International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-10, August 2019

A Bio-inspired Modified PSO Strategy for Effective Web Information Retrieval using RCV1 Datasets

Ramya C, Shreedhara K S

Abstract: ***Information retrieval is a key technology in accessing the vast amount of data present on today's World Wide Web. Numerous challenges arise at various stages of information retrieval from the web, such as missing of plenteous relevant documents, static user queries, ever changing and tremendous amount of document collection and so forth. Therefore, more powerful strategies are required to search for relevant documents. In this paper, a PSO methodology is proposed which is hybridized with Simulated Annealing with the aim of optimizing Web Information Retrieval (WIR) process. Hybridized PSO has a high impact on reducing the query response time of the system and hence subsidizes the system efficiency. A novel similarity measure called SMDR acts as a fitness function in the hybridized PSO-SA algorithm. Evaluations measures such as accuracy, MRR, MAP, DCG, IDCG, F-measure and specificity are used to measure the effectiveness of the proposed system and to compare it with existing system as well. Ultimately, experiments are extensively carried out on a huge RCV1 collections. Achieved precision-recall rates demonstrate the considerably improved effectiveness of the proposed system than that of existing one.***

Index Terms**: *Information Retrieval Systems, Web Information Retrieval, Particle Swarm Optimization, Similarity Functions, Documents Collection.***

I. INTRODUCTION

The exponential growth of information in the World Wide Web (WWW) results in a considerable problem within the academic world, but also for the organization and usability of information in the case of everyday needs. Great advances in information retrieval systems (IRS) have been achieved in the late 1990's related with the WWW and has been recognized as important criteria for the measurement of document relevance.

The research issues pertaining to WIR addressed in the present study are as follow:

1. How efficiently the IRS does accomplish its objectives? i.e., How to obtain reduced response time of the system as less as possible?
2. To what extent the system can retrieve relevant information while prohibiting irrelevant information. i.e., How to retrieve web documents effectively?
3. How a good precision and recall values of the system can be achieved?
4. What are the criteria to conduct the evaluation studies to compare the merits and demerits of two systems?

The main objective of the current study is to explore a WIR, which could retrieve most relevant documents to the users with a reduced response time hence devoting to effectiveness and efficiency of the system. More precisely, we aim at optimizing WIR process using PSO, hybridized with SA, in order to retrieve the documents with a plausible amount of time. The purpose of hybridizing PSO is to overcome the drawbacks of PSO such as its premature convergence and search process will slow down around global optimum. We focus on examining the efficiency of the system developed, using various evaluation metrics concerned to IR, with the intent of examining on a huge RCV1 standard corpuses using various queries.

To meet the above objectives, we develop a system that accepts the user query as input and uses hybridized PSO methodology for optimization of WIR process, hence contributes to the efficiency of the system reducing response time, and achieves good precision and recall metrics in WIR.

The rest of this paper is structured as follows: the next section describes the background of PSO hybridization with SA and similarity functions. Section III firstly gives a brief introduction to proposed system architecture, details of preprocessing, novel SMDR function and then elaborates on hybridized PSO-SA algorithm. Section IV describes experimental details and discusses experimental results. Finally, we conclude our work.

II. BACKGROUND

Our work amplifies a few domains of related work surveyed below, including particle swarm optimization, simulated annealing to hybridize PSO and Similarity functions for effective information retrieval.

Xiaoyu Song et al. [2] adopted PSO to construct the parallel initial solutions of SA to resolve job shop problem (JSP) and improve the quality of searching solutions. Enhanced Simulated Annealing (ESA) algorithm having ability to escape the local optimization is presented wherein shifting bottleneck procedure is introduced.

Manuscript published on 30 August 2019.
***Correspondence Author(s)**
Ramya C, Research scholar, Department of Studies in CS&E, UBDTCE, VTU, Davanagere, Karnataka, INDIA
Dr. Shreedhara K S, Professor, Department of Studies in CS&E, UBDTCE, VTU, Davanagere, Karnataka, INDIA

