ISSN 2278-3091

Volume 10, No.4, July - August 2021

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse121042021.pdf

https://doi.org/10.30534/ijatcse/2021/121042021

Implementation of SVM+PSO Model in Monolingual and Cross Lingual Information Retrieval Ranking

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

[1]Computer Science,Banasthali Vidyapith, Jaipur, India

[2]Computer Science,Banasthali Vidyapith, Jaipur, India

[3]Computer Science,Banasthali Vidyapith, Jaipur, India

shweta.dubey12@gmail.com,mathur.iti@rediffmail.com,nisheeth.joshi@rediffmail.com

ABSTRACT

Now a day’s many research works are going on in the field of Information Retrieval Ranking. Retrieval and Ranking of information from the huge database of internet world is become a most useful and interesting task. Machine learning plays the major role now days. In this paper we have worked for monolingual and cross lingual information retrieval ranking in which the retrieval of document within the same language and retrieval of document in different language has been done. We take English language for monolingual IR ranking and for cross lingual we take Hindi queries and get documents in English. TREC 2008 QA Dataset and FIRE 2011 ADHOC Dataset has been taken in this work respectively. Finally the performance Evaluation has been done for SVM, PSO and SVM+PSO and the results has been compared. The results clearly explain that SVM+PSO model gives best result in compare to other two.

Key words: Information Retrieval, Machine Learning, PSO, SVM, Word2Vec.

1. INTRODUCTION

As in this paper we are talking about Monolingual and Cross Lingual Information Retrieval Ranking with the Machine learning techniques SVM, PSO and its combination SVM+PSO[11] with the help of Word2Vec feature selection technique and afterwards their Performance evaluation measures has been compared in the terms of Precision ,Recall, Accuracy and F1 score. So, let’s discuss one by one all of these terms.

Monolingual IR: In this Information retrieval process both source and destination language will be the same. For example when we give query in English you will get the relevant documents as output in English only. In this paper we have used English for our Monolingual Information Retrieval ranking.

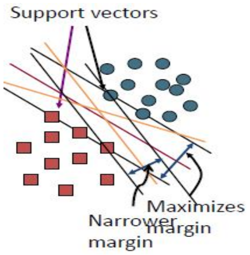
Cross Lingual IR: In this Information retrieval process you will get documents as an output in different language from the query language. For example if you give query in English you will get relevant documents in Hindi. In this paper we have used source language as Hindi and destination language as English.

Ranking: Ranking is the process after the retrieval of relevant documents based on our query in which the most or best matched documents will get the highest ranking score among all the documents. In this paper selection of relevant documents has been done using SVM classifier and after that PSO has been used to get the best optimized result as ranked documents. For Monolingual we have worked on TREC 2008 QA dataset in which queries are given in the term of questions and the document we get in the form of answers. After applying PSO the best answer is displayed.

For Cross Lingual we have taken FIRE 2011 ADHOC dataset in which we have trained our model on 50 queries in Hindi language and for documents retrieval we have worked on the concept of Corpus dictionary through which after preprocessing and word2Vec conversion we got our documents in English .The Corpus of English dataset contains newspaper headings from The Telegraph (2001-2010) and BDnews24.

Machine Learning: It is a part of AI in which machines are trained to predict outcomes for which explicitly they are programmed. As the name defined this term itself as machines learned to perform tasks or to give output based on historical data. In this paper the combination of SVM+PSO algorithm is designed to generate the models which are the machine learning techniques only. Actually Machine learning is basically related to that part of AI in which statistical models are also involved other than the neural network concepts.

SVM: It is the algorithm or model used in the case of supervised learning that is the part of machine learning. It is used as a classifier to differentiate between relevant and irrelevant data. Its name support vector machine is basically gives information as supporting vectors are used for the machine to train or to learn how to classify. In this line or plane or hyper plane is used based on the dimension or features used. The Figure 1 shows clearly how the supporting vectors play their role to differentiate.



.

Figure 1: Support Vector Machine

PSO: It is the algorithm or model based on stochastic behaviour or bird flocking or fish schooling.It is particle swarm optimization algorithm in which the best optimized solution we will get by local best and global best. This algorithm basically based on three factors that are velocity,direction and position.Swarm basically means group and particles basically talks about the solutions.The figure2 below shows clearly the concept of PSO.

