Comparing Performance of PSO and Bat Algorithms for Effective Information Retrieval

Tahera Shaikh[#1], **Shiwani Gupta**[#2]

[#1,2]*Department of Computer Engineering, Thakur College of Engineering and Technology.*
[1]shaikh.tahera.i@gmail.com
[2]shiwani.gupta@thakureducation.org

***Abstract*— Information Retrieval (IR) is discovering records of an unstructured source like content that satisfies information need within large collections generally put away on computers. In view of the semantic disengage amongst query and documents, IR is accountable to return a lot of unwanted items. There is a fast development of the measure of information accessible in electronic libraries, through Internet and undertaking system mediums, propelled techniques for pursuit and data recovery are sought after. Information retrieval systems, intended for storing, seeking vast scale sets of unstructured documents, are the subject of concentrated examination.**

The researchers have introduced various optimization algorithms in the field of Information Retrieval like Swarm Intelligence and Genetic Algorithm. These algorithms are inspired by nature to solve the IR problems. They have used these algorithms on different sets of databases. There is clear interest for fine tuning the performance of IR on Web information. In this paper, Particle Swarm Optimization algorithm (PSO) and Bat algorithm are compared on web database. The PSO and Bat algorithms are used to solve query expansion problem. When the user enters a query keyword, it may not contain sufficient information to retrieve desired information. The query reformulation technique is used to solve this problem, i.e. the relevant keywords are added to the original query and new expanded query is formed. PSO and Bat algorithms are applied to select amongst result based on these expanded query.

We first conduct a preliminary experiment to tune single optimization algorithm parameters. Then, we have compared the results to recent optimization algorithm onto the same dataset. PSO gives better retrieval effectiveness than Bat algorithm and on the other hand Bat has better execution time. The algorithms are compared on accuracy and effectiveness parameters like relevance and execution time.

***Keywords*— Information Retrieval, particle swarm intelligence algorithm, Bat Algorithm, unstructured data, query expansion**

I. Introduction

A. Background

1) Information Retrieval:

During the last decade the information over the web have increased and optimization of information retrieval effectiveness has driven the quality of the results over the web, People are more trusting and preferring web search as a source of information. Information retrieval has come out of academic discipline to become the basis of most preferred and reliable source of information. The field of information retrieval began with scientific library records and scientific publications; it spread rapidly in other domains like journalism, lawyers and medical fields. Information retrieval then spread in web information access. The information retrieval provides solution in finding relevant information in unstructured information [5].

Retrieval effectiveness could be captured through Relevance feedback that involves an iterative process in which users first specify which documents are relevant to them. These specified documents are used by the system to retrieve more or similar documents. The process is then repeated. Relevance feedback may be Explicit where users may be constrained and reluctant to provide feedback or Implicit where user's click-through record provides feedback but it is still under exploration that what interactions should be taken into account and how good they are as relevance feedback or it may be Pseudo which assumes top ranked documents as relevant but doesn't guarantee it. When the user's interactions are removed from this iterative process, this kind of relevance feedback is called PRF or blind relevance feedback.

2) Optimization (Swarm Intelligence)

The researchers have used number of optimization technique in information retrieval domain. There are various models of IR and methods for optimization. Here we are concentrating on one of stochastic optimization technique called swarm intelligence. Swarm intelligence is the study of computational systems inspired by the 'collective intelligence'. Collective Intelligence emerges through the working together of large numbers of similar agents in the environment. Schools of fish, flocks of birds, and colonies of ants are some the examples. The property of swarm intelligence is self-organization, decentralization and distribution throughout the environment. The problems are solved in nature like foraging for food, prey evading, and colony relocation through SI. The information is stored and transferred by the means of agents such as proximity in fish and birds, pheromones in ants and dancing in bees.

B. Information Retrieval / Data Retrieval

TABLE I INFORMATION RETRIEVAL / DATA RETRIEVAL

	Information Retrieval	**Data Retrieval**
Matching	vague	exact
Model	probabilistic	deterministic
Query language	natural	artificial
Query specification	incomplete	complete
Items wanted	relevant	all (matching)
Error handling	insensitive	sensitive

There is clear difference between Information retrieval and Data retrieval. As shown in Table I. traditional deterministic data retrieval process is where we require exact results therefore the query should be complete, this process returns all the results specific to the query.

On the other hand, IR uses probabilistic approach. It returns vague results and uses natural language. Though it is vague it gives more relevant results.

