Soft Computing
https://doi.org/10.1007/s00500-020-05192-0

METHODOLOGIES AND APPLICATION

A novel two-stage optimized model for logo-based document image retrieval based on a soft computing framework

K. Raveendra[1,3] · T. Karthikeyan[1] · Vinothkanna Rajendran[2] · P. V. N. Reddy[3]

© Springer-Verlag GmbH Germany, part of Springer Nature 2020

Abstract
The rapid development of internet helps in organizing documents based on their specific data for large-scale organizations to small-scale organizations. Document retrieval system aims to organize the relevant documents and its information based on specific terms. The availability of information stored by organization requires inexpensive storage, and the searching mechanism needs to get the information-based documents very quickly in real time. This research aims to provide such document retrieval system through logo-based identification model to analyse and organize the documents. A two-stage optimization is implemented to obtain the proposed logo-based document retrieval system using genetic algorithm and inverted ant colony optimization. Utilization of genetic operators in document retrieval classification based on index terms reduces time consumption, and inverted ant colony optimization improves the retrieval efficiency. Parameters such as classification accuracy, precision, retrieval efficiency are observed and compared with existing conventional and hybrid models experimentally to validate the proposed model.

Keywords Two-stage optimization · Logo · Document retrieval · Genetic algorithm

1 Introduction

Data retrieval deals with storage, organization, access and analysis of useful information from database at essential timings. The information system which handles the data helps to manage the documents in an easy way. The key objective of data retrieval is to collect all the relevant information based on the user query. The emphasis on retrieving such information reduces time and cost so that it could be used in enhancing the other features of the system. Present document retrieval systems are based on keyword or similarity-based models. In keyword-based document management model, a particular string is used to search the document from the database. It is a basic information retrieval system that considers the key strings as query and produces all the documents in the database which makes the search purpose useless. It displays entire documents in the database, and it may not relevant to the user who is searching for particular documents. The other difficulty in keyword-based search presents when it retrieves the documents which have the same meaning of keyword and multiple meaning of keywords, while in similarity-based models it produces the similar documents using common keywords. Output of such retrieval system performs search based on degree of relevance which is measured based on closeness to the keywords. The common information retrieval system is classified into three major categories as Boolean, probabilistic and vector models. This Boolean model uses index terms to search and analyse the documents. In the case of probabilistic model, document search is performed based on the probability theory. In vector model, an f-dimensional space is used to represent the documents while searching. Figure 1 depicts the information retrieval system based on the features like user query and user feedback.

Classic models have been proposed as an alternative paradigm for document retrieval systems. Using set

Communicated by V. Loia.

✉ K. Raveendra
raveendrasvew@gmail.com

[1] Konerulakshmaiah Educational Foundation, Guntur, Andhra Pradesh, India

[2] Vivekanandha College of Engineering for Women, Tiruchengode, Namakkal (Dt), Tamilnadu, India

[3] S V Engineering College (Formerly SVEW), Karakambadi road, Tirupati, Andhra Pradesh, India