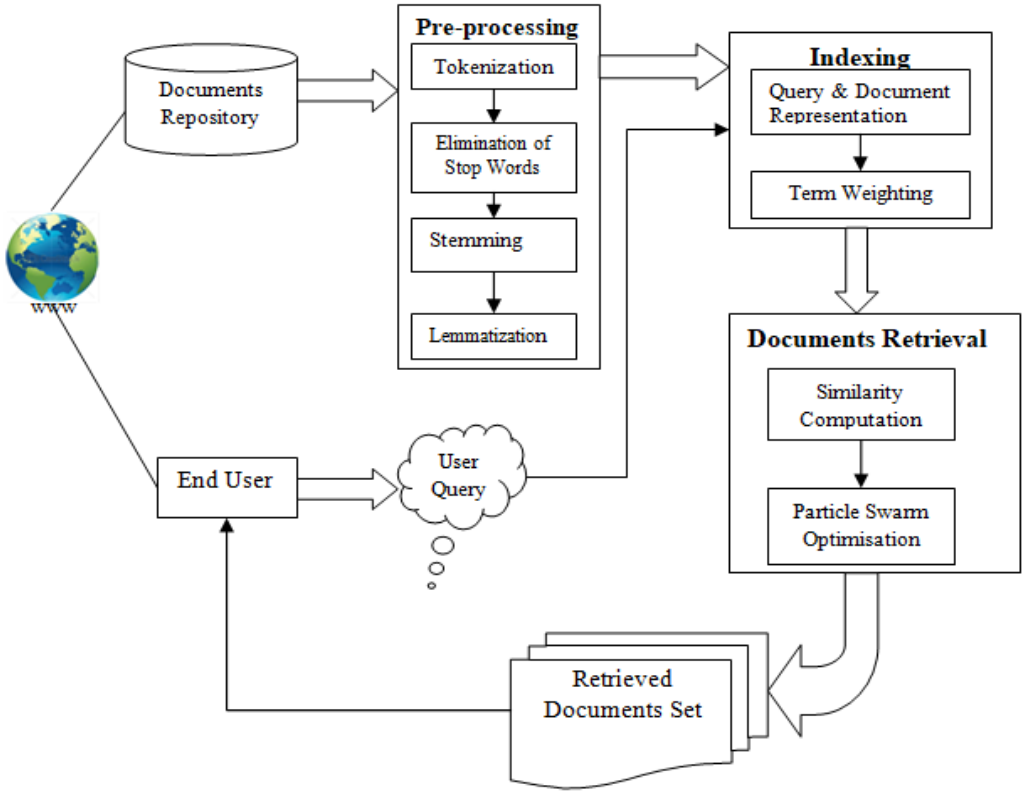

Figure 1 Proposed Architectural Diagram

S. A. Ethni et al. [3] compared the performance of two stochastic search methods PSO and SA when used to detect fault of induction machine stator and rotor winding faults. Experimental Results prove that the PSO better suites for the said problem as it achieves a success rate of 99% whereas SA algorithm scores 60% with substantially improved execution times. Horng-Lin Shieh et al. [4] proposed SA-PSO algorithm, which introduces SA metropolis acceptance rule to enhance the solution quality and to build up convergence rate.

TABLE I. CHARACTERISTICS OF DATASETS

Datasets	RCV1
No. of documents	8,04,414
No. of terms	47,236
Average Document Size (Bytes)	2K

Yan Zichao et al. [5] aim for a hybrid method, which overcomes the drawbacks of PSO such as reduction in the population diversity, appearance of numerous inferior solutions and premature convergence. Xingang Wang et al. [6] proposed a hybrid K-Means algorithm based on PSO-SA for cluster analysis wherein the ability of SA algorithm to jump out of local optima is used. To solve the flowshop problem with coupled-operations in the presence of the time lags, Nadjat Meziani et al. [7] designed a hybrid PSO-SA algorithm.

Similarity function plays immense role in retrieving relevant documents to the user query. It just reflects how similar a document is for the issued query by assigning the rank to it. Selection of similarity measures also has a serious impact on performance of IRSs. Anna Huang [8] in his work, compared and analysed Euclidean distance, Cosine similarity, Jaccard coefficient, Pearson correlation coefficient and averaged Kullback-Leibler divergence for their effectiveness in text clustering.

TABLE II: VARIOUS EVALUATION MEASURES

Query	System	Accuracy (%)	F-Measure (%)	Specificity (%)	DCG (%)
1	Existing	82.20	76.35	72.91	78.65
	Proposed	89.80	86.22	82.44	88.81
2	Existing	77.14	42.88	38.48	44.16
	Proposed	86.44	58.40	51.29	60.16
3	Existing	77.42	46.16	41.69	47.55
	Proposed	86.62	61.62	54.63	63.47
4	Existing	77.14	42.82	38.42	44.11
	Proposed	86.42	58.32	51.19	60.07
5	Existing	76.46	33.16	29.26	34.16
	Proposed	85.9	48	40.91	49.44

Anuradha D. Thakare et al. [9] presented Overall Matching Function (OMF) and Virtual Centre based Matching Function (VCF) to enhance the retrieval performance of IRS using Genetic Algorithm. OMF and VCF used as fitness functions in GA to apply on Iris, Wine and Cancer data sets from UCI machine repository and news articles datasets. G. SureshReddy et al. [10] designed and defined a new similarity measure to compute and judge the similarity values for the two text files/documents. Komal Maher et al. [11] to use the same SMTP for text classification and clustering problems.