C. Conceptual Model for IR

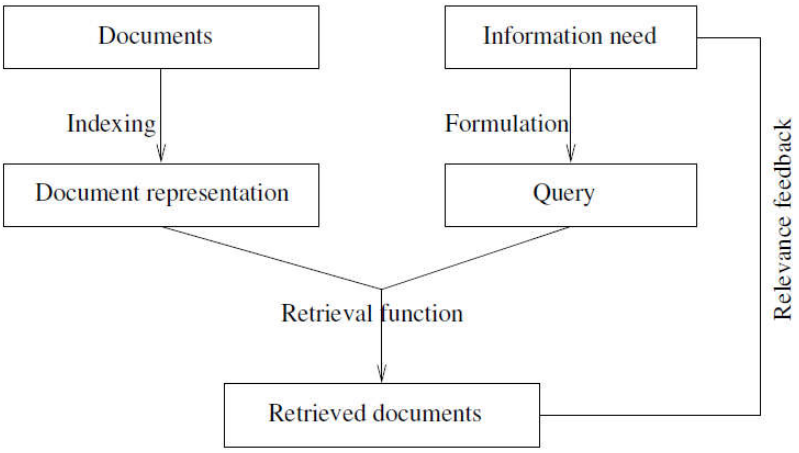


Figure 1. Conceptual Model for IR

Fig 1. Shows the traditional IR model, where documents are indexed according to relevance. Information need give rise to a query. This query is then applied to the documents to retrieve relevant results. The feature of IR is that it can give relevance feedback to the model. The user’s behavioural pattern is relevance feedback, i.e. most selected document.

D. Problem Definition

The model consist of multiple modules. The first module comprise of building an appropriate dataset which can be used for comparison of optimization algorithms. The selection of dataset play a significant role to assess different algorithms. The second module apply the optimization algorithms on the dataset. All the algorithms applied on same dataset to have input consistency. The algorithms will be then evaluated on different criteria such as performance and accuracy, efficiency, time complexity.

II. Literature Survey

A. An accelerated PSO for query expansion in web information retrieval: application to medical dataset

In this paper [1], the author has used swarm intelligence algorithm called Accelerated Particle Swarm Optimization (APSO) to solve problem of Query Expansion (QE). When the query is expanded, the numbers of expanded candidate keywords are generated. Here, to select the best and most relevant keyword is very complex. The author has used APSO to solve this problem. They have used three steps to achieve it which are:

i) Retrieve the pseudo-relevant documents to the initial query
ii) Extract the best expansion keywords from the pseudo relevant documents to constitute the expanded query
iii) Retrieve the relevant documents to the expanded query [1].

B. An ant-colony based approach for real-time implicit collaborative information seeking

In this paper [3], the authors have used multiple techniques to employ Ant colony. They have used Naïve Rank, Random Rank, and Session Rank to implement different functionality of Ant Colony. Ranking results of diverse types of queries depends on users' intent. Queries have been clustered in three distinct categories on the basis of user's intent namely; Informational queries, Navigational queries, Transactional queries and they have used different algorithms to address these queries. The famous family of Ant Colony Optimization (ACO) algorithms is inspired by the Ants swarm intelligence and their use of stigmergic processes [3].

C. Artificial Bee Colony approach for ranking web pages

In this Paper [5], here the model shows collective intelligence of honeybee. It consists of three essential components: food sources, employed foragers, and unemployed foragers. They have used Page Rank algorithm to implement this technique. They have implemented it in three steps as follows:

i) Calculation of user interest
ii) Growth analysis Rate
iii) Total site linking [5]

D. Bat algorithm for efficient query expansion: Application to MEDLINE

In this paper [6], they have used swarm intelligence to solve the problem of query expansion, First, they have calculated rank of the pseudo relevant documents using document scoring function like OKAPI 25, next they have extracted keyword candidate using term scoring function e.g. RSJ, Rocchio, next they have put the best ranked keyword in original query, and finally the document is retrieved by using document scoring function [6].

E. Document clustering with evolved search queries

In this paper [2], they have used genetic algorithm which creates a set of Apache Lucene search queries for text document clustering. The document which is added in only one cluster will add strength of the cluster, whereas the document which is added in more than one cluster will decrease the suitability. Here extra labelling step is not necessary because the final search queries are effortlessly understood and present cluster in uncomplicated manner [2].

F. Distribution separation method using irrelevance feedback data for Information Retrieval

In the field of Information Retrieval, researchers have worked upon few challenges as Relevance modeling and Irrelevance modeling; when relevant documents are less. In [7], authors have come up with approximate true relevance model, developed its framework, theory and Distribution Separation method to separate mixed probability distribution of relevant and irrelevant documents through relevance feedback which is a post query process by building refined query model; to improve the retrieval effectiveness and/or stability of query model estimation in the context of relevance feedback.

III.Design and Methodology

A. Proposed System

The project has been implemented to compare different optimization algorithms i.e. Particle swarm optimization and Bat algorithm in effective information retrieval. The project evaluates different parameters while comparing different algorithms and concluded as why particular algorithm is efficient than others.

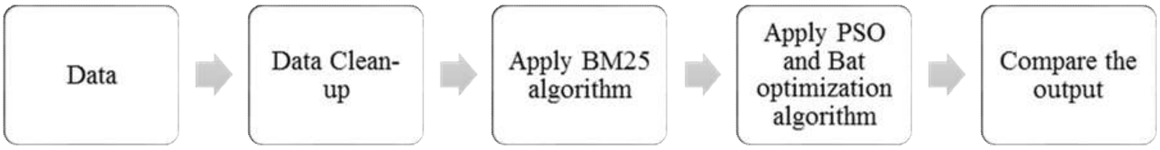


Figure 2. Proposed System

In the proposed framework, is partitioned into different stages:

1. **Phase 1:** The tweets are downloaded from twitter Programming interface which is in .json format. The downloaded tweets should be changed over into .csv arrange by pre-handling it for additionally preparing of the dataset.
2. **Phase 2:** The Tweets are cleaned by removing special characters and hashtags.
3. **Phase 3:** Apply BM25 algorithm using users query and expanded query and store the results.
4. **Phase 4:** Apply PSO and Bat algorithm on the BM25 results, which are nothing but results based on number of keywords present in tweets as well whole database.
5. **Phase 5:** The last stage is to compare the results returned by PSO and Bat.

