Information Retrieval Ranking Using Machine Learning Techniques

Shweta Pandey[1], Iti Mathur[2], Nisheeth Joshi[3]

[1,2,3]Computer Science, Banasthali Vidyapith, Jaipur, India
[1]shweta.dubey12@gmail.com, [2]mathur.iti@rediffmail.com, [3]nisheeth.joshi@rediffmail.com

Abstract: Information retrieval is the research area in which many researcher have been done and many are still going on. The rapidly growing web pages make it very crucial to search up to date documents. In continuation of research works on learning to rank, this research focuses on implication of machine learning techniques for IR ranking. SVM, PSO and hybrid of both are the main techniques implemented for IR ranking. In case of SVM, selecting appropriate parameters is difficult, but it gives potential solutions for the ranking. One of the optimization methods i.e. PSO is easy to implement and has global search capability. Thus to find the fitness function to optimize the ranking of document retrieval Hybrid SVM-PSO model is proposed.

After the comparative study it has been calculated that the ranking parameters gives best result for RankSVM-PSO over RankPSO and RankSVM. The result has been calculated based on single term queries and multi-term queries. The study shows RankPSO gives the better result than RankSVM and RankSVM –PSO gives better result than RankPSO, so it has been concluded that RankSVM-PSO gives best result among the three techniques.

Keywords: Information retrieval, PSO, SVM, Machine Learning

I. INTRODUCTION

1.1 Search: Search is basically to find something by looking or seeking carefully or thoroughly. In continuation of search there is a search engine in the world of Internet that gives a lot of documents related to the specified key words. Web search engine becomes most popular engine [1] as the most vital access technique to the web.IR finds the documents of an unstructured nature usually text among large collections of data stored on computers.

1.2 Ranking: Ranking [2] of query results is the basic problem in IR**.** Among the documents groups related to the query ranking is the problem i.e. to sort the document according to some criterion that are phrased in terms of relevance of documents related to an information need expressed in the query. Ranking is done by statistical ranking in which scores are used as the basis of ranked retrieval. The document with highest score is ranked to be first and so on. The scoring is done as the simple match or by using weighted match which gives better result in ranking. Ranking models are of two types: ranking the query against individual documents and ranking the query against list of related documents. Based on these the ranking models can be categorized as:

Boolean model: Based on set theory this model is the oldest and simplest model of IR. Because it is based on binary concept partial matched documents are not recovered just those documents that matched exactly can be retrieved. So to retrieve documents from group of documents users should have good knowledge in the domain of making queries.

Vector Based Model: Because Boolean model [3] only fetches completely matched documents, so vector model [4] is addressed that basically focuses on weights in place of binary values. The factors that are used to calculate weight is tf(term frequency)and idf(inverse document frequency).Together these two factors or the product of these factors makes the approximated term weights. These factors together are called tf-idf measure.

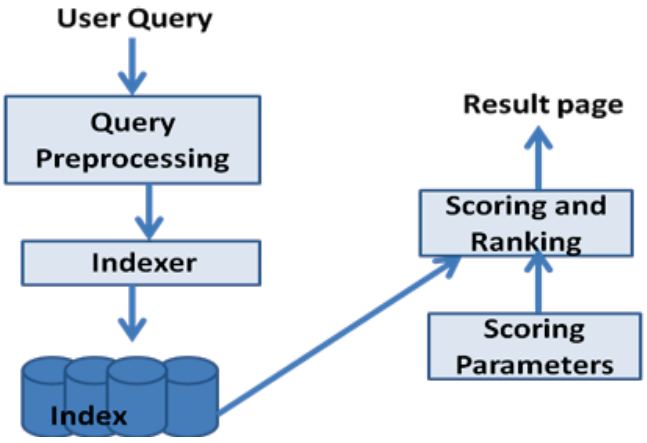

Fig. 1. The classic search model

Probabilistic Model [5]: This model is basically based on probability of documents to be relevant. It finds the probability of estimation of relevant document for the query. The probability depends on the representation of document and query. In it with the help of a set of relevant documents the probability of relevant and non relevant document is calculated.

978-1-5386-9346-9/19/$31.00 ©2019 IEEE

1.3 Problems in ranking: As to get relevant documents from documents that are large in number for a query ranking, ranking is a vital part in internet searcher. But there are many challenges in it:

1. The factors that are considered as a ranking functions are content of the page, link structure etc., so combining all the ranking functions to make a single ranking function is very tough.