A comparative study based on the performances given by Euclidean distance, Cosine similarity and SMTP is presented.

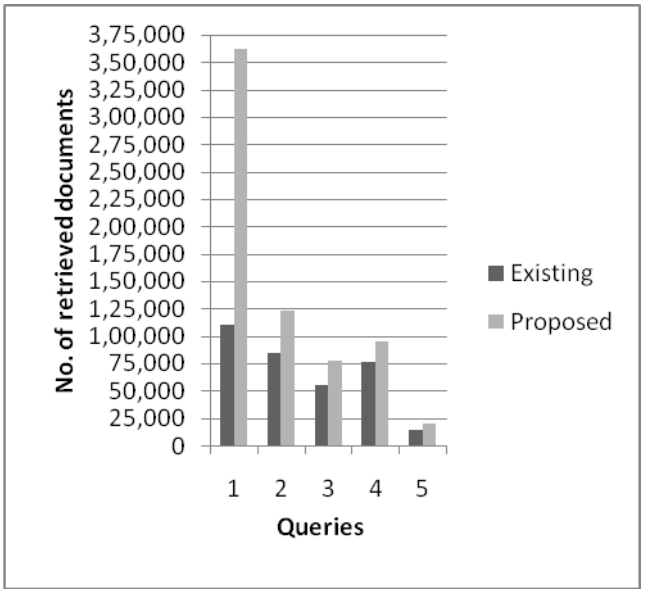

Figure 2 Number of documents retrieved for different queries.

