International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 14 (2022) pp. 028-047

Submitted: 28 October, 2021; Accepted:2 February 2022; Publish: 9 Apri, 2022

Enhanced Elephant Herding Optimization for Large Scale Information Access on Social Media

Yassine Drias[1], Habiba Drias[2] and Ilyes Khennak[3]

[1]University of Algiers,
02 rue Didouche Mourad, Algiers 16000, Algeria
y.drias@univ-alger.dz

[2]LRIA, USTHB,
BP 32 El Alia, Bab Ezzouar, Algiers 16111, Algeria
habiba.drias@usthb.edu.dz

[3]LRIA, USTHB,
BP 32 El Alia, Bab Ezzouar, Algiers 16111, Algeria
Ilyes.khennak@usthb.edu.dz

Abstract: **In this article, we present a novel information access approach inspired by the information foraging theory (IFT) and elephant herding optimization (EHO). First, we propose a model for information access on social media based on the IFT. We then elaborate an adaptation of the original EHO algorithm to apply it to the information access problem. The combination of the IFT and EHO constitutes a good opportunity to find relevant information on social media. However, when dealing with voluminous data, the performance undergoes a sharp drop. To overcome this issue, we developed an enhanced version of EHO for large scale information access. We introduce new operators to the algorithm, including territories delimitation and clan migration using clustering. To validate our work, we created a dataset of more than 1.4 million tweets, on which we carried out extensive experiments. The outcomes reveal the ability of our approach to find relevant information in an effective and efficient way. They also highlight the advantages of the improved version of EHO over the original algorithm regarding different aspects. Furthermore, we undertook a comparative study with two other metaheuristic-based information foraging approaches, namely ant colony system and particle swarm optimization. Overall, the results are very promising.**

Keywords: Information Access, Information Foraging Theory, Swarm Intelligence, Elephant Herding Optimization, Clustering, K-means, Social Media

I. Introduction

Nowadays social media are increasingly being used as an information source and people are becoming more dependent on them in their daily life. They use them to access and share information, which highly contributes to the growth of the volume of online public data. According to the *Digital 2021 Report*, the number of social media users has increased by an average of more than 1.4 million users each day during 2020, which amounts to more than half a billion new users in 12 months [1]. This rapid growth has propelled the total number of active social media users to 4.33 billion by April 2021, which equates to 55% of the world's total population. In fact, social media are being used to seek information about serious topics, such as circulating up-to-the minute information about the Covid-19 pandemic [2]. More generally, these platforms are frequently used by people seeking health information. In the U.S. for instance, 80% of Internet users search for health information online, and 74% of them use social media [3]. In view of this impressive growth in terms of popularity and data volume of social media, the development of new large-scale information access techniques adapted to such platforms is required.

Generally, a person engaged in an information seeking process has one or more goals in mind and uses information access tools to achieve them. Those goals can range quite widely, from finding a specific product to keeping informed about a certain topic. Information foraging is a paradigm related to accessing information online. Usually, when people need an information, they have the opportunity to use the Web to query it. Of course information retrieval helps to get a part of the information, thanks to the existing search engines. However, information foraging is more than just querying a search tool and getting a fragment of information. It consists in exploring the Web while using certain bio-inspired navigation mechanisms as well as Web structure related features. The task of foraging is grounded on the optimal foraging theory (OFT) [4], which paved the way to the information foraging theory (IFT) [5]. The authors of the latter studied the optimal foraging theory to understand how human users search for information. The IFT is based on the assumption that, when searching for information, humans use built-in forag-

ing mechanisms that evolved to help our animal ancestors find food.
Recently, a significant amount of work has been done in the information access field using the *information foraging theory*. Technologies and approaches such as deep learning [6], game theory [7], bio-inspired computing [8], ontologies [9] and multi-agent systems [10] were used for this purpose. On top of that, the IFT was exploited to solve many problems like cyber-attack prediction [11], query auto-completion [12] and recommender systems [13]. Some newer studies also focused on applying the IFT on social media [14].
The aim of this article is to propose a novel bio-inspired approach to large scale information access on social media. Our approach is based on a combination of the information foraging theory and a new enhanced elephant herding optimization that we developed for large scale information access. The main new contributions of the present work can be summarized as follows:

- a detailed formal model for information foraging on social media;
- a new enhanced version of elephant herding optimization with new operators adapted to large scale information access;
- the use of k-means clustering to implement new operators in the enhanced elephant herding optimization;
- a performance evaluation on a dataset of more than 1.4 million tweets;
- a comparative study with other mataheuristic-based information access approaches.

The rest of this paper is organized as follows. In Section II, we discuss related literature that covers some recent works on information foraging and elephant herding optimization. In section III, we present our information foraging model that focuses on social media, while explaining the analogy between animal food foraging and information foraging. Section IV is dedicated to present our adaptation of the elephant herding optimization algorithm to information foraging on social media. In Section V, we explain how we incorporate the territories concept into EHO using clustering. The Enhanced EHO for large scale information foraging is presented in Section VI and the experimental results are detailed in Section VII. Finally, we conclude in Section VIII and discuss some future directions.

II. Related works

This section provides a literature review in two parts.The first one reports the most recent studies on information foraging, while the second summarizes some important efforts on elephant herding optimization.

A. Information foraging

Recently, a number of studies applied the information foraging theory to address issues related to some information access approaches. The authors in [13] explore how changes to the user interface can impact the learning accuracy of recommender systems. They use the information foraging theory to study how feedback quality and quantity are influenced by interface design choices along two axes: information scent and information access cost. To undertake a user study, the authors considered the task of picking a movie to watch. The results obtained from the use of the information foraging theory concepts such as the information scent show that the user interface factors can effectively shape and improve the implicit feedback data that is generated while maintaining the user experience.
In [15], the authors measure the utility and cost of Web search engine result pages using a new measure based on the information foraging theory. According to the authors, the latter provides a number of new dimensions in which to investigate and evaluate user behavior and performance. The analysis of over 1000 popular queries issued to a major search engine show that the proposed foraging based measure provides a more accurate reflection of the utility and of observed behaviors.
The IFT was also used to develop standalone information access systems. In [7] the authors implemented a multi-agent system composed of several self-interested agents with different behaviors. The task of finding relevant information based on an information need introduced by the user was assigned to each agent. The developed system was tested on the Citation Network, which contains scientific publications along with their respective citations. The authors conducted a preprocessing step consisting in classifying the publications using the 2012 ACM ontology. The outcomes of this study demonstrate that introducing such preprocessing step in information foraging can highly contribute in making the system scalable.
Bio-inspired metaheuristics were also exploited in this context. In [16], the authors propose a framework for medical Web information foraging using hybrid ant colony optimization and tabu search. The experimental results on *MedlinePlus* website show that the system is able to locate relevant Web pages related to specific pathologies and diseases thanks to the collaboration and self-organization aspects that characterize ant colony optimization and bio-inspired metaheuristics in general.

B. Elephant herding optimization

The remarkable growth of the size and complexity of optimization problems made the traditional exact algorithms ineffective for solving this kind of problems [17]. Metaheuristic algorithms have proved to be a viable solution to this challenge. These robust algorithms, which are in most cases bio-inspired, are mainly applied to solve NP-hard problems [18, 19]. Elephant Herding Optimization (EHO) is a bio-inspired metaheuristic that takes its origins from the herding behavior of elephants in nature. It was first introduced in [20] to solve hard continuous optimization problems and has since been used to address numerous problems such as numerical optimization problems [21], task scheduling [22], data clustering [23] and smart grid domain for Home Energy Management [24].
Several new EHO variants have been proposed with different improvements. In [25], the authors introduce six individual updating strategies into basic EHO. In each strategy, one, two, or three individuals are selected from the previous

iterations, and their useful information is incorporated into the algorithm updating process. The experimental results on different test functions indicate that the proposed improved EHO version significantly outperformed basic EHO.
A new EHO algorithm with chaos theory to solve unconstrained global optimization problems was introduced in [26]. Two chaotic maps are incorporated into the basic EHO algorithm in order to improve the search quality. The comparison results with standard benchmark functions show that the new proposed algorithm outperforms the basic EHO and PSO in almost all cases.
Except for a very limited number of studies, accessing relevant information on social media based on the information foraging theory has not been addressed in the existing literature. Combining the IFT with an enhanced variant of EHO can substantially contribute in addressing the problem of large scale information access on social media.

III. Analogy between animal food foraging and information foraging on social media

The information foraging process intends to find paths leading to relevant information on the Web. The theory behind it is based on the analogy between animal food foraging behavior and human information seeking behavior. It assumes that when searching for information online, users follow indications and hints that guide them to relevant information, similar to how animals follow the scent of their preys to catch them. Figure 1 and Table 1 present a good illustration of the analogy between information foraging and animals' food foraging.

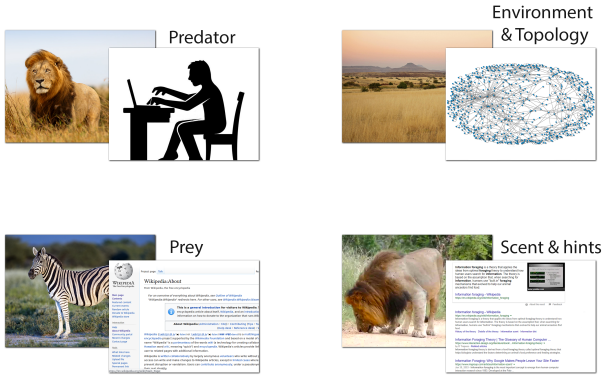


Figure. 1: Analogy between Information Foraging and Food Foraging

Elements	Food Foraging	Information Foraging
Actors	Predator	User
	Prey	Relevant information
Trigger	Hunger	Information needs
Environment	Nature, wilderness	Web structure, social graph
Cues	Scent of the prey	Hyperlinks, icons, titles

Table 1: Food Foraging analogy with Information Foraging

The following subsections describe our proposed model for adapting information foraging theory to social media platforms, as well as the basic notions on which the analogy with animal food foraging is grounded.

A. *Territory: social graph*

In the OFT, it is assumed that each animal operates in a delimited geographical territory, within which it searches for food. In information foraging, the territory corresponds to the search space composed of information sources such as documents, images and Web pages. When it comes to social media, the users' shared content serve as information sources. A social graph [27] is a representation of the users, their shared posts, and their social interactions and relationships.
In this paper, we model a social network as an oriented graph $G(V, E)$, where :

- the set of vertices V represents the social media users,
- the set of directed edges E represents relationships and interactions in the network, such as : a post, a re-post, a friendship, a mention, a reply and a follow.

