The Journal of Supercomputing (2023) 79:2251–2280
https://doi.org/10.1007/s11227-022-04708-9

Hybrid optimization and ontology-based semantic model for efficient text-based information retrieval

Ram Kumar[1] · **S. C. Sharma**[1]

Accepted: 6 July 2022 / Published online: 10 August 2022

Abstract

Query expansion is an important approach utilized to improve the efficiency of data retrieval tasks. Numerous works are carried out by the researchers to generate fair constructive results; however, they do not provide acceptable results for all kinds of queries particularly phrase and individual queries. The utilization of identical data sources and weighting strategies for expanding such terms are the major cause of this issue which leads the model unable to capture the comprehensive relationship between the query terms. In order to tackle this issue, we developed a novel approach for query expansion technique to analyze the different data sources namely WordNet, Wikipedia, and Text REtrieval Conference. This paper presents an Improved Aquila Optimization-based COOT(IAOCOOT) algorithm for query expansion which retrieves the semantic aspects that match the query term. The semantic heterogeneity associated with document retrieval mainly impacts the relevance matching between the query and the document. The main cause of this issue is that the similarity among the words is not evaluated correctly. To overcome this problem, we are using a Modified Needleman Wunsch algorithm algorithm to deal with the problems of uncertainty, imprecision in the information retrieval process, and semantic ambiguity of indexed terms in both the local and global perspectives. The k most similar word is determined and returned from a candidate set through the top-k words selection technique and it is widely utilized in different tasks. The proposed IAOCOOT model is evaluated using different standard Information Retrieval performance metrics to compute the validity of the proposed work by comparing it with other state-of-art techniques.

Keywords Information retrieval system · Query expansion · Semantic information retrieval · Modified Needleman Wunsch · COOT optimization · Aquila optimization

✉ Ram Kumar
ramcseiitr@gmail.com

[1] Electronics and Computer Discipline, DPT, Indian Institute of Technology, Roorkee, India

1 Introduction

In information transfer, electronic documents are one of the primary sources of information. Every day a large number of Web Pages are created and the increase in information makes it complex to retrieve accurate information. The information retrieval system (IRS) is introduced for solving the challenge. Web search engines are one of the common tools of information retrieval and it is utilized for identifying information resources. The IR is the act of identifying materials, be it video, audio, or text between larger collections for satisfying the information requirements [1]. The IRS is employed for providing the information support to search the information block in the exterior resources which are beneficial for synthesizing the new knowledge. The communication environment role is played by the IRS and also assists in knowledge interaction [2, 3].

The search technologies are used for performing operations like determination of relevance, usefulness, significance, and combination [4]. The ontology is the main general knowledge representation design that is utilized for information retrieval and it indicates the knowledge by the terms of information hierarchies of the processable, understandable, and machine-readable. In the information retrieval space, ontology is significant for promoting the information retrieval function with knowledge management [5]. The ontology-based knowledge representations are essential for autonomous robots. Autonomous robots are intelligent agents which are goal-oriented. The ontology-based knowledge representations are detected in the emerging sector of surgical robotics [6]. The keyword query is one of the most important search designs through its wide applications and simplicity [7]. The mechanisms of query update as well as retrieving the relevant documents are completed by using the search engine. The IR is the two-step mechanism that includes matching and indexing. The indexing phase describes how various documents will be saved and the matching phase explains the relevant document is retrieved by the query. The unstructured formats are documents and the query [8].

To allow easy user access, the information must be represented in an orderly manner based on the user's interest. The user requirements need to be always provided in an ordered manner to be transformed into a query and to be efficiently processed by the information retrieval system. The query translation is normally a set of keywords that represents the user query and contains the information associated with the user's area of interest. The query expansion (QE) is an effective method for enhancing the reliability and performance of the document retrieval. Several suitable queries are given for the users for comparing the original or initial queries by adding many expansion keywords. The information filtering, question answering system, and the question answering system are the various applications employed in the QE technique. The automatic query expansion, the active query expansion, and the manual query expansion are the types of the QE [9].

The IR systems are utilized for retrieving the documents which apply to the user intention with larger information space. The systems computed the similarities among documents and search queries, as well as the retrieved documents,

are designed in descending order. More than three or two words are incorporated in the prevalent search queries. An effective approach to dealing with the tern scarcity is query expansions that are prevalent for the queries of web search [10]. The query representation with the user's problems is overcome by using the QE. The user's query quality is improved by the modification of the query [11]. The retrieval efficiency is enhanced by using the query expansion technique [12, 13].

The semantic query is mainly optimized by query transformation rules where the actual query is converted into a syntactically different but semantically same query. The transformed query mainly provides the same results for every input instance and it achieves a tradeoff between the functional dependencies and integrity constraints. In this paper, we design a semantic query optimization technique that aims to find a semantically related query that results in a more optimal query execution plan. The execution cost associated with a large query set is minimized via the word similarity measurement technique introduced. Every query transformation system has a small subset of user-defined constraints. An efficient Information retrieval system should focus on different challenges such as data heterogeneity, keyword expansion technique that finds the semantically related terms, and identifying the relevant outcomes. This paper mainly presents a novel technique for semantic query expansion which overcomes the data heterogeneity problem and also identifies the semantically relevant terms not restricted to the user's vocabulary. The major contributions of this paper are presented as follows:

- The main aim of this paper is to enhance the performance of the text-based information system via novel artificial intelligence techniques.
- A novel Improved Aquila Optimization-based COOT (IAOCOOT) algorithm is proposed to find the candidate query expansion terms from different external sources such as WordNet, Wikipedia, etc. This model mainly overcomes the syntactic and semantic heterogeneity issues in information retrieval systems.
- The proposed IAOCOOT algorithm mainly offers the flexibility for new information updates.
- A Modified Needleman Wunsch algorithm (MNWA) is employed to identify the similarity of the words in both the global and local perspectives. The local information is obtained by the symmetry must hold property and the global information is obtained by a match closer property.
- The experiments are conducted using the Text REtrieval Conference (TREC)—NIST dataset in terms of different performance metrics such as Mean Reciprocal Rank (MRR), precision (P), mean average precision (MAP), recall (R), F-measure, and Normalized Discounted Cumulative Gain, and P-R curve.

The remaining of this paper is arranged as follows: The literature review is provided in Sect. 2. The proposed methodology is explained in Sect. 3. In Sect. 4, the experimentation results are explained. Lastly, the conclusions are described in Sect. 6.

2 Literature survey

Malik et al. [14] investigated a Query Expansion(QE) framework by developing a hybrid method for biomedical literature retrieval. This method was analyzed to ensure Vocabulary Mismatch (VM). The Word Embeddings (WEs) and Clinical Diagnosis Information (CDI) combinations were developed to find the related biomedical literature. The PubMed, WordNet, and Wikipedia datasets were used in this analysis. The experimentation result revealed that the scheme outperformed a lot of improvement in vocabulary mismatch and precision rate through the integration method. Wang et al. [15] suggested a Pseudo Relevance Feedback (PRF) technique with the combination of semantic matching and relevance matching to develop document feedback quality and reduce the semantic gap. Four Text Retrieval Conference (TREC) datasets are used in this method. Bidirectional Encoder Representations from Transformers (BERT) was used for calculating document and query representation. The experimentation result showed that the scheme reached the best retrieval performance based on the combination matching method.

Jafarzadeh et al. [16] proposed a semantic technique for predicting the performance of a query, and the scheme established 3 post-retrieval predictors such as semantic query drift, semantic distinction, and semantic cohesion. ClueWeb12-B, Robust04, and ClueWeb09-B datasets were used for this approach. The experimentation result revealed that the scheme achieved better performance in query prediction than the existing methods. A query expansion technique was developed by Dahir et al. [17], using the features of DBpedia and the topic modeling tech for the prediction of latent features in semantics that was needed in expansion use. Bose–Einstein statistics (Bo1) distribution method was used for rearranging the query documents of the applicant. The related expansion terms were assigned for these documents using the Latent Dirichlet Allocation (LDA) method. The dataset applied here was TREC AP88–90 datasets. The experimentation result demonstrated that the scheme improved the multi-valued attributes. But, it had a limitation of improper precision and recall value in classification.

The Query reformulation technique was designed by Kaur et al. [18] for retrieving the information in semantics based on the domain-specific ontology. In the music domain, String ontology is applied in the Ontology-Based Semantic Information Retrieval Method (OBSIRM) to improve the web search. The experimentation result showed that the scheme achieved a standard Recall value of 0.7 and a precision value of 1.43 which enhanced the accuracy and relevance rate in OBSIRM in comparison with the Google search engine. Semantic indexing based on Medical Subject Headings (MeSH) technique was introduced by Kammoun et al. [19] which involved the improvement of biomedical information. The proper sense was allocated for ambiguous words by using Corpus-based word sense disambiguation (WSD) techniques through a machine learning algorithm. The experimentation result revealed that the scheme reached greater significance in Information Retrieval (IR) process which involved biomedical concepts with sufficient sense.