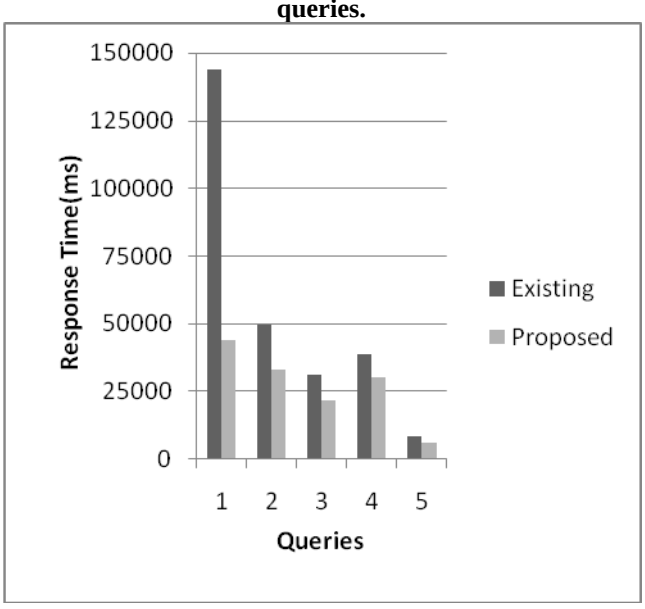

Figure 3 Query Response time

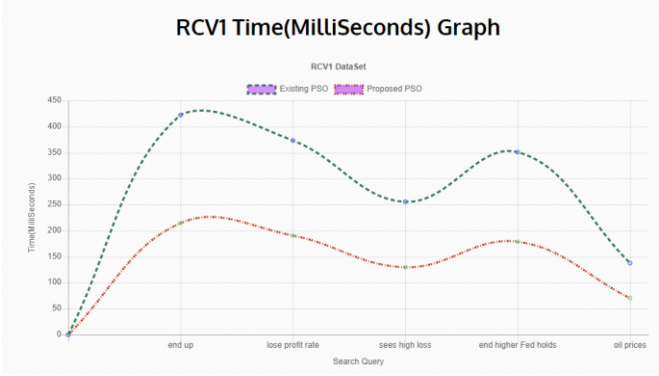

Figure 4 Output plot of response time for different queries

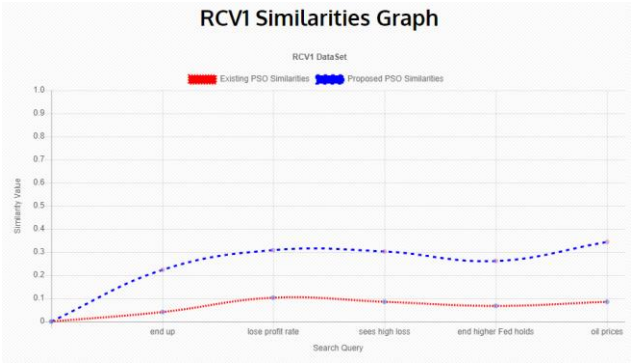

Figure 5 Output plot of similarity values

III METHODOLOGY

A. Proposed System

WIR is ultimately a task of discovering appropriate data from a more significant collection of unstructured data. WIR process runs in the background, uses a huge documents repository and the user query indicating user needs as input and retrieves the most relevant documents at the top as output. A document can be a structured data, text, video, image, sound, musical scores, DNA sequences, etc. A document is normally depicted by a set of keywords/terms contained in it. The user queries and documents must be represented as per a model. Vector space model is the extensively used one where in vectors of term weights depict both documents and queries. Vector space encloses all the terms that the system come across and is constructed during indexing process. Term weight designates the significance of the term in the query or in the document. For the issued query, Similarity values for the documents are computed using a similarity measure. Several techniques are used to rank between documents based on the computed similarity. The top ranked documents are deemed relevant to the query and presented as output. The proposed architecture can be seen in Figure 1.

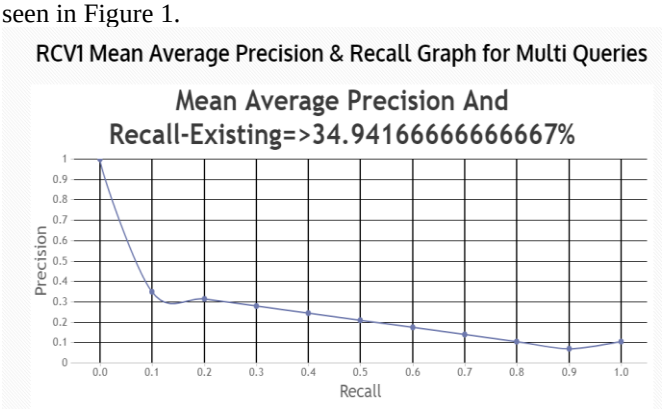

Figure 6 MAP curve of existing system.

B. Preprocessing

Documents preprocessing is a process of transforming text documents to specific terms to be stored in an index. The RCV1 datasets are taken as input to preprocessing. RCV1 is a collection of 8,04,414 XML documents involving archives published by Reuters. These documents are containing totally 47,236 terms in them while the average document size is relatively equal to 2 Kb.

Table I shows the characteristics of RCV1 datasets used. The purpose of preprocessing is to reduce the size of the index extracting distinct terms and to optimize the performance of the further stages of WIR process. The steps of preprocessing are as follows:

- Tokenization
- Elimination of stop words
- Stemming
- Lemmatization

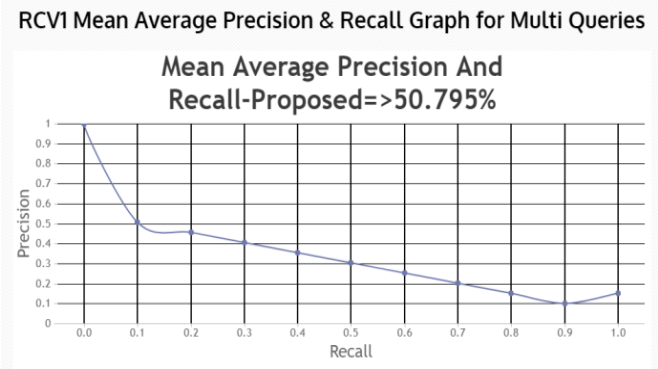

Figure 7 MAP curve of proposed system.