B. Flow of the Proposed Model

As mentioned in the above figure the two algorithms have been implemented. Below figure shows the flow of the overall building of the model.

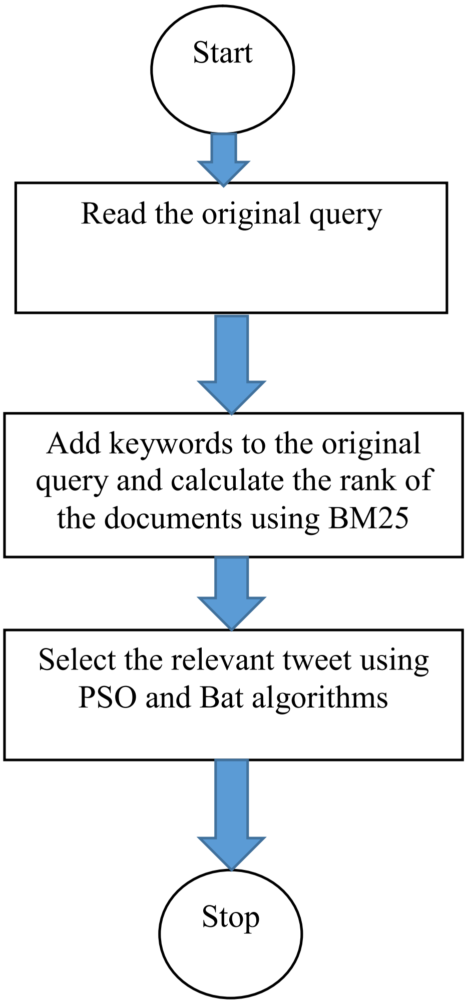

Start
Read the original query
Add keywords to the original query and calculate the rank of the documents using BM25
Select the relevant tweet using PSO and Bat algorithms
Stop

Figure 3. Flow of the Proposed Model

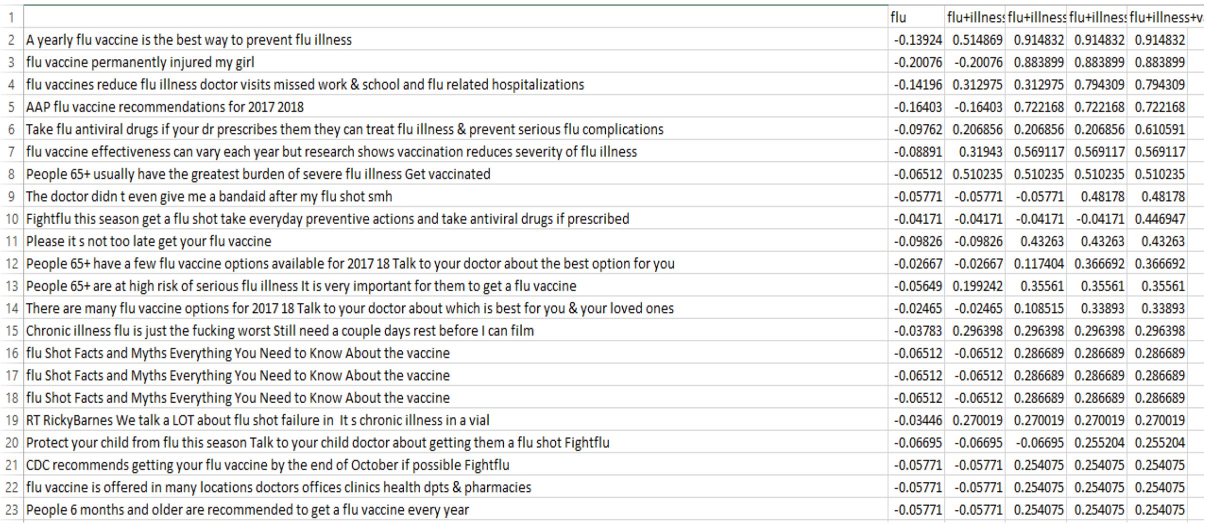
The above figure depicts the flow of the system.

The process starts with extracting tweets in the form of .csv file format. Further, the tweets are cleaned by removing special characters and hashtags. This file is processed by BM25 algorithm. Extra keywords are added to the original query and after adding each extra keyword, a new expanded query is formed and at each iteration the BM25 score is calculated.

After getting all BM25 scores for each expanded query the best document is selected by two proposed PSO and Bat algorithms. Here, we have performed both algorithms on the same dataset to evaluate them based on their effectiveness, relevance and the time.