A simplified social graph structure is shown in Figure 2. The edges that reflect the relations post, re-post, mention, and reply contain the social posts and so represent the information sources. We denote the set of these *content-sharing edges* by $\tilde{E}$ with $\tilde{E} \subseteq E$.

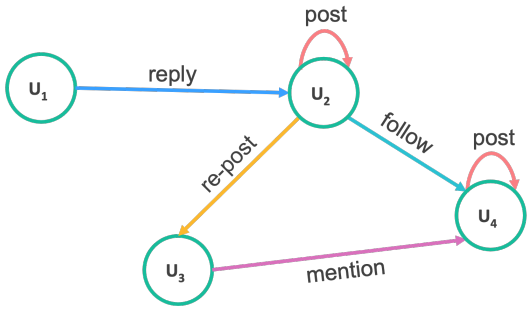


Figure. 2: Social graph structure

B. *Food diet: user's interests*

Each animal in the food chain has its own preferences in terms of food. Wilde animal for instance choose their preys based on their environment, their size and their hunting skills. In information foraging, the animal food diet is translated by the user's information needs that we call the *user's interests*. The information foraging process takes two inputs: a collection of posts represented by a social graph, in addition to the user's thematic interests. These users' interests can be expressed explicitly by the user or inferred implicitly from the user's social media activity (profile and interactions) [14]. The modeling of the user's interests consists in extracting the terms that are the most representative of the user's information needs from the keywords given by the user and/or the information accessible on their profile (biography, previous posts, etc.).
The extraction process includes: tokenization, stop words removal, stemming, and Term Frequency (TF) calculation. The words with the highest TF values are then stored into the user's interests vector I following the bag-of-words model. Figure 3 illustrates the process of modeling the user's topical interests by a vector of terms and using it in information foraging.

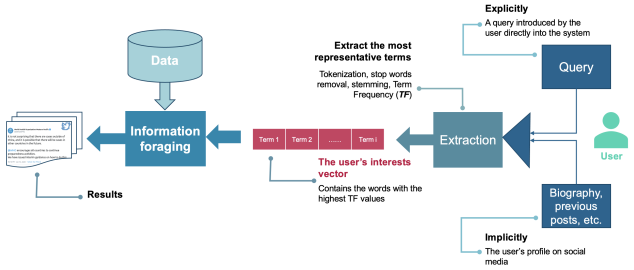


Figure. 3: User's interests extraction

C. *Scent: information scent*

The general goal of information access approaches is to offer mechanisms that can help finding relevant information, while minimizing the time spent doing the search. Likewise, the goal of animals in the wild is to find a decent amount of food whilst spending less energy. To achieve that, animals generally rely on their senses to locate and hunt their preys in an effective way. The authors of the IFT notice that users have a similar behavior when looking for information on the Web. The authors assume that when browsing the Web, users exploit available hints and cues to estimate the information value contained in accessible pages and therefore decide which pages to visit. This can be achieved thanks to the information scent concept [28], which can be seen in real life as the user's estimation of the value that a source of information will deliver to them. This value is primarily computed based on the source's description/content. In the case of the Web, for instance, information sources are Web pages, which are described by a URL, a title, and in certain cases an icon.
The goal in our context is to find relevant posts to satisfy a specific user's information needs. We presume that if a post is related to the user's interests, it will be more appealing to them, and that the information scent value should increase as we get closer to a relevant post and decrease otherwise. We define the information scent as the similarity evolution between the present post being visited at time t with the user's interests vector and the considered post to be accessed in time $t+1$ with the user's interests vector. Formula (1) shows how the information scent is calculated when considering to move from the current post located on the edge $\tilde{e}_i$ to one of its neighbors located on the edge $\tilde{e}_j$.

$$InfoScent(\tilde{e}_j) = Sim(\tilde{e}_j, I) - Sim(\tilde{e}_i, I) \quad (1)$$

Where :

- I is the user's interests vector;
- $\tilde{e}_i$ is the current post;
- $\tilde{e}_j \in N_i$, with N_i being the set of adjacent edges to $\tilde{e}_i$, i.e. $\tilde{e}_i$'s neighborhood;
- $Sim()$ represents the cosine similarity between two vectors.

The main role of the information scent is to guide the foraging, a positive value indicates that we are approaching a relevant post in the social graph, whereas a negative value indicates the opposite.

D. *Trail: surfing path*

While foraging food, animals follow a certain path that allows them to reach food sources in an optimal way according to the OFT. Web users have a similar behavior as they visit Web pages one at a time until reaching a relevant page that satisfies their information needs, constructing there a surfing path composed of one or more Web pages.
The information foraging process starts from an initial post and progresses through each step, attempting to reach a post with more relevant information than its predecessor. A surfing path is built for this purpose, starting with an initial content-sharing edge and then being enriched by adding further edges to create a chain of related posts. This means that at each step of the foraging process, the system should choose one content-sharing edge to visit from the neighborhood of the current post. This choice is made based on Formula (2).

$$P(\tilde{e}_i, \tilde{e}_j) = \begin{cases} 0, & \text{if } InfoScent(\tilde{e}_j) \leq 0 \\ \frac{InfoScent(\tilde{e}_j)}{\sum_{\tilde{e}_l \in Np_i} InfoScent(\tilde{e}_l)}, & \text{otherwise} \end{cases} \quad (2)$$

where :

- $P(\tilde{e}_i, \tilde{e}_j)$ is the likelihood of selecting the edge $\tilde{e}_j$ among the reachable edges from the current edge $\tilde{e}_i$
- Np_i is the set of adjacent content-sharing edges of the edge $\tilde{e}_i$ with a positive information scent value, i.e. $\forall \tilde{e}_l \in Np_i \; InfoScent(\tilde{e}_l) > 0$.

IV. Adapted elephant herding optimization to information foraging

The social structure of elephants is complex, varying by gender and population dynamics. Adult females form a matriarchal societies, while adult males are usually solitary [29, 30]. A herd structure is similar to concentric rings, with the innermost circle comprising a family unit of related female adults. A family unit is formed by the eldest most dominant female called the matriarch as well as her adult daughters, their calves and a number of juveniles. The male calves leave the herd when reaching adulthood, generally between the age of 12 and 15. From this stable core, the groupings widen to include less familiar individuals. A clan is formed when elephants gather in large groups consisting of different herds. The functioning of the elephants society is illustrated by Figure 4.
We decided to combine the information foraging theory with EHO in order to efficiently identify useful information in large social graphs. Each elephant will look for relevant social posts by browsing a section of the graph in this manner, constructing surfing paths that lead to relevant information. The elephants perform the foraging while taking advantage of the hierarchy and organization of their society. This will allow them to collaborate and find relevant information in a more effective and efficient way.
In this section, we adapt the basic Elephant Herding Optimization algorithm to information foraging. Note that the original EHO was developed to address continuous problems, whereas information foraging is a discrete combina-

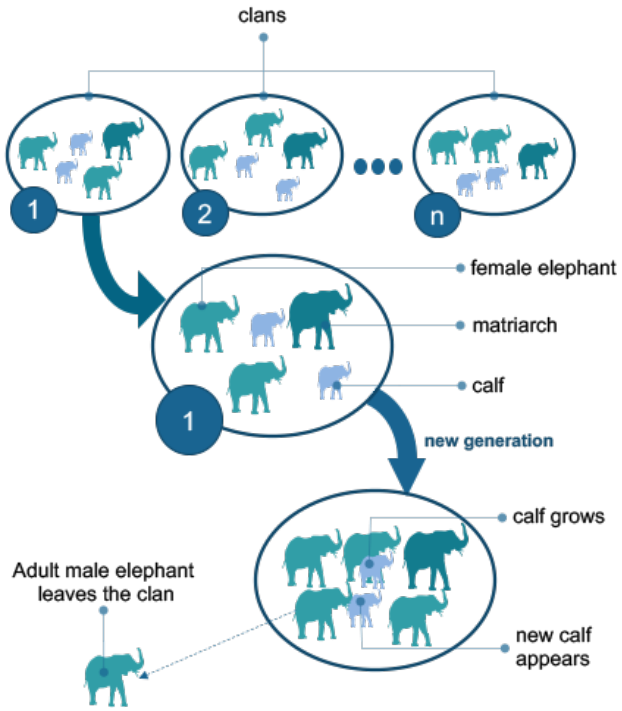


Figure. 4: Elephants society structure

torial problem. Further improvements to basic EHO will be introduced in sections V and VI.

A. *Generating the elephant population and assigning the positions*

To create an elephant population with p clans, we first generate p different elephants with respect to *distClan*, which represents the shortest minimal distance between clans. The rest of the elephants in each clan are then formed using the positions of the p elephants, with respect to *distElephant*, which represents the maximum distance between elephants in the same clan. Figure 5 depicts a population of three clans scattered throughout a social graph.

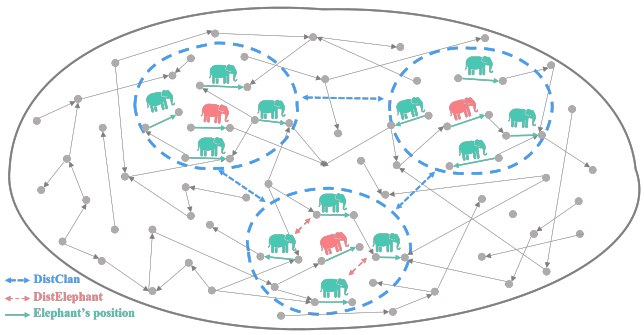


Figure. 5: A population of elephants distributed over a social graph

We assign m different positions to a social graph with m content-sharing edges, one for each edge. These positions are represented by integer values in the interval $[1, m]$. Each elephant j belonging to clan c_i is identified by a unique position denoted by the $x_{c_i,j}$. An elephant's position at time t is the position of the edge it is visiting at that time.

B. *Surfing paths construction*

Each elephant is assigned the duty of building a surfing path leading to relevant information during one iteration of the EHO algorithm. This is accomplished using Algorithm (1).

Algorithm 1 Building a surfing path

Input: $x_{c_i,j}$: elephant's position, I: user's interests, G: social graph;
Output: SP: a surfing path leading to a relevant post;

1: $SP \leftarrow \emptyset$.
2: Locate the edge $\tilde{e}_i$ corresponding to the elephant's initial position $x_{c_i,j}$
3: $SP = SP \cup \{\tilde{e}_i\}$
4: $N_i \leftarrow \emptyset$
5: **for all** adjacent edge $\tilde{e}_j$ to $\tilde{e}_i$ **do**
6: Calculate $InfoScent(\tilde{e}_j)$ using Formula (1)
7: **if** $InfoScent(\tilde{e}_j) > 0$ **then** $N_i = N_i \cup \{\tilde{e}_j\}$
8: **end if**
9: **end for**
10: **if** $N_i = \emptyset$ **then** return SP
11: **else**
12: Select a new content-sharing edge to visit from N_i following Formula (2)
13: Go to 3
14: **end if**

C. *Solutions evaluation*

At the end of each iteration of the algorithm, the solutions (the surfing paths) fetched by the elephants are evaluated according to a fitness function. To do so, we compute the similarity between the user's interests and each solution using Formula (3).

$$f(x_{c_i,j}) = Sim(\tilde{e}_k, I) \tag{3}$$

Where :

- $\tilde{e}_k$ represents the last social post on the surfing path constructed by elephant j in clan c_i;
- I is the user's interests vector;
- $Sim(\tilde{e}_k, I)$ represents the cosine similarity between $\tilde{e}_k$ and I.

