Simulated Annealing Approach to Improve the Effectiveness of Web Information Retrieval Process

Dr. Ramya C
Assistant professor
Department of Information Science and Engineering
CMR Institute of Technology, Bengaluru, India

Dr. Paramesh S.P
Assistant professor
Dept of Computer Science and Engineering
School of Engineering, Central University of Karnataka, Kalaburgi, India

Abstract

The emergence of the internet has revolutionized the process of information retrieval. However, web-based information retrieval systems are faced with several challenges that were not present in traditional setups. To address these challenges and provide efficient and effective retrieval of web documents, a modified particle swarm optimization approach has been proposed. The proposed method incorporates the Simulated Annealing technique, a probabilistic approach that predicts global optimization in large search spaces, in conjunction with PSO. The proposed system for web-based information retrieval utilizes the Simulated Annealing algorithm to process the similarity values of documents. The particles generated through this process are then optimized with the PSO algorithm. The system evaluates the retrieved documents for correctness and accuracy using performance metrics such as precision, recall, and response time. To measure the effectiveness, various measures are used such as accuracy, F-measure, sensitivity, specificity, DCG and n DCG are used. Experimental results demonstrate that the proposed system outperforms the existing system.

Keywords— Simulated Annealing, sensitivity,specificity, Similarity Measure, Precisionand Recall.
DOI Number: 10.48047/NQ.2022.20.7.NQ33509 **NeuroQuantology2022;20(7): 4173-4182**

1. Introduction

Web Information Retrieval (WIR) is a crucial subject that affects our daily lives significantly. It involves the task of retrieving relevant data from a vast collection of unstructured data available on the internet. WIR systems have become the primary means of accessing information on the web, but they face challenges in finding the most appropriate information due to the unstructured nature of the data and the exponential growth of information [1].As a result, users often encounter difficulties in obtaining specific information on the web. There are various factors that contribute to the inefficiency and ineffectiveness of information retrieval. Classic Information Retrieval (IR) approaches are unable to handle the vast collection of web documents. This can lead to the retrieval of both similar and dissimilar documents as results for a query, as even non-relevant documents may be assigned high similarity scores.

A.Motivation

The ultimate aim of WIR is to provide improved systems that retrieve the most relevant information available onthe 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 andefficiency. 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 quite essential to propose amodified version of Particle Swarm Optimization (PSO) algorithm importing the useful characteristics of SAinto it, to provide solutions of good quality, speedy computation and to search for the global optimum solutionand to reduce the query response time of the system, hence contributing to the efficiency of web informationretrieval.

2. Review of Literature

Numerous techniques and methods [2,3] have been recommended for computing the information retrievalaccuracy. Bouadjenek et al. stated that the combination of IR in a social network is contributed by the taxonomyfor the social network and analyzed the association of the IR-social network combination using IR [4]. Theretrieving information in the area of the social network is the major issue. Ahamed et al. worked on theinformation retrieval on the basis of process complexity and measured the fundamental trade off factors for anintelligent framework for web search [5].Thangaraj et al. designed an effective IR in the semantic web byhighlighting that the IR with keywords is less accurate for a vast amount of data [6]. Palomino et al. deliberatedthat the capability of the WIR system to retrieve the data continuously from the vast dynamic source with higherprecision, 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 theirexact information prerequisites [8]. Alloui et al., [9] presented an innovative PSO approach for retrieving theweb information. It uses the detailed feedback to re-compute the user query and therefore enhances the count ofappropriate outcomes. Deo et al., [10] presented a new IR system which uses PSO algorithm for optimization. Inthe 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 algorithmis to measure the similarity between documents and determine the centroids for the group ofdocuments.Abualigah et al. proposed an innovative feature selection technique, viz., feature selection methodusing the PSO algorithm to explain the feature selection problem by creating a new subset of informative textfeatures [12].

Nainika Kaushik [13] proposed a system using support vector machine technique in combination with ParticleSwarm 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 ofdocuments BSO (Bee Swarm Optimization) was chosen. In our approach, PSO is modified and strengthened byadding few best features of SA technique into it. So the Modified Particle Swarm Optimisation with SimulatedAnnealing (MPSOSA) is proposed to overcome the disadvantages of PSO, thus to have efficient and effectiveretrieval of documents hence optimising the WIR process.

