

A Robust Web Information Retrieval using PSO Approach Accelerated by Simulated Annealing

Dr. Ramya C¹, Dr. Hamsha K² and Ms. Savitha Channagoudar³

¹Department of Information Science and Engineering, CMR Institute of Technology, Bengaluru, Karnataka, India
Email: ramya.c@cmrit.ac.in

²Department of Computer Science, NMKRV College for Women, Bengaluru, Karnataka, India
Email: hamsha.nmkrv@rvei.edu.in

³Department of Information Science and Engineering, Sri Taralabalu Jagadguru Institute of Technology, Ranebennur, Karnataka, India
Email: savitha.chennegodar@gmail.com

Abstract—The process of information retrieval has evolved significantly with the usage of web. But unlike the traditional information retrieval set up, web information retrieval system succumbs to several problems. For resolving the issues and to provide an efficient and effective retrieval of web documents, a Modified Particle Swarm Optimization is proposed. This includes the Simulated Annealing technique which is the probabilistic approach that predicts the global optimization in the large search space along with the PSO. The documents similarity values are annealed and processed with the Simulated Annealing algorithm. The particles which are generated through this are then optimized with the PSO algorithm. The documents retrieved are measured for their correctness and accuracy based on several performance metrics that include precision and recall and response time as well. Experimental results clearly exhibit the superior performance of the proposed system over the existing system.

Index Terms— Swarm Optimization, Web Information Retrieval, Similarity Measure, Precision and Recall.

I. INTRODUCTION

Web Information Retrieval (WIR) is a great subject today and moderately closes our daily life. It is the task of discovering appropriate data from a more significant collection of unstructured data. WIR systems currently turn out to be the foremost source of communication with the internet.

WIR faces a challenge in finding more relevant information due to unstructured characteristic of the data and the tremendous growth in the size of the information. As a result, the users in search of information face difficulties in obtaining specific information on the web. There are various causes which affect the efficient and effective retrieval of information. The classic Information Retrieval (IR) approaches fail to handle this abundant collection of web documents [1]. Users could obtain both similar and dissimilar documents as results for their query because of the fact that even non-relevant documents are assigned with high similarity scores.

A. Motivation

The ultimate aim of WIR is to provide improved systems that retrieve the most relevant information available on the Web to better satisfy users information needs. Also, WIR intends to avoid displaying of unwanted documents

focusing on cost reduction of IR systems and query response delay. There by WIR achieves its effectiveness and efficiency. This forms the motivation for the present research study on WIR systems.

To better address the above issues pertaining to the optimisation of WIR process, it is quiet essential to propose a modified version of Particle Swarm Optimization (PSO) algorithm importing the useful characteristics of SA into it, to provide solutions of good quality, speedy computation and to search for the global optimum solution and to reduce the query response time of the system, hence contributing to the efficiency of web information retrieval.

II. LITERATURE REVIEW

Numerous techniques and methods [2,3] have been recommended for computing the information retrieval accuracy. Bouadjenek et al. stated that the combination of IR in a social network is contributed by the taxonomy for the social network and analyzed the association of the IR-social network combination using IR [4]. The retrieving information in the area of the social network is the major issue. Ahamed et al. worked on the information retrieval on the basis of process complexity and measured the fundamental trade off factors for an intelligent framework for web search [5]. Thangaraj et al. designed an effective IR in the semantic web by highlighting that the IR with keywords is less accurate for a vast amount of data [6]. Palomino et al. deliberated that the capability of the WIR system to retrieve the data continuously from the vast dynamic source with higher precision, accuracy and quality would be a difficult process [7].

Chawla explained that information requirements of the user are imprecise or vague and are not aware of their exact information prerequisites [8]. Alloui et al., [9] presented an innovative PSO approach for retrieving the web information. It uses the detailed feedback to re-compute the user query and therefore enhances the count of appropriate outcomes. Deo et al., [10] presented a new IR system which uses PSO algorithm for optimization. In the proposed system, initially, all HTML tags are removed for extracting only the terms from web documents. Djenouri et al. [11] suggests a combined approach of K-means and DCI_Closed algorithms. K-means algorithm is to measure the similarity between documents and determine the centroids for the group of documents. Abualigah et al. proposed an innovative feature selection technique, viz., feature selection method using the PSO algorithm to explain the feature selection problem by creating a new subset of informative text features [12].