D. *Updating Operator*

The elephants' positions are updated using Formula (4) at the end of each iteration of the algorithm, once the new solutions have been evaluated.

$$x_{new,c_i,j} = x_{c_i,j} + \alpha(x_{best,c_i} - x_{c_i,j}) \times r \tag{4}$$

Where:

- $x_{new,c_i,j}$: is the new position of the elephant;
- $x_{c_i,j}$: is the current position of the elephant;
- x_{best,c_i} : is the matriarch's position;

- $\alpha \in [0, 1]$: is an empirical parameter that defines the influence of the matriarch over the new position of the elephant j;
- $r \in [0, 1]$: is a random number, which aims at improving the diversity.

The position of each clan's matriarch is also updated throughout generations by utilizing Formula (5) to calculate the average fitness of each clan. After that, Formula (6) is used to compute the new matriarch's position using the position of the elephant with the closest fitness value to the clan's average fitness.

$$f_{avg,c_i} = \frac{1}{n_{c_i}} \sum_{j=1}^{n_{c_i}} f(x_{c_i,j}) \tag{5}$$

$$x_{newbest,c_i} = x_{avg,c_i} \times \beta \tag{6}$$

Where:

- f_{avg,c_i}: represents the average fitness value of the clan c_i;
- $x_{newbest,c_i}$: represents the new position of the matriarch of the clan c_i;
- x_{avg,c_i}: is the position of the elephant with the closest fitness value to f_{avg,c_i};
- $\beta \in [0, 1]$: is an empirical parameter, which determines the influence of the average position on the matriarch's new position;
- n_{c_i}: represents the number of elephants in the clan c_i;
- $x_{c_i,j}$: is the position of elephant j in clan c_i.

E. Separating Operator

The elephant with the lowest fitness value will leave the clan at the end of each generation. Formula (7) is used to create a new elephant to replace the one that left.

$$x_{worst} = x_{min} + (x_{max} - x_{min} + 1) \times r \tag{7}$$

Where:

- x_{worst} stands for the position of the elephant with worst fitness value;
- x_{min} and x_{max} are the upper and lower bounds of the positions interval;
- r is a stochastic and uniform distribution parameter.

V. Defining territories with clustering

Although elephants are not territorial animals, they utilize specific home areas during particular times of the year. Their home ranges vary from from 15 to 3,700 square kilometers (24 to 5,958 square miles) depending on the population and the habitat. This delimited area helps elephants to better master their environment and remember the location of food and water sources [31].
Implementing this concept and incorporating it into EHO would be of a great benefit to solve large scale problems. In fact, dividing the search space into sub-areas based on some problem-related features can help to limit the search to one of these sub-areas, and thus improve the effectiveness and efficiency of EHO.
When constructing a surfing path during the information foraging process, the choice of the starting point can have a major impact on the outcome and therefore, determine whether or not the path will lead to relevant information. In fact, locating the right post from which to initiate the navigation with regards to the user's interests is of a high importance. This becomes even more obvious and crucial when dealing with large scale data, as skipping posts that are not related to the user's interests could significantly improve the efficiency of the information foraging process. For instance, if a user is interested in information about health, it would be unnecessary to search posts talking about sport.
Consequently and given the above observations, we propose to introduce a new step in the Elephant Herding Optimization algorithm, consisting of dividing the search space into multiple regions in order to explore them more efficiently. In addition to modeling the concept of territories in nature, this will improve the performance of the algorithm, especially when dealing with large scale problems.
There are numerous methods for grouping similar objects together, they can be either supervised or unsupervised depending on whether classes of objects already exist or not. Supervised classification considers classes to insert the objects whereas unsupervised classification generates clusters as outcome. Clustering is the process of organizing objects into groups whose members are similar. A cluster is a collection of objects which are consistent internally, but clearly dissimilar to the objects belonging to other clusters. One of the main advantages of clustering over supervised classification is the fact that it doesn't require predefined classes and It can therefore be performed with data of different sizes without the need of a taxonomy or a training set.
In this paper, we perform the clustering phase using k-means algorithm, which is known for its efficiency with large datasets and its capacity of working with textual data [32, 33]. It intends to automatically group a set of n objects into k clusters, so that objects in a same cluster are similar to one another while objects from different clusters are dissimilar [34]. The grouping decision is based on the distance between the object and each cluster centroid (mean), in a way that each object belongs to the cluster with the nearest centroid. A centeroid serves as a prototype of its corresponding cluster, and is defined as the average of all the objects in that cluster. The number of clusters k can be either predefined or user-defined depending on the problem, the number of objects, and the goal behind clustering.
When it comes to textual data clustering, the idea consists in grouping texts or documents in clusters based on their content similarity. In order to achieve this goal, a proper document representation method is necessary. We use the vector space model (VSM) to represent each social post $\tilde{e}_i$ as a weighted vector of terms $\tilde{e}_i =< w_{i1}, w_{i2}, ..., w_{i|T|} >$ where T is the set of terms or features that occur at least once in the social graph G. The detailed clustering process using k-means is presented in the following subsections.

A. Text preprocessing

This preprocessing phase refers to the set of actions that are applied to the social posts in order to achieve a good statistical representation of the whole collection. This phase is performed before defining the weights of the words using TF-IDF and includes the following actions:

- Tokenization, which consists in dividing the text into individual words.
- Removing special characters related to social media platforms, links, usernames, etc.
- Deleting common words that don't bring any semantic meaning to the text using a stop words list.
- Reducing each word to its root using adequate algorithms such as *Porter Stemming*. As a result, inflected words will be grouped under their word stem, which is referred to as a term.

The result of the textual preprocessing is a collection of posts that are each represented by a set of significant terms. The following subsection explain how each post is afterwards converted into a weighted vector of terms.

B. Feature extraction with TF-IDF

The goal of the feature extraction using TF-IDF is to create a mapping of the textual data into vectors of terms. This vector representation of the social posts is grounded on the term relevance concept. The weight associated to each term should be proportional to its importance, so that terms with high weight values are considered as relevant. This method consists in increasing the weight of a term when it appears many times in a post and lowering it when it is common in many posts. We can summarize the TF-IDF calculation with the two following steps:

1) Term Frequency (TF)

The term frequency $TF(t_i, \tilde{e}_j)$ estimates the importance of a term t_i in a post $\tilde{e}_j$ based on how often t_i appears in $\tilde{e}_j$. The more frequent a term is in a post, the more important it is in its description. We use Formula (8) to compute the term frequency.

$$TF(t_i, \tilde{e}_j) = \frac{freq_{ij}}{\sum_{\forall t_l \in \tilde{e}_j} freq_{lj}} \quad (8)$$

With $freq_{ij}$ being the number of occurrences of term t_i in post $\tilde{e}_j$.

2) Inverse Document Frequency (IDF)

Inverse Document Frequency indicates how commonly a word is used in a collection of documents. A term has an important characterizing power if its frequency is high in a particular social post and low in the rest of the posts. We estimate the inverse document frequency using Formula (9).

$$IDF(t_i) = log(\frac{|\tilde{E}|}{n_i}) \quad (9)$$

Where:

- $|\tilde{E}|$ is the total number of posts in the social graph.
- n_i is the number of posts containing the term t_i.

3) TF-IDF

We compute the weigh of a term t_i in a post $\tilde{e}_j$ using the $TF - ID$ score, which is obtained by Formula (10). The value of the weight increases proportionally to the number of times a term appears in the post and is offset by the number of posts containing that term.

$$w_{ij} = TF - IDF(t_i, \tilde{e}_j) = TF(t_i, \tilde{e}_j) \times IDF(t_i) \quad (10)$$

Table 2 illustrates the vector representation of a social graph containing p posts and n terms using the vector space model with TF-IDF weighting measure.

Terms \ Posts	$\tilde{e}_1$	$\tilde{e}_2$	$\tilde{e}_3$	...	$\tilde{e}_p$
$term_1$	0.015	0.342	0	...	0
$term_2$	0.231	1.164	0.324	...	1.002
$term_3$	0	0.102	0	...	0.076
...	...	...	...	...	...
$term_n$	1.562	0	0.067	...	0

Table 2: Vector representation with TF-IDF

Once the posts are cleaned and represented as weighted vectors of terms, the clustering phase can be launched using k-means algorithm.

C. Clusters initialization

A centroid is assigned to each cluster amongst the posts of the social graph. Each centroid is represented by the weighted terms vector of its corresponding post. The k centroids $m_1, m_2, ..., m_k$ are initialized randomly by choosing k random posts. The pseudo code of the initialization is presented in Algorithm (2).

Algorithm 2 Centroids Initialization

Input: $\tilde{E}$: posts of the social graph G, k: number of clusters;
Output: k centroids;
1: **for** $i \leftarrow 1$ to k **do**
2: $r \leftarrow$ random position
3: $m_i \leftarrow \tilde{e}_r$
4: Insert m_i in $Centroids$
5: **end for**
6: Return $Centroids$

Once the centroids defined, the clusters are populated with posts based on the distance between each post and the clusters centroids. To compute this distance, we use the Euclidian distance measure, which represents the ordinary straight-line distance between two points in Euclidean space. In our case each point is either a post $\tilde{e}_i$ or a centroid m_j, both represented by a weighted terms vector. In an n-dimensional space, the distance is calculated using Formula (11).

$$d(\tilde{e}_i, m_j) = \sqrt{\sum_{l=1}^{n} (\tilde{e}_{il} - m_{jl})^2} \quad (11)$$

Where:

- n : represents the vector's size;
- m_{jl}: represents the weight of term l in the centroid m_j;
- $\tilde{e}_{il}$: represents the weight of term l in post $\tilde{e}_i$.

The pseudo code of the clusters construction is presented in Algorithm (3).

