International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

# A Novel Multi-Parameter Tuned Optimizer for Information Retrieval Based on Particle Swarm Optimization

**Narina Thakur, Deepti Mehrotra, Abhay Bansal, Manju Bala**

***Abstract*: *Tuning multi-parameter and parameter optimization in Information Retrieval has been a huge area of research and development, especially with BM25F scoring functions having a 2F+1 feature with F fields in the documents. The scoring and ranking function conventionally uses multiple input parameters, to augment the quality of results even at the value of huge calculation time. The searching and ranking documents in the medical literature encompass high recall rates, which are difficult to satisfy with multiple input parameters. The performance of the BM25F depends upon the choice of these F parameters. Particle Swarm Optimization (PSO) searches through the solution- space independently and discovers an optimal solution as opposed to improving and optimizing the gradient; henceforth it can straightforward optimize Mean Average Precision (MAP) a non-differentiable function. In this paper, the usage of PSO to tune multi-parameters is proposed to deal with the gaps in BM25Fscoring function. Also, the advantage of the proposed technique by directly optimizing the MAP has been discussed. Experimental results of quantitative performance metrics MAP and Mean Reciprocal Rank of the proposed PSO-optimized BM25F and most recent ranking algorithms have been compared. The performance measure results demonstrate that the proposed PSO-optimized BM25F performance measure outclasses the standard ranking methods for the OHSUMED data set.***

***Index Terms: BM25, BM25F, OHSUMED, Particle Swarm optimizations, Similarity Score.***

## I. INTRODUCTION

Recent years have observed a rapid expansion in the quantity of data in the area of medical, Bioinformatics and furthermore, the size of the data is expected to surge exponentially in the near future. This begets the need for an efficient search, retrieving relevant information from a corpus based on user need and ranking methods. Consequently, Information Retrieval (IR) has therefore emerged as an active, mature and well-understood technology area in the industry, academia, medicine and law. Relevance is a perception and relevant IR and searching are the major challenges in Information technology. IR models mainly consider the webpage or documents usually unstructured text and a similarity/scoring algorithm is usually developed to evaluate the similarity between documents and query. However, the TREC 9 OHSUMED dataset's documents used in this paper are structured, consisting of different field identifiers like title, author, content and metadata descriptors. Best practices and models were developed in the recent past by the research community to address many search viability issues as timely access to accurate information in the different scenario; rapidly expanding data, navigation, and searching techniques. Information is nothing without retrieval and many ranking and retrieval models have been introduced in the past based on set-theoretic, algebraic and probabilistic mathematical models. The examples of set-theoretic models are standard Boolean model (BM) and Extended Boolean M. The BM estimates binary relevance, that is, either the document is relevant or irrelevant for the queries. Salton in 1975 proposed the Vector Space model (VSM) [1], Latent semantic, Topic-based VSM, Balanced topic based VSM, spread activation neural network and Backpropagation neural network are the examples, algebraic models. In VSM the documents and queries are represented by vectors and the relevance and similarity score is the cosine of the angle or inner product [2] between them. The mathematical probabilistic models are Belief Network, Inference Network, Binary Independence, Language Model and retrieval by Logical Imaging. In Probabilistic modeling [3] the relevant probability of a document and user query is calculated.

Okapi BM25 (Best Match) Probabilistic retrieval model uses term frequency (TF), the standard document length normalization and Inverse Document Frequency (IDF). In Bm25 the optimal value for the parameter b and k1 and in BM25F with F fields the free parameters for each fields, k and b has to be carefully chosen. There are many optimization methods have been proposed in literature. "Black-Box" optimization framework [4] addressed the Parameter optimization problem in retrieval function as it is more efficient and reliable as compared to grid optimization method which is computationally inefficient and expensive with many parameters.

Manuscript published on 30 September 2019
* Correspondence Author
**Narina Thakur*,** CSE Department, Bharati Vidyapeeth's College of Engineering New Delhi India and ASET Amity University Uttar Pradesh, Noida, India.
**Deepti Mehrotra**, IT Department ASET Amity University Uttar Pradesh, Noida, India.
**Abhay Bansal**, CSE Department, ASET Amity University Uttar Pradesh, Noida, India.
**Manju Bala,** CS Department, IP College for Women, Delhi University, Delhi.

The Grid search is inefficient with generalization with more parameter in retrieval function. "Black-Box" optimization does not require an investigative objective function.