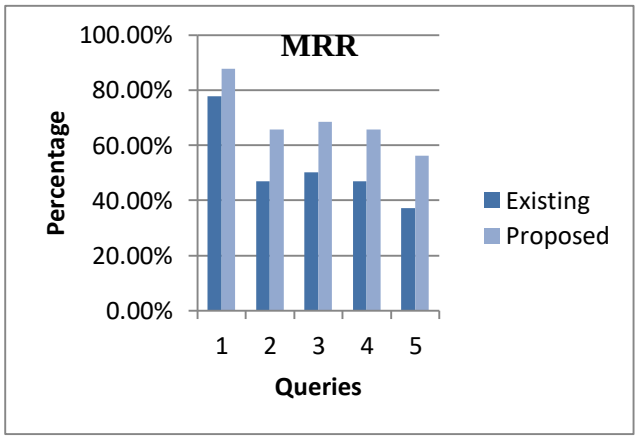

Figure 8 MRR

C. PSO with Simulated Annealing

The most primary contribution of the proposed work was the hybridization of the PSO for improving the optimization process over the document in the web system. The need for the hybridization of PSO is to overcome its convergence error with the vastness of the data and other related aspects. Hence the proposed hybridization of the PSO was accomplished through the integration of the simulated Annealing (SA) algorithm. In the present model, we employed the SA algorithm initially over the obtained similarity measures from the SMDR over the document in the data set. The similarity values among the query and the document terms were annealed with the SA algorithm and optimized to provide the similarity document list. The annealed similarity list was fed into the PSO algorithm in which the optimization was performed through the concept of particle velocity. The algorithm which describes the hybridization of PSO with SA is as follows:

1. Initialize the group of particles by assigning Document Similarity List.
2. Initialize the swarm size, processer size and maximum iteration size.
3. Set Similarity for Task Execution.
4. Initialize resource allocation using number of particles to perform task scheduling.
5. Form the unique particles and transform particles then assign locations and new randomly produced solution from the similarity list.
6. Compute the velocity and update the best solution for best particle.
7. Evaluate the fitness values of group of particles.
8. Repeat the step 7 for every particle satisfied the fitness function.
9. Repeat until the maximum number of iterations is reached.

In step 6 of the algorithm, velocity of the particles is computed by the below equations:

$vel_{bc}=insf*vel_{bc}+randpart*(part_{bc}-newvel_{bc})+randglob*(bss-newvel_{bc})$
$newvel_{bc}=(newvel_{bc}+vel_{bc})mod(totdocscount)+1$

where,
randpart=randval ,0 <randval<((insf+1)*2)/2
randglob=randval ,0 <$randval$<((insf+1)*2)/2

and the best solution for the particle is updated by the below equations:
$newvel_{bc}>f(part_{bc}) \rightarrow part_{bc}=newvel_{bc}$
$f(part_{bc})>f(bss) \rightarrow bss=part_{bc}$

To gauge the similarity between query and the documents, a novel similarity measure, called SMDR is used. SMDR satisfies the basic desirable properties to qualify as a similarity function. It significantly used to rank the documents based on their attained similarity values and thereby helps to retrieve the most similar documents from the repository of documents. It is given as:

$$SMDR(q,t)=\frac{\sum_{i=0}^{n-1} tc_i qhc_i q_i t_i}{\sqrt{\sum_{1=0}^{n-1}(\frac{tc_i}{qhc_i})^2}+\sqrt{\sum_{1=0}^{n-1}(q_i)^2}*\sqrt{\sum_{1=0}^{n-1}(t_i)^2}}$$

Where, SMDR= Novel Similarity, q= user query, t=terms, tc=terms count and qhc=query hit terms count. This novel SMDR has been used as a fitness function to compute fitness values for the particles in the proposed PSO-SA algorithm.

IV. EXPERIMENTAL RESULTS

A. Setting up the experiment

The novel similarity measure SMDR with hybridized PSO with SA is coded in Java and have gone through several experiments. Numerous queries are executed to test the systems.

For RCV1 datasets proposed PSO-SA takes 300 number of iterations and the number of particles 80. These empirical parameters are initially set with the said values to yield the best quality results.

Retrieval Number: J89030881019/19©BEIESP
DOI: 10.35940/ijitee.J8903.0881019
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

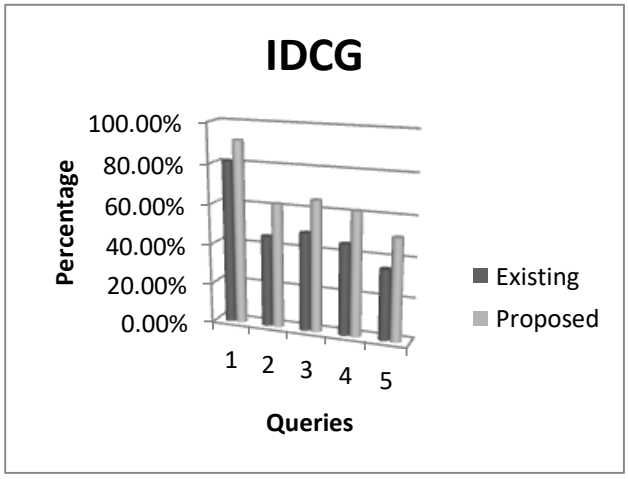

Figure 9 IDCG

B. Evaluation metrics

Precision and recall are the two prominent evaluation measures used in IRS. Precision is the potential to retrieve the top-ranked documents that are substantially relevant. Recall is the capability to search for all the relevant documents in the repository. These are calculated using Eq.s (1) and (2) respectively.