Algorithm 3 Clusters construction

Input: $\tilde{E}$: posts of the social graph G, $Centroids$: centroids vectors;
Output: S: set of clusters;
1: **for all** $\tilde{e}_i \in \tilde{E}$ **do**
2: $min_{dist} = \infty$
3: $cluster_{id} \leftarrow -1$
4: **for all** $m_j \in Centroids$ **do**
5: compute the distance $d(\tilde{e}_i, m_j)$ using Formula (11)
6: **if** $d(\tilde{e}_i, m_j) < min_{dist}$ **then**
7: $cluster_{id} \leftarrow j$
8: **end if**
9: **end for**
10: Insert $\tilde{e}_i$ in the right cluster $S_{cluster_{id}}$
11: **end for**
12: Return S

D. Centroids update

The centroids are updated throughout the iterations of the k-means algorithm. Their positions are recalculated and moved towards the center of their respective clusters at the end of each iteration. For this purpose, the mean weighted terms vector μ is computed for each cluster. This vector represents the average weight of each term in all the posts belonging to the cluster, and will be used to define the new cluster's centroid. After generating the vector μ_j of the cluster S_j, the distance between this vector, which doesn't constitute a real social post, and each post of the cluster S_j is calculated in order to select the nearest post to μ_j and set it as the new centroid m_j. This process is detailed in Algorithm (4).

E. K-means algorithm for territories definition

K-means algorithm is launched prior to the information foraging process, in order to define search territories and therefore divide the search space into multiple clusters based on the content of the social posts. Once the centroids are initialized, the algorithm enters a loop composed of two main steps. The first takes in charge the clusters creation by assigning each post of the social graph to the cluster whose centroid is the nearest. The second step defines a new centroid for each cluster, based on the mean of the weighted terms vectors of all posts assigned to that cluster. The stop condition of the loop is either convergence or the reach of a maximum number of iterations. The territories definition process using k-means clustering is presented in Algorithm (5).
Note that the time complexity of the algorithm is estimated to $\mathcal{O}(n*m*k*l)$, with n being the total number of posts in the social graph, m the size of the vectors, k the number of clusters and l the number of iterations. Although the complexity is linear, it can require a significant time especially with large scale social graphs. Nevertheless, this will not affect the information foraging performance, since the clustering is performed offline and only once.

Algorithm 4 Centroids update

Input: S_j: a cluster;
Output: m_j: the new cluster's centroid;
1: $\mu_j \leftarrow [0, 0, ..., 0]$
2: **for all** $\tilde{e}_i \in S_j$ **do**
3: $\mu_j \leftarrow \mu_j + \tilde{e}_i$
4: **end for**
5: $\mu_j \leftarrow \frac{1}{|S_j|}\mu_j$
6: $min_{dist} = \infty$
7: $post_{id} \leftarrow -1$
8: **for all** $\tilde{e}_i \in S_j$ **do**
9: compute the distance $d(\tilde{e}_i, \mu_j)$ using Formula (11)
10: **if** $d(\tilde{e}_i, \mu_j) < min_{dist}$ **then**
11: $post_{id} \leftarrow i$
12: **end if**
13: **end for**
14: $m_j \leftarrow \tilde{e}_{post_{id}}$
15: Return m_j

Algorithm 5 K-means for territories definition

Input: $\tilde{E}$: posts of the social graph G;
Output: S: set of k clusters;
1: initialize k centroids randomly using Algorithm (2)
2: **repeat**
3: create k clusters using Algorithm (3)
4: update the centroids using Algorithm (4)
5: **until** convergence or max_itererations
6: return S

VI. Enhanced EHO for large scale information foraging (EEHOLSIF)

Addressing large scale information foraging can be tricky and time consuming. The worst case complexity of information foraging corresponds to the case when the social graph is a complete graph and is estimated to $\mathcal{O}(\prod_{i=1}^{p} n - i)$, with n being the total number of posts in the graph and p the surfing depth [16]. In this paper, we propose a new bio-inspired approach to information access based on enhanced elephant herding optimization using the concepts presented in sections III, IV and V. We introduce a new enhanced Elephant Herding Optimization variant to improve the performance of the original algorithm and adapt it to large scale information foraging. Our contribution focuses on several aspects, from the initialization of the algorithm to the clans' structure. The main aspects are detailed in the following subsections.

A. Semantic position assignment

The concept of territories introduced in section V allows to considerably optimize the foraging process by delimiting the search area. We exploit the clustering results to assign a nu-

merical position to each post on the social graph based on the posts' content. Following the clustering process, each post of the social graph is given an integer identifier, which will serve as a position in EEHOLSIF algorithm, in a way that posts belonging to the same cluster have neighbor positions. These positions are sorted according to the Euclidian distance between posts and the centroid of the cluster, so within the same cluster if a position i is less than another position j, this means that the post associated with i is closer to the centroid than the post associated with j. For instance, let us consider a social graph with 4000 posts, and 3 clusters with S_1 having 1500 posts, S_2 having 1500 posts and S_3 having 998 posts. The positions in each cluster will be distributed as follows:

- $S_1 = \bigcup_{i=1}^{1500} \tilde{e}_i$
- $S_2 = \bigcup_{i=1501}^{3001} \tilde{e}_i$
- $S_3 = \bigcup_{i=3002}^{4000} \tilde{e}_i$

B. Initialization of the algorithm

Territories definition and semantic positions assignment will play a major role in the initialization phase of EEHOLSIF. Indeed, unlike the original version of the algorithm, where the initialization is performed in a complete random way, the clustering and the semantic positions assignment permit to target the cluster containing the posts that are the closest to the user's interest and then set the initial elephants' positions accordingly. The minimal distance between clans *distClan* and the maximal distance between elephants of the same clan *distElephant* become more representative since the positions are assigned based on the content of the posts. In fact, this ensures that the elephants of the same clan are browsing pots that have similar content while elephants of different clans are located on dissimilar posts with regards to their content. This will result in a better distribution of the elephants on the search space and a better coverage of the potential solutions. In order to initialize the clans, we first compute the euclidean distance between the user's interests vector I and each centroid vector m_j. Then, we launch the different clans either on the cluster with the nearest centroid or on a cluster chosen according to a uniform distribution probability. For this purpose, we introduce q, a random variable uniformly distributed in $[0, 1]$ and $q_0 \in [0, 1]$ a tunable parameter. We propose the pseudo random proportional rule for choosing a territory highlighted by Formula (12).

if $q < q_0$ then

$$P(c_i, S_j) = \begin{cases} 1 & \text{if } j = argmin(d(I, m_j)) \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

else

$$P(c_i, S_j) = \frac{d(I, m_j)}{\sum_{l=1}^{k} d(I, m_l)}$$

Where:

- $P(c_i, S_j)$ is the probability to place clan c_i on cluster S_j;
- $d(I, m_j)$ is the euclidean distance between the user's interests I and the centroid of cluster S_j

C. Solution construction

Another improvement is related to the construction of the solution, which is in our case a surfing path. Unlike in the original algorithm where the elephants consider ready-made solutions, we give the task of constructing a solution to each elephant starting from its initial position. For this purpose, we incorporate information foraging concepts including the information scent (Formula (1)) and the surfing decision rule (Formula (2)). The solution construction process is given by Algorithm (1).
Note that during the solution construction, an elephant can leave its territory if the surfing path leads it towards a post located on a neighbor territory. We consider two territories as neighbors if they share adjacent edges. Two edges are adjacent if they are both incident with a common vertex.

D. Clan migration

In nature, an elephant clan might separate from the larger herd in response to limited food supplies encountered during a dry season. If food sources are scarce, it is more efficient for elephants to travel as individual clans, rather than large herds.
We incorporate this natural phenomenon in the enhanced EHO for large scale information foraging as a stagnation prevention mechanism. We believe there is a strong analogy between the lack of food sources in nature on one side and the inability of the elephants to improve their solution after several generations on the other. We introduce a migration parameter t_0, which serves as a threshold that controls the maximum number of generations a clan can spend without improving its best solution. If a clan exceeds this threshold, it migrates towards a new territory with the hope of finding better solutions. The migration is performed by choosing a new cluster randomly and defining the migrating clan positions within that cluster.
The enhanced elephant herding optimization for large scale information foraging is presented in Algorithm (6).

VII. Experiments

This section is organized in four subsections, first we describe the dataset we use in the evaluation. We then present the results obtained with the adapted EHO for information foraging. Next, we follow up with the results of the enhanced EHO for large scale information foraging. Finally, we finish by doing a comparative study with other approaches from the literature.
All the experiments were implemented using *Java* programming language and were conducted on a laptop running Windows 10 with an Intel Core i5-4300M CPU at 2.60GHz and 6GB of RAM.

Algorithm 6 Enhanced EHO for large scale information foraging

Input: I: the user's interests, G: the social graph;
Output: a list of surfing paths ranked by relevance;

Divide the search space into k different territories using Algorithm (5)
Set the generations counter $t \leftarrow 1$, the solutions list $sols \leftarrow \emptyset$ and the stagnation counter for each clan $g_{c_i} \leftarrow 0$
Initialize empirical parameters α, β, q_0, t_0, maximum generations $MaxGen$, number of clans $nClans$, population size and number of elephants in each clan n_{c_i}.
Initialize the positions of the elephants according to the user's interests I using Formula (12) and with respect to $distClan$ and $distElephant$
while $t \leq MaxGen$ **do**
    for $i \leftarrow 1$ to nClans **do** ▷ for all clans in elephant population
        for $j \leftarrow 1$ to n_{c_i} **do** ▷ for all elephants in clan c_i
            Build the elephant's j surfing path using Algorithm (1)
            Calculate the elephant's fitness using Formula (3).
        end for
        if clan c_i improved its best solution compared to the previous generation **then**
            Update $bestSol_{c_i}$
        else
            $g_{c_i} \leftarrow g_{c_i} + 1$
        end if
        if $g_{c_i} \geq t_0$ **then**
            Migrate clan c_i towards a new territory chosen randomly
            $g_{c_i} \leftarrow 0$
        else
            Update the positions of the elephants $x_{c_i,j}$ using Formula (4).
            Update the matriarch's position using Formula (6).
            Locate the worst elephant to leave clan c_i according to the fitness function.
            Generate a new elephant in the clan c_i using Formula (7).
        end if
    end for
    Append the best surfing paths found in generation t to $sols$
    Update the generation counter, $t \leftarrow t + 1$.
end while
Return the best surfing paths ranked relevance.

A. Dataset description

We tested our algorithm on *Twitter*, which is one of the most popular social networks and microblogging platforms. We constructed a dataset composed of 1 410 246 tweets that we grouped in one big social graph. The data crawling was performed using NodeXL [35] and took place during the end of 2020. Figure 6 and Table 3 showcase the main topics covered by the dataset.