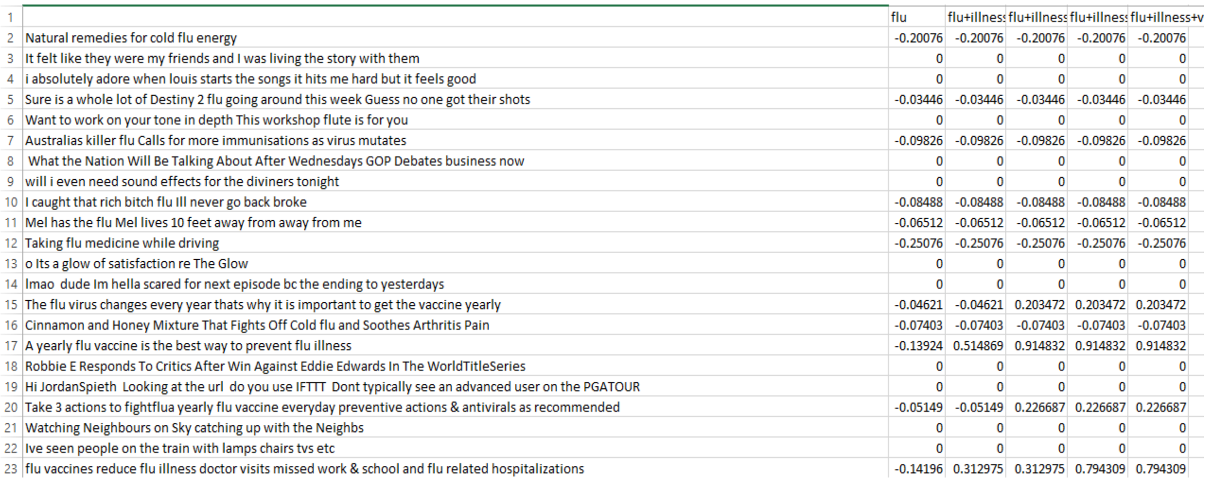
IV. Results and Discussion

The following are the tweets snippets with expanded queries and the BM25 score. It is observed that BM25 scores calculates the relevance based on keywords. It is a probabilistic measure, and it give values between -1 to 1. If the keyword is not present in the tweet than the value is 0.



1		flu	flu+illnes	flu+illnes	flu+illnes	flu+illness+v
2	A yearly flu vaccine is the best way to prevent flu illness	-0.13924	0.514869	0.914832	0.914832	0.914832
3	flu vaccine permanently injured my girl	-0.20076	-0.20076	0.883899	0.883899	0.883899
4	flu vaccines reduce flu illness doctor visits missed work & school and flu related hospitalizations	-0.14196	0.312975	0.312975	0.794309	0.794309
5	AAP flu vaccine recommendations for 2017 2018	-0.16403	-0.16403	0.722168	0.722168	0.722168
6	Take flu antiviral drugs if your dr prescribes them they can treat flu illness & prevent serious flu complications	-0.09762	0.206856	0.206856	0.206856	0.610591
7	flu vaccine effectiveness can vary each year but research shows vaccination reduces severity of flu illness	-0.08891	0.31943	0.569117	0.569117	0.569117
8	People 65+ usually have the greatest burden of severe flu illness Get vaccinated	-0.06512	0.510235	0.510235	0.510235	0.510235
9	The doctor didn t even give me a bandaid after my flu shot smh	-0.05771	-0.05771	-0.05771	0.48178	0.48178
10	Fightflu this season get a flu shot take everyday preventive actions and take antiviral drugs if prescribed	-0.04171	-0.04171	-0.04171	-0.04171	0.446947
11	Please it s not too late get your flu vaccine	-0.09826	-0.09826	0.43263	0.43263	0.43263
12	People 65+ have a few flu vaccine options available for 2017 18 Talk to your doctor about the best option for you	-0.02667	-0.02667	0.117404	0.366692	0.366692
13	People 65+ are at high risk of serious flu illness It is very important for them to get a flu vaccine	-0.05649	0.199242	0.35561	0.35561	0.35561
14	There are many flu vaccine options for 2017 18 Talk to your doctor about which is best for you & your loved ones	-0.02465	-0.02465	0.108515	0.33893	0.33893
15	Chronic illness flu is just the fucking worst Still need a couple days rest before I can film	-0.03783	0.296398	0.296398	0.296398	0.296398
16	flu Shot Facts and Myths Everything You Need to Know About the vaccine	-0.06512	-0.06512	0.286689	0.286689	0.286689
17	flu Shot Facts and Myths Everything You Need to Know About the vaccine	-0.06512	-0.06512	0.286689	0.286689	0.286689
18	flu Shot Facts and Myths Everything You Need to Know About the vaccine	-0.06512	-0.06512	0.286689	0.286689	0.286689
19	RT RickyBarnes We talk a LOT about flu shot failure in It s chronic illness in a vial	-0.03446	0.270019	0.270019	0.270019	0.270019
20	Protect your child from flu this season Talk to your child doctor about getting them a flu shot Fightflu	-0.06695	-0.06695	-0.06695	0.255204	0.255204
21	CDC recommends getting your flu vaccine by the end of October if possible Fightflu	-0.05771	-0.05771	0.254075	0.254075	0.254075
22	flu vaccine is offered in many locations doctors offices clinics health dpts & pharmacies	-0.05771	-0.05771	0.254075	0.254075	0.254075
23	People 6 months and older are recommended to get a flu vaccine every year	-0.05771	-0.05771	0.254075	0.254075	0.254075