Published online: 18 July 2020

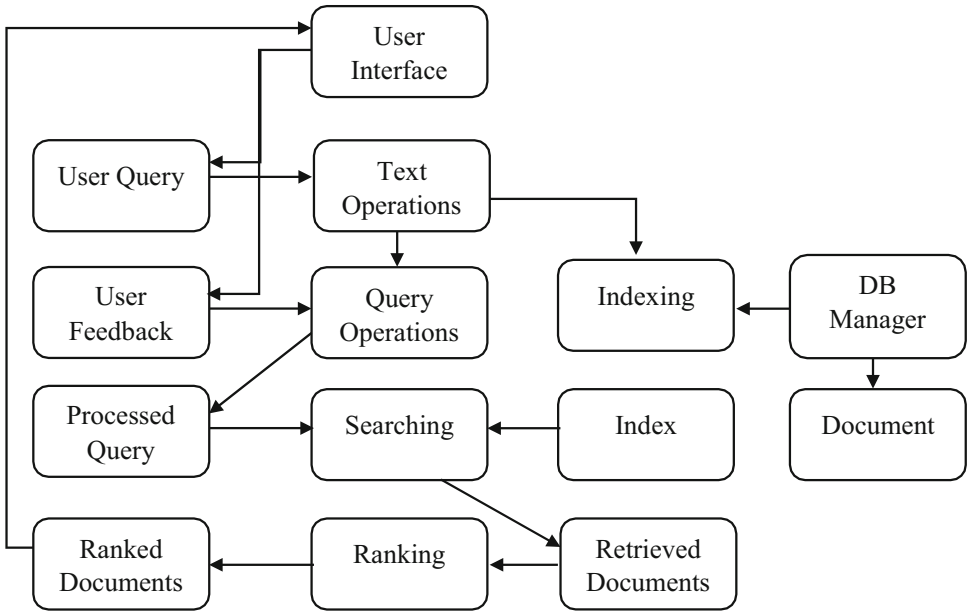


Fig. 1 Illustration of information retrieval model

theoretic models, the documents are distinguished and organized. Similarly, fuzzy-based models help to organize the data like Boolean models using their logical functions. Alternative probabilistic models use neural networks and belief models to enhance the performance of document retrieval systems. Once the documents are pre-processed, the documents are indexed using indexing module. So that it would be useful while retrieving the documents in searching operation. Prioritizing the documents based on the relevancy is essential, and it is performed through ranking operation. The user interaction and their functionality with interface module are related to query and result management systems. The results are displayed as ranked documents so that user could access the organized data. The feedback mechanism is used to improve the results by obtaining query and restarting the search process. Pre-processing the documents is to remove the unnecessary characters and tags. The essential phrases are comprised of multiple words identified based on the selection.

Instead of using keywords and similarity strings logo-based document retrieval system is familiar nowadays. Logo represents the symbolic representation of an organization. It adds ease identification of their products and firms to public and other competitive firms. The documents process in an organization has their logo in it to validate the contents to be approved. Since logo attains more importance in documents, the retrieval system using logo becomes more popular and the research still continues to improve the performance of the system. Detailed description of simple logo-based document retrieval is depicted in Fig. 2. Based on the user request, the information is pre-processed.

The database pre-processes the documents based on their logo by selecting the features. Based on the selected features, the logo is segmented and it is stored in the logo database. The user query was a logo, and it is given to the search module. Once the request is received, the features of given query are extracted and are compared with the logos present in the database. Once it features are matched, then the relevant documents are retrieved and presented to the user through interface module. The common issues in documents retrieval model are listed as 1. It requires proper structural form arrangements to maintain a database. 2. It requires the type of data extracted to obtain relevant information. 3. Pre-processing the documents is essential, and it requires a special rules and characteristics. 4. Classification and discrimination based on the information requires suitable models. 5. Image processing required to process the user query consumes more operational time and computation cost. The above-mentioned points are motivated this research to develop an optimized logo-based document retrieval model. Classification accuracy and improved efficiency are the main objectives of this research; based on this, two-stage optimization is preferred. Optimization algorithms help to improve the performance in terms of classification, efficiency and accuracy. In particular, the proposed model is developed with genetic algorithm in the first stage of optimization which improves the classification performance and inverted ant colony optimization in the second stage which improves the retrieval efficiency and accuracy (Fig. 3).