3. Proposed MPSOSA Algorithm

The proposed system is formulated using a vector space model (VSM) [15], wherein both the queries anddocuments are expressed as vectors in the vector space. After preporcessing, all the distinct terms of queries anddocuments are assigned with term weights based on tf × idf scheme. The proposed MPSOSA is intended to learnthe significance of query terms in the form of term weights over the documents in the collection. It explores thevariety of zones of the documents space, thereby finding the

most relevant documents which best match thequery according to the user need.

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

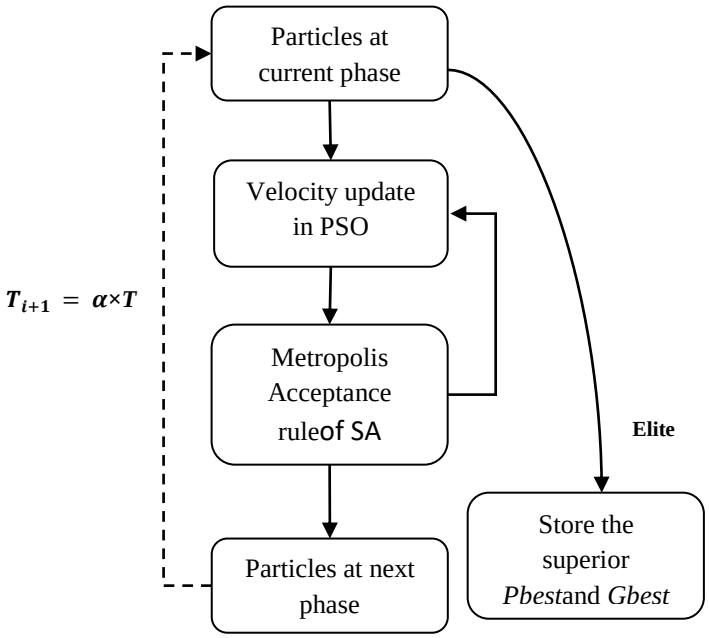


Figure 1. MPSOSA Notional Diagram

Besides, the process explores for the solutions in the direction of *Pbest*and *Gbest*because the metropolisacceptance rule either to consider or deny a new solution is formulated on the parameters such as the differencein fitness values and the present temperature. When a possible solution is not accepted by the metropoliscriterion, 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 newposition 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 computationtime can be achieved exploiting one of the features of PSO i.e. parallel processing.