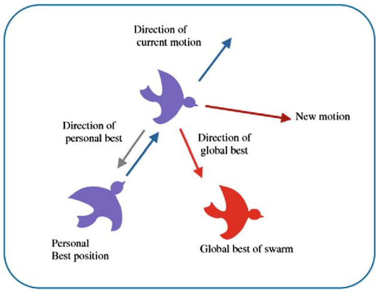


Figure 2: Particle Swarm Optimization

Word2Vec: It is the part of word embeddings. In this the sentences or the documents are converted in the vector form or in simple language it converts plain text or strings from their raw form to numbers to implement machine learning algorithms.

Performance Evaluation Measures: There are many performance evaluation measures but basically in machine learning the measures which are calculated or evaluated are Precision, Recall, Accuracy and F1 score.

Precision: In simple language we can define it as among all the predictions how many of your predictions are correct. So, we can formulate it as Precision=Correct positive predictions (TP)/All the predictions (TP+FP).

Recall: We can define recall as among all the truth samples how many of yours are right. So, we can formulate it as Recall=Correct predictions (TP)/All the truth samples (TP + FN).

Accuracy: It is defined as how many predictions are correct among all the samples. So, Accuracy =correct positive as well as negative predictions (TP+TN)/All the samples (TP+TN+FP+FN).

F1- score: It is just the Harmonic mean of precision and recall. It gives the overall performance of the model.

Objective of this work: The main objective of this research is to introduce the hybrid of SVM+PSO algorithm for IR ranking to get the most optimized result.

The other objective is among so much of features in the field of Information retrieval how to select the one or how to combine them into one. Here word 2vec concept is used in place of combining all the features and get the optimized one. The main purpose is to get the best ranking score performance evaluation values in comparison to other algorithms as SVM and PSO. This research gives the scope for further research in IR field using deep learning concept.

2. LITERATURE SURVEY

In 2001 , Djoerd Hiemstra[1] introduced the concept of relevance feedback with the help of two probabilistic models. The first one statistical based and the other was binary independence model. They did their experimental setup on TREC collection DATASET.

In 2006 Talvensaari et al.[2] focused on that corpus-based CLIR retrived more relevant documents. They created a Finnish-Swedish comparable corpus and used it as a source of knowledge for query translation. Finnish test queries were translated into Swedish and run against a Swedish test collection.

In 2007 Mandal et al. [3] described their experiment based on two cross-lingual and one monolingual English text retrievals at CLEF in the ad-hoc track. The cross-language task includes the retrieval of English documents in response to queries in two most widely spoken Indian languages, Hindi and Bengali.

In 2012, P.Sudhakar [4] focused on content and keywords in place of keyword and page ranking that search engines generated for the ranking of relevant documents. They used the advantage of full word matching against dictionary; user request is processed for search engine to obtain the results. Search results are extracted and sent for pre-processing.

In 2012 Sarkar et al. [5] presented a performance evaluation and an analysis of Bengali monolingual information retrieval task based on some selected models. Two models, TF-IDF and the Okapi BM25 model have been taken for their experiment. The developed IR models were tested on FIRE ad hoc retrieval data sets released for different years from 2008 to 2012 and the obtained results have been explained in their paper.

In 2014, Maryam yassi et al. [6] performed selection of appropriate and efficient features using Fisher ranking method considering classification power. Optimization of SVM classifier parameters including penalty factor C and Radial Basis function (RBF) parameter υ is performed utilizing chaotic accelerated Particle Swarm Optimization (CAPSO).

In 2015 Katris et al. [7] calculated the performance of two SMT (Statistical Machine Translation) systems within the CLIR pipeline. One is based on cloud and the other is open source software. They used two corpora from the medical domain.

In 2016,Kakde and Gulhane[8] introduced a combined technique that contains Particle Swarm Optimization and Support Vector Machine. This technique was introduced for Devnagari script recognition system.

In 2018, Gaurav Pandey et.al [9] focused on document ranking problem and concluded it as the feature extraction problem in information retrieval. They gave the idea of Life Rank algorithm for Ranking. In this algorithm they used the concept of matrix and its transformation.

