number of relevant records in the database. These metrics are typically expressed as percentages and are fundamental in assessing the effectiveness of search strategies.

$$\text{Precision} = \frac{t_p}{t_p + f_p} \qquad \text{Recall} = \frac{t_p}{t_p + f_n}$$

$$\text{Accuracy} = \frac{t_p + t_n}{t_p + t_n + f_p + f_n}$$

where, t_p = true positive,

t_n = true negative,

f_p = false positive,

f_n = false negative.

The findings indicate that the standard particle swarm optimization algorithm achieved an accuracy rate of 87.95%. In comparison, combining the normalized variant of PSO and a support vector machine model improved accuracy by 93.68% (Figure 15). This suggests that the normalized variant of PSO outperforms the standard PSO implementation in terms of accuracy.

In summary, the proposed system leverages a normalized variant of the particle swarm optimization algorithm in conjunction with a support vector machine model to enhance the relevance and ranking of search results.

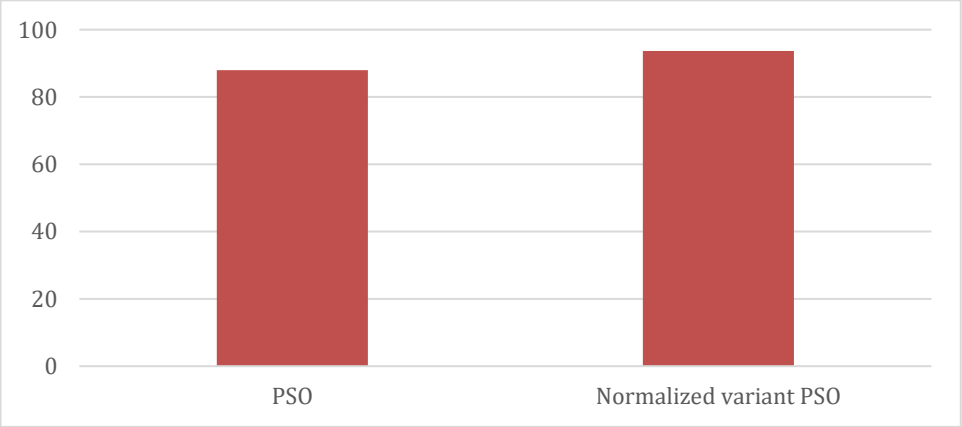


Figure 15: Comparing Accuracy Of standard PSO algorithm against the normalized variant PSO

REFERENCES

1. Manning, C D., Raghavan, P., & Schütze, H. (2008, July 7). Support vector machines and machine learning on documents. Cambridge University Press, 293-320. https://doi.org/10.1017/cbo9780511809071.016
2. Naidu, K B., Dhenge, A., & Wankhade, K. (2014, April 1). Feature Selection Algorithm for Improving the Performance of Classification: A Survey. https://doi.org/10.1109/csnt.2014.99
3. Prakash, K B., & Raman, A R. (2018, July 4). Data Engineered Content Extraction Studies for Indian Web Pages. Springer Nature, 505-512. https://doi.org/10.1007/978-981-10-8055-5_45
4. Xu, Q., Wu, T., & Wei-wei, W. (2021, January 1). Nonlinear Dissipative Particle Swarm Algorithm and Its Applications. Institute of Electrical and Electronics Engineers, 9, 158862-158871. https://doi.org/10.1109/access.2021.3131167
5. Kaushik, N., & Bhatia, M. K. (2020). Information Retrieval from Search Engine Using Particle Swarm Optimization. Advances in Computing and Intelligent Systems, 127.
6. Kaushik, N., Bhatia, M. K., & Rastogi, S. (2020). SVM and cross-validation using R studio. Int. J. Eng. Adv. Technol, 10, 46-54.