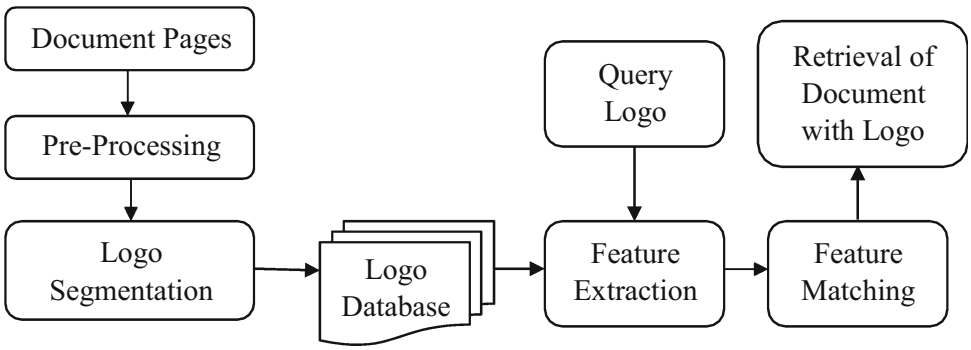


Fig. 2 Logo-based document retrieval system

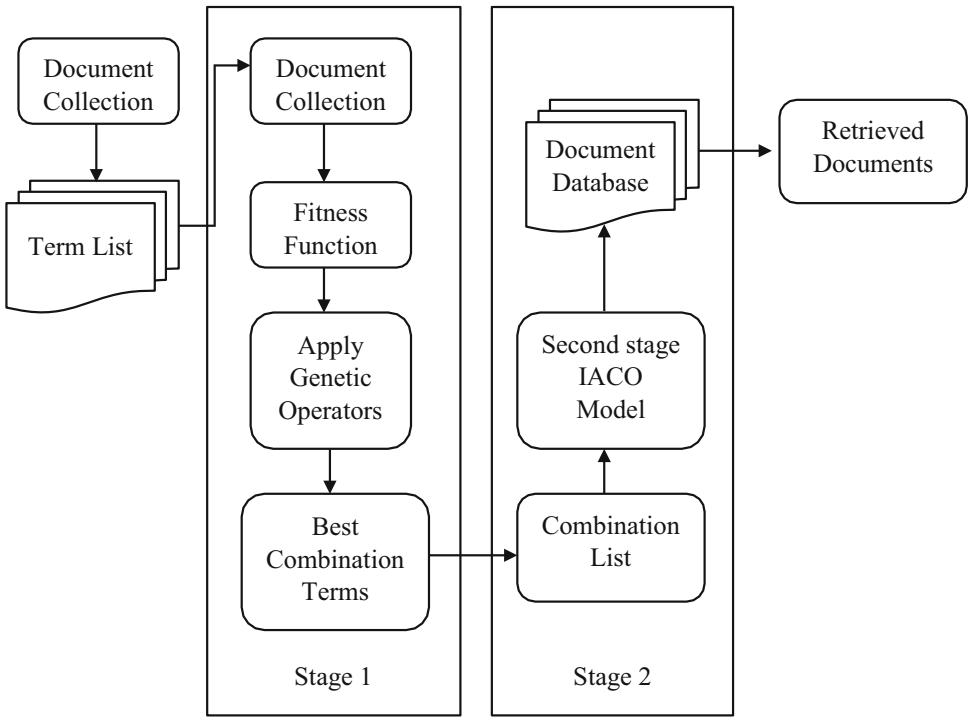


Fig. 3 Proposed two-stage optimized logo-based document retrieval model

2 Related works

The advance of technology in organizing the documents in an organization is common based on the salient features like stamps, seals and logos. Various detection algorithms are available to perform document retrieval. A vast survey to classify the algorithms and identify the pros and cons of the existing system is performed in this section. Logo-based document algorithms are presented as survey in the literature (Alaei et al. 2016). The state of art of research provides a consolidation of various logo- and seal-based document retrieval systems with their advantages and shortcomings. It is helpful in further way to increase the efficiency of the retrieval system based on the limitations. Global characteristics (Kumar et al. 2016) are used in document retrieval process such as colour, texture and shape. Combination of these characteristics or fusion of all these is used to categorize the documents. K-nearest neighbour classifier is used in the classification process to evaluate the accuracy, precision, f-measure and accuracy of the model through confusion matrix.

Hyperspectral document retrieval system is reported in the literature (Qureshi et al. 2019), and it retrieves information based on document images and paintings. Analysis of such documents and its limitations are presented, and it is overcome by commercial HIS system to solve the

challenges. The evaluation of document retrieval system through automated processing techniques (Cristani et al. 2018) improves the classification accuracy. It emerges clear picture to solve the issues incorporated to document processing. Cloud computing technologies are also involved in these automated processing technologies to perform document retrieval. Retrieval based on optical characters (Nagy 2016) underlies most of the document processing applications. It uses microprocessors and optical sensor arrays to classify the documents. Using suitable algorithmic approaches, the scripts are analysed and applicable in real-time systems like postal and other systems. Digital documents and its management process through smartphones are used to bring forward the retrieval process.