Figure 4. Dataset and output after applying BM25 (a)



1		flu	flu+illnes	flu+illnes	flu+illnes	flu+illness+v
2	Natural remedies for cold flu energy	-0.20076	-0.20076	-0.20076	-0.20076	-0.20076
3	It felt like they were my friends and I was living the story with them	0	0	0	0	0
4	i absolutely adore when louis starts the songs it hits me hard but it feels good	0	0	0	0	0
5	Sure is a whole lot of Destiny 2 flu going around this week Guess no one got their shots	-0.03446	-0.03446	-0.03446	-0.03446	-0.03446
6	Want to work on your tone in depth This workshop flute is for you	0	0	0	0	0
7	Australias killer flu Calls for more immunisations as virus mutates	-0.09826	-0.09826	-0.09826	-0.09826	-0.09826
8	What the Nation Will Be Talking About After Wednesdays GOP Debates business now	0	0	0	0	0
9	will i even need sound effects for the diviners tonight	0	0	0	0	0
10	I caught that rich bitch flu Ill never go back broke	-0.08488	-0.08488	-0.08488	-0.08488	-0.08488
11	Mel has the flu Mel lives 10 feet away from away from me	-0.06512	-0.06512	-0.06512	-0.06512	-0.06512
12	Taking flu medicine while driving	-0.25076	-0.25076	-0.25076	-0.25076	-0.25076
13	o Its a glow of satisfaction re The Glow	0	0	0	0	0
14	lmao dude Im hella scared for next episode bc the ending to yesterdays	0	0	0	0	0
15	The flu virus changes every year thats why it is important to get the vaccine yearly	-0.04621	-0.04621	0.203472	0.203472	0.203472
16	Cinnamon and Honey Mixture That Fights Off Cold flu and Soothes Arthritis Pain	-0.07403	-0.07403	-0.07403	-0.07403	-0.07403
17	A yearly flu vaccine is the best way to prevent flu illness	-0.13924	0.514869	0.914832	0.914832	0.914832
18	Robbie E Responds To Critics After Win Against Eddie Edwards In The WorldTitleSeries	0	0	0	0	0
19	Hi JordanSpieth Looking at the url do you use IFTTT Dont typically see an advanced user on the PGATOUR	0	0	0	0	0
20	Take 3 actions to fightflua yearly flu vaccine everyday preventive actions & antivirals as recommended	-0.05149	-0.05149	0.226687	0.226687	0.226687
21	Watching Neighbours on Sky catching up with the Neighbs	0	0	0	0	0
22	Ive seen people on the train with lamps chairs tvs etc	0	0	0	0	0
23	flu vaccines reduce flu illness doctor visits missed work & school and flu related hospitalizations	-0.14196	0.312975	0.312975	0.794309	0.794309

Figure 5. Dataset and output after applying BM25 (b)

The Figure 6 graph shows the number of positive values verses the number of keywords for BM25 results. It is observed that as number of keywords increases the BM25 score move towards positive values, also when a keyword is repeated in the documents many times the value comes negative.

It concludes that as the relevant keywords are added into the expanded query the BM25 score move towards the positive value. The more positive value BM25 score, the document is more relevant. Initially when the user fires a query this score would be negative as the same keyword may occur many times in a document. As we expand the query more relevant words gets added to the query this score becomes positive, meaning it is more relevant documents amongst many.