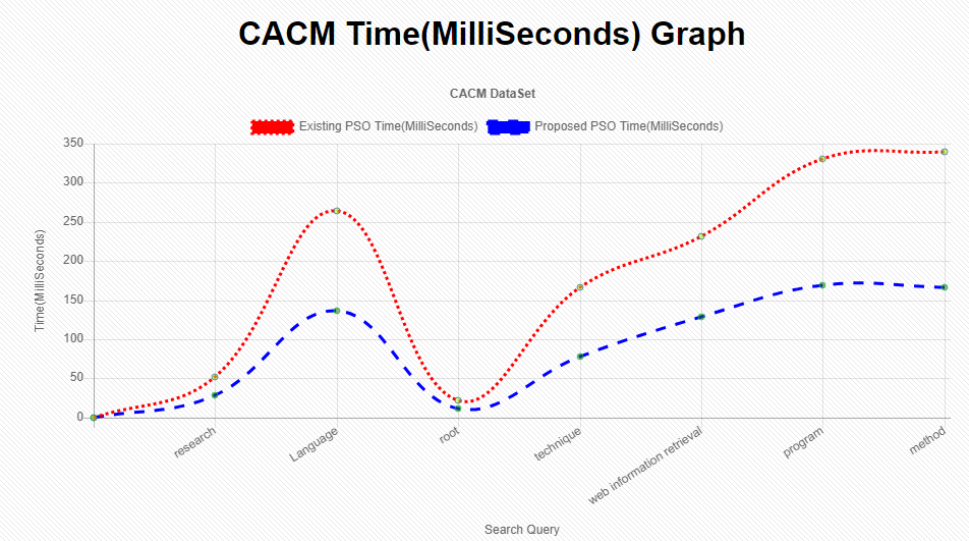
4. Experimental Results and Discussion

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

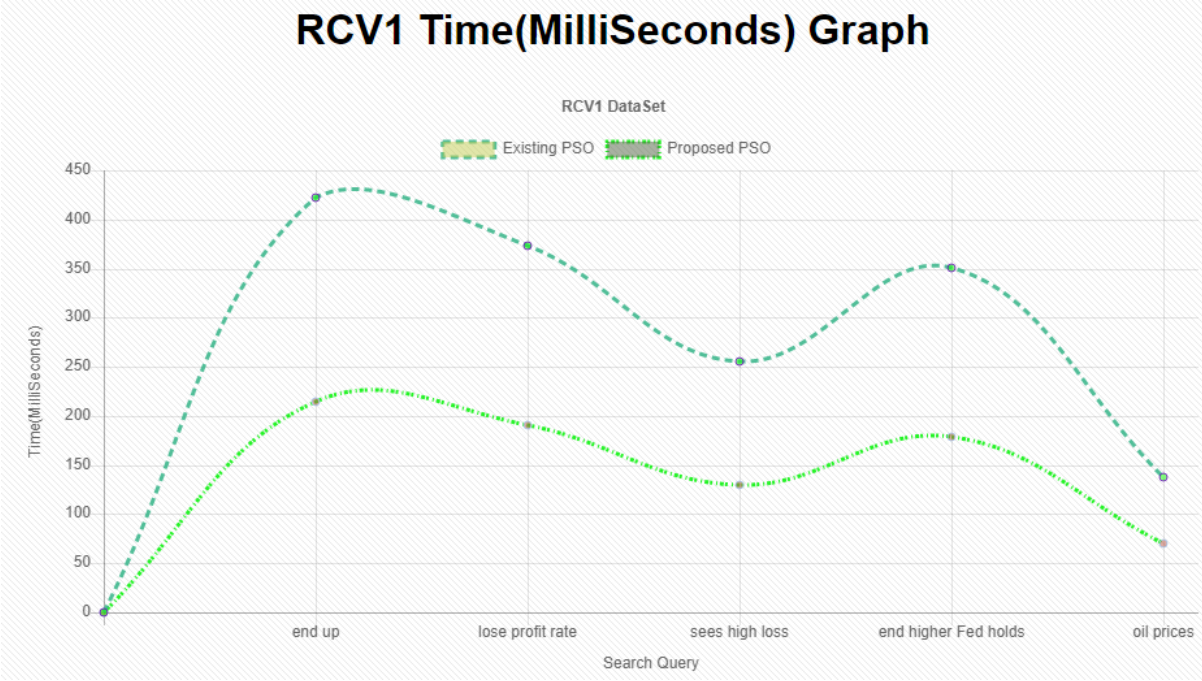
A. Efficiency Analysis

In the domain of WIR, Many works are available in the literature which employ PSO as the tool to optimise WIRprocess for effective and efficient retrieval. Hence to build an existing system, the bio inspired PSO is consideredas 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 retrieveddocuments. 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 thegiven set of queries and there is a reduction of 154.9 ms of response time by proposed system for CACMcollection. Whereas for RCV1 collection, proposed system takes 300.9 ms and existing system takes 540.4 ms toretrieve the documents for the given set of queries and there is a reduction of 239.5 ms of response time byproposed system. It can be clearly noticed that proposed system consumes less time even though it retrievesmore number of documents than existing system.



(a)



(b)