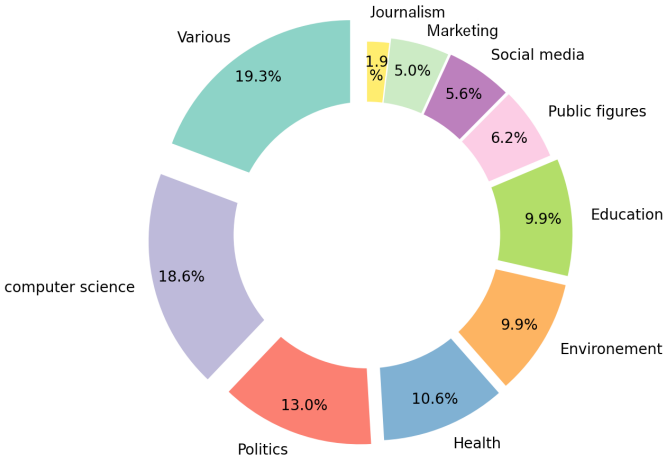


Figure. 6: Topics covered by the dataset

B. Adapted EHO for information foraging results

1) Empirical parameters setting

We conducted extensive tests for the sake of tuning the empirical parameters to values that ensure the best results in terms of relevance and response time, i.e. maximizing the similarity between the user's interests and the surfing path while minimizing the execution time. It is important to note that the stochastic aspect of the EHO algorithm requires to test each parameter value multiple times, to get stable outcomes. For that purpose, we run the tests at least 100 times for each parameter.
First, we started with parameters $\alpha \in [0, 1]$ and $\beta \in [0, 1]$. Recall that α is a scale parameter that determines the influence of the matriarch's position on rest of the elephants of the same clan, while β determines the influence of the average position of the clan on the matriarch's position. To select the best values of both parameters, we combined each value of α in the range $[0, 1]$ with all possible values of β also in the same range. Figure 7a shows the similarity score results, while Figure7b displays the response time results in seconds. The 3D representation gives a good visualization of the similarity and time evolution with the variation of the parameters. With respect to the results showed in both figures we set α to 0.9 and β to 0.4.
Another important combination of parameters is the number of clans and the number of elephants in each clan. A proper number will help to visit different parts of the social graph and therefore get closer to relevant posts. The results shown in Figure 8a and Figure 8b allow to determine the adequate number of clans and the number of elephants in each clan. The number of clans is set to 8, with 90 elephants in each clan.
The number of generations is the parameter that allows the algorithm to evolve a sufficient amount of time so it can reach better results and approach the global optimum. We can observe from Figure 9 that the best number of generations would be 40, since it maximizes the similarity and minimizes the response time.

2) Foraging results

Table 4 presents some examples of the adapted EHO for information foraging results with 7 different users' interests (column one) generated for evaluation purpose. The surfing path with the most relevant tweet is displayed in column two, the similarity value between the surfing path and the user's interests is shown in column three, and the response time in seconds alongside the length of the surfing path are displayed in columns four and five, respectively. Note that when the surfing depth is greater than 1, the entire surfing path is displayed in chronological order of access, as in the case of the user's interest "diabetes type 2, intermittent fasting," for example.
We observe that in almost all cases, the system is capable of finding relevant tweets. However, the response time is relatively long, mainly because of the big size of the social graph and the fact that the foraging process happens exclusively online. We can also notice that the surfing depth is to a certain extent small, which can be explained by the fact that the social graph is not strongly connected. Moreover, during the construction of the surfing path, a tweet is only inserted if it is more relevant than the tweets that were accessed before it in the same path.

C. Enhanced EHO for large scale information foraging results

Although we were able to reach relevant posts using our first attempt based on the adaptation of the original EHO algorithm to information foraging, the results showed some limitations related to the efficiency, especially when it comes to big social graphs. To cope with this issue, we proposed in Section VI a novel approach consisting in an enhanced version of EHO for large scale infomration foraging.

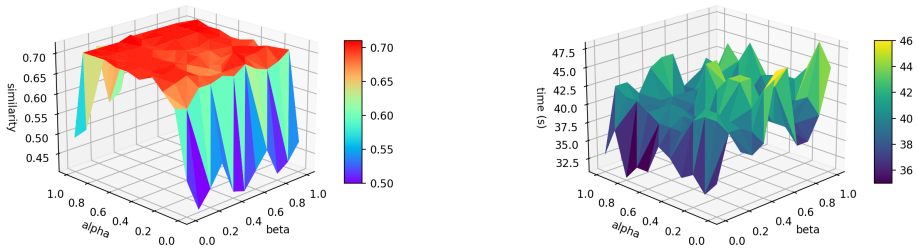
1) Empirical parameters setting

The first parameter to define is the number of territories, i.e. the number of clusters k. For this purpose, we tested the k-means algorithm with different values of k in the interval $[1, 80]$. For each fixed number of clusters k, we use Formula (13) to compute the total Within Cluster Sums of Squares (WSS), which measures the average distance between the posts and their corresponding centroids for each cluster [36, 37].

$$WSS = \sum_{i=1}^{k} \sum_{\tilde{e} \in S_i} d(\tilde{e}, m_i) \tag{13}$$

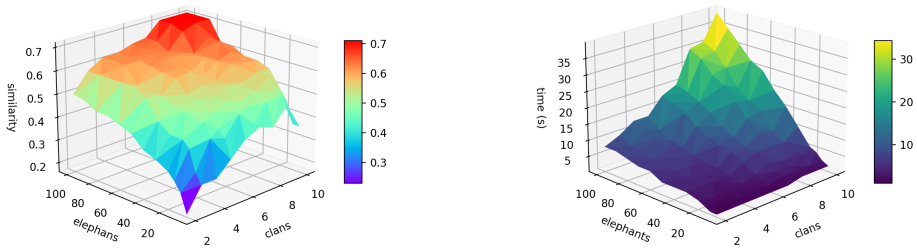
Where:

- k: is the number of clusters



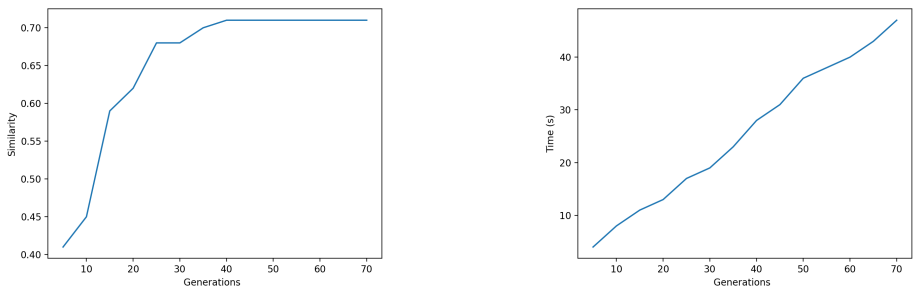
(a) Similarity variation (b) Time variation

Figure. 7: Setting α and β parameters based on Time and Similarity variation



(a) Similarity variation (b) Time variation

Figure. 8: Setting the number of clans and elephants based on Time and Similarity variation



(a) Similarity variation (b) Time variation

Figure. 9: Setting the number of generations based on Time and Similarity variation

Main topic	Subtopics
Computer science	Machine learning, Deep learning, Artificial intelligence, Big data, Graph database, Open Data, IoT, 5G, Social graph, Cyber security, Cyber-attack, Blockchain, Bitcoin, Hack, IBM, Data science, Power BI, Robotics, Smart city, Smart Home, Digital Predictive Analytics, Mathematics, Cisco, self-driving cars, VMware, Virtual reality, Web, domains, TensorFlow.
Politics	American express, Free speech, Black lives matter, Time is up, Immigration, Twitterstorian, Brexit UK, Vote, President Trump, Democracy, Breaking News, Democrats, Racism, white supremacy.
Health	Covid-19, flatern the curve, Vaccine, Cholera, intermittent fasting, Sugar free diet, Healthcare, HealthTech, Hemophilia, Malaria, Paludism, World Mustiquo day. Pregnancy, Abortion, protest against Exams in covid, diabetes, cannabis, AIDS, C-Section, Hydroxychloroquine, personalized Medicine.
Environment	Biodiversity, Food security, Climate change, Dogs, Dogs lovers, Fosil oil, fuels, Global warming, CO2, climate Strike.
Education	Books, E-books, QuickBooks, science teachers, Distance learning, Homeschooling, School closing, School reopening, Online learning, Book awards celebration.
Public figures	Bernie Senders, Joe Biden, Donald Trump, Michelle Obama, Snowden, Michael Cohen, Liam Payne, Bill Gates, SpaceX.
Social Media	Social Media, Social Media Marketing, Blogging, Tik Tok, Twitter, delete Facebook, YouTube, Podcast.
Marketing	Marketing, Social Media Marketing, Digital Marketing, public relations
Journalism	Journalism, sociologist, Forbs, Articles.
Various	Motor Trend, Unilever, Post Master, boycott whole food, Toxic masculinity, Feminism, Sexual harassments, save your children, lets chat, social pulse summit, youth day.

Table 3: Topics and subtopics covered the dataset

User's interests	*Most relevant surfing path*	*Score*	*Time (s)*	*Surfing depth*
Machine Learning, IA, Python	Python for Machine Learning and Data Mining #DeepLearning #datamining #learning via https://t.co/qcC4wrx6m6 https://t.co/kLvs68HzEQ	0.71	26	1
American Express, free speech, democracy	American Express https://t.co/o9suYDdsV3	0.64	27	1
Public Relations, communication	A public relations strategy is critical now more than ever #PublicRelations https://t.co/KZmGOD2Xjg	0.54	29	1
COVID19 immunity transmission	@CoocoLa_Vrej WHO is still not sure if those who recovered from COVID 19 develop a certain immunity that they will not get COVID 19 virus again.	0.57	26	1
diabetes type 2, intermittent fasting	Can intermittent fasting make you diabetic?	0.74	28	3
	Does anyway here do intermittent fasting? How do you do it?			
	Intermittent fasting has proven to help cure Type II diabetes			
Digital marketing, business, social media	RT @V2M2Group: Get Social: The Power of Social Media for Marketing Your Business? #business #digitalmarketing #marketing #smallbusiness #SocialMedia #GuernseyBusinesses https://t.co/8wnlALHXsh	0.73	29	1
Bitcoin prices market	Bitcoin price within about 3% of gold price https://t.co/GwjcMSB9Jp	0.67	24	1
Joe Biden and Bernie Senders	@LyndaMo85130479 @BugOffDear Biden positions are literally just copy/pasted from Bernie Sanders	0.46	26	3
	Folks mention Biden's past plagiarism True			
	But who believes Joe had anything to do with deciding this, or preparing the doc? Who is in charge? https://t.co/x8RBCz4H6D			
Smart City, 5G, IoT	Samsung IoT Smart City https://t.co/Xnf5JHnOq9 via @YouTube @_funtastic5_ #TelkomFuntastic5 #RWSTREG5 #smartcity	0.70	25	1

Table 4: Information Foraging Results

- S_i: is a cluster
- $\tilde{e}$: is a post
- m_i is the centroid of cluster S_i
- $d(\tilde{e}, m_i)$: is the euclidean distance between the post and its associated centroid

Once the calculations are finished, we plot the curve of WSS according to the number of clusters k. The location of a bend (knee) in the plot is generally considered as an indicator for the proper number of clusters. The results shown in Figure

10, indicate that the best number of clusters is $k = 55$.