Document retrieval system using face recognition is reported in the literature (Dixit and Shirdhonkar 2018). Documents containing sensible information such as passport, identity card, voter ID, certificates and property documents is need to be processed securely. Face-recognition-based document retrieval system useful in organizations expects high security and government offices. The process in this system is based on detecting face image from query document, then it extracts the features of face image, and finally it retrieves all the documents similar to that face image. Three types of similarity metrics are used in the literature (Zhou et al. 2019) to retrieve the document. It helps in preserving the semantic similarity and manifold data points in hamming space. The proposed model is performed through different sized sliding window to extract the features from raw data, and hash function is used in the feature extraction phase. Finally, deep neural network model is used to train the extracted features and produce faster results.

Literature (Haijiao et al. 2019) reports the document retrieval system through multi-modal objects (Dang et al. 2018) to improve the training process. Leveraging the unlabeled and labelled objects enhances the performance of retrieval system, and it is obtained using semi-supervised multi-concept deep learning model. Regardless of multi-modal convolution neural network concept, this classifier classifies the images and texts to annotate the images. Literature (Deng et al. 2018) discusses about fabric image retrieval system which is applicable in cloth searching in online shopping and modelling fields. Using convolutional neural networks, this significant process is achieved and it minimizes the cost attribute and provides efficient optimization. Similar deep learning neural network model for logo recognition is presented in the literature (Bianco et al. 2017). The performance of the proposed model is evaluated and compared with convolutional neural network on Flickr logos- 32 database. Investigating the benefits of training choices based on sample weight, class balancing and explicit the background class are performed in the proposed model.

In logo (Bai et al. 2018)-based document retrieval system, feature extraction and similarity measurements (Yang et al. 2019) are essential, and they are performed using classical deep convolutional neural network. The proposed model improves the feature extraction ability and optimizes the process in three aspects as pooling layer, fully connected layer and hidden layer. It is applicable in image-based document retrieval system which has large-scale database. Similar image-based recognition is presented in the literature (Hao et al. 2018a, b; Alaei et al. 2019) to achieve realistic recognition. Traditional convolution neural network models have not considered the weights of learning instances and reduced the image recognition accuracy. To solve this issue, optimized CNN is proposed in this research model. First, it selects the bottom-up region to retrieve the target regions of each learning instance and then it enhances the model through optimization to make use of full learning instances.

Learning based on hashing technologies is presented in the literature (Joby 2020; Ge et al. 2019), and it is widely adopted in multimedia retrieval systems. It extracts the semantic information for high-dimensional data, and the limitation present in the model is that it requires hashing codes which makes the optimization hard process. In this proposed model, a supervised learning-based discrete hashing function is used to learn the binary codes in deep learning framework. It achieves an optimized direction of multipliers with efficient iteration to obtain better image-based document retrieval system. Similar image-based model using optimized deep learning is presented in the literature (Hao et al. 2018a, b) to improve the retrieval quality of document. The limitation in high-quality image retrieval needs much human resource in selection process, and also it consumes more sources for its computation. This proposed model uses joint entropy-based learning model (Tang and Peng 2017; Costa et al. 2018) to overcome this issue and reduces the learning instances through optimization. Preselection is performed thorough watershed segmentation, and joint entropy mode is used to reduce the possibility of double and useless instances. This increases the accuracy of the retrieval system.

To solve the issues in fundamental document retrieval literature, Djenouri et al. (2018) proposed a data mining technique and optimization model. Two-stage investigation uses data mining process in the pre-processing stage to split the document collection into similar clusters. This is achieved using k-means algorithm. The second phase of the proposed model extracts the relevant terms based on bee swarm optimization to explore the document deeply. From the survey, it is evident that document retrieval

systems perform better in optimization models compared to conventional neural network and fuzzy-based models (Chaieb et al. 2017). The proposed model is designed based on the limitations of the existing works to perform efficient document retrieval process based on logo using two-stage optimization process.