Esposito et al. [20] established a hybrid query expansion (HQE) approach based on word embeddings and lexical resources for the question answering

system. The question answering system was predicted from the information retrieval process and the questions were obtained from the MultiWordNet, then contexts were given to the document collection. The result was performed in the Word2Vec model and the parameters like accuracy and MRR were used to obtain good performance. Selvalakshmi et al. [21] proposed a new ontology-based semantic information retrieval (NOSIR) approach by using feature selection and classification. This information retrieval method was mainly used for fast information retrieval of big data. To enhance the feature selection and classification scores, the new fuzzy rough set-based feature selection, and latent Dirichlet Allocation-based semantic information retrieval algorithms were used. The parameters like relevancy accuracy and relevancy score were used to predict the outstanding performance. The agent-based communication did not support fast information retrieval.

3 Proposed model

Query expansion is an information retrieval process that selects and adds the relevant terms to the user query. Initially, the user submits the actual query to the search engine. In our proposed model, the query terms are similar to the documents present in the TREC dataset. Initially, the user query is preprocessed and the documents are ranked. The in-links and out-links in the text are also extracted from the texts during this process. The similarity between the documents is identified via the Modified Needleman Wunsch algorithm (MNWA) algorithm which follows a pseudo-relevance feedback mechanism. It mainly performs top-k ranking to identify the relevant documents. The documents which do not make it to the top-k are discarded. The main aim of the IAOCOOT MODEL is to minimize the query and relevant information mismatch by addressing the syntactic heterogeneity problem. In this way, the IAOCOOT model improves information retrieval performance. The IAOCOOT algorithm mainly expands the user queries by processing the information yielded from the previous steps. At last, the expanded query is utilized to return the new results that accurately match the user's needs and the documents are re-ranked once again and the information is provided to the user. Figure 1 presents the overall architecture of the proposed model.

3.1 Preprocessing of the initial query

In this step, each query is merged by using Brill's tagger and in the query for each word different part of speech (POS) is assigned. Individual words and phrases in queries are identified by POS information and also POS tagging is performed in the queries. Whereas, these individual words and phrases are utilized in the following steps of QE. Normally, phrases provide lesser uncertainty and better context. For an individual word, the phrases have a particular meaning other than the cumulative meaning thereby in the query higher priority is for phrases than individual words. To remove phrases verbs, nouns and adjectives are considered. An appropriate phrase

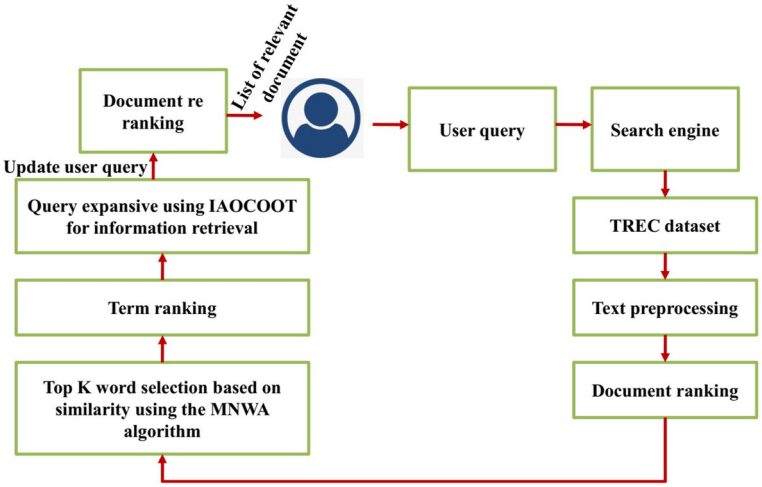

Fig. 1 Overall architecture of the proposed model

should have a cardinal number with two or more verbs, nouns, or adjectives. More result shows that better results are obtained using queries rather than term-to-term relationship.

3.2 Extraction of in-links and out-links

Computation of term frequency and extraction of in-links are two sub-steps involved. Titles of all Wikipedia articles are comprised in in-links. Computation of term frequency of an initial query is considered as the term frequency of initial query and its meaning is achieved from WordNet. For example, consider the initial query as "cat" and "tail" is its in-links then the "cat" in the article, the term frequency of "cat" is "tail" and the meaning for the word "tail" is collected from WordNetThe query in Wikipedia has a hyperlink and these hyperlinks are removed thereby the out-link of a query is obtained. For example, consider the initial query as "cat" so the hyperlink in the body of the article "cat" is derived as out-links.

3.3 Modified Needleman Wunsch algorithm (MNWA) for top-k word selection based on similarity

The most frequently used approach is the knowledge base for top-k selection. For instance, the word lexical database, WordNet collects nouns, adjectives, verbs, and adverbs into synonym sets known as synsets. All those synsets are interconnected by means of semantic as well as lexical relationships namely meronym, hypernym, hyponym, etc. Some approaches rely on edge-based counting strategies to compute the similarity distance among each word using WordNet [20] or other databases. If the distance is longer, the similarity between the words will be less. Unlike these

approaches, some measures the word similarities on the basis of common features present in the knowledge base namely definitions, relationships, synonyms, etc. Despite many benefits in the utilization of the knowledge base approach, there arises the possibility of performance degradation by several techniques [12]. As an illustration, the common features generally necessities manual annotation, when a single word has different meanings used in different fields, and changing the meaning of words over time are the reasons for the diminishing performance of the approach. In this section, to overcome the above drawback we are using a Needleman–Wunsch algorithm [22, 23]. To perform these modifications, Needleman–Wunsch metrics are utilized as it changes the calculation more effectively. This modified metric aimed to apply in clustering process data and the three fundamental properties (a) symmetry must hold (b) a point obtains itself 0 (c) match closer is signified by smaller value should be preserved.

3.3.1 Variables

The first change required is not all the activities are permitted to swap each other. Because contact with different care centers and pathways from a resource planning perspective makes all the activities swap each other to form various pathways considered to be the same. This is carried out by explaining the no-swap variable. Also, the algorithm should figure out the permitted activities which are eligible to swap with each other thereby group of activities is introduced and they are swapped if assigned in the same group.

3.3.2 Weightings

In this algorithm, adding weightings makes it more complex and difficult to evaluate by hand. The way to designate the weighting in the activities and merge them with the algorithm is described. Initially, the activities in the domain experts are arranged from higher to least important. The activities which occur repeatedly are considered as more important and rank 0 is assigned whereas the activities which is not much occurred is considered as low important. Based on the ranking, they are switched to weighting where the least ranked activity is given a weight of 1 thereby for each activity $1/(N-1)$ of increment is given, where the number of activities is denoted as N. Also, a weight of 2 is assigned to the higher important activities.

3.3.3 Groupings

The group activities that occur at some point in the pathway are asked to assign to the same group. Because, at a similar point in the pathway, if two different patients play different activities it would be better to assign them to one group than a separate group as they would be viewed as similar to each other. And this will enable better meaning to the pathway where the values don't look unique. For different case studies, grouping is tabulated in the Table 1.

Table 1 Various grouping for each activity

Different group	Activities
0	g,h,i,j
1	c,q
2	m,k,e
3	a,b,l
4	r,s
5	d
6	f,m,o
7	p,n,t

Both weighting and grouping are merged together with the algorithm. The match equation is expressed as in Eq. (1), with the multiplication of p parameter the match equation is improved and the initial 0 is allowed to propagate.

$$M = Y[k-1]\big[q-1\big] * \left(p + \frac{1}{Y[k-1]\big[q-1\big] + t_k}\right) \tag{1}$$

The previous matrix value is need to be added to the denominator to manage the magnitude and to ensure that the match value shouldn't reach more than 1. A match is a positive event and also the higher important activity has low impact than the lesser important activity. The swapped and no-swap match equation is written as,

$$M = Y[k-1]\big[q-1\big] + c + \mathrm{ABS}\big(t_k - t_{\mathrm{q}}\big) \tag{2}$$

$$M = Y[k-1]\big[q-1\big] + \mathrm{nc} + \big(t_k - t_{\mathrm{q}}\big) \tag{3}$$

This equation shows that the no-swap match equation value is higher thereby this is not chosen in the matrix. Gap equations are modified by adding particular weighting of that direction and are given in Eqs. (4) and (5). Finally, in the Needleman–Wunsch algorithm, the value of M, G, and H are selected as a minimum.

$$H = Y[k-1]\big[q\big] + e + t_k \tag{4}$$

$$G = Y[k]\big[q-1\big] + e + t_{\mathrm{q}} \tag{5}$$

3.3.4 Penalty value

Based on the literature around the Needleman–Wunsch algorithm, the user has the ability to determine the appropriate value for swap, match, and gap penalty and they have no surrounding guidelines. While choosing the variables, the following equation is created to ensure the condition MATCH < SWAP < GAP < NO-SWAP.

$$1 < e \tag{6}$$

$$1 < c < g \tag{7}$$

$$\mathrm{nc} = 2e + 1 \tag{8}$$

$$p = 1 \tag{9}$$

The value of p is set to be 1 where the match equation performs only multiplication and it is not required that the nc value should be higher than $2e+1$ thereby no swap is needed. The minimal possible values are $p=1$, $e=2$, $c=2$, $\mathrm{nc}=5$. As the penalty values are modified, the Needleman–Wunsch algorithm obtained different distances. The MNWA has certain features and they are listed below: The unpredictable distance between two pathways, point to itself is 0, minimum values are achieved in the swap of activities that are closer in ranking, matches among higher importance activities create minimum distance, for string until the first non-match, the distance score is 0, the gap in lower importance activities is smaller than the higher importance activities, and already occurred match in the string is higher than the earlier match. In the original query, the additional term in $Q_{\mathrm{u}}{}^{+}$ is ranked by the MNWA algorithm. The mathematical expression for query sorted expansion is given in Eq. (10). The relevance ranking is denoted as x, expansion word is represented as d_x and S_x denotes the similarity to query.