Nainika Kaushik [13] proposed a system using support vector machine technique in combination with Particle Swarm Optimization (PSO) algorithm for web content search to obtain the best results. Relan Simran Vinod used [14] K-Mean algorithm to split the collection of documents into similar clusters, and to explore the cluster of documents BSO (Bee Swarm Optimization) was chosen. In our approach, PSO is modified and strengthened by adding few best features of SA technique into it. So the Modified Particle Swarm Optimisation with Simulated Annealing (MPSOSA) is proposed to overcome the disadvantages of PSO, thus to have efficient and effective retrieval of documents hence optimising the WIR process.

III. PROPOSED MPSOSA ALGORITHM

The proposed system is formulated using a vector space model (VSM) [15], wherein both the queries and documents are expressed as vectors in the vector space. After preprocessing, all the distinct terms of queries and documents are assigned with term weights based on $tf \times idf$ scheme. The proposed MPSOSA is intended to learn the significance of query terms in the form of term weights over the documents in the collection. It explores the variety of zones of the documents space, thereby finding the most relevant documents which best match the query according to the user need.

SA metropolis acceptance rule is applied to determine whether to accept the newly found position or compute another possible position again. This acceptance of position is based on the difference between the fitness values of the new and old particle positions and metropolis acceptance rule as well. This enables the solution to leap out at local optima and lowers the vibration near the end of locating a solution, thus favours to boost the degree of convergence and the quality of the solution. The notional diagram of MPSOSA is shown in Fig. 1.

Besides, the process explores for the solutions in the direction of P_{best} and G_{best} because the metropolis acceptance rule either to consider or deny a new solution is formulated on the parameters such as the difference in fitness values and the present temperature. When a possible solution is not accepted by the metropolis criterion, then a new position is computed using PSO.

This process will be persistently performed as far as the upper limit of disturbances is reached or the new position is accepted by the metropolis acceptance rule. Thusly, the algorithm could explore solutions in many

tracks by increasing the probability of obtaining global optimum solution. Furthermore, reduction in computation time can be achieved exploiting one of the features of PSO i.e. parallel processing.

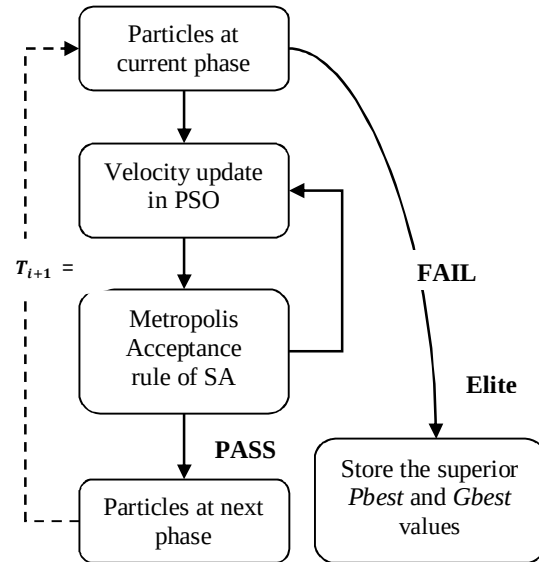


Figure 1. MPSOSA Notional Diagram

IV. RESULTS AND DISCUSSION

In order to verify the efficiency and effectiveness of the proposed system, the two different data collections CACM and RCV1 are used. CACM is an assemblage of 3,204 number of articles abstracts in HTML that were distributed in the ACM journal Communications between 1958 and 1979. Reuters Corpus Volume I (RCV1) is an accumulation of more than 8,04,414 XML documents having newswire stories. The system is tested over 100 different queries.

A. Efficiency Analysis

In the domain of WIR, Many works are available in the literature which employ PSO as the tool to optimise WIR process for effective and efficient retrieval. Hence to build an existing system, the bio inspired PSO is considered as a retrieval algorithm. Cosine similarity function is considered for query-document matching.

The goal is to reduce the response time of the system without compromising in the quality of the retrieved documents. The corresponding response time taken by the systems can be seen in Fig. 2 both for CACM in (a) and RCV1 collections in (b).