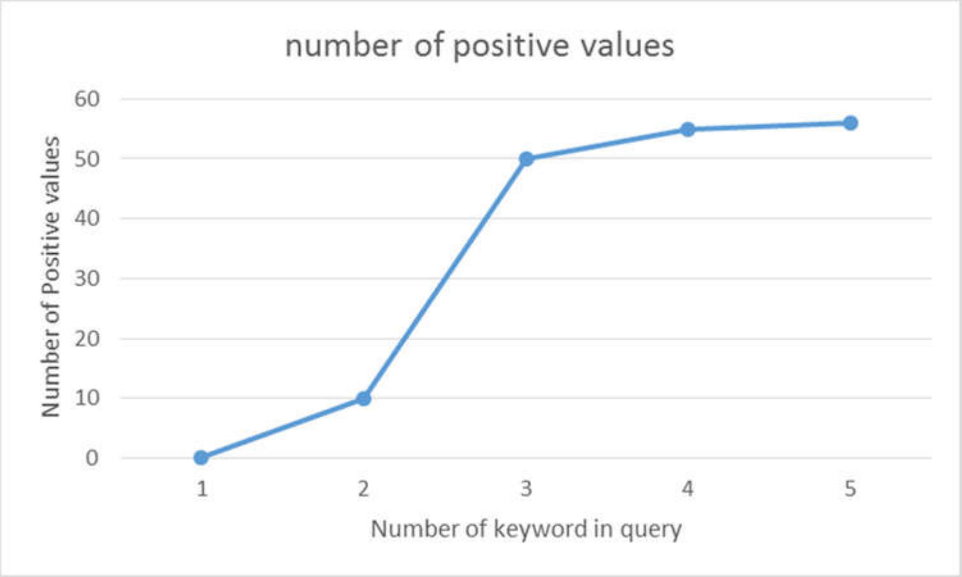
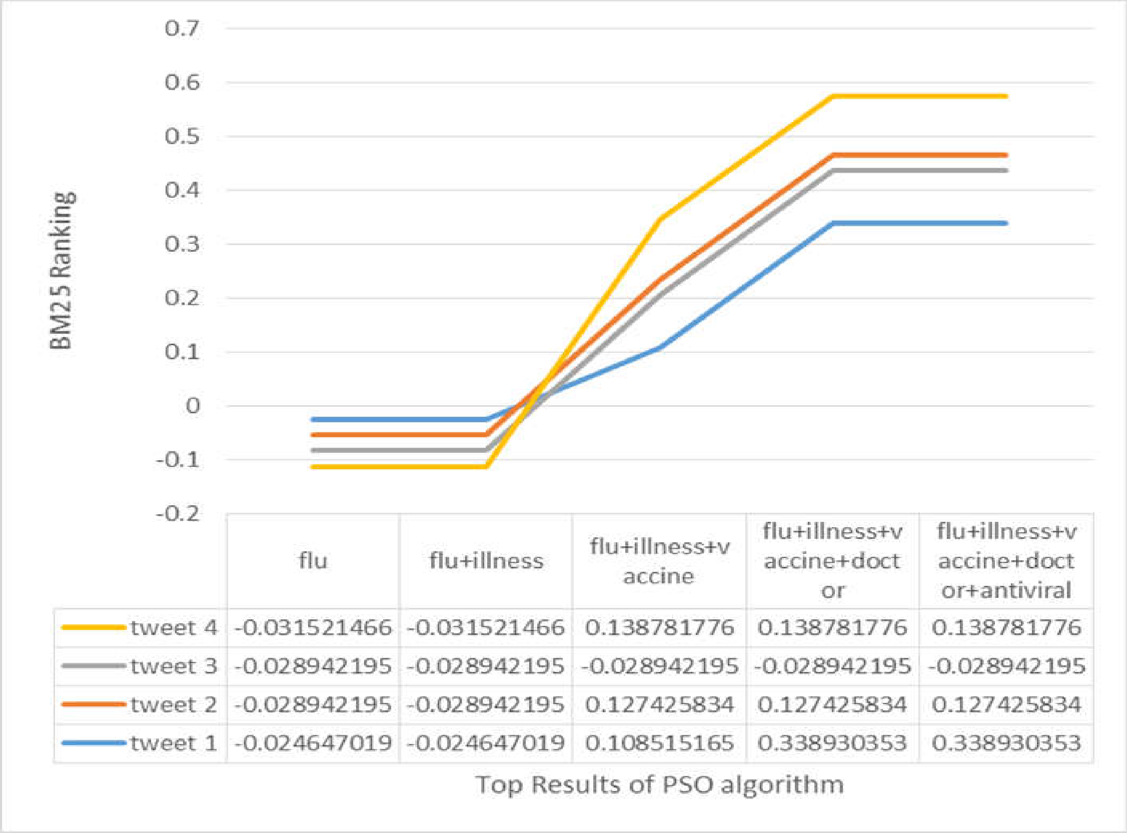
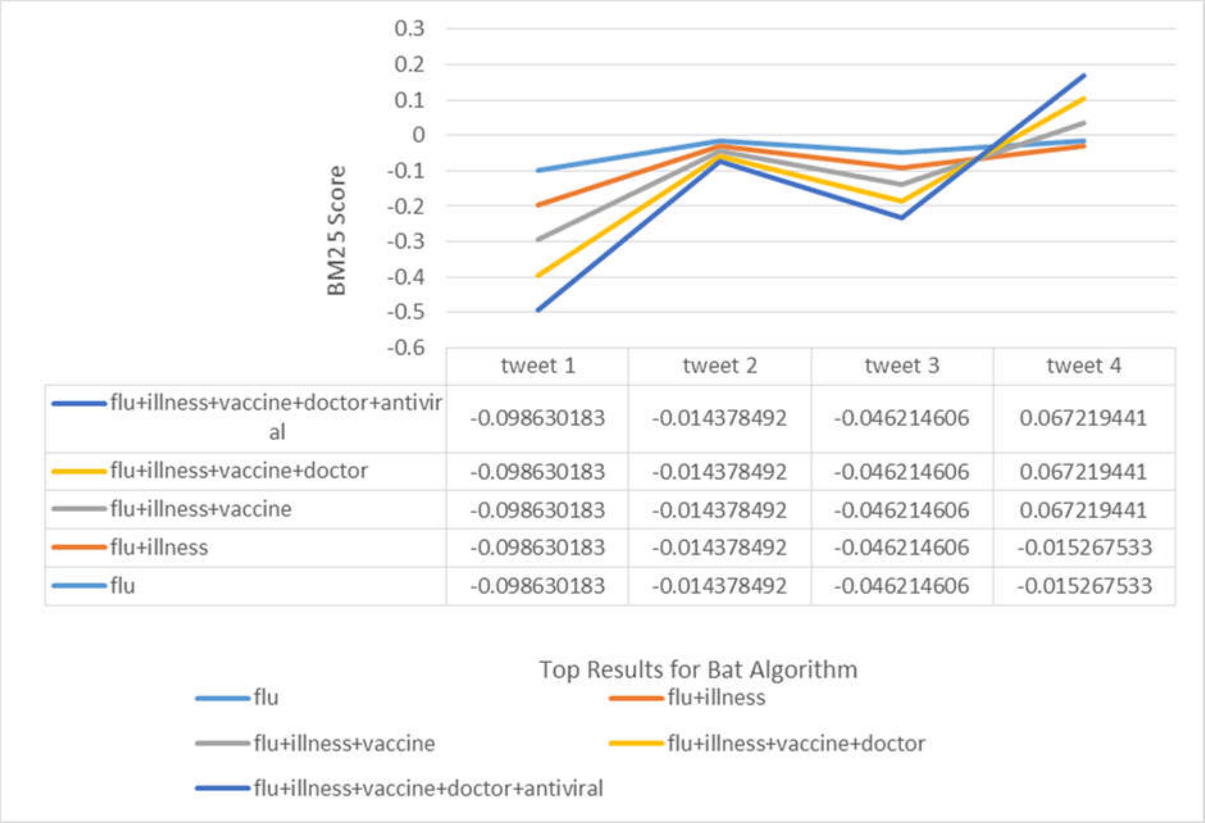