$$M_{\mathrm{SORTED}}^{(\mathrm{Q})} = \left\{ \left(d_1, S_1\right), ..., \left(d_x, S_x\right), ..., \left(d_y, S_y\right) \right\} \tag{10}$$

where,

$$S_x = \cos\left(p_{d_x}, \mathrm{pw}\right) \tag{11}$$

The reweighting function is expressed as in Eq. (12). Here,x shows the ranking, size of an additional term is denoted as y and the average of the term's similarity is represented as S_{AVG}.

$$F(x) = \left(\left[y/2 \right] - x \right) * \left| S_x - S_{\mathrm{AVG}} \right| + S_{\mathrm{AVG}} \tag{12}$$

3.4 Word embedding

The word embedding approach exemplifies all the words w in vocabulary v as a D-dimensional vector $\omega \in \Re^{\delta}$. The word embedding approach using Word2Vec displays a robust baseline. In order to generate high quality word embeddings, the Word2Vec approach utilizes high computation efficient log-linear model. In this, the sliding window passes over the text corpus where the central word is considered as a target word; meanwhile, the context is formed by the remaining words. The Word2Vec approach incorporates two models namely the Continuous Bag of Words Model (CBOW) and skip-gram. These models minimize computational issues

considerably and increase the ability of model to learn word embeddings from a large-scale dataset. In this, the CBOW approach utilizes mean of contextual words as input to determine the targeted word while the Skip gram utilizes target word as an input to determine the context words. Assume, the target word as w_{p}, context window as $\left\{w_{\mathrm{p-r}}, ..., w_{\mathrm{p-1}}, w_{\mathrm{p}}, w_{\mathrm{p+1}}, ..., w_{\mathrm{p+r}}\right\}$ and the remaining are context words. The below expression is formulated to determine the target word by means of the CBOW model.

$$\Gamma_{\mathrm{CBOW}} = \mathrm{Arg}_{\mathrm{MAX}} \sum_{\mathrm{p}=1} \log \wp\left(\omega_{\mathrm{p}} \middle| \omega_{\mathrm{context\ t}}\right) \tag{13}$$

The term ω_{p} indicates vector of target word and $\omega_{\mathrm{context}}$ signifies the mean vector obtained for each context word. The mathematical representation of probability with a softmax function is defined as,

$$\wp\left(\omega_{\mathrm{p}} \middle| \omega_{\mathrm{context}}\right) = \frac{\exp\left(\omega_{\mathrm{p}} \cdot \omega_{\mathrm{context}}\right)}{\sum_{w \in V} \exp\left(\omega_{\mathrm{p}} \cdot \omega_{\mathrm{context}}\right)} \tag{14}$$

From the above equation, V implies vocabulary. On contrary to CBOW, the skip gram model determines all the context words as an input to target word. The main goal of this model is to enhance the log probability which is numerically defined as follows.

$$\Gamma_{\mathrm{Skip\_Gram}} = \mathrm{Arg}_{\mathrm{MAX}} \sum_{\mathrm{p}=1} \sum_{-\mathrm{r} \le \mathrm{C} \le \mathrm{r},\, \mathrm{C} \ne 0} \log \wp\left(\omega_{\mathrm{p+C}} \middle| \omega_{\mathrm{p}}\right) \tag{15}$$

In addition, the probability is represented with a softmax function for skip gram model is given by,

$$\wp\left(\omega_{\mathrm{p+C}} \middle| \omega_{\mathrm{p}}\right) = \frac{\exp\left(\omega_{\mathrm{p+C}} \cdot \omega_{\mathrm{p}}\right)}{\sum_{\mathrm{w} \in \mathrm{V}} \exp\left(\omega_{\mathrm{p+C}} \cdot \omega_{\mathrm{p}}\right)} \tag{16}$$

The objective function of these models is optimized by means of a negative sampling strategy. Subsequent to the optimization process, all the vocabulary words are mapped to less dimensional real-valued vectors. Finally, the word similarity is measured using the vector similarities in word embeddings based on the metrics such as Euclidean distance or cosine similarity.

3.5 Hybrid IAOCOOT formation

This section demonstrates the integration of the Aquila optimization and COOT algorithm via a.

3.5.1 Aquila optimization (AO) algorithm

AO is a metaheuristic optimization approach, inspired by four kinds of hunting activity of Aquila bird namely high soar with vertical stoop (expanded exploration), contour flight with sort glide attack (narrowed exploration), low flight with

slow descent attack (expanded exploitation) and walking and grabbing prey (narrowed exploitation) [24]. It has the ability to swift the attacking strategy based on the type of prey. The two main phases of this algorithm are exploration and exploitation. The step-by-step mathematical formulation of this AO algorithm is described as follows.

- Exploration phase

In the expanded exploration process, the Aquila explores the search dimension broadly in search of prey by flying very high over the ground surface. The Aquila dives speedily toward the prey at once when it finds the exact location of prey on the ground. This hunting activity is numerically illustrated as,

$$Z_1(x+1) = Z_{\mathrm{Bst}}(x) \times \left(1 - \frac{x}{S}\right) + \left(Z_{\mathrm{p}}(x) - Z_{\mathrm{Bst}}(x) \times R_1\right) \tag{17}$$

$$Z_{\mathrm{p}}(x) = \frac{1}{n} \sum_{\mathrm{t}=1}^{\mathrm{n}} Z_{\mathrm{t}}(x) \tag{18}$$

From the above equation, $Z_{\mathrm{Bst}}(x)$ signifies the best location on the search space, $Z_{\mathrm{p}}(x)$ implies mean location of Aquila in the present iteration x, S denotes total number of iterations, n depicts population size and the random integer R_1 lies within the range [0, 1]. In narrowed exploration process, the Aquila utilizes a short glide attacking strategy to catch the prey. It descends the flying movement and flies around to attack the prey. This kind of strategy is commonly used by the Aquila and is represented as,

$$Z_2(x+1) = Z_{\mathrm{Bst}}(x) \times \mathrm{Levy}_{\mathrm{f}}(d) + Z_{\mathrm{r}}(x) + (b-a) \times R_2 \tag{19}$$

Here, $Z_{\mathrm{r}}(x)$ signifies random location of Aquila, d denotes search dimension, $\mathrm{Levy}_{\mathrm{f}}$ represents levy flight function and R_2 indicates random value which lies between 0 and 1. The terms a and b depicts spiral space exploration and are measured as follows.

$$a = R \times \sin\vartheta;\ b = R \times \cos\vartheta \tag{20}$$

where, $R = R_3 + 0.00565 \times d_1$ and $\vartheta = -\varpi \times d_1 + {}^{3\pi}/_{2}$. In this, the terms R_3, d_1 and ϖ signifies total number of search cycles with range [1, 12], an integer lies within [1, d] and constant value (0.005), respectively. Moreover, the levy flight function is measured based on the below expression.

$$\mathrm{Levy}_{\mathrm{f}}(d) = c \times \frac{p \times \mu}{|q|^{\frac{1}{\alpha}}} \tag{21}$$

The terms p and q indicates random numbers and are set as 0 and 1; c and α signifies constants and are set as 0.001 and 1.5, respectively, and the term μ is computed as,

$$\mu = \left(\frac{\Gamma(1+\alpha) \times \sin\left(\frac{\pi\alpha}{2}\right)}{\Gamma\left(\frac{1+\alpha}{2}\right) \times \alpha \times 2^{\left(\frac{\alpha-1}{2}\right)}} \right) \tag{22}$$

- Exploitation phase

In expanded exploitation, the Aquila determines the location of prey roughly and descends towards prey position to start a preliminary attack. Aquila slows its speed and moves down to catch the prey which is mathematically described as,

$$Z_3(x+1) = \left(Z_{\text{Bst}}(x) - Z_{\text{p}}(x) \times \beta - R_4 + \left((ub - lb) \times R_5 + lb\right) \times \eta\right) \tag{23}$$

From the above equation, the exploitation adjustment parameters are represented as β and η, in which their values are fixed as 0.1; ub and lb depicts the upper bound and lower bound values of optimization problem; R_4 and R_5 represents random numbers and are fixed as [0, 1]. In narrowed exploitation process, Aquila tracks and attacks the prey on the ground surface. It is numerically expressed as,

$$Z_4(x+1) = \text{Quality}_{\text{f}} \times Z_{\text{Bst}}(x) - \left(m_1 \times Z(x) \times R_6\right) - m_2 \times \text{Levy}_{\text{f}}(d) + R_7 \times m_1 \tag{24}$$

$$\text{Quality}_{\text{f}} = x^{\frac{2 \times \Re() - 1}{(1-S)^2}} \tag{25}$$

From the above two equations, $\text{Quality}_{\text{f}}$ depicts quality function value which helps to control the search individuals;$Z(x)$ represents current location of Aquila; $m_1 = 2 \times R_8 - 1$ and $m_2 = 2 \times (1 - {}^{x}/_{S})$ indicates movement regulating parameter and flight slope of Aquila that are ranged between [− 1,1] and reduces straightly from 2 to 0, respectively; R_6, R_7 and R_8 denotes random integers lies in the range [0, 1]. In addition to this, the AO algorithm transforms smoothly from exploration to exploitation phase using diverse attacking strategies; according to the condition $x \leq {}^{2}/_{3} * S$, the exploration search is carried out, or else exploitation is performed.