2. Speed is the most important challenge in ranking of documents as millions of documents are defined.

3. Mean average precision (MAP) and Normalized Discounted Cumulative Gain (NDCG) are the measures to evaluate ranking algorithm in IR which are non-convex. So it's difficult for optimization by optimization tool that are conventional.

4. The main problem in ranking is that bulk of irrelevant information is retrieved.

5. Different encoding schemes and types of documents with same fingerprints.

1.4 Evaluation measures [6]: Evaluation is to calculate the goodness of a system i.e. how well it meets the user's information need. Traditional evaluation for Boolean retrieval or Top-K retrieval includes the following four measures:

Precision: Retrieved documents to the relevant documents ratio.

Precision: P = TP/TP+FP

Precision=relevant retrieved documents/ (relevant+ non relevant) documents that are retrieved.

Recall: Relevant documents in collection to the retrieved documents ratio.

Recall: R = TP/TP+FN

Recall=relevant retrieved documents/ (relevant documents that are retrieved+ relevant documents that are not retrieved)

F-measure: It is the weighted harmonic mean of precision and recall. It is commonly denoted by F1 or F.

F1-Score: F1 = 2*Precision x Recall/ Precision + Recall

Accuracy is a commonly used evaluation measure in machine learning classification work but it is not a very useful evaluation measure in IR. The reason behind this is that in all the conditions the data is extremely skewed.

Accuracy is the fraction of the correct classifications.

Accuracy: ACC = TP+TN/TP+TN+FP+FN

1.5 PSO

A global optimization called Particle swarm optimization (PSO) is a algorithm in which best solution can be represented as a point or surface in an n-dimensional space for dealing with problems. In this space Hypotheses are seeded and plotted with an initial velocity, and also it is a communication channel between the particles.

1.6 SVM

A discriminative classifier formally characterized by an isolated hyperplane is a Support Vector Machine (SVM). It can also be defined as a supervised learning as it has labeled training data, the algorithm outputs an ideal hyperplane which sorts new models. This hyperplane separates a plane in two sections where in each class lay in either side in two dimensional spaces.

II. RELATED WORKS

In 2008, Zhai et.al.[7] focused to set appropriate parameters of SVM algorithm, SMO (sequential Minimal optimization) is an effective training algorithm belonging to SVM i.e. LS_SVM. On the basis of the above integrated algorithms they introduced PSO and utilized an example to certify its validity.

In 2009, Jun et al. [8] had proposed PSO+SVM for intrusion detection problem. They used standard PSO for the determination of free parameters of SVM and binary PSO is for optimum feature subset.

In 2012, Indrajit and Sairam[9] proposed a strategy of feature selection based on SVM that is optimized by PSO.They took into consideration the leave one out strategy and 20 classifiers performance were evaluated using performance metrics.

In 2014,Lu et.al.[10] considered prediction for Bankruptcy that has been calculated by data mining techniques,since this issue is a critical issue in the finance and accounting field. They proposed switching particle swarm optimization (SPSO) and support vector machine (SVM) together as a new hybrid algorithm to solve the bankruptcy prediction problem.

In 2015,Dong and Jian[11] proposed a SVM parameter selection based on CPSO keeping in mind the end goal to decide the ideal parameters of the Support vector machine rapidly and proficiently. SVMs are new techniques being created, in view of factual learning hypothesis. Preparing a SVM can be defined as a quadratic programming issue. The parameter determination of SVMs must be done before illuminating the QP (Quadratic Programming) issue.

In 2016,Kakde and Gulhane[12] proposed a technique based on particle swarm optimization and support vector machine for Devnagari script recognition system. Devanagari script is

generally utilized as a part of the Indian subcontinent in a few noteworthy dialects, for example, Hindi, Sanskrit, Marathi and Nepali.

In 2018, Gaurav Pandey et.al [13] addressed a document ranking problem that is the feature extraction problem in information retrieval.They also proposed a linear feature extraction algorithm called as Life Rank algorithm for Ranking.

In 2018, Kehinde Agbele et.al[24] proposed predictive document ranking model that computed measures of individual search in their domain of knowledge.The query context determines the relevance of retrived information.The technique that they referred was DROPT technique.