Overall, proposed system takes 159.8 ms and existing system takes 314.7 ms to retrieve the documents for the given set of queries and there is a reduction of 154.9 ms of response time by proposed system for CACM collection. Whereas for RCV1 collection, proposed system takes 300.9 ms and existing system takes 540.4 ms to retrieve the documents for the given set of queries and there is a reduction of 239.5 ms of response time by proposed system. It can be clearly noticed that proposed system consumes less time even though it retrieves more number of documents than existing system.

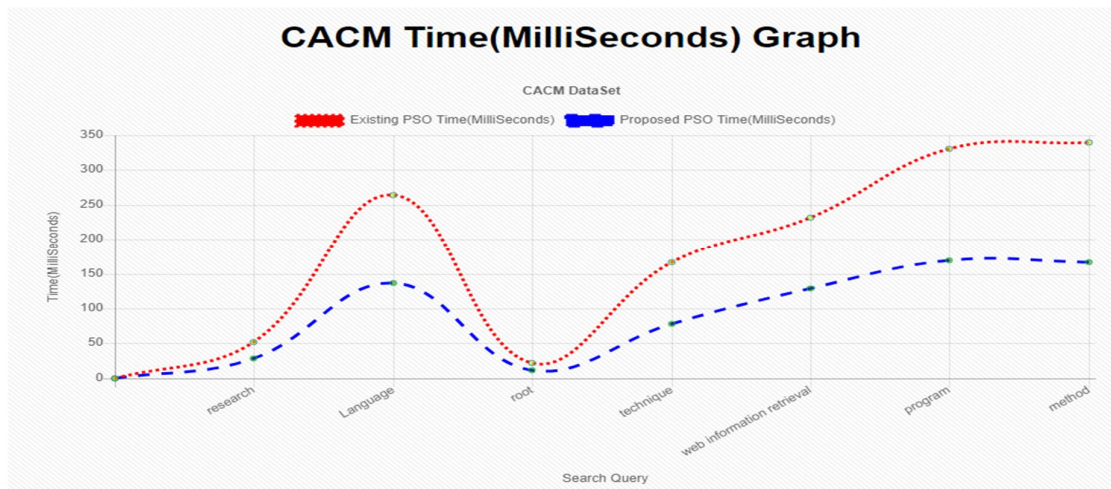
B. Effectiveness Analysis

To measure the effectiveness of the system, the following various parameters are used.

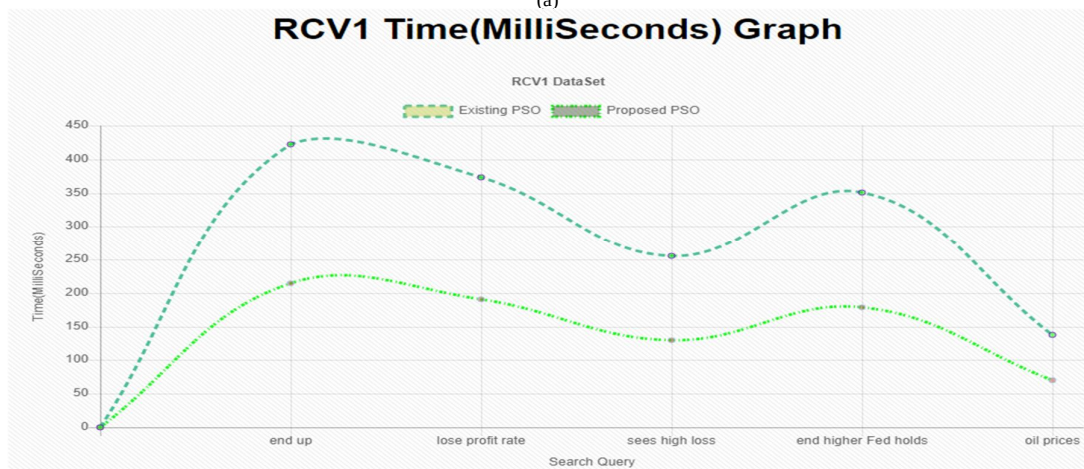
Precision and recall

Precision and recall quantitatively evaluate both the quality of the overall resultant documents and the breadth of the retrieval algorithm used, thus widely used as standard evaluation strategy in information retrieval [16]. Figs 3 and 4 show the Plot of precision and recall for single queries for both the systems over CACM and RCV1 respectively.