3.5.2 Coot optimization algorithms

The little aquatic bird belongs to the rail family known as Coots. It belongs to the genus Fulica, which means “coot” in Latin [25]. Extensive research on habitat, migratory, and breeding behavior has been published in the literature on American coot behavior. On the water, coots exhibit a variety of movements and behaviors. To develop an optimization approach coot bird’s activity on the water’s surface is utilized.

3.5.2.1 Mathematical model and algorithm The process begins with an initial random population $\left(\vec{y}\right) = \left\{\vec{y}_1, \vec{y}_2, \ldots \vec{y}_{\text{m}}\right\}$. The target function evaluates the random population and it is done repeatedly and $\left(\vec{Q}\right) = \left\{Q_1, Q_2, \ldots Q_{\text{m}}\right\}$ is the target value.

From the optimization method's basis, it is enhanced by a set of rules. Utilizing the below equation the population produced at random is visible space.

$$C_{\text{oot}}P_{\text{os}}(j) = \Re(1, c) \cdot {}^{*}(\text{upperbound} - \text{lowerbound}) + \text{lowerbound} \tag{26}$$

Coot position is represented by $C_{\text{oot}}P_{\text{os}}(j)$, the problem dimensions/number of variables is denoted by c, search space's lower bound is represented as lowerbound, and search space upper bound is represented as upperbound. A lower bound and upper bound problem exists for each variable.

$$\text{lowerbound} = \left[\text{lowerbound}_1, \text{lowerbound}_2, ..., \text{lowerbound}_{\text{c}}\right], \tag{27}$$

$$\text{upperbound} = \left[\text{upperbound}_1, \text{upperbound}_2, ..., \text{upperbound}_{\text{c}}\right] \tag{28}$$

The starting population and the determination of all agent's positions after generating, and all solution's fitness is determined utilizing the objective function $Q_j = e(\vec{y})$.

3.5.2.2 The random movement to this side and that side A random position from the below equation in the search space and move the coot to this random point in the search space.

$$O = \Re(1, c) \cdot {}^{*}(\text{upperbound} - \text{lowerbound}) + \text{lowerbound} \tag{29}$$

The Coot movement investigates search space. When the algorithm becomes trapped in the local optimal, then to escape from the local optimal the movement will cause the algorithm. The position of the coot is derived using the below equation.

$$C_{\text{oot}}P_{\text{os}}(j) = C_{\text{oot}}P_{\text{os}}(j) + B \times \text{ran2} \times \left(O - C_{\text{oot}}P_{\text{os}}(j)\right) \tag{30}$$

Range from 0 to 1 the random number is denoted as ran2, utilizing the below equation B is evaluated.

$$B = 1 - K \times \left(\frac{1}{\text{iteration}}\right) \tag{31}$$

The current iteration is represented by K, maximum iteration is denoted as iteration.

3.5.2.3 Chain movement Chain movement is implemented by taking the two coot's average positions. To evaluate the vector's distance between the two coots first and after that by half the vector's distance, the coot is moved towards the other coot for implementing a chain movement another method is utilized. In the below equation evaluated the first method and coot's new position is given

$$C_{\text{oot}}P_{\text{os}}(j) = 0.5 \times \left(C_{\text{oot}}P_{\text{os}}(j-1) + C_{\text{oot}}P_{\text{os}}(j)\right) \tag{32}$$

The second coot positon is represented as $C_{\text{oot}}P_{\text{os}}(j-1)$.

3.5.2.4 Adjusting the position based on the group leaders A few coots on the group's front lead the group, and the remainder of the coots must change their positions and migrate toward the group's leaders. The leader's average position is considered and based on the average position the coot is updated. Premature convergence is caused by the average position. To select the leader a mechanism is implemented utilizing the below equation.

$$L = 1 + \left(j \,\mathrm{MOD}\, N_{\mathrm{L}}\right) \tag{33}$$

Current coot's index number is represented as j, the leader's number is denoted by N_{L} and the index number of the leader is denoted as L.

$$C_{\mathrm{oot}}P_{\mathrm{os}}(j) = \mathrm{Leader}_{P_{\mathrm{os}}}(l)| + 2 \times \Re 1 \times \cos\left(2\Re\pi\right) \times \left(\mathrm{Leader}_{P_{\mathrm{os}}}(l) - C_{\mathrm{oot}}P_{\mathrm{os}}(j)\right) \tag{34}$$

The coot's current position is denoted by $C_{\mathrm{oot}}P_{\mathrm{os}}(j)$, the selected position of the leader is represented by $\mathrm{Leader}_{P_{\mathrm{os}}}(l)$, random number ranges from 0 to 1 is denoted as $\Re 1$ 3.14 pie value is denoted as π and random number ranges from -1 to 1 id denoted as $\Re$.

3.5.2.5 Leading the group by the leaders towards the optimal area (leader movement) Leaders must update their position toward the target in order to direct the group toward a goal. The current optimal point for better positions the formula is given below. To find the best position from the current optimal position sometimes leaders have to go away. The below-given formula provides the best way optimal location to get closer and away from it.

$$\mathrm{Leader}_{P_{\mathrm{os}}}(j) = \left\{A \times \Re 3 \times \cos\left(2\Re\pi\right) \times \left(h_{\mathrm{best}} - \mathrm{Leader}_{P_{\mathrm{OS}}}(J)\right) + h_{\mathrm{best}} \quad \Re 4 < 0.5\right\} \tag{35}$$

$$= \left\{A \times \Re 3 \times \cos\left(2\Re\pi\right) \times \left(h_{\mathrm{best}} - \mathrm{Leader}_{P_{\mathrm{os}}}(j)\right) - h_{\mathrm{best}} \quad \Re 4 >= 0.5\right\} \tag{36}$$

The best position ever found is denoted as h_{best} the random number ranges from 0 to 1 it is denoted by $\Re 3$ and $\Re 4$, A is evaluated according to the below equation

$$A = 2 - K \times \left(\frac{1}{\mathrm{iteration}}\right) \tag{37}$$

$2 \times \Re 3$ larger random motions are made to keep the algorithm from becoming attentive to the local optimum. Throughout the exploitation, phase is also conducting exploration. $\cos\left(2\Re\pi\right)$ searches with different radiuses around the best search agent to discover a better place around this search agent.

3.5.3 Random opposition based learning (ROBL)

In order to enhance the convergence speed of the metaheuristic algorithms, the opposition based learning (OBL) strategy is utilized. It estimates the fitness of a search agent and its relevant opposite measurement is computed to achieve an

optimal candidate solution. In this work, the traditional OBL strategy is enhanced to prevent the solutions from local optima that is termed as ROBL strategy and is mathematically described as,

$$\widehat{y}_i = ib_i + ub_i - \Re \times y_i, i = 1, 2, 3, \ldots, N \tag{38}$$

The term $\widehat{y}_i$ indicates opposite solution; lb_i and ub_i depicts lower bound and upper bound values of ith dimension, respectively. The opposite solution achieved by the above equation is more random compared to traditional OBL and it efficiently prevents the algorithm from local optimum issues.

3.5.4 Proposed IAOCOOT algorithm

The exploring phase of the AO algorithm is inspired by the hunting activity of Aquila that catches the fast-moving prey from the broad search space. This exploring capability makes AO have better global search as well as faster convergence. But, the chosen search dimension is not completely explored through exploitation. Moreover, the impact of the levy flight function is fairly low and it causes premature convergence. In short, the AO algorithm is best suited to perform exploration that possesses fast convergence and good randomness while it is exposed to local optima when performing exploitation. On the contrary, the COOT algorithm better performs exploitation search while the exploration process is unsatisfactory. Therefore, to enhance the searching ability of the algorithm, the exploration phase of the AO and exploitation phase of the COOT algorithm are integrated to form a new algorithm called the improved AO-based COOT (IAOCOOT) algorithm. It retains the algorithm with better searching capability, and faster convergence speed, and prevents local optima. Furthermore, the adoption of the ROBL strategy in the exploitation phase increases the capability of the algorithm to jump out from local optimum problems. The implementation of all these mechanisms improves the overall performance of the optimization approach effectively. Figure 2 presents the formation of the IAOCOOT algorithm. The description of the different symbols used in the algorithm is presented in Table 2.

3.6 Query expansion using IAOCOOT

This section provides a brief description of the query expansion process via IAOCOOT and the different steps involved.