III. OBJECTIVES OF CURRENT RESEARCH

Keeping in view of the above shortcomings in the existing research, we propose to address the following objectives:

1. To compare the developed ranking system with the state of the art.
2. Feature extraction from monolingual documents.
3. To implement a ranking system for information retrieval using combination of support vector machine and particle swarm optimization.

IV. PROPOSED METHODOLOGY

4.1 Query formulation and preprocessing: For query formulation relevance feedback is one of the techniques that can be implemented.

Different relevancy tools are used but some work for long queries and some work for short queries. Operations related to the single term queries are:

- Synonyms
- Thesaurus
- User term in the title
- Stemming

When a user enters longer queries some of the operations that can be applied are:

- Many
- Phrase
- Near/Proximity
- "Like" or "Accrue"

4.2 Document Feature extraction: Features that are to be extracted in query document pair are as follows:

i) Low-level Content Features

It includes tf, idf and dl and their combination, for each combination of four is to be taken for body, anchor, title and URL.

ii) High-level Content Features

It includes the outputs of BM25 and LMIR algorithms. There are total nine language model features considered including different smoothing methods (DIR, JM and ABS).

iii) Hyperlink Features

It includes pagerank, HITS and their variations. Total 7 hyperlink features are considered.

iv) Hybrid Features

Hybrid features refer content and hyperlink information, including "hyperlink-based relevance propagation" and "sitemap-based relevance propagation".

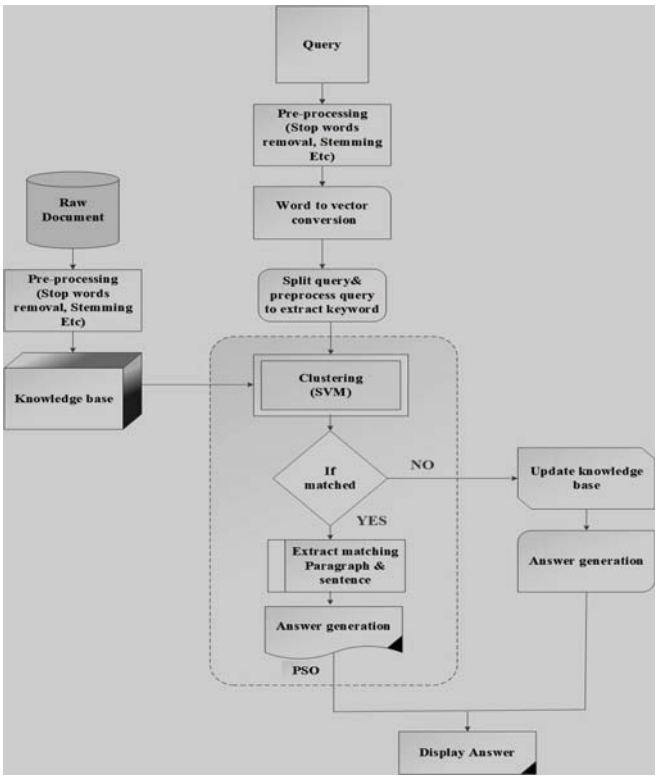

Fig. 2. Structure of information retrieval ranking using SVM+PSO

Initially the raw document is pre-processed by removing stop words, stemming etc., then it is given to the knowledge base. Same as the document, the query also pre-processed and then it undergoes to the conversion of word to vector. Then the query is needed to be splitted and pre-processed the query to extract

keywords. After that, the knowledge base and the splitted query are given to the clustering process. By using the SVM classifier, it classifies the document. Then the condition is applied, if it is matched, and then extracts the matching paragraph and the sentence and also the answer is generated. If it doesn't match, then update the knowledge base and then generate the answer. Finally, by using the PSO optimization, the answers are ranked and display the best answer.