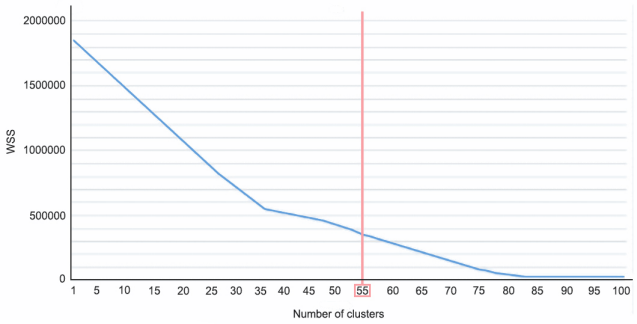


Figure. 10: Within Cluster Sums of Squares plot

Figure 11 displays the distribution of the 1 410 246 tweets over the 55 clusters. We observe that the smallest cluster contains 7 821 tweets while the largest one groups a total of 30 144 tweets with a median of 18 267 tweets per cluster.

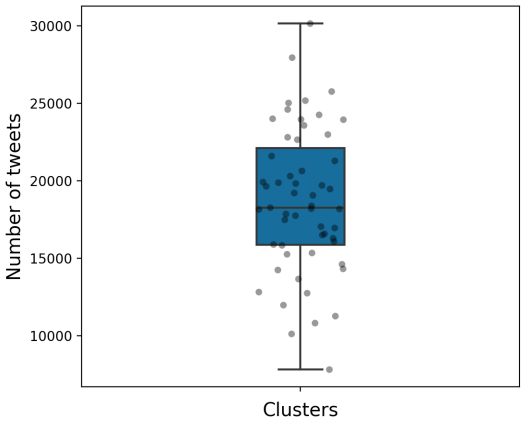


Figure. 11: Boxplot of clustering results

Given that the social graph is now divided into 55 territories, the rest of the parameters needs to be tuned again accordingly. To do so, we conducted extensive tests following the same steps of subsection VII-B.1. We also performed the tests at least 100 times for each parameter. Figure 12 shows the tests we undertook to set the empirical parameters α, β, number of clans, number of elephants and the maximum number of generations, while Table 5 shows the optimal values of these parameters.

Parameter	Value
α	0.5
β	0.5
$nClans$	5
n_{c_i}	50
$MaxGen$	25
q_0	0.75
t_0	6

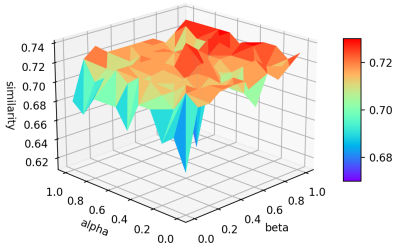
Table 5: Empirical parameters values for enhanced EHO for large scale information foraging

2) Foraging results

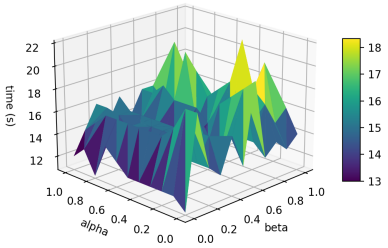
The comparison results between adapted EHO for information foraging (EHOIF) and enhanced EHO for large scale information foraging (EEHOLSIF) are reported in Table 6. The first column represents the user's interest, while the rest of the columns provide for each approach the most relevant surfing path, its similarity with the user's interests, the response time and the surfing depth.
We observe that both approaches are capable of finding relevant tweets that can potentially satisfy the user's interests. However, the main difference between the two approaches resides in the score and the response time. In fact, we can see that EEHOLSIF can achieve a higher score in almost all cases. Furthermore, its response time is considerably faster. We believe that this gain in performance is the result of the improvements we brought to the algorithm to make it able to undertake large scale information foraging and in particular the territory concept and the migration mechanism.
Figure 13 displays the comparison results between EHO for information foraging and enhanced EHO for large scale information foraging in terms of relevance score, surfing depth, convergence, and response time. This comparison was made by testing 70 different users' interests, generated randomly, with both approaches. Figure 13a shows that EEHOLSIF is able to achieve better relevance scores with an average of 0.77 against 0.65 for EHOIF. Moreover, EEHOLSIF can reach very high scores exceeding 0.9, while EHOIF is limited to 0.74. The opposite can be said regarding the surfing depth, which is generally higher with EHOIF as shown in Figure 13b. This is due to the fact that in EHOIF the surfing process is initialized in a complete random way, without taking into consideration the content of the tweets. On the other hand, the territories concept in EEHOLSIF allows to target tweets similar to the user's interest, which helps better guiding the surfing and thus shortening the surfing paths. As for convergence, we remark that EEHOLSIF has a faster convergence rate while reaching a higher optimum compared to EHOIF. We can see from Figure 13c that EEHOLSIF converges after 25 generations achieving a relevance score of about 0.77. EHOIF on the flip side, converges after 40 generations with a score of 0.65. This results confirm that the migration mechanism introduced in EEHOLSIF helps preventing stagnation and hence allows the algorithm to reach better solutions rapidly. Figure 13d exhibits the run time results, which show that EEHOLSIF is remarkably faster in all cases with an average response time of 0.9 seconds against 26.5 seconds for EHOIF.

D. Comparative study

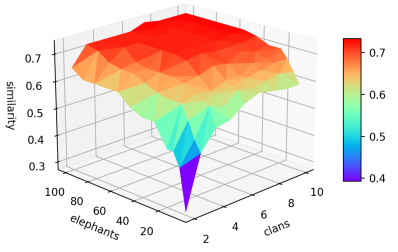
A crucial step to validate our work is to compare it to other metaheuristic-based information foraging approaches from the literature. In this section, we pay particular attention to two approaches.
The first approach we consider is based on *Ant Colony System (ACS)* and was already used to address Web information foraging in [16]. First, we implemented ACS and adapted it to tackle information foraging on social media. Then, we conducted a series of tests to set ACS empirical parameters with the aim of maximizing the system's performance. Table 7 indicates the parameters' values we fixed following these



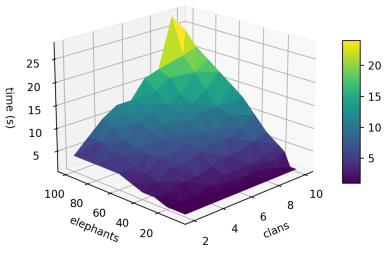
(a) Similarity variation regarding α and β



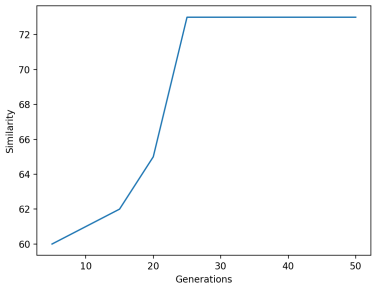
(b) Time variation regarding α and β



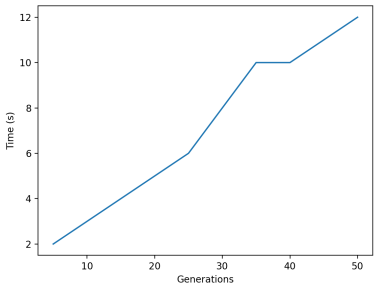
(c) Similarity variation regarding the number of clans and elephants



(d) Time variation regarding the number of clans and elephants



(e) Similarity variation regarding the number of generations

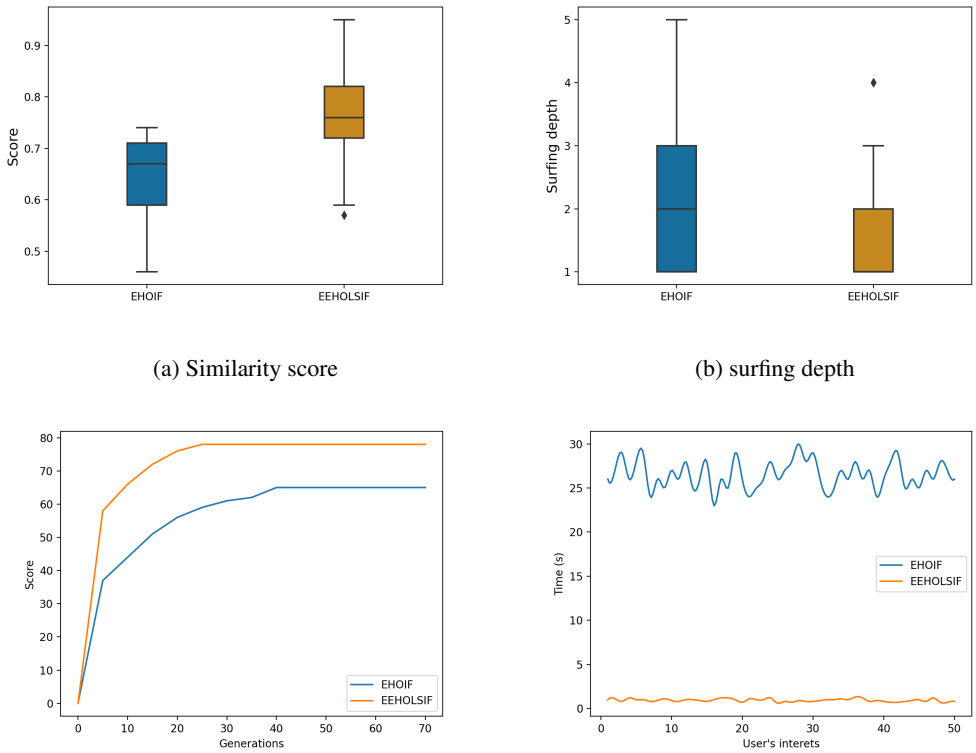


(f) Time variation regarding the number of generations

Figure. 12: Setting the empirical parameters for enhanced EHO for large scale information foraging