Figure 6. Graph for number of positive values verses the number of keywords for BM25 results



	flu	flu+illness	flu+illness+v accine	flu+illness+v accine+doct or	flu+illness+v accine+doct or+antiviral
tweet 4	-0.031521466	-0.031521466	0.138781776	0.138781776	0.138781776
tweet 3	-0.028942195	-0.028942195	-0.028942195	-0.028942195	-0.028942195
tweet 2	-0.028942195	-0.028942195	0.127425834	0.127425834	0.127425834
tweet 1	-0.024647019	-0.024647019	0.108515165	0.338930353	0.338930353

Figure 8. PSO Top Results



	tweet 1	tweet 2	tweet 3	tweet 4
flu+illness+vaccine+doctor+antiviral	-0.098630183	-0.014378492	-0.046214606	0.067219441
flu+illness+vaccine+doctor	-0.098630183	-0.014378492	-0.046214606	0.067219441
flu+illness+vaccine	-0.098630183	-0.014378492	-0.046214606	0.067219441
flu+illness	-0.098630183	-0.014378492	-0.046214606	-0.015267533
flu	-0.098630183	-0.014378492	-0.046214606	-0.015267533

Figure 9. Bat Top Results

The Figure 8 and 9 shows the Top results from Particle swarm optimization algorithm and bat algorithm. The PSO does row wise analysis and Bat does column wise analysis. It is observed that PSO gives more relevant results as compared to Bat.

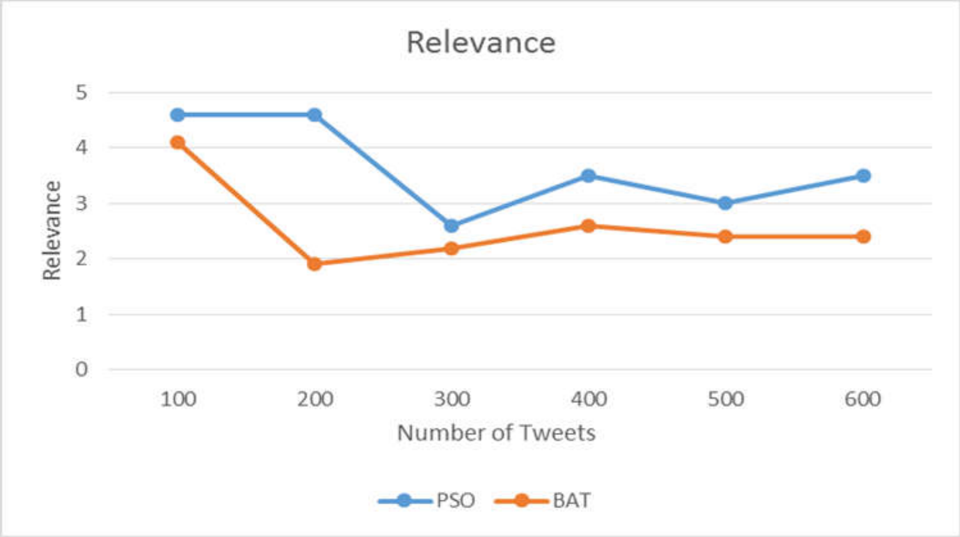


Figure 10. PSO and Bat algorithm with respect to relevance and number of tweets

The top results from PSO and Bat algorithms are captured with respect to 100, 200, …, 600 tweets. The relevance of the tweets with the keywords is measured based on the score 1 to 5. 1 is assigned for least and 5 for most relevant of the results.

As shown in the figure 10, it is observed that PSO performs better as compared to Bat algorithm. PSO performs better because it returns more number of positive values.