Precision= (TP/ (TP+FP)) (1)

Recall= (TP/ (TP+FN)) (2)

Fig.s 10-13 show the precision-recall curves for the mentioned queries (single). The P/R curves clearly show the effectiveness of the proposed system over the existing system. Fig.s 14 and 15 demonstrate the output of Precision-Recall curves for multiple queries for both proposed and existing systems respectively. The arithmetic mean of average precision (MAP) is used to estimate the retrieval accuracy. MAP at various recall points drawn is exhibited in Fig.s 6 and 7 respectively. Undoubtedly, proposed system has obtained better MAP values than existing one.

Accuracy is the portion of true results among the entire documents in the repository. It is calculated using the expression as in Eq. (3). Similarly, specificity is calculated by Eq. (4). F- Measure provides single measurement for a system considering both precision and recall together and the mathematical expression of the same is given in Eq. (5). Table II shows the superior performance of proposed system over the existing system in terms of these measures.

Accuracy= ((TP+TN)/ (TP+TN+FP+TN)) (3)

Specificity= (TN/ (TN+FP)) (4)

F-Measure=2*((Precision*Recall)/ (Precision + Recall)) (5)

Where,

TP- True Positive

TN- True Negative

FP- False Positive

FN- False Negative

Discounted Cumulative Gain (DCG) considers the top ranked retrieved documents. DCG is generally calculated over a set of queries using the Eq. 6.

$$DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{log_2 i} \quad (6)$$

Where, rel – relevant document and p – particular rank. It is generally normalized using the ideal DCG, that is IDCG. IDCG is defined as the ordered documents in the decreasing order of relevance which is shown in Fig. 9. Mean Reciprocal Ranking (MRR) used for ranking the best site or item being searched and is computed using the Eq. 7. Fig. 8 shows the MRR for both the systems.

$$R = \frac{\sum_{q=1}^{n} \frac{1}{rankq}}{n} \quad (7)$$

C. Discussion

When the user issues a query, the time taken by the IRS to search for and retrieve the relevant documents is its response time. In Fig. 3 significant reduction in the response time of proposed system can be observed. Fig. 4 presents the snapshot of the same, thus accelerating the efficiency of WIR process. SMDR described in section III. C, helps retrieving the top-ranked most similar documents than that of Cosine similarity function used in the existing system. It can be observed in Fig. 5. Fig. 2 shows proposed system retrieves more number of relevant documents than that of existing system, thereby demonstrating the impact of PSO-SA approach used.

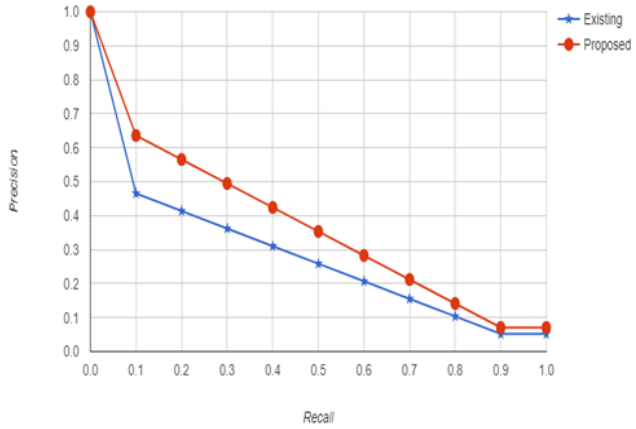

Figure 10 PR graph for the query "endup"

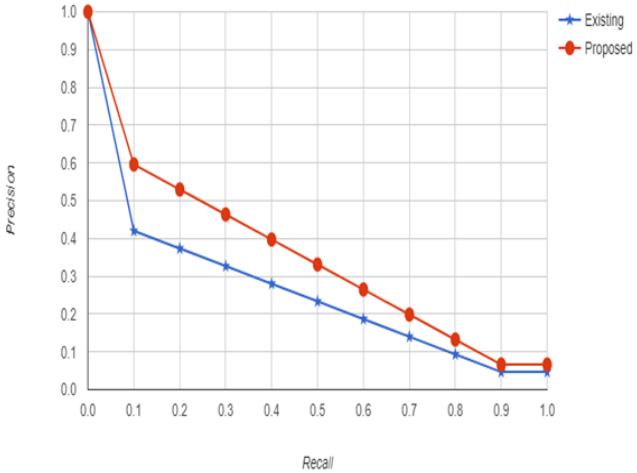

Figure 11 PR graph for the query "end higher fed holds"