In 2019, Jyoti Mor et.al [13] focused on designing of an effective crawler by ranking of URL, considering network load and number of retrieved pages.

In 2020, Othma and Faiz [10] enhanced the performance of existing QASs based on relevance feedback of the number of increasing correct answers for the particular question. They proposed a novel approach for retrieving and re-ranking passages based on n-grams and SVM models. The main concept behind it was that first they calculated the dependency degree using n-gram first then with the help of SVM improved the passage ranking with help of various similarity measures as lexical, syntactic etc.

3. PROPOSED METHODOLOGY

In figure 3 the first step is the preprocessing of the raw data by removing stop words, stemming etc., after the preprocessing it is transferred to knowledge base. Like the document preprocessing the queries or questions also preprocessed before conversion from word to vector. For the extraction of keywords the query is preprocessed and splitted. After that, the knowledge base and the splitted query are given to the clustering process. After the clustering process the SVM plays its role to classify the documents. To extract the matched documents the condition is applied. If it doesn't match, then update the knowledge base and then generate the answer. After the classification of the relevant documents the PSO plays its role finally to get the most optimized answer or document .The workflow of the MLIR (Monolingual Information Retrieval from English to English language is explained below:

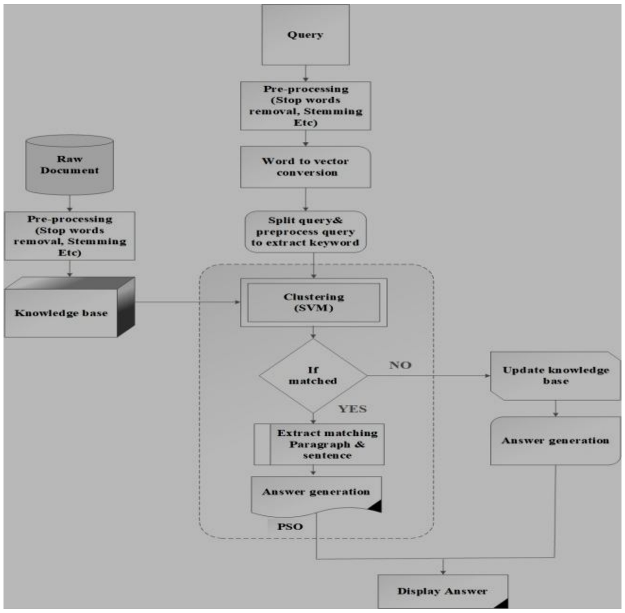

Query
Pre-processing (Stop words removal, Stemming Etc)
Raw Document
Word to vector conversion
Pre-processing (Stop words removal, Stemming Etc)
Split query& preprocess query to extract keyword
Knowledge base
Clustering (SVM)
If matched
NO
Update knowledge base
YES
Extract matching Paragraph & sentence
Answer generation
Answer generation
PSO
Display Answer

Figure 3: Processing of Information Retrieval Ranking (Monolingual) using SVM+PSO

In figure 4 Query is in Hindi language which is processed using Corpus Dictionary. After translation with the help of dictionary the preprocessing of the query took place like stop word removal, stemming tokenization etc. After preprocessing the query is converted from word to vector. After word2vector the query is splitted to extract keywords.
Now the extracted keywords and the raw document (after preprocessing) is clustered through SVM (Support Vector Machine) classifier.SVM classifier separated the relevant and non relevant documents using hyperplane. The relevant documents have been listed out, if there will not be any matching document then the knowledge base will be updated and the answer or sentences generated.
Among the list of relevant documents with the help of PSO (Particle Swarm Optimization) the most relevant document or answer will be displayed. The workflow of the CLIR from Hindi to English language is explained below:

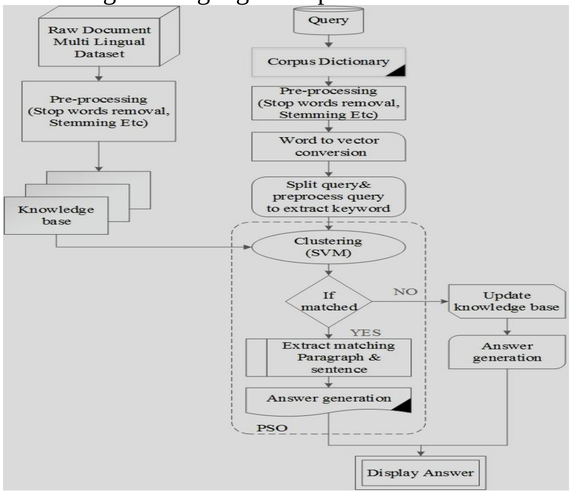

Raw Document Multi Lingual Dataset
Query
Corpus Dictionary
Pre-processing (Stop words removal, Stemming Etc)
Pre-processing (Stop words removal, Stemming Etc)
Word to vector conversion
Split query& preprocess query to extract keyword
Knowledge base
Clustering (SVM)
If matched
NO
Update knowledge base
YES
Extract matching Paragraph & sentence
Answer generation
Answer generation
PSO
Display Answer

Figure 4: Processing of Information Retrieval Ranking (Crosslingual) using SVM+PSO

4. EXPERIMENTAL SETUP

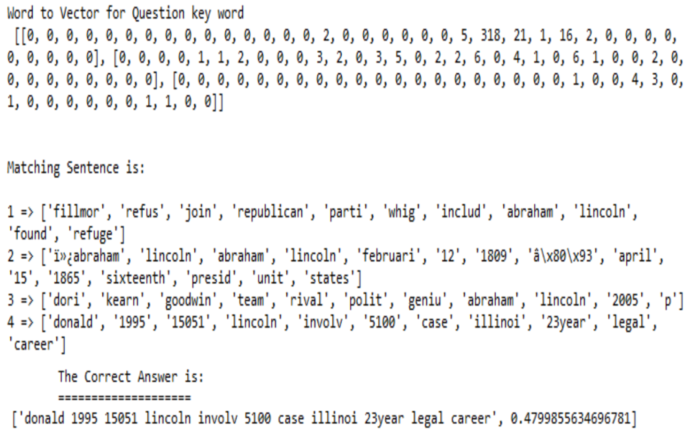
The experiments were conducted on a PC with 2.13 GHz Intel (R) pentium(R) P6200 CPU with 2GB RAM and the implementation was in Python version 3.6.5 Spyder version 3.2.8 under Windows 7 Enterprise Edition. Package name Version nltk 3.3 numpy 1.14.3 sklearn 0.19.1 matplotlib 2.2.2.

4.1 Result analysis for Monolingual Information Retrieval Ranking:

For Monolingual Information Retrieval the TREC 2004 QA DATASET has been used:

Question: How long was Lincoln's legal Career?

Take a query-How long was Lincoln's legal career?



```
Word to Vector for Question key word
 [[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 5, 318, 21, 1, 16, 2, 0, 0, 0, 0,
0, 0, 0, 0, 0], [0, 0, 0, 0, 1, 1, 2, 0, 0, 0, 3, 2, 0, 3, 5, 0, 2, 2, 6, 0, 4, 1, 0, 6, 1, 0, 0, 2, 0,
0, 0, 0, 0, 0, 0, 0, 0], [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 4, 3, 0,
1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0]]


Matching Sentence is:

1 => ['fillmor', 'refus', 'join', 'republican', 'parti', 'whig', 'includ', 'abraham', 'lincoln',
'found', 'refuge']
2 => ['ï»¿abraham', 'lincoln', 'abraham', 'lincoln', 'februari', '12', '1809', 'â\x80\x93', 'april',
'15', '1865', 'sixteenth', 'presid', 'unit', 'states']
3 => ['dori', 'kearn', 'goodwin', 'team', 'rival', 'polit', 'geniu', 'abraham', 'lincoln', '2005', 'p']
4 => ['donald', '1995', '15051', 'lincoln', 'involv', '5100', 'case', 'illinoi', '23year', 'legal',
'career']
        The Correct Answer is:
        ====================
 ['donald 1995 15051 lincoln involv 5100 case illinoi 23year legal career', 0.4799855634696781]
```

Answer is 23 year.
Best=0.4799855634696781

Table 1: Single term queries

Techniques/parameters	SVM			PSO			SVM+PSO		
	Types 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

Table 1 shows the performance evaluation values calculated from the three machine learning techniques for monolingual IR system for single term queries.