The "Black-Box" optimization framework has been applied in Bm25 to tune b and k1 free parameter and BM25F with f fields where 2f + 1 free parameters are tuned: k1, plus free parameters for each fields, bs and vs. The "Black-Box" optimization paradigm is effective with many free parameters and the objective function is unknown. In BM25 there is no distinction between its document fields, which generally leads to an overestimation of the significance of terms in ascertaining the scores through fields. The main limitation of BM25 is that it ranks long documents that do not match the query term, higher rather than similar as compared to a short document that does not contain the query word at all. Hence the proposed research work uses BM25F rather than Bm25 retrieval function. BM25F uses diverse length normalizing factor b and distinctive weight, w for all the fields like title, the body of text, headlines, and anchor text are best suited for solving its limitations. These 2F+1parameters influence the contribution of the TF, field length, and width. However, the effectiveness of BM25F depends upon the estimation of optimal value estimation for these 2F+1parameters. Simple searching of the entire solution space works well in Optimization with the small numbers of parameters and choosing the values which optimize the performance of the selected evaluation measure [5] .The optimization time required surges exponentially about the parameters count, hence Grid search, despite having the capacity to discover rational parameter set for specified target cost can be too slow when scaled to substantial large data collection and dimensionalities. Despite the fact that Gradient descent is significantly speedier, yet it cannot optimize the objective parameters for a particular target cost measure, and the outcomes acquired do not come out to be any superior to grid search [6]. The main emphasis of this research presented here is free parameter's value estimation in BM25F. The aim is to review various optimization methods and implement the proposed PSO-tuned BM25F retrieval algorithm for an efficient retrieval. In traditional searching the agent may get stuck in local minima rather than global maxima, if the used function is not continuous, in that case, its derivative doesn't exist and hence the gradient technique fails. PSO can be utilized to tackle such issues. PSO is a Bio-inspired algorithm inspired by bird flocking social behavior. It has turned out to be a standout amongst the most robust, stochastic parallel and optimized metaheuristic searching techniques. It depends on movement, knowledge or intelligence of the swarm. It is capable to find the optimal solution in multidimensional solution spaces. For OHSUMED medical corpus experiment has been carried out and the obtained results of the proposed PSO-optimized BM25F method has been compared with BM25 and other IR methods for standard IR evaluation measure. This paper is segregated into five sections as: Research gaps are identified in Section 2 Related work, Section 3 proposed PSO-tuned BM25 model, followed by the Experiment setup in Section 4. Results of the Experiment and its analysis have been elaborated in section 5, and finally the paper is concluded last section.

## II. RELATED WORK

This section presents the review of literature in IR optimization in the context of free parameter value estimation for retrieval functions. The various existing free parameter estimation optimization methods that are used in IR in the area of retrieval uses standard optimization mentioned like ant colony optimization (ACO), PSO, Direct optimization. Table I elaborates the various Optimization methods used in IR for estimating the values for free parameters. BM25 and BM25F models and similarity score formulae have been discussed in equation (1) to (4).

BM25 is commonly used in IR ranking function due to its consistently high retrieval accuracy. BM25 [7] is based on the probabilistic retrieval model developed by Stephen E Robertson in 1970 and 1980. It has evolved from the BM approximations to the 2-Poisson model. The formulae for the Okapi BM25 similarity score function for the t term in d document is as:

$$BM25(t,d)=\sum_{t\in q} Log\left(\frac{N-n_t}{n_t}\right)\times\frac{(k_1+1)\,fd,t}{K+fd,t}\times\frac{(k_3+1)\,fd,q}{k_3+fd,q} \quad (1)$$

where

N: documents count in the corpus,
$n_t$ : Total count of documents that contain t term.
q: query
fd, t: t term occurrences count in d document i.e. t term, tf in the current document.
fd, q: Occurrence count of t term in d document

$$k=k_1.\left((1-b)+\frac{b.dld}{avl}\right) \quad (2)$$

where
dld: terms count in document d.
avl: document Average length

Let the tuning parameter k1, k3 and b can have values as 1.2, 1000, 0.75 respectively.

$$BM25(d,q)=\sum_{t\in q}\frac{log\left(1+freq(t,d)\right)}{\log\left(1+avg\left(freq(t,d)\right)\right)}\times\left(1+log\frac{numDocs}{docFreq(t)+1}\right) \quad (3)$$

$$\times boost(t,d)\times\sqrt{0.8\,avg\left(\#uniqueterms\right)+0.2\,\#uniqueterms(d)}$$

where
avg : average of freq
Boost (t. field, d): boosting factor assigned in indexing for the t term field in document d.

The normalized similarity score used for document d with the search term and tf-idf weight [8] with q query is as in equation (6).
The Tuning parameters enable us to control the length normalization thereby improved retrieval results.

The optimal value of these parameters can be determined for the test corpus using documents, queries, and judgments and optimize effective retrieval metric like Mean Average Precision.

(BM25 retrieval function fails through lengthy documents and hence, BM25F a variant of BM25 consider the structure of the documents and other metadata fields like anchor tags, author, title and abstract as important fields in score calculation. A BM25F score formula is as in equation (4).

$$BM25F(t,d)=\sum_{t\in q\cap d}\frac{tf(t,d)}{k_1+tf(t,d)}\times\frac{N-n_t+0.5}{n_t+0.5} \quad (4)$$

Where

C : fields in d document

N: documents count in the corpus,

$n_t$ : Total count of documents that contain t term.

$tf_c(t,d)$ : Term-Frequency of t Term in f Fields of d document and the formulae is as in equation (5).

$w_c$ :document each field Weight/boost factor as described by the equation (6).