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

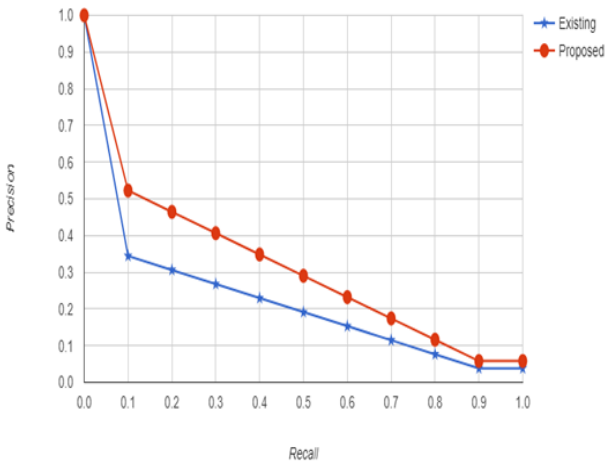

Figure 12 PR graph for the query "see high loss"

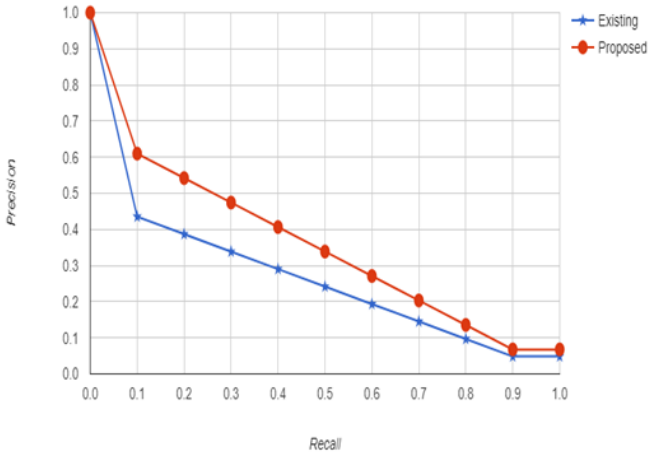

Figure 13 PR graph for the query "lose profit rate"

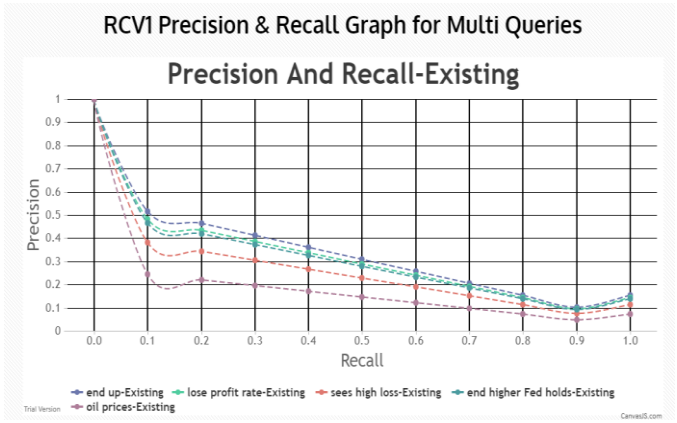

Figure 14 PR curve of existing system for multiple queries.

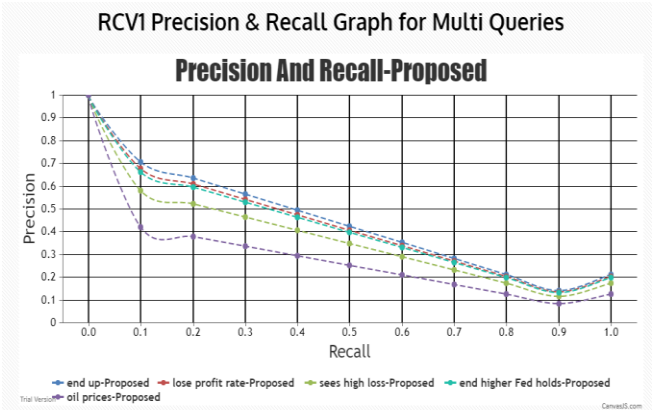

Figure 15 PR curve of proposed system for multiple queries.