3.6.1 Candidate expansion terms extraction and contextualization

This process is designed to find the set of optimal expansion terms for all terms $Q_{\mathrm{u}}^{\mathrm{Ex}} \in q^{\mathrm{Ex}}$ and contextualize the text in terms of document collection D_{C}. More particularly, consider the set of vocabulary words v_{n} incorporated in n, the function Ex_{n} represents that it returns the set $\mathrm{T}_{\mathrm{Qu}} = \left\{t_1^{\mathrm{u}}, t_2^{\mathrm{u}}, .., t_{\mathrm{Nu}}^{\mathrm{u}}\right\}$ containing N_{u} terms in the vocabulary. The following criteria are made to operate this function. Initially, the set $\delta^{\mathrm{u}} := \left\{\delta_1^{\mathrm{u}}, \delta_2^{\mathrm{u}}, .., \delta_{\mathrm{Lu}}^{\mathrm{u}}\right\}$ comprising Lu synsets corresponding to each

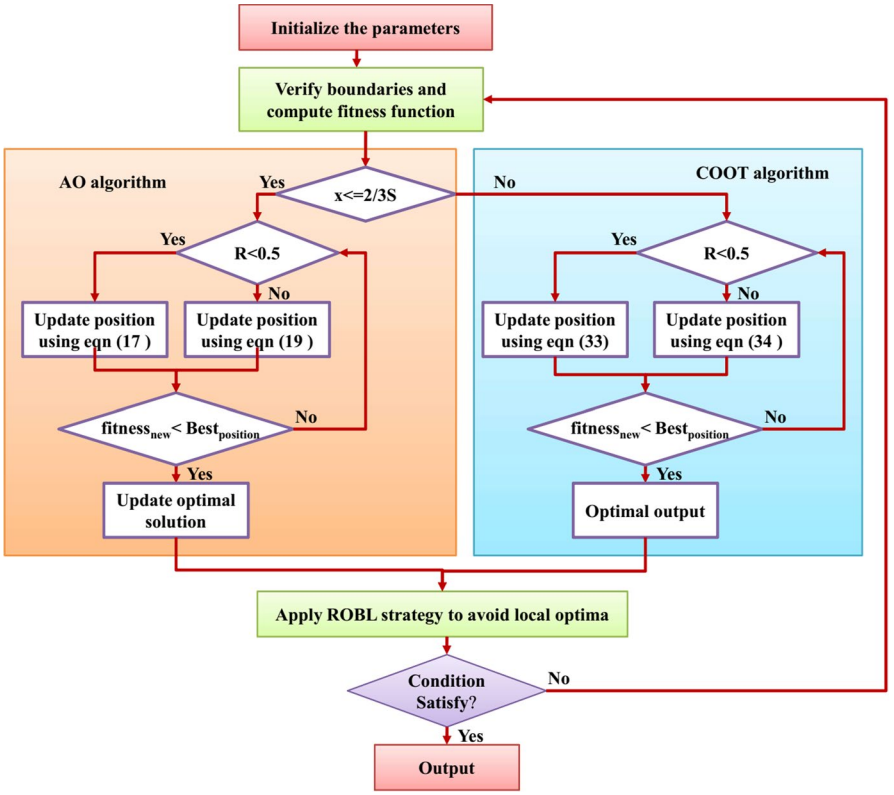

Fig. 2 Flowchart of the IAOCOOT algorithm

word is extracted from all the terms Q_u^{Ex}, in which the synset $\delta_v^u := \{\delta_1^u, \delta_2^u, .., \delta_{wv}^u\}$ includes w_v synonyms δ_v^u. After that, the synset h_v^u is straightly connected to the synset δ_v^u are extracted using hypernym relation. Due to this process, it generates the set $h^u := \{h_1^u, h_2^u, \ldots, h_{Lu}^u\}$ in which the set $h_v^u := \{h_1, h_2, \ldots, h_{pv}\}$ is comprised of p_v hypernyms $h_v^u \in v_n$; where the value of p_v differs from one synset to the other. At last, the function provides the set $T_{Qu}^{Ex} := \{t_1^u, t_2^u, \ldots, t_{Nu}^u\} := \{\delta_1^u, \delta_2^u, \ldots, \delta_{Lu}^u\} \cup \{h_1^u, \delta_2^u, \ldots, \delta_{Lu}^u\} := \{\delta_1^u, \delta_2^u, \ldots, \delta_{w1}^u\} \cup \{\delta_1^u, \delta_2^u, \ldots, \delta_{w2}^u\} \cup \ldots \cup \{\delta_1^u, \delta_2^u, , \delta_{wLu}^u\} \cup \{h_1^u, h_2^u, \ldots, h_{p1}^u\} \cup \{h_1^u, h_2^u, \ldots, h_{p2}^u\} \ldots \cup \ldots \cup \{h_1^u, h_2^u, \ldots, h_{pLu}^u\}$ with $N_u \leq w_1 + w_2 + \ldots + w_{Lu} + p_1 + p_2 + .. + p_{Lu} \; \forall \; Q_u^{Ex} \in q^{Ex}$; thereafter, the duplicate terms which belong to two or more synsets are eliminated. To better demonstrate the choice of MultiWordNet and also the model setup made initially has been utilized to choose optimal expansion terms. To achieve its goal, the similarity terms from the set $\delta^u := \{\delta_1^u, \delta_2^u, .., \delta_w^u\}$ derived using the MNWA algorithm are extracted. Here, the similarity measurement is carried out by estimating the penalty value between Q_u^{Ex} and the similar terms in the vector dimension. Subsequent to the determination of set of optimal expansion terms for all query terms

Table 2 Symbol description

Symbol	Description
x	Ranking
y	Additional term size
S_{AVG}	Average value of the term similarity
d_{x}	Expansion word
S_{x}	Query similarity
w_{p}	Target word
$\omega_{\mathrm{context}}$	Mean vector obtained for each context word
V	Vocabulary
$Z_{\mathrm{Bst}}(x)$	Best location on the search space
$Z_{\mathrm{p}}(x)$	Mean location of Aquila in the current iteration x
S	Total number of iterations
n	Population size
R_1,R_2,R_4 and R_5	Random integer in the range [0, 1]
$Z_{\mathrm{r}}(x)$	Random location of Aquila
d	Dimensionality
$\mathrm{Levy}_{\mathrm{f}}$	Levy flight function
R_3	Total number of search cycles
d_1	Integer that lies within [1, d]
ϖ	Constant value (0.005)
p,q	Random numbers that are set as 0 and 1, respectively
c	Constant with a value of 0.001
α	Constant with a value of 1.5
ub and lb	The upper and lower bound values
$\mathrm{Quality}_{\mathrm{f}}$	Quality function value
$C_{\mathrm{oot}}P_{\mathrm{os}}(j)$	Coot position
$\mathrm{Leader}_{P_{\mathrm{os}}}(l)$	Selected position of the leader

$Q_{\mathrm{u}}^{\mathrm{Ex}} \in q^{\mathrm{Ex}}$, it is further minimized to prevent the terms that appear in the document collection. To provide real-time responses to all the queries, the initial retrieval approach on the basis of pseudo-relevant feedback is considered rather than considering the whole vocabulary terms.

3.6.2 Candidate expansion terms ranking and filtering

This step aims to select the most appropriate terms by reducing the mismatching query documents. Let us consider the question term $Q_{\mathrm{u}}^{\mathrm{Ex}} \in q^{\mathrm{Ex}}$ and the expansion term set as $\mathrm{T}_{\mathrm{Q}}^{\mathrm{CEx}}u := \left\{ t_1^{\mathrm{u}}, t_2^{\mathrm{u}}, .., t_{\mathrm{pu}}^{\mathrm{u}} \right\}$ which is obtained from the earlier step. A rank function is assigned to the terms based on their appropriateness with respect to semantic utilization. It is numerically formulated as,

$$\forall Q_{\mathrm{u}}^{\mathrm{Ex}} \in q^{\mathrm{Ex}}, \mathrm{T}_{\mathrm{Qu}}^{\mathrm{REx}} := \left\{ \left(t_{\mathrm{v}}^{\mathrm{u}}, app_{\text{terms all}} \left(t_{\mathrm{v}}^{\mathrm{u}}, Q_{\wp} \right) \right) : t_{\mathrm{v}}^{\mathrm{u}} \in \mathrm{T}_{\mathrm{Q}}^{\mathrm{CEx}} u \right\} \tag{39}$$

From the above equation, both the terms $\mathrm{app}_{\mathrm{terms}}$ and $\mathrm{app}_{\mathrm{all}}$ indicates two diverse modalities in which they exploit vector magnitudes and their angles in order to estimate the appropriateness score to allocate for all the candidate expansion terms. The pre-processed questionnaires Q_{p} as well as the expansion terms $\mathrm{T}_{\mathrm{Q}}^{\mathrm{CEx}} u$ of the question term $Q_{\mathrm{u}}^{\mathrm{Ex}} \in q^{\mathrm{Ex}}$ are described in the vector dimension constructed in the model. The output of the rank function is allowed to pass through the filter function that reduces the term list by adopting a selection approach. In this regard, we utilize a fundamental selection mechanism that selects the top candidate terms from the whole ordered list corresponding to each question term.

3.6.3 Query expansion (QE) in wikipedia

When the pre-processing is completed in the initial query, the phrases and individual words are considered as keywords for expanding the initial queries by using Wikipedia. The in-links, out-links, and Wikipedia titles are needed for selecting the candidate expansion terms (CET).

3.6.3.1 Wikipedia representation Wikipedia is defined as the perfect information source which is used for query expansion and it is represented by a directed graph $H(B, M)$. Here, M,B depicts links and articles, respectively. Each and every article has a short view of the entire paper and gives the links to other users for browsing other relevant articles. In this section, two types of links are utilized that are out-links and in-links.