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

B. Effectiveness Analysis

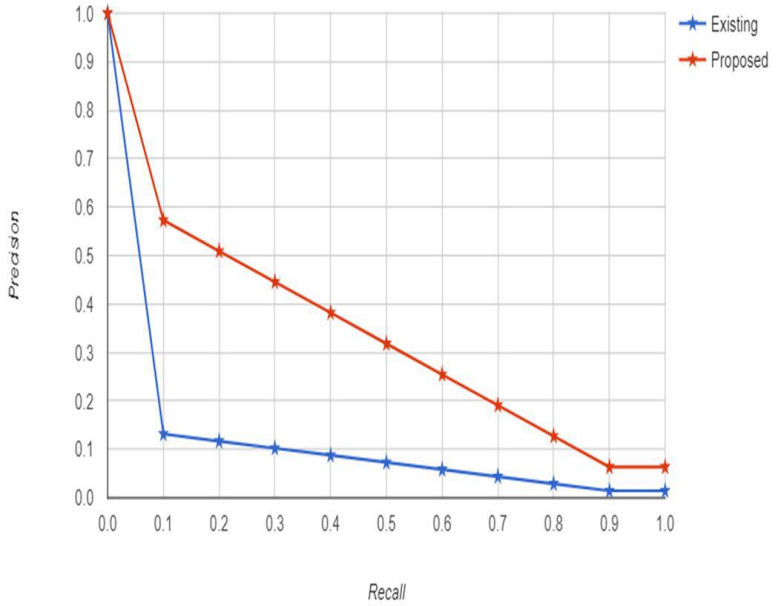
To measure the effectiveness of the system, the following various parameters are used.

Precision and recall

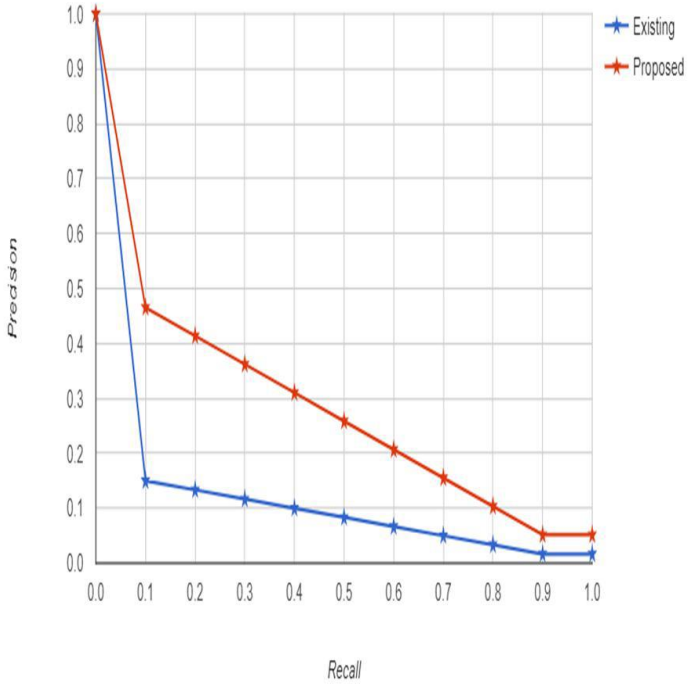
Precision and recall quantitatively evaluate both the quality of the overall resultant documents and the breadth ofthe retrieval algorithm used, thus widely used as standard evaluation strategy in information retrieval [16]. Fig.s3 and 4 show the Plot of precision and recall for single queries for both the systems over CACM and RCV1respectively.Experimental results showed that the retrieval performance obtained by the proposed system could constantlyoutperform the existing approach on different data collections in terms of the parameters response time andprecision and recall. Hence the proposed system is significantly effective as it retrieves the most similar documents to the user and efficient as it covers most of the documents in the collection and retrieves morenumber of documents in considerably reduced response time.

Other Measures

Accuracy measures the portion of relevant documents that are retrieved among the entire documents in the repositoryand also measures the irrelevant documents that are not retrieved. Table 1 clearly demonstrates the resultant documents given by the proposed system are more accurate than that of existing system over CACM collection. It is also true in case of RCV1 collection as it can be seen in Table 2.Sensitivity is called as the true positive rate which measures the amount of actual relevant documents that are correctly retrieved by the system. Specificity is called as the true negative rate which measures the portion of non relevant documents that are not retrieved.F-measure presumes a high value only when both recall and precision are high.DCG and nDCG are evaluated on the basis of the two points: Extremely relevant documents are more applicable than slightly relevant documents in the collection. A relevant document is of less use when it possesses the least ranked position.Table 1 and 2 clearly demonstrates the increased values of proposed method for each evaluation measures for various queries incase of CACM and RCV1 data collection both.