$$tf(t,d)=\sum_{c\in d}w_c\times tf_c(t,d) \quad (5)$$

$$w_c=\frac{1}{(1-b)+b\frac{\sum_{t\in d}tf(t,d)}{avgdl}} \quad (6)$$

**Table I. Literature review on Parameter value estimation using some standard IR optimization methods**

| Reference No. | Description |
|---|---|
| [9] | Wang has proposed a PSO based Document classifier for TREC and Reuter dataset which optimize the parameters implicit in the documents. |
| [10] | The choice of initial cluster centroid is sensitive in calculating the clustering results. It may converge and get stuck in a local optimal solution. Cui et al. has proposed a Hybrid PSO with K-means document clustering algorithm that performs fast document clustering by incapacitating the gap of the standard K means document clustering algorithm to solve the convergence and optima problem. |
| [11] | Learning to Rank model in Information Retrieval aims at ranking documents according to user Preferences. In machine learning coalescing models is a widespread; often implemented by averaging the predictions of models. A novel Sim-Multi Leave gradient that uses the reference document similarities for ranking that addresses the speed concerns in aspect and quality trade and gathers quickly henceforth, provides an improved user experience has been proposed along with a cascading Online Learning to Rank (C-OLTR) approach, called as Cascading Multilevel Gradient Descent (C-MGD) which uses, a fast simple model and a slower complex model in online Learning to Rank has been presented. |
| [12] | Direct Optimization algorithm is straightforward to implement by calculating the parameter values that satisfy the metric criteria. The metric satisfying criteria can be maximization or minimization. The proposed research work has tuned the parameters for BM25F by minimizing the cost function of RankNet and maximizing NDCG score. |
| [13] | This paper presents an effective learning function for documents ranking by investigating significant and exclusive documents features by using Genetic Programming in evolutionary computation methods. The optimization criteria used in Swarm Rank learns method is a combination of content and hyperlink features which directly maximizes the widely used evaluation measure Mean Average Precision. |

PSO search algorithm is based on population [14][15], put forward by James Kennedy and Russell C. Eberhart which employs the social behavior of bird's simulation within a flock. Various machine learning and data mining applications [16-21] optimize using PSO. Work in [22], applies a novel hybrid genetic algorithm based computational approach to text clustering optimization. Authors of [23] have considered an adaptive inertia weight method on Bayesian technique and Cauchy mutation to overcome the defects of PSO's searching ability. The work in [24], addresses the Optimization by Artificial Bee Colony (ABC) query paraphrasing method. The voting multiple hyperplanes for document retrieval for learning to rank LETOR has been proposed in research paper [25], which employs multiple hyperplanes, based Ranking SVM with base rankers and vote strategically for estimating the relevance score. It uses Swarm ranking to tune content and hyperlink features for LETOR data sets. A differential evolutionary inspired based RankDE method [26] was proposed to rank the documents. The obtained results from their experiments have demonstrated that evolutionary-based RankDE is better than PSO and Genetic Programming (GP) for LETOR data set. "Eberhart and Kennedy in 1995", coined PSO as stimulated by the communal activities of biological organisms. The problems of PSO was initially addressed to outline the capacity of gatherings or creatures to work in finding desired positions in a given boundary or territory for e.g. a group of birds looking for food. Swarm a populace of specialists, is used in PSO calculation that travel through the multi-dimensional problem space for finding an optimization solution to the problem i.e. food in the case that is presently being considered, refreshing their speed as indicated by the data gathered by the particles or bird. Each particle location or position represents a solution to the problem in the problem space. With every movement, another arrangement is acquired that can be considered as another solution and is assessed by the criteria

(minimization or maximization) of the fitness function which gives a measurable estimation of the result. Every particle is looking for the best result that it has established and its neighborhood particle has established. PSO calculations are global optimization non-linear algorithms that won't require cost approximation as compared to gradient [9].

PSO is an intelligent Metaheuristic classed optimization algorithm inspired by social behavior developed by James Kennedy and Russell C. Elberhart in 1995. PSO can be applied to many fields in the area of Image Processing, Searching and Operational research. In PSO SWARM intelligence behavior is created by means of some agent's fish and birds. The level of intelligence achieved in PSO cannot be reachable by any agent individually but can be reached by cooperation. The main principle used in PSO is communication and learning of better. PSO contains a population of candidate solution called as Swarm. In PSO the Swarm of particles is considered and every particle is a candidate solution to the optimization problem in the solution space. Any particle has a position in the search space optimization problem which is the set of possible solution in the search space and the aim is to find the best solution in search space as shown in the Figure 1.

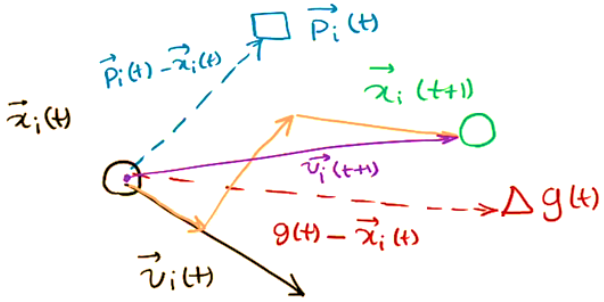


**Figure1. Geometrical representation of particle in Particle Swarm Optimization [27]**

Consider the mathematical of PSO with i particle where i is the index of particle and x is a position vector let $\vec{x_1}(t)$ be the particle i vector with $\vec{v_1}(t)$ velocity vector with dimensions 1. and direction as x. This particle is a member of swarm and a particle is learning from each other and neighbors. The particle got its personal best location $\vec{p_1}(t)$ for every particle and a global best g(t). The particle moves parallel to the personal vector, new velocity and global best. Every particles movement is based on inertia term, cognitive component and social components. Every agent position is jointly influenced by the agents own position, velocity, best position and swarm experience. For updating the velocity, position of particle parallel to personal position, personal best and global best as represented in the equation (7) and (8).

$$v_{ij}(t+1) = w \times v_{ij}(t) + c_1 \times r_1 \times \left(p_{ij}(t) - x_{ij}(t)\right) + c_2 \times r_2 \times \left(g_j(t) - x_{id}\right) \quad (7)$$

$$x_{ij}(t+1) = x_{ij}(t) + v_{ij}(t+1) \quad (7)$$

where

$r_1$ and $r_2$ are the uniformly distributed random number in the range of 0 to 1.