Table 2: Multi -term queries

Techniques/parameters	SVM			PSO			SVM+PSO		
	Types 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

Table 2 shows the performance evaluation values calculated from the three machine learning techniques for monolingual IR system for multiple term queries.

4.2 Result analysis for Crosslingual Information Retrieval Ranking:

The analysis has been done using FIRE AD HOC RETRIEVAL 2011 DATASET

Query 1: Give the Query== माइकल जैक्सन की असामयिक मृत्यु

Matching Sentence is:
1 => ['profil', 'champion', 'leagu', 'overshadow', 'uefa', 'cup', 'complet', 'day', 'untim', 'exit', 'milan', 'valencia', 'help', 'matters']
2 => ['nakagawa', 'untim', 'departur', 'major', 'blow', 'increasingli', 'unpopular', 'prime', 'minist', 'taro', 'aso']
3 => ['untim', 'strain', 'first', 'three', '2009', 'ash', 'test', 'angri', 'miss', 'final', 'match', 'seri', 'stake']
4 => ['comreut', 'bizarr', 'life', 'michael', 'jackson', 'complex', 'person', 'affair', 'take', 'stranger', 'twist', 'death', 'report', 'tuesday', 'question', 'parentag', 'children', 'caus', 'untim', 'demise']

The Correct Answer is:
===============
['comreut bizarr life michael jackson complex person affair take stranger twist death report tuesday question parentag children caus untim demise', 0.5145600658722396]

Table 3.Hindi terms in query and the related English terms in documents are given in below table:

Query term in Hindi	Meaning in English	Translation using corpus dictionary and word2vec
मृत्यु	A person or living thing is no more live	Death, demise
माइकल	Name of a person	michael
जैक्सन	Surname of a person	jackson
असामयिक	Without any time,untimely	untim

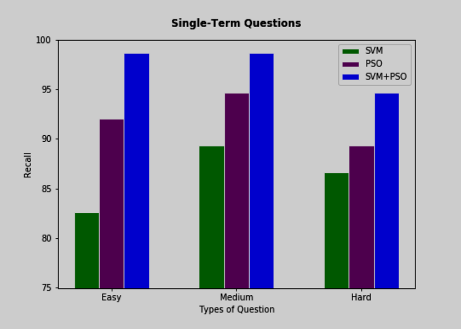
Table 4: Performance Evaluation for three techniques (CLIR)

Techniques/ Parameters	SVM	PSO	SVM+PSO
Accuracy	87.2	91.8	96
Recall	86	90.4	95.19
Precision	88.11	93	96.74
F1 score	87.04	91.68	95.96

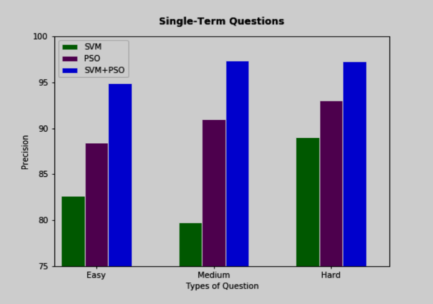
Table 3 shows an example of Hindi terms in query and the related English terms in documents that retrieved after running the query on our platform.

Table 4 shows the result of performance evaluation of three techniques SVM, PSO and SVM+PSO implemented on Crosslingual IR system.

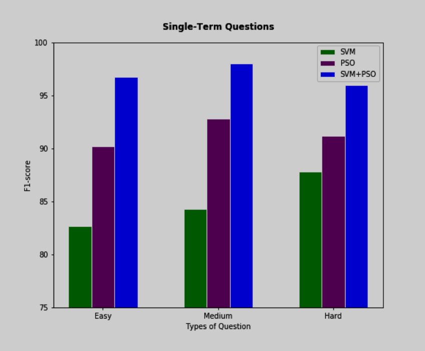
The graphs for single term and multiple term queries in the case of monolingual Information retrieval are defined below:



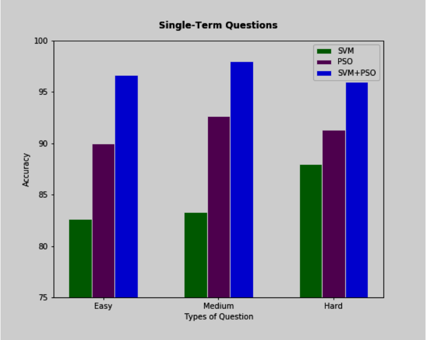
Recall



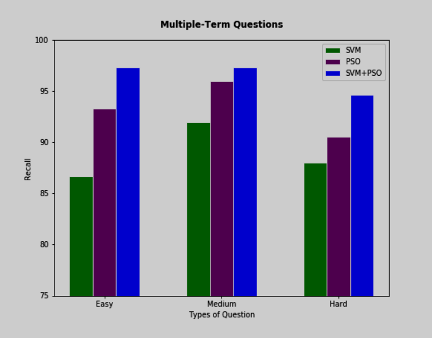
Precision



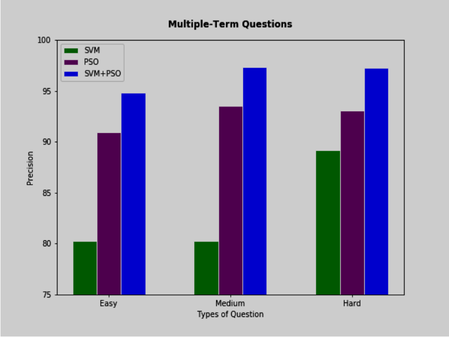
F1-score



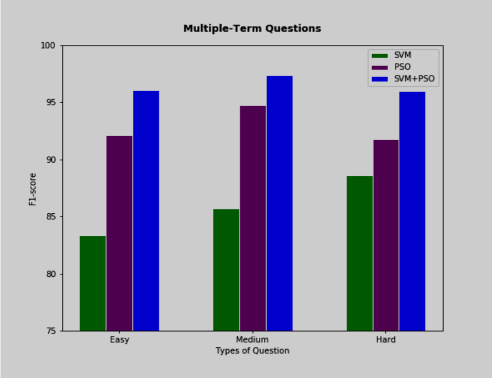
Accuracy



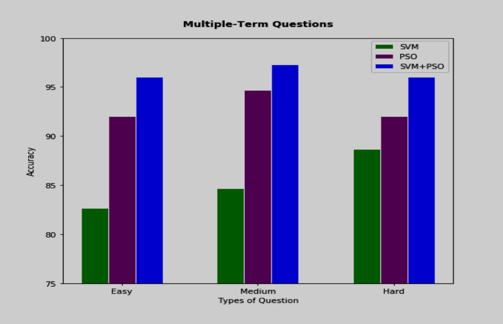
Recall



Precision

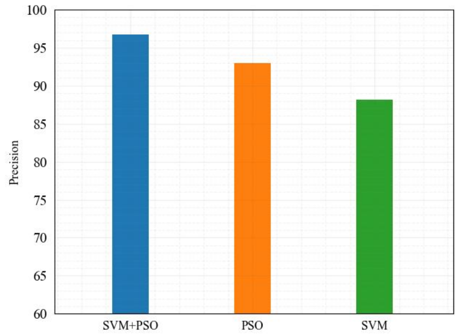


F1-score

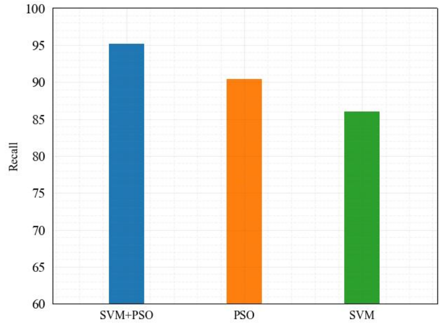


Accuracy

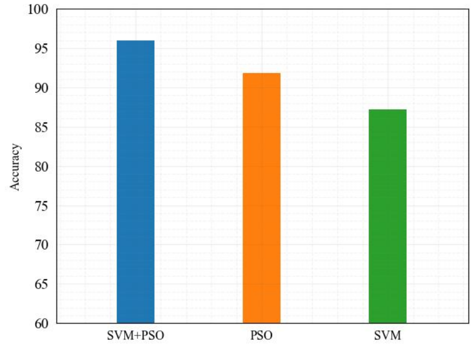
Based on the performances for SVM+PSO,PSO and SVM following graphs[12] has been generated for Crosslingual IR in which light blue color represents SVM+PSO, light yellow color represents PSO and green color represents SVM.