3.6.3.2 In-links The set of articles in in-links is represented by $J(y)$ and is expressed below equation,

$$J(y) = \left\{ y_j \middle| \left(y_j, y \right) \in M \right\} \tag{40}$$

For example, consider the article title "wind energy". The article in-links are given to the Wikipedia page and then the hyperlink contains the article title "wind energy" in their main body or text.

3.6.3.3 Out-links The set of articles in in-links are represented by $P(y)$ and is expressed in below equation,

$$P(y) = \left\{ y_j \middle| \left(y, y_j \right) \in M \right\} \tag{41}$$

For example, assume the article title as "wind energy". The out-links are hyperlinks in the Wikipedia page of the article title "wind energy" (https://en.wikipedia.org/wiki/Wind_power). Wikipedia has a page "redirect" which is another path for reaching the correct articles to reduce the query terms. For example, the query

"USA" redirects to the article "The United States of America". This paper follows some steps to expand the query using Wikipedia is given below,

- In-link's extraction
- Out-link's extraction
- Assigning an in-link score to expansion terms
- Top terms are selected as expansion terms
- Expansion terms rebalancing

3.6.3.4 Assigning in-link scores to expansion terms When the in-link and out-link extraction process is completed, the expansion terms can be chosen on the basis of semantic resemblances in out-links. The semantic resemblances are evaluated depending upon the scores of in-links. Here, s,s_1 denotes query term and CET. If the below-mentioned two conditions are satisfied, two articles are semantically similar in Wikipedia, (1)s_1 has high scores of in-link(2)s_1 has both in-link and out-links. The in-link score is calculated as,

$$S_{\text{core}}\left(J\left(s_1\right)\right) = \text{sg}\left(s, s_1\right) \cdot \text{jeg}\left(s_1, V_{\text{E}}\right) \tag{42}$$

where $\text{sg}\left(s, s_1\right)$ is the frequency of s and its synonyms,$\text{jeg}\left(s_1, V_{\text{E}}\right)$ denotes the inverse document frequency of s_1. The inverse document frequency is given below,

$$\text{jeg}\left(s_1, V_{\text{E}}\right) = \log \frac{O}{\left|\left\{e \in V_{\text{E}} : s_1 \in e\right\}\right|} \tag{43}$$

where O represents the total number of articles in Wikipedia and $\left|\left\{e \in V_{\text{E}} : s_1 \in e\right\}\right|$ denotes the number of articles. Assume the in-link score can capture the total similarity between the initial query term and the expansion term and the entire necessary information is given by using the expansion term regarding query expansion. The term frequency can be used for providing the semantic similarities between the expansion term and the initial query term. In Wikipedia articles, the stop words have the lowest priority and both the expansion term and common terms are the query term article's hyperlinks. If the in-link scores are assigned to every expanded term, the top terms can be picked based on their scores of in-link which is considered one part of the expanded query. The other part is predicted from WordNet. The description of the different symbols used in query expansion is presented in Table 3.

3.6.4 Query expansion (QE) in wordnet

When the pre-processing is completed in the initial query, the phrases and individual words considered as keywords are searched for the query expansion in WordNet. In the extraction of semantic similarity terms, the phrases have high priority compared with individual terms. Phrases are mainly concentrated in the WordNet for expansion and the phrases are predicts the semantically similar term which is obtained from WordNet. In the query of semantically similar terms from WordNet,

Table 3 Symbol description

Symbol	Description
$Q_{\mathrm{u}}^{\mathrm{Ex}} \in q^{\mathrm{Ex}}$	Query expansion terms
D_{C}	Document collection
Ex_{n}	Function that returns N_{u} words in the vocabulary
v_{n}	Vocabulary words included in n
Lu	Synsets
p_{v}	Hypernyms
$\mathrm{T}_{\mathrm{Q}}^{\mathrm{CEx}}u$	Expansion term set
$H(B, M)$	Directed graph
s,s_1	Query term and CET
$\mathrm{sg}(s, s_1)$	Frequency of s and its synonyms
$\mathrm{jeg}(s_1, V_{\mathrm{E}})$	Inverse document frequency of s_1
O	Total number of articles on Wikipedia

the hyponym and synonym set of queries are the CET. The hyponyms and synonyms are fetched in two levels and after completing this process, a broad range of semantically similar terms are received. The terms are ranked and are given below,

$$S_{\mathrm{core}}(s_1) = \mathrm{sg}(s_1, s) \cdot \mathrm{jeg}(s_1, V_{\mathrm{E}}) \tag{44}$$

where s_1,s are denoted as the expanded term and initial query term. V_{E} is the Wikipedia dump. The expanded terms are ranked based on the predicted score and then the top terms are gathered which is considered as an intermediate expanded query. The intermediate query terms are reweighted depending on their expanded scores and the top terms can be predicted from the correlation score which is denoted as the second portion of the expanded query. The first portion of the expanded query can be obtained from Wikipedia.

3.6.5 Ontology-based query expansion

S_{p} and S_{r} are the expansion of the variable in a semantically related set of terms used for translating the keywords in semantic expansion. Through external resources, a multilingual knowledge graph R_{b} is produced for this function completion. A prefiltering process is done in the enterprise data graph R_{f} terms to establish the important R_{b} terms. Based on the scheme's actual use, there are 3 properties in multilingual knowledge graph nodes: the *language* property for the term language contained in the *lang* list, the *weight* property for the term weighting, and the *label* property for the term name. The *type* property of relationship present in the multilingual knowledge graph expresses the type of relationship among the terms. (Translation, Synonym, Abbreviated by, In relation) values are obtained in this *type*. The multilingual knowledge graph is expressed as follows,

$$R_{\mathrm{b}} = \{t, u\} \leftrightarrow \left\{ : \text{Knowledge Graph}\{'\text{label}' : \text{word}, '\text{language}' : \text{lang}, '\text{weight}' : r\}, (t) - [: \text{type}] \rightarrow (v) \right\} \tag{45}$$

In a multilingual knowledge graph, the term weighting is done by taking the term's real use. Hence, the user's language and vocabulary are adapted to the scheme through the following queries. Then, the below rule is used: other language synonyms S_{p} and S_{r} is utilized at first and introducing another similar term to the user. weight = weight + 1, while selecting and using the terms.weight = weight − 1, while deselecting the terms.S_{p} and S_{r} are related with proximity terms 1 of the synonymous type that has various languages, after presenting a query by the user. If its weight is greater than or equal to 1, they are also related to another proximity term 1.

4 Result and discussion

In this section, various methods such as WordNet, Wikipedia, HQE, NOSIR, and proposed IAOCOOT are employed for the performance analysis. The performance metrics like recall, precision, MRR (mean Reciprocal Rank), MAP (Mean Average Precision), F-measure, and NDCG (Normalized Discounted Cumulative Gain) are applied for predicting the best performance rate. The TREC dataset is used and the text processing is implemented in R-language. Terrier IR platform tool is utilized for large-scale text collection. The entire sentences or expressions and entities are interpreted through the semantic types of the information and also candidate answers with the correct type will focus.

4.1 Terrier IR platform tool

Terrier is used in large-scale text collection and it is extremely valuable, efficient, and flexible in search engines open source. The retrieval functions and index state-of-the-art are implemented by Terrier and for the quick estimation and improvement in large-scale retrieval use, it offers an ideal platform. At Glasgow University inside the Computing Science School, the information retrieval group introduced Terrier and it is used in Java.

In-text retrieval experimentation and research, Terrier is a flexible, comprehensive, transparent, and open-source platform.

4.2 TREC dataset

In 1992, the National Institute of Standards and Technology (NIST) and the U.S. Defense Department co-sponsored the Text REtrieval Conference (TREC) [26] in the context of the TIPSTER text program. Maintaining research inside the information retrieval community was its main aim through supplying the needed

infrastructure for the text retrieval methodology in large-scale estimation. The objectives in the series of TREC workshops are given below:

- Research ideas are replaced via open forum generation to improve the communication between academia, government, and industry;
- In academia and industry, to improve the usefulness of proper estimation methods, along with the present system that makes use of improvements in the latest evaluation methods;
- Under the high number of test collections, to boost up the information retrieval research; and
- During the difficulties in the real world, to accelerate the technology transformation to commercial products from research labs by showing the large retrieval methodologies development.

The initial large-scale estimations in non-English (Chinese and Spanish) document retrieval were sponsored by TREC. The estimations are initiated by TREC for digital video retrieval in content-based and answering the questions in open-domain. The operational settings are practically designed because of sufficient test collections in TREC.