$c_1$ and $c_2$ are acceleration coefficients.

## III. PROPOSED PSO-TUNED BM25F RETRIEVAL MODEL

This section presents the proposed PSO tuned BM25F model with b PSO in (2F+1) dimensional space representation and its algorithm. A PSO optimization method is used to tune the BM25F retrieval function as discussed in Section 2 with maximizing the objective function i.e. retrieval efficiency **"Mean Average Precision"**. The objective of the proposed method is to tune or optimizes the parameter values in BM25F using Particle Swarm Optimization technique. BM25F offers supplementary flexibility to user to enhance the efficiency of the IR model. The parameter values can be tuned with a wide array of free parameters and hence efficiency can be significantly improved. The most retrieval functions have "Free parameters" which significantly affect the performance of the retrieval model. The free parameters must be set before the retrieval, and the iteration is chosen as 50 and population of 100 particles. The efficiency of the retrieval function is affected generally by an inefficient Parameter optimization method when repeated optimization is performed. Mean Average Precision is the objective function selected in this proposed PSO based BM25F retrieval model. Let "**m**" represent the numbers of documents present in the corpus and ith particle vector can be mathematically represented as a vector with **"d"** dimension as [free parameter : <k1>; field weights for f fields : < w1, w2... wf >; b values for f fields <b1, b2... bf>]. The aim of the proposed PSO tuned BM25F retrieval model is to maximize the MAP for 63 batched queries in the given topic file. In the proposed PSO tuned BM25F model title and content fields are considered, i.e. F=2 hence 2F+1 = 5, hence the particle position vector will be 5 dimensional.

**Algorithm1.** PSO tuned BM25F retrieval model

**Input:** OHSUMED Topics, Relevance Judgements for OHSU topics, qrels.

Initialization

Initialize the ith particle with $x_i$ position and velocities $v_i$ in (2F+1) dimensional parameter solution space, where free parameters considered are k1 used in equation (4) in section 2 and considered field parameters $x_i$ [1: nvar], nvar=5 number of unknown variable or the parameters [k1, tw, cw, tb, cb].

- Initialize the particle structure with position $x_i$ , velocity $v_i$ personal best position $p_i$ and population best $p_g$ .
- Initialize the inertia, and acceleration parameter w=1, $c_1 = 2$ and $c_2 = 2$ .
- Initialize the Maximum iteration MaxIt=50 and Population size nPop= 100. The output is the optimal value of the free parameters k1, w and b for the selected fields.
- Initialize

r1=rand(varSize) and r2=random(varSize)

1. Set $p_i$ of each particle of xi
2. $p_k = x_j$ , such that $MAP(x_j)$ for all particle j in swarms.
3. Two field parameters considered are content and title their weight and b value are assigned to the particle $x_i$ is 5 dimensional array $x_i\ [1:5] \rightarrow [k1,\ tw,\ cw,\ tb,\ cb]$
4. //Create Population Array
   For i = 1: nPop
   $x_i.best.MAP = x_i.MAP$ //generate particle solution
   //update Global best
   if $x_i.Best.MAP > Global.Best.MAP$
   $Global.Best.MAP = x_i.Best.MAP$
   End if
   End For
2. For it =1:MaxIt
   For i = 1: nPop
   // update particle velocity

$$v_i \rightarrow w \times v_i + c_1 \times r_1 \times (p_i - x_i) + c_2 \times r_2 \times (p_g - x_i)$$

   // update particle location
   $x_i \rightarrow x_i + v_i$
   //calculate the MAP for particle $x_i$
   $x_i.MAP$

   if $x_i.MAP > x_i.Best.MAP$
   //update particle best MAP and location
   $p_i = x_i$
   $x_i.Best.MAP = x_i.MAP$
   if $x_i.Best.MAP > Global.Best.MAP$
   /update Global best
   $Global.Best.MAP = x_i.Best.MAP$
   End
   End
   End
   End
3. For it =1:MaxIt
   For i = 1: nPop
   $x_{it}.Best.MAP = Global.Best.MAP$
   Print $(it, x_{it}.Best.MAP)$
   End
   End

---

The proposed model is implemented using whoosh library in Python for TREC 9 OHSUMED data set as shown in the Figure2.