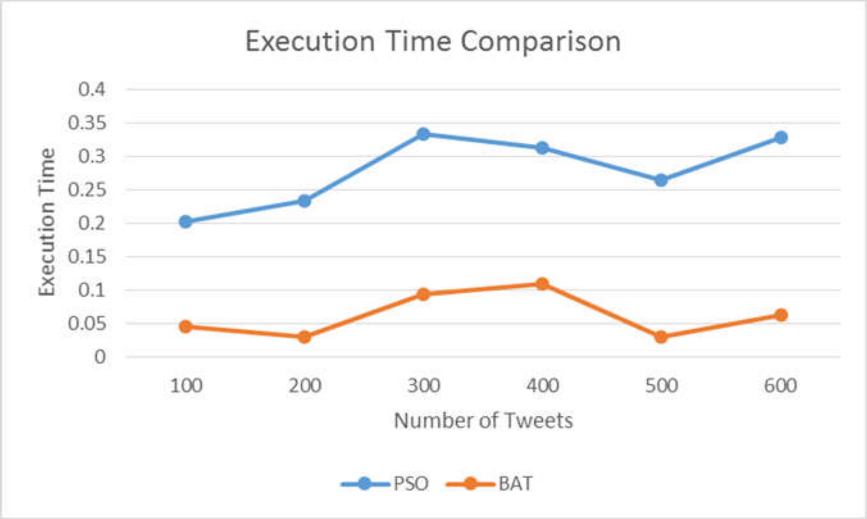


Figure 11. Comparing PSO and Bat algorithm with respect to Execution Time

The figure 11 shows the execution time for both algorithms with respect to number of tweets. Here, Bat performs better as compared to PSO.

V. Conclusions

There are various techniques to solve the problem of information retrieval and optimization algorithms like swarm intelligence algorithms and Bat Algorithm.

This work comprises of various researches on Particle Swarm Optimization algorithm and Bat algorithm, along with their comparison with respect to relevance and execution time. The work shows that PSO performs better with large dataset as it captures positive rank values, on the other hand Bat has better execution time.

Optimization approaches are useful for us to take care of numerous issues. Optimization algorithm has various types of models and structures, and it has had numerous applications in numerous angles. For these, we can see that PSO has an incredible advancement potential.

In future, it is predictable that Optimization algorithms could solve many problems. In the meantime, its capacities of unsupervised learning will be upgraded since there is many information on the web updated every second and information retrieval needs to be optimized. The optimization algorithms can be further enhanced and can be used to solve other real life problems.

References

[1] Ilyes Khennak, Habiba Drias, "An accelerated PSO for query expansion in web information retrieval: application to medical dataset", Springer Science + Business Media, pp 1-16, International journal of Applied Intelligence, 2017.

[2] L.Hirsch, A. D. Nuovo, "Document clustering with evolved search queries" IEEE congress on Evolutionary Computation 2017, Spain.

[3] Malizia, K. A. Olsen, T. Turchi, P. Crescenzi, "An ant-colony based approach for real-time implicit collaborative information seeking", Information Processing and Management. Vol 53(3) May 2017. 608-623, ScienceDirect, Elsevier. DOI: 10.1016/

[4] S. Lohar, P. Howale, S. Pradhan, M. Redekar, "Implementation of improved IR System through Swarm Intelligence Technique", IJRITCC, Oct 2015.

[5] G. Anuradha, G. Lavanya, "Artificial Bee Colony approach for ranking web pages", IJCA, Aug 2014.

[6] Khennak I, Drias H, "Bat algorithm for efficient query expansion: Application to MEDLINE", Proceedings of the 4th World Conference on Information Systems and Technologies, pp 113–122. Springer, 2016.

[7] D. Carmel, L. L. Eytan, A. Libov, Y. Maarek, A. Raviv," The Demographics of Mail Search and their Application to Query Suggestion", ACM IW3C2, Apr 2017, Perth, Australia.

[8] C. L. Smitha, J. Gwizdka, H. Field, "The use of query auto-completion over the course of search sessions with multifaceted information needs" Science Direct, Elsevier, May 2017.

[9] Query Expansion Techniques for Information Retrieval: a Survey. Hiteshwar Kumar Azad, Akshay Deepak. s.l. : arXiv.org, 2017.

[10] Spatio-Temporal Pseudo Relevance Feedback for Scientific Data Retrieval. Shin'ichi Takeuch, Yuhei Akahoshi, Komei Sugiura, Koji Zettsu. s.l. : IEEJ Transactions, Vol. 131

[11] Balaneshinkordan S., Kotov A. (2016) An Empirical Comparison of Term Association and Knowledge Graphs for Query Expansion. In: Ferro N. et al. (eds) Advances in Information Retrieval. ECIR 2016. Lecture Notes in Computer Science, vol 9626. Springer, Cham

[12] Lv C., Qiang R., Fan F., Yang J. (2015) Knowledge-Based Query Expansion in Real-Time Microblog Search. In: Zuccon G., Geva S., Joho H., Scholer F., Sun A., Zhang P. (eds) Information Retrieval Technology. Lecture Notes in Computer Science, vol 9460. Springer, Cham