V. CONCLUSION

The paper describes the hybridized PSO-SA approach which optimizes the WIR process by considerably reducing the query response time of the system, thereby contributing to the efficient retrieval of web documents. A novel SMDR function used as a fitness function significantly contributes in retrieving the most similar documents, thereby contributing to the effective retrieval of web documents. Various system evaluation measures such as precision, recall, MRR, MAP, DCG, accuracy, specificity and F-measure are used to evaluate both proposed and existing systems. Experimental results clearly prove the superior performance of proposed system over the existing one.

ACKNOWLEDGMENT

The authors acknowledge and express the gratitude to the Department of studies in Computer Science and Engineering, UBDTCE, Davanagere for providing needed conveniences to accomplish the research work. The Authors also express their thankfulness to anonymous reviewers and referees for thoughtful comments and criticism.

REFERENCES

1. Habiba Drias, "Web Information Retrieval using using particle Swarm Optimization Approaches", in proceedings of International Conference on Web Intelligence and Intelligent Aget Technology, 2011, pp. 37-39.
2. Xiaoyu Song, Yang Cao and Chunguang Chang. "A Hybrid Algorithm of PSO and SA for Solving JSP". Fifth IEEE International Conference on Fuzzy Systems and Knowledge Discovery. 2008, Pp. 111-115. DOI 10.1109/FSKD.2008.430
3. S. A. Ethni, B. Zahawi, D. Giaouris and P. P. Acarnley. "Comparison of Particle Swarm and Simulated Annealing Algorithms for Induction Motor Fault Identification". 7th IEEE International Conference on Industrial Informatics (INDIN 2009). 2009, pp. 470-474. DOI: 10.1109/INDIN.2009.5195849
4. Horng-Lin Shieh, Cheng-Chien Kuo and Chin-Ming Chiang. "Modified particle swarm optimization algorithm with simulated. annealing behavior and its numerical verification". Applied Mathematics and Computation 218, 2011, pp. 4365–4383. doi:10.1016/j.amc.2011.10.012
5. Yan Zichao and Luo Yangshen. "A Particle Swarm Optimization Algorithm Based on Simulated Annealing". Advanced Materials Research Vols. 989-994. 2014, pp. 2301-2305. doi:10.4028/www.scientific.net/AMR.989-994.2301.

6. Xingang Wang and Qi Sun. "The Study of K-Means Based on Hybrid SA-PSO Algorithm". 9th IEEE International Symposium on Computational Intelligence and Design. 2016, Pp. 211-214. DOI 10.1109/ISCID.2016.162
7. Nadjat Meziani, Mourad Boudhar and Ammar Oulamara. "PSO and simulated annealing for the two-machine flowshop scheduling problem with coupled-operations". European J. Industrial Engineering, Vol. 12, No. 1. 2018, Pp. 43-66. DOI: 10.1504/EJIE.2018.089877
8. Anna Huang. "Similarity Measures for Text Document Clustering". Proceedings of the New Zealand Computer Science Research Student Conference. 2018, pp. 49-56. doi=10.1.1.332.4480
9. Anuradha D. Thakare and C.A. Dhote. "An Improved Matching Functions for Information Retrieval Using Genetic Algorithm". IEEE International Conference on Advances in Computing, Communications and Informatics (ICACCI). 2013, pp. 770-774.
10. G. SureshReddy, T.V.Rajinikanth and A. Ananda Rao. "Design and Analysis of Novel Similarity Measure for Clustering and Classification of High Dimensional Text Documents". International Conference on Computer Systems and Technologies - CompSysTech'14, Ruse, Bulgaria. 2014, pp. 194-201. http://dx.doi.org/10.1145/2659532.2659615.
11. Komal Maher and Madhuri S. Joshi. "Effectiveness of Different Similarity Measures for Text Classification and Clustering". International Journal of Computer Science and Information Technologies, Vol. 7 (4). 2016, pp. 1715-1720.

AUTHORS PROFILE

Ramya C has received M.Tech. in Computer Science and Engineering from Davanagere University in 2011 and pursuing Ph.D. in the Department of studies in Computer Science & Engineering, University B.D.T College of Engineering, Davanagere from Visvesvaraya Technological University, Belagavi, Karnataka, India. Her research interests lie in Information Retrieval, Data Mining, Soft Computing and Neural Networks.

Shreedhara K S has received M.Tech in System Analysis from NITK, Surathkal, Mangalore University in 1997 and Ph.D. from Manipal University, Karnataka, India in 2008. He is a professor in the Department of studies in Computer Science & Engineering, University B.D.T College of Engineering, Davanagere, Karnataka, India. His research interest includes Data Mining, Machine Learning, Pattern Recognition, Soft Computing, Computer Graphics and Image and Video Processing.