V. EXPERIMENTAL RESULTS

This section basically sums up experimental results. The experiment is done in spider (python 3.7) and the dataset used in this is TREC 2004 QA DATA referred in the link https://trec.nist.gov/data/qa/t2004_qadata.html. The comparison tables based on the performance parameters are as given below:

TABLE 1: Single term queries

Techniques/ parameters	SVM			PSO			SVM+PSO			
	Type of questions									
	Easy	Medium	Hard	Easy	Medium	Hard	Easy	Medium	Hard	
Precision	81.33	89.34	86.66	92.0	94.67	89.33	98.66	98.67	94.66	
Recall	82.43	79.76	89.04	88.46	91.02	93.05	94.87	97.36	97.26	
F1 score	81.87	84.27	87.83	90.19	92.81	91.15	96.73	98.01	95.95	
Accuracy	82.0	83.33	88.0	90.0	92.67	91.33	96.67	98.0	96.0	

TABLE2: Multi -term queries

Techniques/parameters	SVM			PSO			SVM+PSO			
	Type of questions									
	easy	medium	hard	easy	medium	hard	easy	medium	hard	
precision	86.66	92.0	88.0	93.33	96.0	90.54	97.33	97.34	94.66	
Recall	80.24	80.23	89.18	90.90	93.50	93.05	94.80	98.64	97.26	
F1 score	83.34	85.71	88.59	92.10	94.73	91.78	96.05	97.98	95.94	
Accuracy	82.67	84.67	88.67	92.0	94.67	92.0	96.0	98.0	96.0	

VI. PERFORMANCE EVALUATION

This section contains the performance evaluation of three algorithms that is SVM, PSO and HybridSVM-PSO based on single term queries and multiple term queries. A learned ranking function is applied in ranking of documents for a user query.The performance is calculated based on easy,medium and hard questions as categorised in dataset which has been taken in this work.

This algorithm is presented for the collection of user queries and their corresponding retrieved documents. Human annotations help in relevance judgment. In it ranking score is assigned to each document based on its relevancy. The document ranked at the top which contains a higher ranking score. A corpus of 1000 documents will be considered out of which 70% will be used for training or learning phase and 30% will be used for the testing purpose. A query set of 300 queries will be considered out of them 150 will be taken as single term queries and 150 will be taken as multiple term queries. It will be clear that our system is feasible in ranking of documents among multiple documents and will provide a response time of a few milliseconds within an acceptable memory cost. As our framework is a model created with the one of a kind motivation behind doing research and evaluation, the calculation's execution could be significantly progressed.

as shown below using graphs:

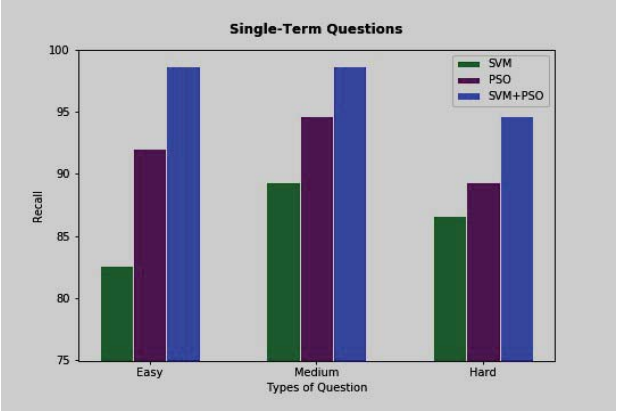

Fig. 3. Recall for single term questions

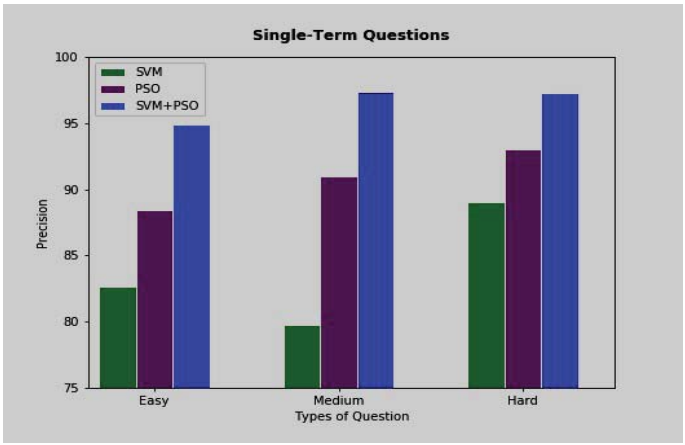

Fig. 4. Precision for single term questions

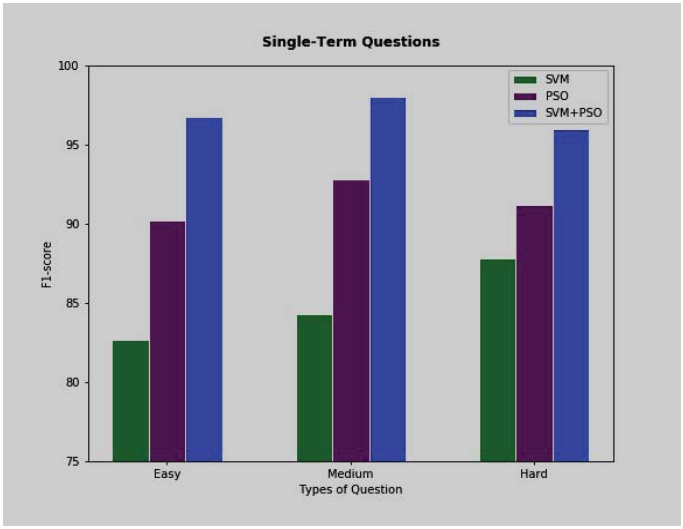

Fig. 5. F1-score for single term questions

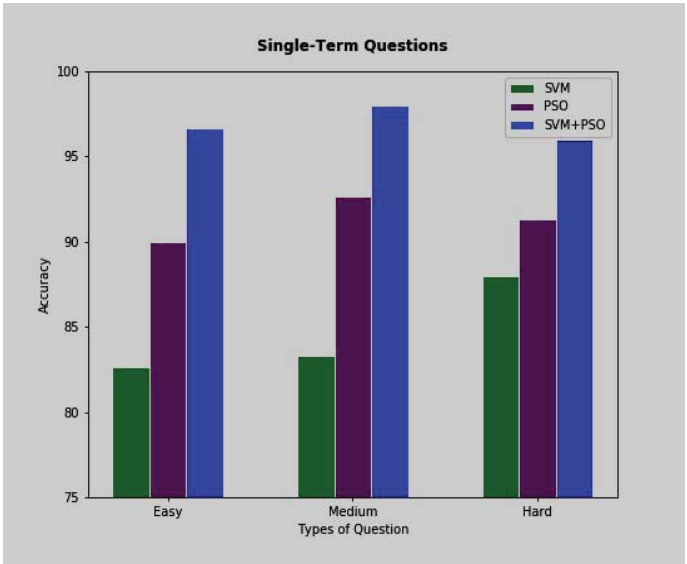

Fig. 6. Accuracy for single term questions

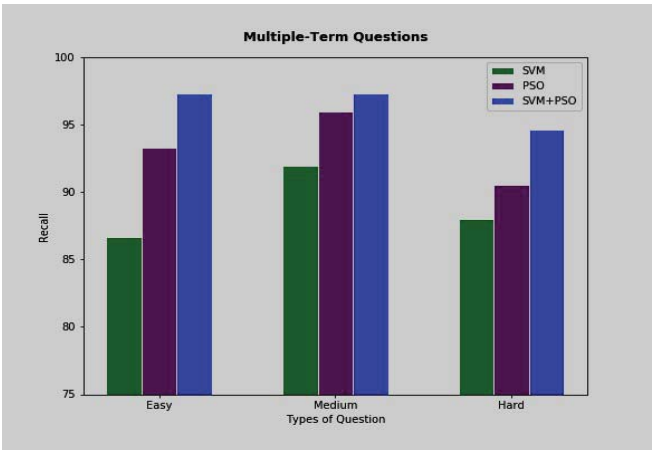

Fig. 7. Recall for multiple term questions

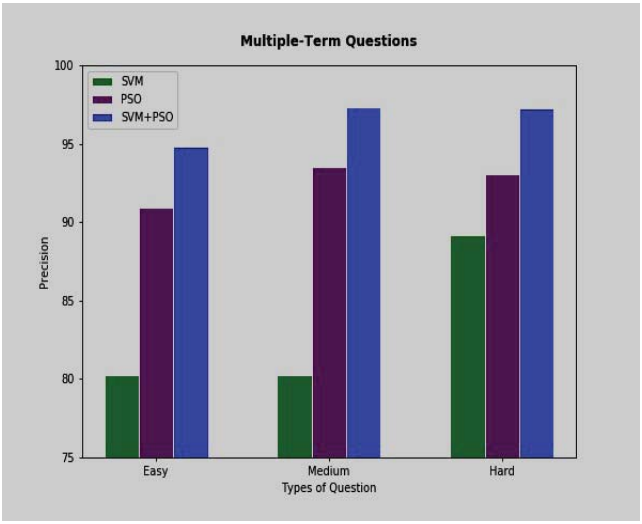

Fig. 8. Precision for multiple term questions

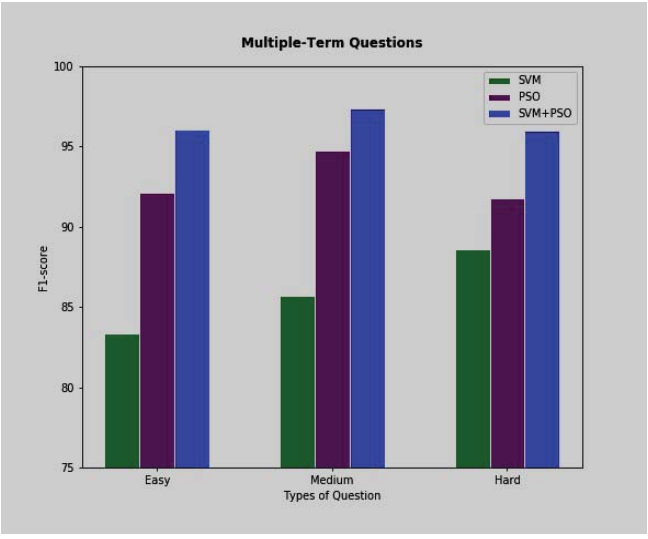

Fig. 9. F1-score for multiple term questions

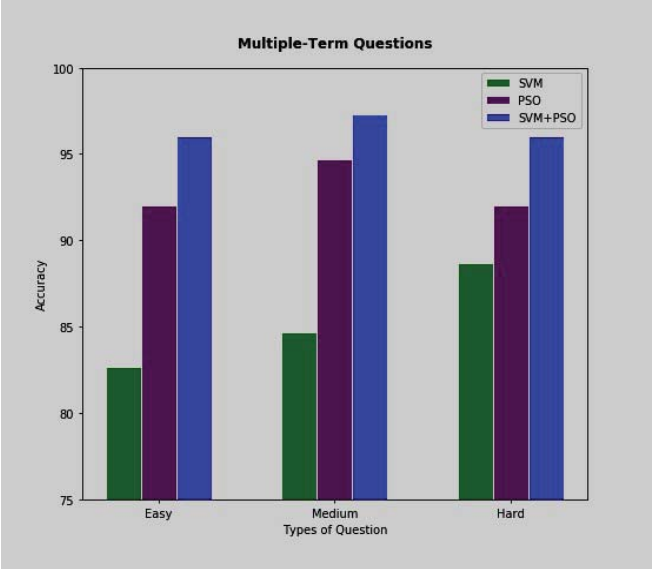

Fig. 10. Accuracy for multiple term questions

VII. CONCLUSION AND FUTURE SCOPE

As our system is the hybridization of both SVM and PSO, it overcomes all the previous shortcomings in ranking of information retrieval and improves the performance of the ranking system as we have seen in our evaluation tables and graphs. This paper contains the ranking system for monolingual only using SVM and PSO. The research can be further done for cross lingual and for real time retrieval system.

REFERENCES

[1] Broder, Andrei. "A taxonomy of web search." ACM Sigir forum. Vol. 36. No. 2. ACM, 2002.

[2] Cao,Yunbo "Adapting ranking SVM to document retrieval." Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2006.

[3] Lee, Joon Ho. "Properties of extended Boolean models in information retrieval." Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval. Springer-Verlag New York, Inc., 1994.

[4] Lee, Dik L., Huei Chuang, and Kent Seamons. "Document ranking and the vector-space model." IEEE software 14.2 (1997): 67-75.

[5] Jones, K. Sparck, Steve Walker, and Stephen E. Robertson. "A probabilistic model of information retrieval: development and comparative experiments: Part 2." Information Processing & Management 36.6 (2000): 809-840.

[6] Järvelin, Kalervo, and Jaana Kekäläinen. "IR evaluation methods for retrieving highly relevant documents." Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2000.