Experimental results showed that the retrieval performance obtained by the proposed system could constantly outperform the existing approach on different data collections in terms of the parameters response time and precision and recall. Hence the proposed system is significantly effective as it retrieves the most similar

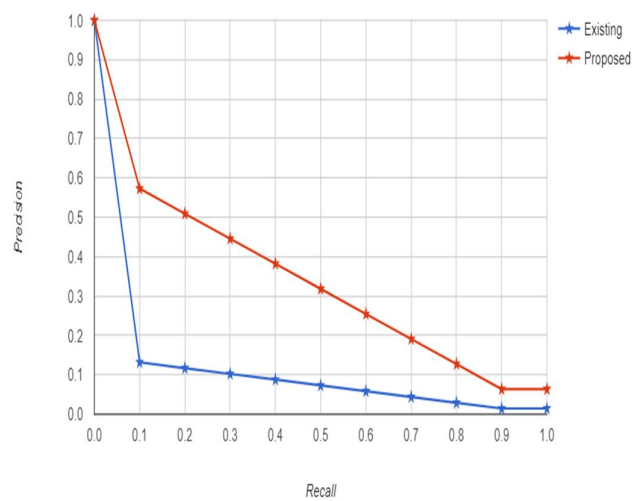


(a)

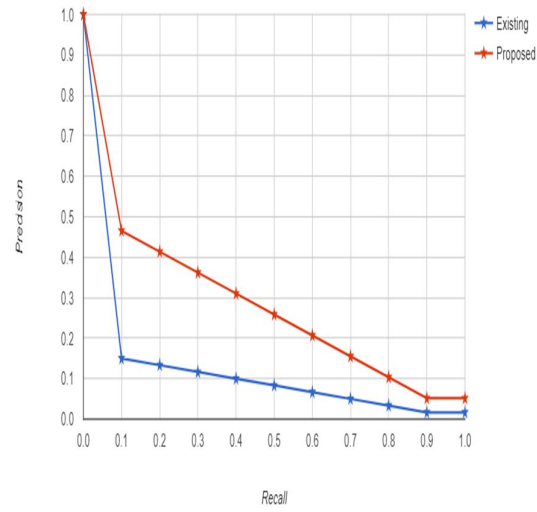


(b)

Figure. 2. Plot of query response time over (a) CACM and (b) RCV1 collections

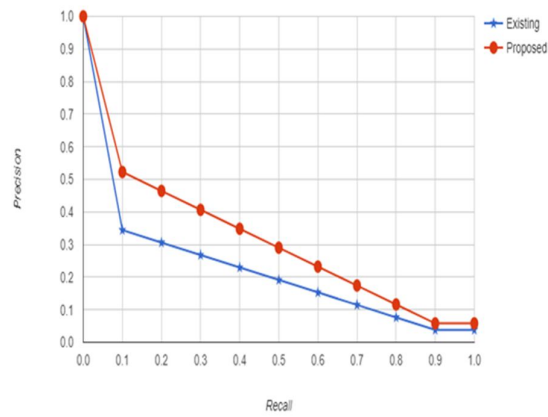


(a) Query: "product"

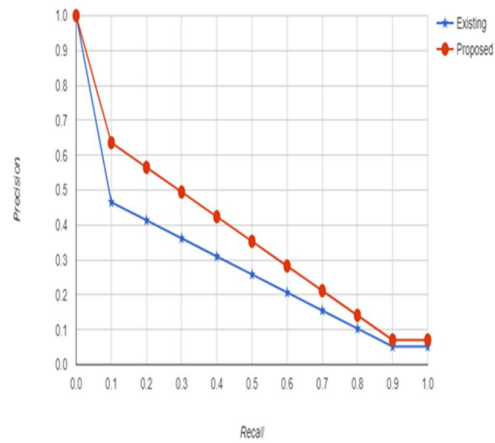


(b) Query: "root"

Figure. 3 Plot of precision and recall for single queries a) and b) for both systems over CACM



(a) Query: "see high loss"



(b) Query: "end up"

Figure. 4 Plot of precision and recall for single queries a) and (b) for both systems over RCV1

documents to the user and efficient as it covers most of the documents in the collection and retrieves more number of documents in considerably reduced response time.