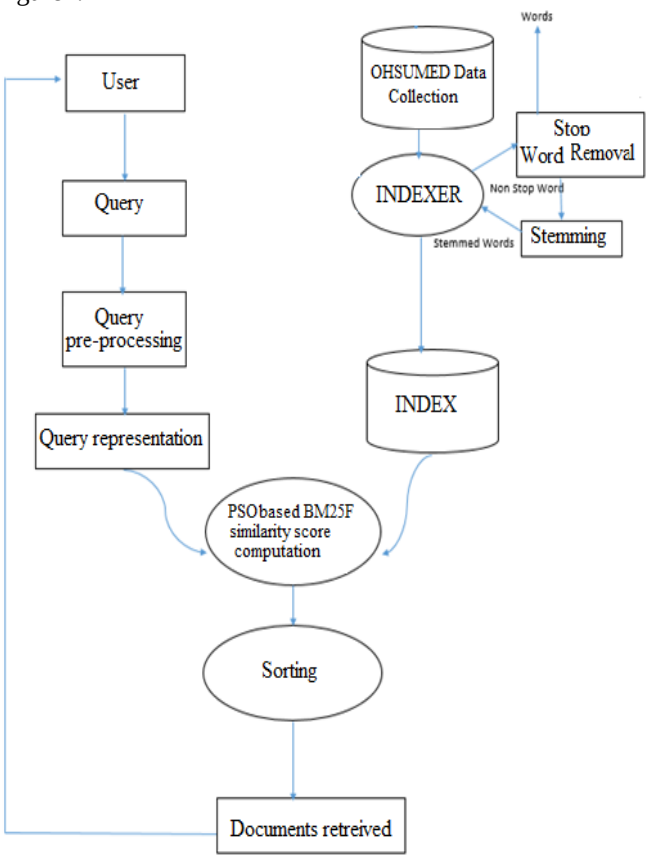


**Figure2. Proposed PSO tuned BM25F Model**

Initially, the test QRELS file path is set in the search python script and documents are retrieved from the OHSUMED repository using the following steps as:

1) The OHSUMED repository contains the meta-data attributes of the documents. The documents stored in the corpus needs to be indexed for fast search and retrieval. The documents are preprocessed with stop word removal and stemming followed by document indexing using inverted index keywords for the schema objects.

2) The full-text search index is employed that is an inverted index is created on every string field or keywords of the document collection.

3) The keywords present in the batched queries are mined or parsed with the Query Parser and the document objects are retrieved by the searcher from the index which is matched with the QRELS keywords.

4) Each retrieved document or results object got a similar score for the QREL. The score is given on the basis of the document term frequency and PSO based BM25F ranking.

5) Ranking module tries to find more important result document. These documents or results object is retrieved and returned in reverse sorted order of document similarity score.

## IV. EXPERIMENTAL SETUP

In this section TREC-9 OHSUMED dataset used for the experiment has been illustrated with the performance evaluation methodology.

### A. Data set

The experiment was performed on the OHSUMED TREC-9 dataset [28]. The Text Retrieval Conference was started in 1992. The details of TREC OHSUMED dataset is as in Table II. The dataset is collected from 1987 to 1991, with a typical testing set size of 54710 documents [29]. Figure 3 illustrate a snapshot of sample OHSUMED training document.

```
.I 1
.U
87049087
.S
Am J Emerg Med 8703; 4(6):491-5
.M
Allied Health Personnel/*; Electric Countershock/*; Emergencies;
Emergency Medical Technicians/*; Human; Prognosis; Recurrence; Support,
U.S. Gov't, P.H.S.; Time Factors; Transportation of Patients; Ventricular
Fibrillation/*TH.
.T
Refibrillation managed by EMT-Ds: incidence and outcome without paramedic
back-up.
.P
JOURNAL ARTICLE.
.W
Some patients converted from ventricular fibrillation to organized
rhythms by defibrillation-trained ambulance technicians (EMT-Ds) will
refibrillate before hospital arrival. The authors analyzed 271 cases of
ventricular fibrillation managed by EMT-Ds working without paramedic
back-up. Of 111 patients initially converted to organized rhythms, 19
(17%) refibrillated, 11 (58%) of whom were reconverted to perfusing
rhythms, including nine of 11 (82%) who had spontaneous pulses prior to
refibrillation. Among patients initially converted to organized rhythms,
hospital admission rates were lower for patients who refibrillated than
for patients who did not (53% versus 76%, P = NS), although discharge
rates were virtually identical (37% and 35%, respectively). Scene-to-
hospital transport times were not predictively associated with either the
frequency of refibrillation or patient outcome. Defibrillation-trained
EMTs can effectively manage refibrillation with additional shocks and are
not at a significant disadvantage when paramedic back-up is not
available.
.A
Stults KR; Brown DD.
```

**Figure3. A Sample OHSUMED training document**

The document fields that are considering for the experiment are title and contents.

**Table II. TREC OHSUMED Dataset description**

| Sr. No. | Descriptor | Value |
|---|---|---|
| 1. | Dataset Name | TREC-9 |
| 2. | Track Name | Filtering Track |
| 3. | References | 348,566 |
| 4. | Type of Data Source | online Medical information database |
| 5. | MEDLINE FIELDS | .I : Document identifier,<br>.U: unique identifier,<br>.T: title,<br>.W: abstract,<br>.M: MeSH terms,<br>.A: author,<br>.S: source<br>.P: publication type |
| 6. | MSH/ MSH-SMP terms | 4904 MeSH terms with definitions along with 500 sub-sets of the MeSH terms. |
| 7. | Medical journals | 270 |
| 8. | Duration of dataset collection | 5 years |
| 9. | Size of Training set | 54710 documents |

### B. Evaluation Methodology

The core standards for IR evaluation analysis are Mean average precision (MAP) and recall. The BM25F retrieval model uses the proposed PSO parameter tuning method as discussed in section 3, to estimate the value of k, cw, tw, cb and tb for the two fields i.e. title and content. The optimization iteration considered was 50, for m=100 particles. The proposed model was then executed on the test OHSUMED data set and the results are compared with the other models. Precision and recall are the measure for accuracy of an IR system. They measure the accomplishment of the retrieval methods. Here we are using he standard Precision, Mean average Precision (MAP) and Recall measures. The main problem with the Precision and Recall measure is that it requires knowledge of all the documents stored in the corpus for the proper estimation of recall.