	EHOIF				EHO-KM-IF			
User's interests	**Most relevant surfing path**	**Score**	**Time (s)**	**Surfing depth**	**Most relevant surfing path**	**Score**	**Time (s)**	**Surfing depth**
Machine Learning, IA, Python	Python for Machine Learning and Data Mining #DeepLearning #datamining #learning via https://t.co/qcC4wrx6m6 https://t.co/kLvs68HzEQ	0.71	26 s	1	Introduction To Machine Learning with Python #MachineLearning #deeplearning #learning via https://t.co/IWfQGVjKXK https://t.co/ZobXDvNWRO	0.71	1 s	1
American Express, free speech, democracy	American Express https://t.co/o9suYDdsV3	0.64	27s	1	American Express https://t.co/bGaqvvZp69	0.64	1.1s	1
Public Relations, communication	A public relations strategy is critical now more than ever #PublicRelations https://t.co/KZmGOD2Xjg	0.54	29 s	1	RT: The Relevance Of Public Relations & Communication In Fashion https://t.co/YrOhuv0mwX #fashion #Fashionista #bloggerstr	0.76	0.8s	1
COVID19 immunity transmission	@CoocoLa_Vrej WHO is still not sure if those who recovered from COVID 19 develop a certain immunity that they will not get COVID 19 virus again.	0.57	26 s	1	@CoocoLa_Vrej WHO is still not sure if those who recovered from COVID 19 develop a certain immunity that they will not get COVID 19 virus again.	0.57	1.2s	1
diabetes type 2, intermittent fasting	Can intermittent fasting make you diabetic? Does anyway here do intermittent fasting? How do you do it? Intermittent fasting has proven to help cure Type II diabetes	0.74	28s	3	RT @EvolveHolistic: How to Intermittent Fast and Which Type of Fasting Is Right for You https://t.co/sx9iNQr312 https://t.co/v2QVQVVRcv	0.73	1s	1
Digital marketing, business, social media	RT @V2M2Group: Get Social: The Power of Social Media for Marketing Your Business ? #business #digitalmarketing #marketing #smallbusiness #SocialMedia #GuernseyBusinesses https://t.co/8wnlALHXsh	0.73	29s	1	Learn How to Market Your Business on Social Media – Affiliate or Network Marketing on Social Media https://t.co/iK9SrVzuLM #OnlineBusiness #SocialMedia https://t.co/45waDGLoGh	0.76	1s	1
Bitcoin prices market	Bitcoin price within about 3% of gold price https://t.co/GwjcMSB9Jp	0.67	24 s	1	Bitcoin Average - bitcoin price index - ($ 9638.9) - https://t.co/z6cbnPDdmv #bitcoin https://t.co/0PoQwUAU1a	0.59	0.8s	1
Smart City, 5G, IoT	Samsung IoT Smart City https://t.co/Xnf5JHnOq9 via @YouTube @_funtastic5_ #TelkomFuntastic5 #RWSTREG5 #smartcity	0.70	25s	1	Getting Around Smart Cities #SmartCities via https://t.co/yXaZMpRqm9 https://t.co/2HSDlUh7WN	0.74	0.9s	1
Joe Biden and Bernie Senders	@LyndaMo85130479 @BugOffDear Biden positions are literally just copy/pasted from Bernie Sanders Folks mention Biden's past plagiarism True But who believes Joe had anything to do with deciding this, or preparing the doc? Who is in charge? https://t.co/x8RBCz4H6D	0.46	26 s	3	@JoeBiden has become Bernie Sanders 2.0!!!! @JoeBiden	0.81	1.1s	1
Global warming, climate change	So is global warming	0.67	27s	1	@Ilhan Climate change or global warming?	0.95	0.9s	1
Food security and Agriculture	With food security on the rise, do what you can to help another #foodsecurity #food #endhunger https://t.co/3bLuC6daiN	0.59	26s	1	RT Moreover Food security and Agricultural self-sufficiency #foodsecurity #Agricultural_sufficiency #Yemen	0.82	0.8s	1
Black lives matter and gunshots	#Facebook groups are falling apart over Black Lives Matter posts https://t.co/g0eff0hcLC #Socialmedia https://t.co/R3dov8nCAq	0.59	28s	1	Black Lives STILL Matter, just in case you forgot. And ALL Lives won't matter until Black Lives do. https://t.co/bULE7LLNk7	0.68	1s	1

Table 6: Information Foraging Results: EHOIF vs. EEHOLSIF



(a) Similarity score

(b) surfing depth

(c) Convergence

(d) Response time

Figure. 13: EHOIF vs. EEHOLSIF

tests.

Parameter	Value
α	0.2
β	0.4
ρ	0.8
q_0	0.8
Number of ants	50
Number of generations	50

Table 7: ACSIF empirical parameters setting

The second approach we implemented is based on *Particle Swarm Optimization (PSO)*. To our best knowledge this is the first time PSO is used to address information foraging. Just like with ACS, we carried out several tests to find the optimal values of the empirical parameters, which are presented in Table 8. We denote this approach as PSOIF.

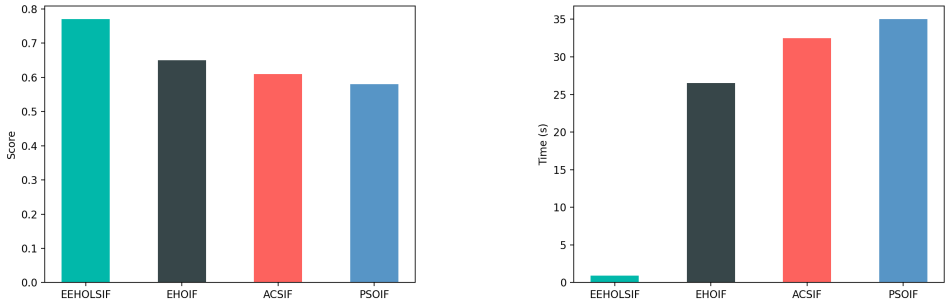
Parameter	Value
c_1	1.5
c_2	0.4
Number of particles	600
Number of generations	90

Table 8: PSOIF empirical parameters setting

Once we fixed the empirical parameters for both approaches, we conducted extensive experiments to make a comparative study between EHOIF, EEHOLSIF, ACSIF and PSOIF. Figure 14a exhibits the difference in relevance sore between the four approaches. We can see that EEHOLSIF achieves the highest score followed by EHOIF then ACSIF and finally PSOIF. We also notice that the difference in score is quite significant between EEHOLSIF and the other approaches. When it comes to response time, Figure 14a shows that EEHOLSIF is the fastest with an average lower than 1 second. The three other approaches have an average response time between 25 and 35 seconds. This means that EEHOLSIF is more than 25 times faster than the other approaches.
We assume that the main problem faced by EHOIF, ACSIF and PSOIF resides in the social graph's size. For instance, ACS is well adapted to work on graphs, since it was first developed to solve the traveling salesman problem. It was also applied to information foraging and gave good results both in terms of score and response time. However, it was only tested on limited size web graphs, which contain less than 2000 web pages [16]. Once we increase the number of the web pages or in our case social posts, the complexity of the problem causes a noticeable slow down in terms of response time and a decrease in the relevance score.

VIII. Conclusion and perspectives

A novel bio-inspired approach to large scale information foraging using enhanced elephant herding optimization and clustering was proposed in this paper. First, we adapted the Elephant Herding Optimization algorithm to information foraging on social media. EHO was originally proposed to solve continuous optimization problems, so to make it able to work on combinatorial problems and more precisely information foraging, we undertook modifications on some important aspects including the elephants' positions implementation, the solution construction and the solution evaluation. To our best knowledge, this is the first attempt to use EHO to address in-



(a) Similarity

(b) Response time

Figure. 14: Comparison with ant colony optimization

formation foraging. The first results show that our approach based on the adaptation of EHO has the ability to find relevant information on social graphs. However, the response time can be very slow especially when the social graph is big.

To overcome this issue and better handle large scale information foraging, we proposed a new enhanced version of EHO. We introduced several new concepts to the algorithm including two natural phenomena related to elephants' behavior, namely territories delimitation and clan migration. Clustering and more precisely k-means algorithm was used for the implementation of the territories delimitation. Thanks to this phase, we were able to define a better representation of the elephants positions taking into account the real distance between the potential solutions on the search space. Furthermore, dividing the search space into multiple small territories and bounding the foraging to just a few of them, using a newly introduced pseudo random proportional rule, helped to substantially reduce the problem complexity. In addition to that, the clan migration phenomenon prevents the algorithm from stagnation and premature convergence.

In order to evaluate the proposed approach, we built a dataset containing more that 1.4 million tweets covering different topics. We conducted extensive experiments to test both the adapted EHO for information foraging and the enhanced EHO for large scale information foraging. The results showcase the advantages of EEHOLSIF compared to EHOIF in diffrent aspects including relevance score, response time, convergence and surfing depth. They also demonstrate that the new concepts introduced in EEHOLSIF contribute in boosting the performance considerably.

Finally, we did a comparative performance analysis with two metaheuristic-based information foraging approaches. The first one being Ant Colony System and the second one being Particle Swarm optimization. The outcomes show that our approach has an much better performance. As far as we know, this is the first time an information foraging approach based on elephant herding optimization is proposed and one of the few destined to work on social media, which gives this work a remarkable originality.

Further work will focus on the parallel implementation of EEHOLSIF using GPUs, as well as a dynamic territories definition process using deep learning for a real time clustering. Another intersting direction would be to integrate a dictionary such as WordNet to better cover the semantic features of the tweets in the vector space representation.

Acknowledgement

This study is funded in part by the General Directorate of Scientific Research and Technological Development (DGRSDT), under Grant No. C0662300.

References

[1] We Are Social, Digital 2021 April Global Statshot Report. 2021. https://wearesocial.com/blog/2021/04/60-percent-of-the-worlds-population-is-now-onli

[2] S. Laato, N. Islam, M. Islam, E. Whelan. "What drives unverified information sharing and cyberchondria during the COVID-19 pandemic?". European Journal of Information Systems, 29:3, pp. 288-305, 2020. https://doi.org/10.1080/0960085X.2020.1770632

[3] C.L. Ventola. "Social media and health care professionals: benefits, risks, and best practices". P&T: a peer-reviewed journal for formulary management. 39(7), pp. 491-520, 2014. PMID: 25083128; PMCID: PMC4103576.