[7] Juan M. Fernández-Luna, Juan F. Huete, Óscar AlejoDirect Optimization of Evaluation Measures in Learning to Rank using Particle Swarm.

[8] Jun Wang, "Mean-Variance Analysis: A New Document Ranking Theory in Information Retrieval".M. Boughanem et al. (Eds.): ECIR 2009, LNCS 5478, pp. 4–16, 2009.Springer-Verlag Berlin Heidelberg 2009.

[9] Indrajit Mondal, Sairam.N"SVM-PSO based Feature Selection for Improving Medical Diagnosis Reliability using Machine Learning Ensembles" Natarajan Meghanathan, et al. (Eds): SIPM, FCST, ITCA, WSE, ACSIT, CS & IT 06, pp. 267–276, 2012.© CS & IT-CSCP 2012 DOI : 10.5121/csit.2012.2326

[10] Yang Lu, Nianyin Zeng, Xiaohui Liu,, and Shujuan Yi, "A New Hybrid Algorithm for Bankruptcy Prediction Using Switching Particle Swarm Optimization and Support Vector Machines",2014 Hindawi Publishing Corporation Discrete Dynamics in Nature and Society Volume 2015, Article ID 294930.

[11] Huang Dong, Gao Jian, Parameter Selection of a Support Vector Machine, Based on a Chaotic Particle Swarm Optimization Algorithm, Cybernetics and Information Technologies • Volume **15**, No 3 Print ISSN: 1311-9702; Online ISSN: 1314-4081.

[12] Prashant M. Kakde, Dr. S.M. Gulhane, A comparative analysis of particle swarm optimization and support vector machines for devnagri character recognition: an android application, Procedia Computer 1877-0509 © 2016 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license Peer-review under responsibility of the Organizing Committee of ICCCV 2016 doi: 10.1016/j.procs.2016.03.044 Science Direct 7th International Conference on Communication, Computing and Virtualization 2016. Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, pages 64-71, 2004.

[13] Gaurav Pandey, Zhaochun Ren, Shuaiqiang wang, Jari Vaijilainen, Maarten De Rijke"Linear feature extraction for ranking" Information Retrieval journal in Springer link, Volume 21, Issue 6, pp 481–506, December 2018.

[14] G. Salton. "The SMART Retrieval System: Experiments in automatic document processing". Prentice-Hall, Englewood Cliffs, NJ, 1971.

[15] J. Ponte and W. B. Croft. "A language model approach to information retrieval". Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, pages 275-281, 1998.

[16] Djoerd Hiemstra and Arjen P. de Vries," Relating the new language models of information retrieval to the traditional retrieval models", Published as CTIT technical report TR-CTIT-00-09, May 2000.

[17] S. Robertson and D. A. Hull. The TREC-9 Filtering Track Final Report. Proceedings of the 9th Text Retrieval Conference, pages 25-40, 2000.

[18] Y. Freund, R.I., R. Schapire, Y. Singer: "An efficient boosting algorithm for combining preferences", In Proceedings of JMLR 2003.

[19] Massimo Melucci, "On Rank Correlation in Information Retrieval Evaluation", ACM SIGIR Forum Vol.41 No. 1 June 2007.

[20] Jun Xu, Hang Li," AdaRank: A Boosting Algorithm for Information Retrieval", *SIGIR'07,* July 23–27, 2007, Amsterdam, The Netherlands.

[21] Ronan Cummins, Colm O'Riordan," Analysing Ranking Functions in Information Retrieval Using Constraints". In *SIGIR '09:* Proceedingsof the 32nd international ACM SIGIR conference on Research and development in information retrieval (pp. 251–258). New York, NY, USA: ACM.

[22] Tie-Yan Liu1, Jun Xu, Tao Qin, Wenying Xiong, and Hang Li, "LETOR: Benchmark Dataset for Research on Learning to Rank for Information Retrieval".

[23] Jun Wang, Xu Hong, Rong-rong Ren, Tai-hang Li," A Real-time Intrusion Detection System Based on PSO-SVM", 2009 ACADEMY PUBLISHER AP-PROC-CS-09CN004.

[24] Kehinde Agbele, Eniafe Ayetiran, Olusola Babalola," A Context-Adaptive Ranking Model for Effective Information Retrieval System". Copyright © 2018 The Author(s). Published by Scientific & Academic Publishing.