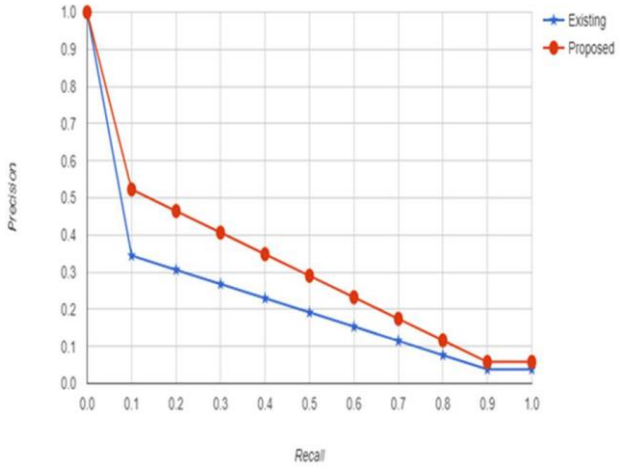


(a) Query: “product”

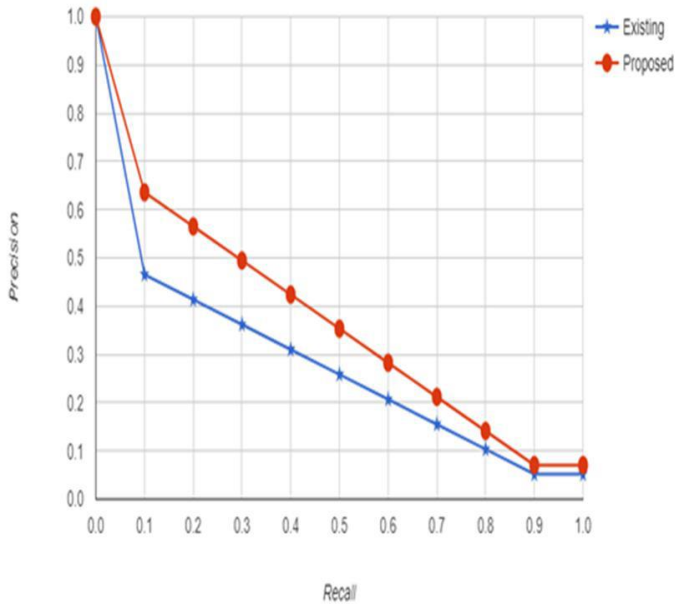


(b) Query: “root”

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

Table 1. Evaluation measures for various queries using CACM data collection

Sl. No.	Query	Existing (E) & Proposed (P) Systems	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	Recall (%)	F-Measure (%)	DCG (%)	nDCG
1	Research	E	97.3558	83.5556	97.382	34.6774	91.5556	50.3024	16.9811	0.91
		P	99.1261	90.4356	99.177	62.3188	93.5556	74.8073	24.5015	0.98
2	Technique	E	88.9455	26.9423	96.5846	63.5294	95.9423	45.9104	29.0754	0.5827
		P	91.1049	30.8223	98.8881	83.9378	98.9423	51.3047	62.8137	0.5842
3	Language	E	91.2074	64.8383	92.3312	38.587	70.9583	49.9897	39.4742	0.6293
		P	92.4782	70.6695	93.5302	42.4779	73.7895	53.9175	59.9182	0.6405
4	Program Schemes	E	86.7792	44.223	96.1648	79.6095	89.343	61.6807	34.8364	0.656
		P	88.9825	48.13	97.9332	88.1549	95.25	67.9274	84.8971	0.6663
5	Its Technology and Economics A survey is offered of techniques	E	81.474	67.015	94.1377	82.1012	91.298	55.0735	75.7246	0.6685
		P	85.5581	70.8971	99.0159	93.75	95.681	69.1771	98.898	0.6718