[4] E. Werner, D. Hall. "Optimal foraging and the size selection of prey by the bluegill sunfish (lepomis macrochirus)". Ecology Journal, 55(5), pp. 1042–1052, 1974. https://doi.org/10.2307/1940354

[5] P. Pirolli, S. Card. "Information foraging". In Psychological Review, 106(4), pp. 643–675, 1999. https://doi.org/10.1037/0033-295X.106.4.643

[6] X. Niu, X. Fan. "Deep Learning of Human Information Foraging Behavior with a Search Engine". In Proceedings of the International Conference on Theory of Information Retrieval, Santa Clara, CA, USA. pp. 185-192. ACM, 2019. https://doi.org/10.1145/3341981.3344231

[7] Y. Drias, S. Kechid. "Dynamic Web information foraging using self-interested agents: Application to scientific citations network", Journal of Concurrency and Computation: Practice and Experience, Wiley, 31-(22), 2019. https://doi.org/10.1002/cpe.4342

[8] Y. Drias, S. Kechid, G. Pasi. "Bee Swarm Optimization for Medical Web Information Foraging", Journal of Medical Systems, 40-(2), Springer, 2016. https://doi.org/10.1007/s10916-015-0373-5

[9] V. Nguyen, G. Rabby, V. Svátek, O. Corcho. "Ontologies Supporting Research-related Information Foraging Using Knowledge Graphs: Literature Survey and Holistic Model Mapping". In Proceedings of the 22nd International Conference on Knowledge Engineering and Knowledge Management, EKAW, Bolzano, Italy, Springer, 2020. https://doi.org/10.1007/978-3-030-61244-3_6

[10] Y. Drias, S. Kechid. "Dynamic Web information foraging using self-interested agents". In Recent Advances in Information Systems and Technologies - WorldCIST'17, Springer, 569, pp. 405-415, 2017. https://doi.org/10.1007/978-3-319-56535-4_41

[11] A. Dalton, B. Dorr, L. Liang, K. Hollingshead. "Improving cyber-attack predictions through information foraging". In Proceedings of the International Conference on Big Data, Boston, MA, USA, IEEE Computer Society, pp. 4642-4647, 2017. https://doi.org/10.1109/BigData.2017.8258509

[12] A. Jaiswal, H. Liu, I. Frommholz. "Utilising Information Foraging Theory for User Interaction with Image Query Auto-Completion". In Advances in Information Retrieval, ECIR, Springer, pp. 666-680, 2020. https://doi.org/10.1007/978-3-030-45439-5_44

[13] T. Schnabel, P. Bennett, T. Joachims. "Shaping Feedback Data in Recommender Systems with Interventions Based on Information Foraging Theory". Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (WSDM '19), Association for Computing Machinery, New York, NY, USA, pp. 546–554, 2019. https://doi.org/10.1145/3289600.3290974

[14] Y. Drias, G. Pasi. "Credible Information Foraging on Social Media", Trends and Innovations in Information Systems and Technologies - WorldCIST'20, pp. 415-425, Springer, 2020 https://doi.org/10.1007/978-3-030-45688-7_43

[15] L. Azzopardi, P. Thomas, N. Craswell. "Measuring the Utility of Search Engine Result Pages: An Information Foraging Based Measure". In Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. Association for Computing Machinery, New York, NY, USA, pp. 605–614, 2018. https://doi.org/10.1145/3209978.3210027

[16] Y. Drias, S. Kechid, G. Pasi. "A Novel Framework for Medical Web Information Foraging Using Hybrid ACO and Tabu Search". Journal of Medical Systems. 40-(1), Springer, 2016. https://doi.org/10.1007/s10916-015-0350-z

[17] V. Santucci, M. Baioletti, A. Milani. "An algebraic framework for swarm and evolutionary algorithms in combinatorial optimization". Swarm and Evolutionary Computation Journal, 55, 2020. https://doi.org/10.1016/j.swevo.2020.100673

[18] S. Alirezaa, H. Ashkanb. "A bibliography of metaheuristics-review from 2009 to 2015", International Journal of Knowledge-based and Intelligent Engineering Systems, vol. 22, no. 1, pp. 83-95, 2018. https://doi.org/10.3233/KES-180376

[19] N.I. Anuar, M.H.F. Md Fauadi. "A Study on Multi-Objective Particle Swarm Optimization in Solving Job-Shop Scheduling Problems". International Journal of Computer Information Systems and Industrial Management Applications, Volume 13, pp. 051-061, 2021.

[20] G. Wang, S. Deb, X. Gao, L. Coelho. "A new metaheuristic optimisation algorithm motivated by elephant herding behaviour". International Journal of Bio-Inspired Computing, 8-(6), pp. 394-409, 2016. https://doi.org/10.1504/IJBIC.2016.10002274

[21] W. Li, G. Wang, A. Alavi. "Learning-based elephant herding optimization algorithm for solving numerical optimization problems". Knowledge-Based Systems Journal, Volume 195, 2020. https://doi.org/10.1016/j.knosys.2020.105675.

[22] R.M. Sahoo, S.K. Padhy. "Elephant Herding Optimization for Multiprocessor Task Scheduling in Heterogeneous Environment". Computational Intelligence in Pattern Recognition, pp. 217-229, Springer, 2020. https://doi.org/10.1007/978-981-15-2449-3_18

[23] E. Tuba, D. Dolićanin Dekic, R. Jovanovic, D. Simian, M. Tuba. "Combined Elephant Herding Optimization Algorithm with K-means for Data Clustering". In Proceedings of Information and Communication Technology for Intelligent Systems, Volume 2, 2019. https://doi.org/10.1007/978-981-13-1747-7_65.

[24] S. Muhammad Mohsin, N. Javaid, S.A. Madani, S.M. Akber, S. Manzoor, J. Ahmad. "Implementing Elephant Herding Optimization Algorithm with different Operation Time Intervals for Appliance Scheduling in Smart Grid". In Proceedings of 32nd International Conference on Advanced Information Networking and Applications Workshops (WAINA), pp. 240-249, 2018. https://doi.org/10.1109/WAINA.2018.00093.

[25] J. Li, L. Guo, Y. Li, C. Liu. "Enhancing Elephant Herding Optimization with Novel Individual Updating Strategies for Large-Scale Optimization Problems".

Mathematics Journal, 2227-7390, 7(5), 2019. https://doi.org/10.3390/math7050395

[26] E. Tuba, R. Capor-Hrosik, A. Alihodzic, R. Jovanovic, M. Tuba. "Chaotic elephant herding optimization algorithm". In IEEE 16th World Symposium on Applied Machine Intelligence and Informatics (SAMI), pp. 213-216, 2018. https://doi.org/10.1109/SAMI.2018.8324842.

[27] Facebook F8. "Facebook Unveils Platform for Developers of Social Applications". Facebook F8 Event, May 24, 2007, San Francisco, USA. https://about.fb.com/news/2007/05/facebook-unveils-platform-for-developers-of-social-applications/

[28] R. Budiu, C. Royer, P. Pirolli. "Modeling Information Scent: A Comparison of LSA, PMI and GLSA Similarity Measures on Common Tests and Corpora". In Proceedings of the 8th International Conference on Computer-Assisted Information Retrieval, Pittsburgh, PA, USA. 31-(22), 2007.

[29] R.A. Charif, R.R. Ramey, W.R. Langbauer, K.B. Payne, R.B. Martin, L.M. Brown. "Spatial relationships and matrilineal kinship in African savanna elephant (Loxodonta africana) clans", Behavioral Ecology and Sociobiology Journal, Springer, 57, pp. 327–338, 2005. https://doi.org/10.1007/s00265-004-0867-5

[30] H. Fritz. "Long-term field studies of elephants: understanding the ecology and conservation of a long-lived ecosystem engineer". Journal of Mammalogy, Oxford Academic, 98-(3), pp. 603–611, 2017. https://doi.org/10.1093/jmammal/gyx023

[31] African Elephant. "IUCN, International Union for Conservation of Nature", 2021. https://www.iucn.org/ssc-groups/mammals/african-elephant-specialist-group/faq

[32] H. Wang, C. Zhou, L. Li. "Design and Application of a Text Clustering Algorithm Based on Parallelized K-Means Clustering". Revue d'Intelligence Artificielle. 33-(6), pp. 453-460, 2019. https://doi.org/10.18280/ria.330608

[33] J. Dobsa, D. Mladenic, J. Rupnik, D. Radosevic, I. Magdalenic. "Cross-language Information Retrieval by Reduced k-means". International Journal of Computer Information Systems and Industrial Management Applications, Volume 10, pp. 314-322, 2018.

[34] J. MacQueen. "Some methods for classification and analysis of multivariate observations". In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Lucien M. Le Cam, Jerzy Neyman, 5-(1), pp. 281-297, 1967.

[35] M. Smith, A. Ceni, N. Milic-Fraylin, B. Shneiderman, E. Mendes Rodrigues, J. Leskovec,C. Dunne. "NodeXL: a free and open network overview discovery and exploration add-in for Excel 2007/2010/2013/2016". The Social Media Research Foundation, 2010. https://www.smrfoundation.org

[36] A. Fauzan, G.W. Sasmito, S.K. Dini. "Application of Clustering Algorithm and Spatial Analysis for Industrial Optimization". Proceedings of the 2nd International Seminar on Science and Technology, Advances in Social Science, Education and Humanities Research, volume 474, pp. 165-171, 2019. https://dx.doi.org/10.2991/assehr.k.201010.024

[37] H. Estiri, B. Abounia Omran, S. Murphy. "Kluster: An Efficient Scalable Procedure for Approximating the Number of Clusters in Unsupervised Learning". Big Data Research Journal, 13, pp. 38-51, 2018. https://doi.org/10.1016/j.bdr.2018.05.003

Author Biographies

Yassine Drias received his Ph.D. degree in Computer Science from the University of Milano-Bicocca in 2017. Prior to that, he prepared his Master's degree in Intelligent Informatics Systems at USTHB University, Algeria in collaboration with ENSMA, France and graduated in 2013. He is an Assistant Professor at the University of Algiers. He is currently working on topics including Bio-inspired Computing,Web Information Foraging, Multi-Agent Systems, Machine Learning and Data Mining. His works have appeared in computer science journals and international conferences proceedings.

Habiba Drias received the M.S. degree in computer science from CWRU Cleveland OHIO USA and the Ph.D. degree in Computer Science from USTHB/Paris6, Algiers. She is a full professor at USTHB University and the head of the Laboratory of Research in Artificial Intelligence (LRIA). She has published more than 200 papers in the domain of artificial intelligence, e-commerce, satisfiability problem, multi-agent systems, meta-heuristics and large scale information retrieval and data mining in well recognized international conference proceedings and journals. She has also directed 25 Ph.D. theses, 38 master theses and 31 engineer projects. In 2013, she won the Algerian Scopus award in computer science and in 2015, she was selected by a jury of international academicians as a founding member of the Algerian Academy of Science and Technology (AAST).

Ilyes Khennak received his Ph.D. degree in Computer Science from USTHB University, Algiers in 2017 and the Master's degree in Intelligent Informatics Systems at USTHB University, Algiers in 2011. He is currently an Assistant Professor at USTHB University and his research interests include Metaheuristics, Information Retrieval, Bio-inspired Computing and Data Mining. His works have appeared in computer science journals and international conferences proceedings.