V. CONCLUSION

The novel and effective framework for retrieving the document from the web is accomplished with the modified optimization algorithm MPSOSA. The datasets of CACM and RCV1 are employed to analyze the performance of the proposed framework. The input datasets are initially pre-processed and indexed. The query and documents are represented using VSM and all the terms are weighted. This MPSOSA is used as a retrieval algorithm in proposed system to optimise WIR process. The evaluation measures such as precision and recall are used to test the effectiveness of the system and to measure the efficiency response time for both the systems are compared. We can observe the superior performance of the proposed system in both the cases.

REFERENCES

- [1] Djenouri, Y., Belhadi, A., Fournier-Viger, P., and Lin, J. C. W., "Fast and effective cluster-based information retrieval using frequent closed itemsets", *Information Sciences*, Vol. 453, pp. 154-167, 2018.
- [2] Broder A, Kumar R, Maghoul F, and Ragavan P, "Graph structure in the Web source", *International Journal of Computation in Telecommunications Network*, Vol. 33, 2000.
- [3] Arasu A, Cho J, Garcia-Molina H, Paepcke A, and Raghavan S., "Searching the Web", *ACM Transactions on Internet Technology*, 2001.
- [4] Bouadjenek, M. R., Hacid, H., and Bouzeghoub, M., "Social networks and information retrieval, how are they converging? A survey, a taxonomy and an analysis of social information retrieval approaches and platforms", *Information Systems*, Vol. 56, 2016, pp. 1-18.
- [5] Ahamed, B. B., and Ramkumar, T., "An intelligent web search framework for performing efficient retrieval of data", *Computers & Electrical Engineering*, Vol. 56, pp. 289-299, DOI:10.1016/j.compeleceng.2016.09.033
- [6] Thangaraj, M., and Sujatha, G., "An architectural design for effective information retrieval in semantic web", *Expert Systems with Applications*, Vol. 41(18), pp. 8225-8233, 2014. DOI:10.1016/j.eswa.2014.07.017.
- [7] Palomino, M. A., Vincenti, A., and Owen, R. "Optimising web-based information retrieval methods for horizon scanning", *Foresight*, Vol. 15(3), 2013, DOI:10.1108/fs-10-2011-0045.
- [8] Chawla, S., "Effective Personalization of web search based on Fuzzy Information Retrieval", *International Journal of Computer Science and Information Technologies*, Vol. 6(3), pp. 2831-2837, 2015.
- [9] Alloui, T., Boussebough, I., and Chaoui, A. "A Particle Swarm Optimization Algorithm for Web Information Retrieval: A Novel Approach", *In Information Retrieval and Management: Concepts, Methodologies, Tools, and Applications*, pp. 1200-1216, 2018.
- [10] Deo, A., Gangrade, J., and Gangrade, S., "A Survey Paper On Information Retrieval System", *International Journal of Information retrieval*, Vol. 9(1). 2018.
- [11] Djenouri, Y., Belhadi, A., and Belkebir, R. "Bees swarm optimization guided by data mining techniques for document information retrieval", *Expert Systems with Applications*, Vol. 94, pp. 126-136, 2018.
- [12] Abualigah, L. M., Khader, A. T., and Hanandeh, E. S. "A new feature selection method to improve the document clustering using particle swarm optimization algorithm". *Journal of Computational Science*, Vol. 25, pp. 456-466, 2018. DOI:10.1016/j.jocs.2017.07.018.
- [13] Nainika Kaushik and Manjot Kaur Bhatia, "Information Retrieval from Search Engine Using Particle Swarm Optimization", *In book: Advances in Computing and Intelligent Systems*, pp.127-140, January 2020. DOI:10.1007/978-981-15-0222-4_11
- [14] Relan Simran Vinod, " Bee Swarm Optimization for Document Information Retrieval using Data Mining Technique", *International Journal of Modern Trends in Engineering and Research (IJMTER)* Volume 07, Issue 03, March -2020. DOI:10.21884/IJMTER.2020.7018.MYJ7H
- [15] R. Baeza-Yates and B. Ribeiro-Neto, "Modern Information Retrieval", *ACM Press*, ISBN: 0-201-39829-X, 2009.
- [16] Xie, Y., and O'Hallaron, D., "Locality in search engine queries and its implications for caching", *Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies*, Vol. 3, pp. 1238-1247, 2002.