6	Programming communication	E	92.0994	86.4898	91.8541	45.0575	91.6098	60.4052	78.7364	0.6492
		P	95.6355	87.4898	96.3638	73.0769	94.6098	79.1996	98.7041	0.6699
7	Glossary of computer engineering	E	88.1809	63.4972	92.3325	69.6418	85.6172	70.0969	66.1254	0.6983
		P	89.5443	65.8194	93.5928	78.9616	92.9394	83.9367	82.6611	0.7056
8	Web Information Retrieval	E	95.986	86.1132	95.8319	63.0303	95.1132	75.8174	45.3492	0.6159
		P	96.3795	92.9932	96.2567	75	96.1132	87.5524	79.7858	0.6347
9	Preliminary execution of programs	E	91.2392	52.9945	96.1639	61.2468	73.1145	64.8091	84.5467	0.6347
		P	93.2896	55.3161	98.2182	75.4599	81.4361	71.4836	85.6407	0.6399
10	Techniques of matrix program schemes	E	79.4839	45.0375	89.8876	68.75	80.1575	58.0002	46.7932	0.7063
		P	81.7104	49.1327	91.4124	82.5627	85.2527	62.7356	78.3394	0.7154

Table 2. Evaluation measures for various queries using RCV1 data collection

Sl. No.	Query	Existing (E) & Proposed (P) Systems	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	Recall (%)	F-Measure (%)	DCG (%)	nDCG
1	competition	E	82.201	91.786	86.443	80.15	72.91	83.9367	67.76	0.7834
		P	89.803	96.324	87.4625	90.37	82.44	89.465	75.876	0.8634
2	washington	E	77.142	81.456	87.4156	48.42	38.48	92.2365	77.189	0.8486
		P	86.443	89.465	93.2356	67.81	51.29	96.456	89.567	0.9754
3	Planet hollywood	E	87.425	90.987	95.9843	51.71	41.69	91.456	51.71	0.6934
		P	92.626	96.345	99.9895	70.66	54.63	99.454	79.345	0.7843
4	Chains may raise	E	89.147	91.987	81.4346	48.36	38.42	79.481	85.194	0.9145
		P	96.423	96.345	89.475	67.76	51.19	88.129	96.645	0.9934
5	End up	E	91.461	88.976	92.2365	38.27	29.26	79.568	96.345	0.9532
		P	95.963	92.345	98.209	58.06	40.91	87.595	99.343	0.9912
6	Lose profit rate	E	91.8991	89.987	89.195	75.678	66.895	83.324	70.66	0.8124

		P	96.5343	93.234	96.412	86.453	74.456	89.512	81.931	0.9234
7	Sees high loss	E	85.4509	87.456	81.498	57.442	48.11	95.893	86.323	0.8967
		P	90.9843	93.234	89.465	79.481	68.29	99.127	92.678	0.9578
8	End higher fed holds	E	95.986	90.823	90.945	63.71	52.87	85.7874	93.234	0.9212
		P	99.985	96.842	96.345	80.66	64.98	93.2356	98.432	0.9854
9	Oil prices	E	92.765	82.890	77.189	78.36	58.42	60.344	49.115	0.6198
		P	98.2642	89.393	86.323	97.76	85.19	79.9773	67.987	0.7409
10	aggressive marketing campaigns	E	83.8765	93.23	92.065	54.97	46.98	85.9104	68.29	0.7834
		P	91.7112	98.830	97.2687	78.43	59.91	91.3047	84.765	0.8712

5. Conclusions and Future Work

The novel and effective framework for retrieving the document from the web is accomplished with the modifiedoptimization algorithm MPSOSA. The datasets of CACM and RCV1 are employed to analyze the performanceof the proposed framework. The input datasets are initially pre-processed and indexed. The query and documentsare represented using VSM and all the terms are weighted. This MPSOSA is used as a retrieval algorithm inproposed system to optimise WIR process. The evaluation measures such as precision and recall are used to testthe 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 retrievalusing 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 ofComputation 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 InternetTechnology*, 2001.

[4] Bouadjenek, M. R., Hacid, H., and Bouzeghoub, M., “Social networks and information retrieval, how are theyconverging? A survey, a taxonomy and an analysis of social information retrieval approaches andplatforms*”. 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”, *ExpertSystems 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 horizonscanning”, *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 ofComputer 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 ofInformation retrieval,* Vol. 9(1). 2018.
[11] Djenouri, Y., Belhadi, A., and Belkebir, R. "Bees swarm optimization guided by data mining techniques for documentinformation 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 documentclustering 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 SwarmOptimization ", *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 AnnualJoint Conference of the IEEE Computer and Communications Societies,* Vol. 3, pp. 1238-1247, 2002.