4.3 Performance metrics

The different performance metrics used in this paper are presented as follows:

4.3.1 Precision

It is defined as the proportion of the overall quantity of the retrieved documents that are relevant to the number of retrieved documents with the query and it is calculated as;

$$P_{\mathrm{re}} = \frac{\tau_{\mathrm{rd}}}{\tau_{\mathrm{d}}} \tag{46}$$

4.3.2 Mean average precision (MAP)

It is defined as the average value of the precision scores with the retrieved documents through the κ query groups. Then the MAP is expressed and calculated as;

$$\mathrm{M}_{\mathrm{A}\rho} = \frac{1}{\kappa} \sum_{\kappa} \mathrm{A}\rho_{\kappa} \tag{47}$$

4.3.3 Recall(R)

It is defined as the proportion of the overall amount of retrieved documents to the number of relevant documents in the database and it is calculated as;

$$\mathfrak{R} = \frac{\tau_{\mathrm{rd}}}{\tau_{\mathfrak{R}}} \tag{48}$$

4.3.4 F-measure

The F-measure is considered by the harmonic average of recall and precision, and then it is expressed and calculated as;

$$F_{\mathrm{meas}} = \frac{2\rho\mathfrak{R}}{\rho + \mathfrak{R}} \tag{49}$$

4.3.5 Normalized discounted cumulative gain

The NDCG (normalized discounted cumulative gain) is defined as the proportion of the DCG to the IDCG (ideal discounted cumulative gain). Then, it is calculated and expressed as;

$$\mathrm{N}_{\mathrm{dCG}} = \frac{d_{\mathrm{CG}}}{\mathrm{I}d_{\mathrm{CG}}} \tag{50}$$

4.3.6 P-R Curve

The precision-recall (P-R) curve is defined as the tradeoff between recall and precision for various thresholds.

4.3.7 Mean reciprocal rank (MRR)

The MRR is defined as the information retrieval (IR) calculated the rank reciprocal with the primary relevant documents are retrieved. Then, it is expressed as;

Table 4 Comparative analysis of MAP (mean average precision) for the TREC dataset

Dataset used	Various methods	MAP
TREC dataset	WordNet	0.2901
	Wikipedia	0.3166
	HQE	0.3387
	NOSIR	0.3596
	Proposed IAOCOOT	0.3945

Table 5 Comparative analysis of different expansion terms using the TREC dataset

Various methods	10	20	30	40	50	60
WordNet	0.2867	0.3045	0.3439	0.3402	0.3398	0.2853
Wikipedia	0.3065	0.2973	0.3667	0.3576	0.3521	0.3274
HQE	0.3398	0.3566	0.3512	0.3589	0.3464	0.3563
NOSIR	0.3361	0.3332	0.3698	0.3521	0.3497	0.3599
Proposed IAOCOOT	0.3334	0.3498	0.3675	0.3690	0.3587	0.3412

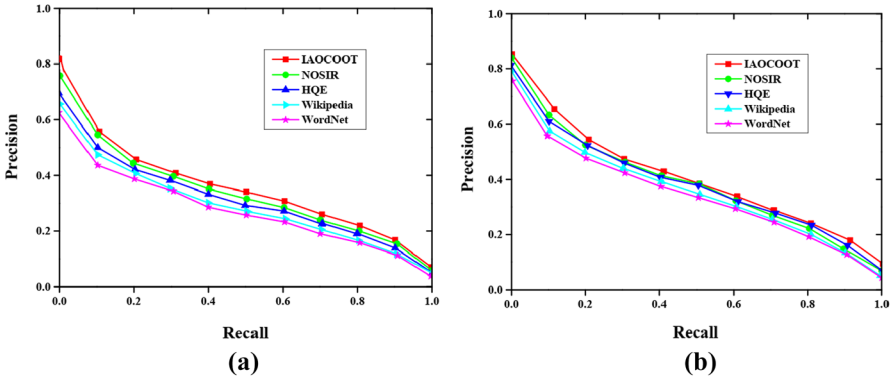

Fig. 3 Comparative analysis of precision-recall curve **a**without query expansion **b**withquery expansion

$$M_{\Re\Re} = \frac{\sum_{e=1}^{\kappa} \frac{1}{\Re AN\kappa d}}{\kappa} \tag{51}$$

4.4 Performance evaluation

Table 4 shows the mean average precision (MAR) for various methods by using the TREC dataset. The various methods like WordNet, Wikipedia, HQE (Hybrid Query Expansion), NOSIR (New Ontology-based Semantic Information Retrieval), and proposed IAOCOOT (Improved Aquila Optimization-based COOT algorithm) are employed to get the separate MAP values and find the superior values from all methods. The table explains the MAP rate of 0.2901 for WordNet, 0.3166 for Wikipedia, 0.3387 for HQE, 0.3596 for NOSIR, and 0.3945 for the proposed IAOCOOT method. The proposed IAOCOOT method has a very high MAP rate compared with other methods.

The comparative analysis of different expansion terms is described in Table 5 by using the TREC dataset. For comparative analysis, various methods such as WordNet, Wikipedia, HQE, NOSIR, and proposed IAOCOOT are employed. From this table, the proposed IAOCOOT method has a very high MAP rate of 0.3690.

Figure 3 represents the P-R curve (precision-recall curve) for various methods like WordNet, Wikipedia, HQE, NOSIR, and proposed IAOCOOT. Figure 3a shows the P-R curve without query expansion. In the P-R curve without query expansion, the proposed IAOCOOT method has a very high-performance rate of 0.82 and the WordNet has a very low-performance rate of 0.61 when compared with other methods. The precision rate is rapidly increased with the increment of the recall rate. Figure 3b denotes the P-R curve with query expansion and the proposed IAOCOOT method has a very high-performance rate of 0.85 and the WordNet has a very low-performance rate of 0.76 when compared with other methods. The precision rate is rapidly increased with the increment of the recall rate.

The comparative analysis of various performance metrics like F-measure, MRR, and NDCG is explained in Fig. 4. Different methods such as WordNet, Wikipedia, HQE, NOSIR, and proposed IAOCOOT are used for the comparative analysis. In F-measure, the proposed IAOCOOT method has a high-performance rate of 0.3 and WordNet has a low-performance rate. In MRR, the proposed IAOCOOT method has a high-performance rate of 0.3529 and WordNet has a low-performance rate. In NDCG, the proposed IAOCOOT method has a high-performance rate of 0.3812 and WordNet has a low-performance rate.

Table 6 describes the query expansion terms for various methodologies by utilizing the WordNet, Wikipedia, and TREC dataset. Figure 5 shows the comparative analysis of mean average precision (MAP) and the methods such as WordNet, Wikipedia, HQE, NOSIR, and proposed IAOCOOT are applied for getting a good performance rate. The proposed IAOCOOT method provides a good performance rate of 0.3945 and the WordNet has the lowest performance rate compared with other methods.

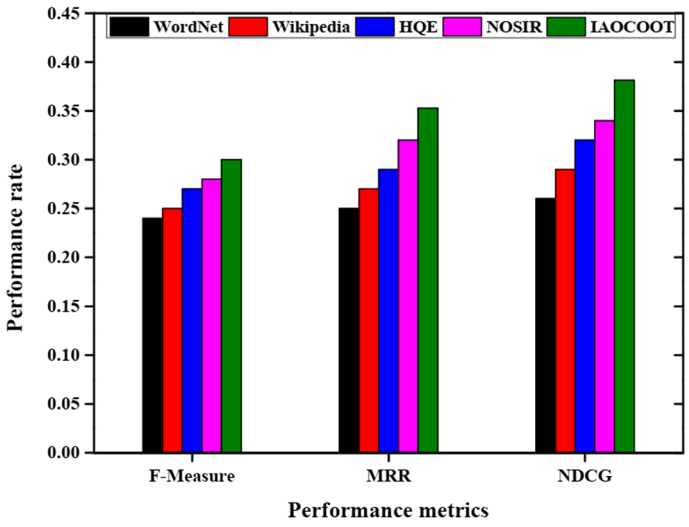

Fig. 4 Comparative analysis of various performance metrics

Table 6 Query expansion terms were obtained by using different methodologies in the TREC dataset

Query ID	Original query	Expansion terms acquired from WordNet	Expansion terms acquired from wikipedia	Expansion terms acquired from the proposed model
132	Covid19 vaccine	Immunogenicity, vitro diagnostics, outweigh risks, invading germs, aspirin, immunization, pre-clinical development	CDC's covid-19 booster tool, vaccination card, COVAX global vaccine, acetaminophen, immunization, PubMed, aspirin	Pfizer, Moderna, 2-dose series, immunocompromised, booster dose, mRNA covid-19 vaccine, ibuprofen, metformin, allergic reactions, mammogram,
141	Ukraine disaster	Turbine hall, combustible material, burning reactor, fission products, nuclear accident, coal miners, core catchers	Relief valves, turbine generation, RBMK control rods, steam explosion, steam boiler, radioactive fallout, ionized airglow	RBMK-type nuclear reactor, decontamination, Chornobyl nuclear power plant, acute radiation, ionizing radiation, radioactive decay, control rods
153	Ukraine refugee relief	762 project, psychological counseling, GoFundMe Campaign, Swiss-based organization, troop buildup, sunflower of peace	Project hope, UNICEF, world central kitchen, voices of children, the international committee of children, hygiene kits	CARE, Convoy of hope, Doctors without borders, international medical corps, Internews, humanitarian aid, Kyiv independent