Let r be the count of relevant document retrieved, R be the count of relevant documents in the corpus and T be the document retrieved.

#### *1) Precision (P):*

Precision [30] is an evaluation measure the retrieved documents are relevant. It is the portion of relevant documents count to the document count retrieved as described in equation (9).

$$\text{Precision} = \frac{r}{T} \tag{8}$$

Precision measures effectiveness over a set of queries processed in a batch mode. The relevance judgement Precision oriented metric (T9P) is considered for TREC-9 OHSUMED data-set. It is the fraction relevant retrieved documents to the maximum 50 retrieved documents as described in equation (10).

$$\text{T9P} = \frac{r}{\max\left(number\ retrieved\ ,50\right)} \tag{9}$$

#### *2) Recall (R'):*

Recall is the numbers of relevant document are retrieved and it gives the coverage of the result. It is the fraction of count of documents retrieved to the corpus relevant document count as described in equation (11).

$$\text{Recall} = \frac{r}{R} \tag{10}$$

100 percent recall level can be achieved when all the relevant documents are retrieved in the answer set.

#### *3) Mean Average Precision (MAP):*

It is the mean of precision [31] values of all the queries as described in equation (12).

$$\text{MAP} = \frac{\sum_{q=1}^{\text{number of Queries}} \text{Precision}(q)}{\text{Number of queries}} \tag{11}$$

## V. RESULTS & ANALYSIS

Retrieval results specify that the proposed PSO-tuned BM25F model effectively optimizes the BM25F model for the given OHSUMED data set,

surpassing the standard model un-optimized BM25F on standard IR metrics. The main benefit is the computational efficiency that has been achieved from optimization. The optimal values of the free parameters (k1, title weight (tw), content weight (cw), (tb), (cb)) obtained from the proposed PSO tuned BM25F ranking function are as in Table III.

**Table III. PSO-optimized BM25F optimal free parameter values after applying the PSO algorithm**

| Sr. No. | Parameter Name | PSO Optimized parameter value |
|---|---|---|
| 1. | k1 | 1.5 |
| 2. | Tw | 2.0 |
| 3. | Cw | 0.5 |
| 4. | Tb | 1.0 |
| 5. | Cb | 0.5 |

The proposed PSO-optimized BM25F performance measure's with respect to MAP, MRR and R-precision have been compared to BM25, Pl2, and TF-IDF standard models

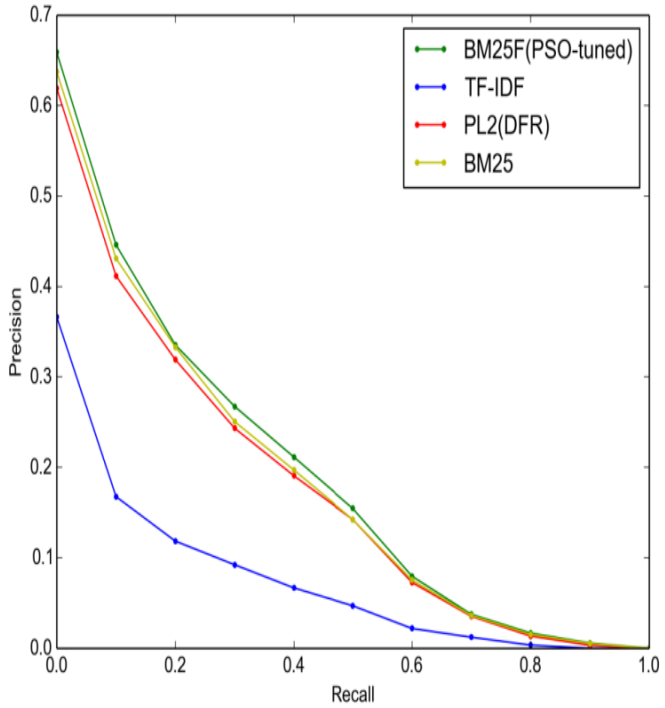


**Figure4. Interpolated Precision-Recall Curves for proposed PSO tuned BM25F and UN-optimized BM25F.**

The Quantitative Evaluation measures of the PSO tuned BM25F model has been compared with other models as in Table IV.

**Table IV. PSO-optimized BM25F model Performance with BM25, Pl2, and standard TF-IDF models.**

| Model | MAP | MRR | R-precision |
|---|---|---|---|
| **PSO-optimized BM25F** | 0.19 | 0.62 | 0.26 |
| **BM25** | 0.17 | 0.58 | 0.23 |
| **PL2(DFR)** | 0.16 | 0.56 | 0.22 |
| **TF-IDF** | 0.06 | 0.32 | 0.11 |

The precision-recall curve [32] plots Precision score on x-axis and Recall on y-axis. Instead of using precision and recall on at each rank position, the curve is normally uses interpolated precision and recall, P@0%, P@10%, p@ 20%..... P@100% and R'@0%, R'@10%, R'@ 20%..... R'@100level. Figure4 depicts the Interpolated Precision-Recall curve. The proposed PSO tuned BM25F performing better for the lower recall measure where as it gradually decreases for higher recall measures. The findings clearly indicates that the proposed PSO tuned BM25F with two fields content and title demonstrates the highest values precision at all the recalls values whereas the difference between MAP , MRR and R-precision is marginal.