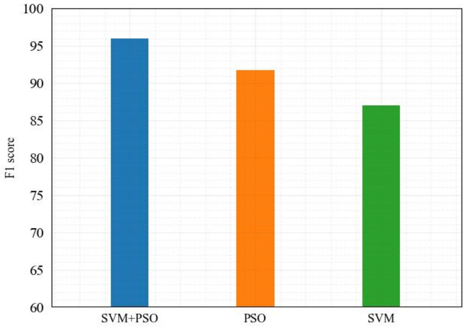
Precision comparison between SVM+PSO, PSO and SVM



Recall comparison between SVM+PSO, PSO and SVM



Accuracy comparison between SVM+PSO, PSO and SVM



F1 Score comparison between SVM+PSO, PSO and SVM

REFERENCES

1. 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.
2. Tuomas Talvensaari, Martti Juhola and Jorma Laurikkala, Kalervo Järvelin,” Corpus-based CLIR in retrieval Of highly relevant documents”, 2006, From ResearchGate.
3. Debasis Mandal, Sandipan Dandapat, Mayank Gupta, Pratyush Banerjee, Sudeshna Sarkar,” Bengali and Hindi to English Cross-language Text retrieval under Limited Resources”, Cross Language Evaluation Forum. http://clef-campaign.org ,2007, From Research Gate.
4. P.Sudhakar, G.Poonkuzhali, R.Kishore Kumar, “Content Based Ranking for Search Engines”, Member IAENG Proceedings of the international multiconference of engineers and computer scientists 2012 vol1, IMECS 2012 MARCH 14-16,2012 HONG, KONG.
5. Kamal Sarkar,Abhishek Gupta,” An Empirical Study of Some Selected IR Models for Bengali Monolingual Information Retrieval”, www. arxiv.org, 2012.
6. Maryam yassi, Mohammad Hossein Moattar,” Optimal SVM Parameters Estimation Using Chaotic Accelerated Particle 16 Swarm Optimization for Genetic Data Classification”, First International Congress on Technology, Communication and Knowledge (ICTCK 2014)November, 26-27, 2014 - Mashhad Branch, Islamic Azad University, Mashhad, Iran.
7. Nikolaos Katris,” Evaluation of Two Statistical Machine Translation Systems within a Greek-English Cross-Language Information Retrieval Architecture”,2015, UNIVERSITY of LIMERICK.
8. 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.Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, pages 64-71, 2004.

9. 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.
10. Nouha Othman,Rim Faiz," Question Answering Passage Retrieval and Re-ranking Using N-grams andSVM", ComputacionySystemas vol.20 no.3 México jul./sep. 2016 https://doi.org/10.13053/cys-20-3-2470
11. Shweta Pandey, Iti mathur, Nishith Joshi,"Hybrid model with word2vector in Information Retrieval Ranking",2020https://ujw.com.pl/wp-content/uploads/2020/06/Final_Springer_Proceeding.pdf
12. Shweta PANDEY,ITI MATHUR,NISHITH JOSHI," HYBRID OF SVM+PSO IN CLIR RANKING USING WORD2VECTOR",SOLID STATE TECHNOLOGY VOL. 63 NO. 5 (2020).
13. JYOTI MOR , NARESH KUMAR , DINESH RAI," AN IMPROVED CRAWLER BASED ON EFFICIENT RANKING ALGORITHM", INTERNATIONAL JOURNAL OF ADVANCED TRENDS IN COMPUTER SCIENCE AND ENGINEERING HTTPS://DOI.ORG/10.30534/IJATCSE/2019/04822019, VOLUME 8, NO.2, MARCH - APRIL 2019, ISSN 2278-3091.