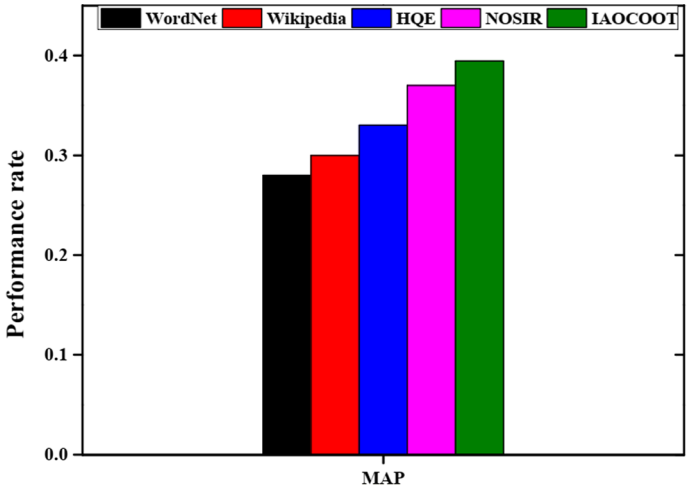

Fig. 5 Comparative analysis of mean average precision

Table 7 Query expansion for various metrics

Performance metrics	Q_1	Q_2	Q_3
Precision	0.53	0.50	0.54
Recall	0.49	0.46	0.50
F-measure	0.46	0.40	0.45
MAP	0.42	0.38	0.40

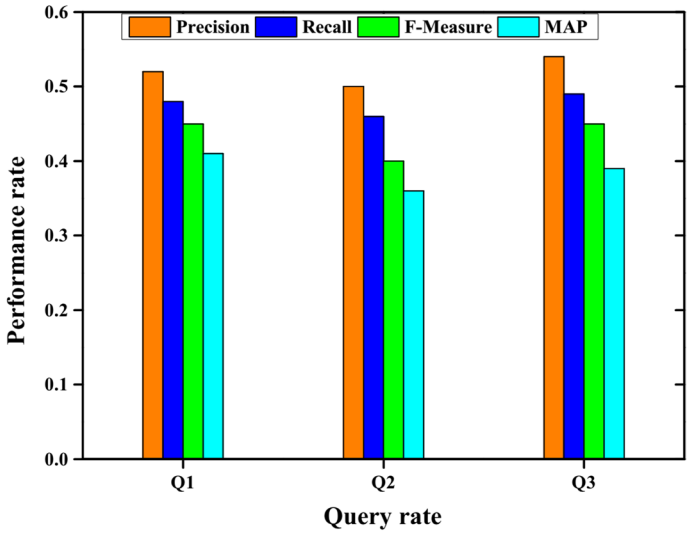

Fig. 6 Query expansion analysis

Table 7 illustrates the results for various queries concerning the precision, recall, F-measure, and MAP. The Q_1results showed that the various metrics like precision, recall, F-measure, and MAP acquired the values of 0.53, 0.49, and 0.46,042. The value obtained for Q2 concerning precision is 0.50, recall is 0.46, F-Measure is 0.40, and MAP is 0.40. The results of Q_3 indicated that the various parameters such as precision, recall, F-measure, and MAP achieved higher values of 0.54, 0.50, 0.45, and 0.40. Figure 6 illustrates the graphical analysis of query expansion. The experimentation results of Q_3 showed that the various performance metrics such as precision, recall, F-measure, and MAP acquired better results than Q_1 and Q_2.

5 Conclusion

This paper presents a novel IAOCOOT algorithm for optimal query expansion in text-based information retrieval by overcoming the syntactic heterogeneity issues. When the pre-processing is completed in the initial query, the phrases and individual words are considered as keywords for expanding the initial queries by using Wikipedia. The in-links, out-links, and Wikipedia titles are needed for selecting the candidate expansion terms (CET). When the pre-processing is completed in the initial query, the phrases and individual words considered as keywords are searched for the query expansion in WordNet. In the extraction of semantic similarity terms, the phrases have high priority compared with individual terms. The MNWA algorithm is used to identify the top-k similarity terms by eliminating the less relevant terms and helping the users rapidly analyze the information needed. In this paper, various external resources such as WordNet, Wikipedia, etc. are analyzed using the proposed IAOCOOT model. The performance metrics like recall, precision, MRR (mean Reciprocal Rank), MAP (Mean Average Precision), F-measure, and NDCG (Normalized Discounted Cumulative Gain) are applied for predicting the best performance rate. In the P-R curve without query expansion, the proposed IAOCOOT method has a very high-performance rate of 0.82 and the WordNet has a very low-performance rate of 0.61. In the P-R curve with query expansion, the proposed IAOCOOT method has a very high-performance rate of 0.85 and the WordNet has a very low-performance rate of 0.76. The proposed IAOCOOT method has a performance rate of 0.3 for F-measure, 0.3529 for MRR, 0.3812 for NDCG, and 0.3945 for MAP.

Author contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by [RK] and [SCS]. The first draft of the manuscript was written by [RK] and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Data availability Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Human and animal rights This article does not contain any studies with human or animal subjects performed by any of the authors.

Informed consent Informed consent was obtained from all individual participants included in the study.

References

1. Kumar R, Sharma SC (2018) Information retrieval system: an overview, issues, and challenges. Int J Technol Diffus (IJTD) 9(1):1–10
2. Sundararaj V, Selvi M (2021) Opposition grasshopper optimizer based multimedia data distribution using user evaluation strategy. Multimed Tools Appl 80(19):29875–29891
3. Sundararaj V (2019) Optimal task assignment in mobile cloud computing by queue based ant-bee algorithm. Wireless Pers Commun 104(1):173–197
4. Maksimov N, Golitsina O, Monankov K, Gavrilkina A (2020) Knowledge representation models and cognitive search support tools. Procedia Comput Sci 169:81–89
5. Oyefolahan IO, Aminu EF, Abdullahi MB, Salaudeen MT (2018) A review of ontology-based information retrieval techniques on generic domains.
6. Manzoor S, Rocha YG, Joo SH, Bae SH, Kim EJ, Joo KJ, Kuc TY (2021) Ontology-based knowledge representation in robotic systems: a survey oriented toward applications. Appl Sci 11(10):4324
7. Yang D, Shen DR, Yu G, Kou Y, Nie TZ (2013) Query intent disambiguation of keyword-based semantic entity search in dataspaces. J Comput Sci Technol 28(2):382–393
8. Jain S, Seeja KR, Jindal R (2021) A fuzzy ontology framework in information retrieval using semantic query expansion. Int J Inf Manag Data Insights 1(1):100009
9. Sharma DK, Pamula R, Chauhan DS (2019) A hybrid evolutionary algorithm-based automatic query expansion for enhancing document retrieval system. J Ambient Intell Human Comput. https://doi.org/10.1007/s12652-019-01247-9
10. Raza MA, Mokhtar R, Ahmad N, Pasha M, Pasha U (2019) A taxonomy and survey of semantic approaches for query expansion. IEEE Access 7:17823–17833
11. Afuan L, Ashari A, Suyanto Y (2019) A study: query expansion methods in information retrieval. J Phys Conf Series 1367(1):012001 (**IOP Publishing**)
12. Azad HK, Deepak A (2019) A new approach for query expansion using wikipedia and WORDNET. Inf Sci 492:147–163
13. Torjmen-Khemakhem M, Gasmi K (2019) Document/query expansion based on selecting significant concepts for context based retrieval of medical images. J Biomed Inform 95:103210
14. Malik S, Shoaib U, Bukhari SAC, El Sayed H, Khan MA (2022) A hybrid query expansion framework for the optimal retrieval of the biomedical literature. Smart Health 23:100247
15. Wang J, Pan M, He T, Huang X, Wang X, Tu X (2020) A pseudo-relevance feedback framework combining relevance matching and semantic matching for information retrieval. Inf Process Manage 57(6):102342
16. Jafarzadeh P, Ensan F (2022) A semantic approach to post-retrieval query performance prediction. Inf Process Manag 59(1):102746
17. Dahir S, El Qadi A (2021) A query expansion method based on topic modeling and DBpedia features. Int J Inf Manag Data Insights 1(2):100043
18. Kaur N, Aggarwal H (2021) Query reformulation approach using domain specific ontology for semantic information retrieval. Int J Inf Technol 13(5):1745–1753
19. Kammoun H, Gabsi I, Amous I (2022) Mesh-based semantic indexing approach to enhance biomedical information retrieval. Comput J 65(3):516–536
20. Esposito M, Damiano E, Minutolo A, De Pietro G, Fujita H (2020) Hybrid query expansion using lexical resources and word embeddings for sentence retrieval in question answering. Inf Sci 514:88–105

21. Selvalakshmi B, Subramaniam M (2019) Intelligent ontology-based semantic information retrieval using feature selection and classification. Clust Comput 22(5):12871–12881
22. Liu Q, Huang H, Xuan J, Zhang G, Gao Y, Lu J (2020) A fuzzy word similarity measure for selecting top-*k* similar words in query expansion. IEEE Trans Fuzzy Syst 29(8):2132–2144
23. Aspland E, Harper PR, Gartner D, Webb P, Barrett-Lee P (2021) Modified Needleman–Wunsch algorithm for clinical pathway clustering. J Biomed Inform 115:103668
24. Wang S, Jia H, Abualigah L, Liu Q, Zheng R (2021) An improved hybrid aquila optimizer and harrishawks algorithm for solving industrial engineering optimization problems. Processes 9(9):1551
25. Naruei I, Keynia F (2021) A new optimization method based on COOT bird natural life model. Expert Syst Appl 183:115352
26. Data-English documents. Text REtrieval conference (TREC) english documents. (n.d.). Retrieved from https://trec.nist.gov/data/docs_eng.html. Accessed 27 May 2022