## VI. CONCLUSION

In this paper, various IR tuning and parameters value estimation using optimization methods have been presented and discussed. The Bio-inspired PSO optimization method has been proposed to tune BM25F function in estimating the free parameters values. The experiments have been conducted on OHSUMED TREC-9 Medical data-set. The proposed model has been assessed by the standard IR performance metrics like Precision, Mean Average Precision and Recall. Results from experimentation demonstrate that the proposed PSO-optimized BM25F model with optimization criteria MAP and the content and title fields' considered outclass the recent ranking methods. Forthcoming the ranking research work can further be extended by exploring the use of other optimization algorithms, identification and selection of a suitable fitness function and other Machine learning algorithms to tune the retrieval function parameters.

## REFERENCES

1. G. Salton, A. Wong and C. Yang, "A vector space model for automatic indexing", Communications of the ACM, vol. 18, no. 11, 1975. Available: 10.1145/361219.361220, pp. 613-620.
2. Narina Thakur, Deepti Mehrotra, Abhay Bansal, and Manju Bala, "Analysis and Implementation of the Bray Curtis Distance-based Similarity Measure for Retrieving Information from the Medical Repository." International Conference on Innovative Computing & Communication Lecture Notes in Networks and Systems, Print ISBN978-981-13-2353-9, Online ISBN978-981-13-2354-6, vol. 56, Springer, Singapore, 2019, pp. 117-125.
3. Robertson, S. E., & Jones, K. S., "Relevance weighting of search terms", Journal of the Association for Information Science and Technology, vol. 27, no. 3, 1976 pp. 129-146.
4. A. Costa, E. Di Buccio, M. Melucci and G. Nannicini, "Efficient Parameter Estimation for Information Retrieval Using Black-Box Optimization", IEEE Transactions on Knowledge and Data Engineering, vol. 30, no. 7, 2018. Available: 10.1109/tkde.2017.2761749, pp. 1240-1253
5. Sanderson, Mark. Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze, "Introduction to Information Retrieval", in Cambridge University Press. 2008, ISBN-13 978-0-521-86571-5, xxi+ 482 pages, Natural Language Engineering16.1, 2010, pp. 100-103.
6. Svore KM, Burges CJ., "A machine learning approach for improved BM25 retrieval", in proceedings of the 18th ACM Conf. Information and knowledge management 2009 Nov 2, pp. 1811-1814.
7. T. Roelleke, "Information Retrieval Models: Foundations and Relationships", Synthesis Lectures on Information Concepts, Retrieval, and Services, vol. 5, no. 3, 2013. Available: 10.2200/s00494ed1v01y201304icr027, pp. 1-163.

8. T. Reddy, B. Vardhan and P. Reddy, "Profile specific Document Weighted approach using a New Term Weighting Measure for Author Profiling", International Journal of Intelligent Engineering and Systems, vol. 9, no. 4, 2016. Available: 10.22266/ijies2016.1231.15, pp. 136-146.
9. Wang, Z., Zhang, Q., & Zhang, D, "A PSO-based web document classification algorithm", Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing , SNPD 2007, IEEE, Vol. 3, 2007, July pp. 659-664.
10. Metzler D, "Using gradient descent to optimize language modeling smoothing parameters", in SIGIR 2007, 30th annual international ACM SIGIR conference on Research and development in information retrieval, Jul 23, Vol. 7, pp. 687-688 .
11. Cui X, Potok TE, Palathingal P., "Document clustering using particle swarm optimization", InProceedings 2005 IEEE Swarm Intelligence Symposium, 2005. SIS 2005. 2005 Jun 8, pp. 185-191.
12. Oosterhuis H, de Rijke M., "Balancing speed and quality in online learning to rank for information retrieval", InProceedings of the 2017 ACM on Conference on Information and Knowledge Management, 2017 Nov 6, ACM, pp. 277-286.
13. Huston S, Croft WB, "A comparison of retrieval models using term dependencies", In Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management 2014 Nov 3, ACM, pp. 111-120.
14. Diaz-Aviles, E., Nejdl, W, Schmidt-Thieme, L, "Swarming to rank for information retrieval", In Proceedings of the 11th Annual conference on Genetic and evolutionary computation ACM, July 08-12, 2009, ACM New York, pp. 9-16.
15. Sadafi, M. H., R. Hosseini, Hamed Safikhani, A. Bagheri, and M. J. Mahmoodabadi. "Multi-objective optimization of solar thermal energy storage using hybrid of particle swarm optimization and multiple crossover and mutation operator", IJE Transactions B: Applications Vol. 24, No. 3, December 2011, pp. 367-377.
16. RC. Eberhart, Y. Shi and J. Kennedy "Swarm Intelligence", The Morgan Kaufmann Series in Evolutionary Computation", Morgan Kaufmann Publishers 1st Ed; 2001.
17. Van der Merwe DW, Engelbrecht AP, "Data clustering using particle swarm optimization", in the 2003 Congress on Evolutionary Computation, 2003, CEC'03, 2003 Dec 8, vol. 1, IEEE, pp. 215-220.
18. MG Omran, A Salman, AP Engelbrecht, "Dynamic clustering using particle swarm optimization with application in image segmentation", Pattern Analysis and Applications, 2006 Feb 1, vol. 8, no. 4, pp. 332.
19. C. Grosan, A., Abraham, M. Chis, "Swarm intelligence in data mining", in Swarm Intelligence in Data Mining Studies in Computational Intelligence, vol 34, Springer, Berlin, Heidelberg, 2006, pp. 1-20.
20. CL Huang, JF Dun, "A distributed PSO–SVM hybrid system with feature selection and parameter optimization", Applied soft computing, 2008 Sep 1, vol. 8, no. 4, pp. 1381-91.
21. W Song, Y Qiao, SC Park, X Qian, "A hybrid evolutionary computation approach with its application for optimizing text document clustering", Expert Systems with Applications. 2015 Apr 1, vol. 42, no. 5, pp. 2517-24.
22. L Zhang, Y Tang, C Hua, X Guan, "A new particle swarm optimization algorithm with adaptive inertia weight based on Bayesian techniques", Applied Soft Computing, 2015 Mar 1, vol. 28,pp. 138-49.
23. A Al-Dayel, M. Ykhlef, "Query paraphrasing enhancement using artificial bee colony" In Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics, 2013 Jun 12, ACM pp. 27.
24. HL Sun, BQ Feng, JB Huang, "Learning to rank with voted multiple hyperplanes for documents retrieval", In2008 3rd International Conference on Intelligent System and Knowledge Engineering, IEEE 2008 Nov 17, Vol. 1, pp. 572-577.
25. D Bollegala, N Noman, H Iba, "Rankde: Learning a ranking function for information retrieval using differential evolution", InProceedings of the 13th annual conference on Genetic and evolutionary computation, ACM, 2011 Jul 12, pp. 1771-1778.
26. S. E. Robertson and D. A. Hull, "The TREC-9 filtering track final report", in TREC, 2000, pp. 25-40.
27. Particle Swarm Optimization in MATLAB - Yarpiz Video Tutorial and "Mastering MATLAB Intelligent Algorithms", Blog.csdn.net, 2019. [Online]. Available: https://blog.csdn.net/zhaohaibo_/article/details/82465789.
28. National Institute of Standards and Technology. 2000. "Text REtrieval Conference: Overview". http://trec.nist.gov/overview.html
29. TREC-9 filtering track collections, http://trec.nist.gov/data/t9_filtering.html.
30. Olsen, D.L.; Denlen, D., "Advanced Data Mining Techniques"; Springer: New York, NY, USA, 2008; p. 138.
31. E. M. Voorhees, "Overview of trec 2002", In Proceedings of the 11th Text Retrieval Conference TREC 2002, NIST Special Publication 500-251, 2002, pp. 1–15.
32. C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall, and F-score, with implication for evaluation", In ECIR '05: Proceedings of the 27th European Conference on Information Retrieval, 2005, pp. 345–359.


## AUTHORS PROFILE

**Narina Thakur,** Email: narinat@gmail.com is presently an Associate Professor of Computer Science Department in Bharati Vidyapeeth's College of Engineering, New Delhi (GGSIPU). She is a B. Tech gold medalist and pursued M.Tech from Punjab Technical University with exceptional career Record and pursuing PHD from Amity University Noida. Apart from 14 years of teaching experience, she has also authored many books in the areas of Algorithms and programming and Operating system. She is an active member of Research and Development team in Bharati Vidyapeeth and working closely in the areas of Algorithms, Information Retrieval, Data mining, and Embedded Systems. She has also authored many papers, published in IEEE International Conferences, ACM SIGARCH, Elsevier and International Journals and completed successfully one DRDO DESI-DOC lab consultancy cum research project and currently doing one DST research project.

**Dr. Deepti Mehrotra** Email: dmehrotra@amity.edu gold medalist of Lucknow University. She received M.Sc.(Physics), MCA and Ph.D. from Lucknow University. Currently, she is a Professor in Department of IT Amity School of Engineering & Technology, Amity University Noida. She has more than 17 years of experience in teaching, research and content writing. Over 50 papers in International refereed Journals and conference Proceedings. Dr. Mehrotra is a member of many committees like a member of the BOS of Mahamaya Technical University and Amity University Rajasthan, RDC of Amity University, Uttar Pradesh and reviewer for many referred journals and conferences. She is regularly invited as recourse person for FDPs and invited talks at national and international conference. She is currently guiding 7 Ph.D. Scholar and guided many M.Tech Scholars. Dr. Mehrotra is being Author of more than 10 books.

**Dr. Abhay Bansal** Email: abhaybansal@hotmail.com is BE (CS), ME (IT), MBA and PhD. He is Joint head Amity School of Engineering & Technology, Amity University Noida, and Professor & HOD (CSE). With over 17 years of Industry and Academic Experience. Dr. Bansal is guiding 8 students for Ph.D. and also a member of the DRC of Amity University, Noida. Dr. Bansal is also a member of the examination committee of several of the university. Dr. Bansal is Fellow, The Institution of Engineering and Technology (U.K), Sr. Member, International Association of Computer Science and Information Technology, Member, International Association of Engineers, External Member of RDC of Northern India Textile & Research Association (Ministry of Textile, Govt. Of India) and Sr. Life Member of Computer Society of India.

**Dr. Manju Bala** Email: manjugpm@gmail.com is B. E (CSE) from Maharishi Dayanand University Rohtak, Haryana, M. Tech (CSE) and obtained her Ph.D. in Computer Science from School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi. She is currently Assistant professor in department of Computer Science, I. P.

College for Women, University of Delhi, New Delhi, India. Her main areas of research interests are pattern recognition, and Data mining. Dr. Manju Bala authored two International books, published thirteen papers in International conferences and eight in National conference papers and three referred International journal papers. She also authored one book chapter.