3 Proposed work

The proposed model is comprised of two-stage double optimization which has genetic algorithm for the first-stage optimization and inverted ant colony optimization for final stage optimization. The resultant logo-based systems are derived in this section mathematically. In the field of data and information management, genetic algorithm is famous for its genomic behaviour-based operation to produce best optimized results. Managing large space documents through genetic algorithm specifically attain more importance due to its operational results. It is a powerful searching mechanism and performs quick search in high-dimensional space. In the proposed model, genetic algorithm used to generate the best set of document collection is performed in the first stage. GA considers three factors while designing the system model such as representing the problems in the form of chromosome, uses fitness evaluation to optimize the performance and chooses the parameters similar to population size, structure and genetic operators. This research work is essentially designed to produce best combinations to bring the relevant documents from a set of database through genetic algorithm. Later, the search results are again optimized using inverted ant colony optimization.

A set of documents are indexed in the initial process based on the logo, and the index terms are stored as list. Suitable frequency measure is given for each index term in the list. In order to make this search efficient, a combination of index and term list is introduced and this combination helps in retrieving the relevant information. Using genetic algorithm, a set of best index terms are identified, and if both logos are in the same frequency, then a random frequency is assigned for that logo. This reduces time consumption and retrieves relevant information in short duration.

The chromosome representation in genetic algorithm is based on the keyword which is repeated number of times. This is related to document collection in terms of high frequency for more number of search appearance and low frequency for less number of search appearance. The mean frequency value is calculated, and it is used as threshold value to group the documents accordingly. GA cycle is interconnected as four elements. In the manipulation process, genetic operators are used to produce a set of new individuals and their offspring. In the manipulation, this genetic information is processed as a pair to reproduce. Then the information is stored as strings which describe the individuals.

Two operators are used in GA such as crossover and mutation. These generate the older population, and it is repeated until maximum number of cycles is reached. Figure 4 represents the crossover process. It takes two chromosomes and swaps the genetic information to produce new chromosome. Selection process associated with recombination is performed by crossover that assures the genetic structures called building blocks which is used for future generation. It represents the fitness function of genetic structure to explore the search space.

The mutation operator helps to identify the new genetic structures in the algorithm using local minima traps. The modification is random in nature, and it is not related to previous genetic structures. It creates structures based on other sections in the search space. Figure 5 depicts the mutation process implemented occasionally by altering a random bit from a string which is applied to fifth element of chromosome.

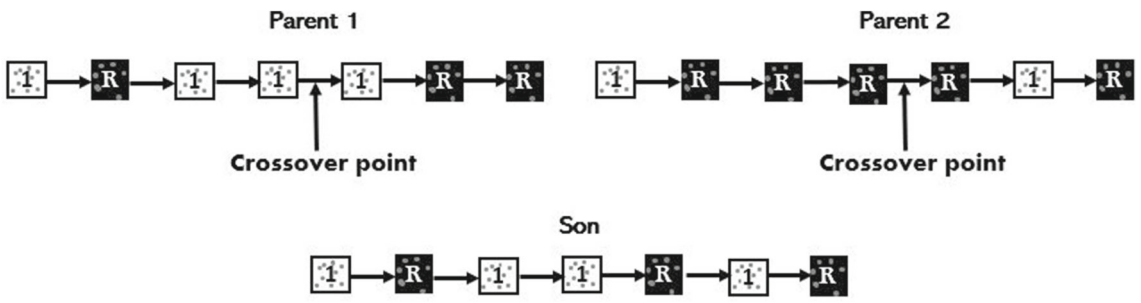


Fig. 4 Crossover process

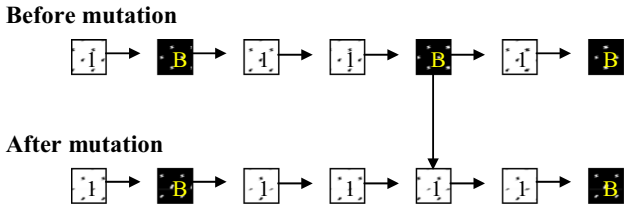


Fig. 5 Mutation

Other than crossover and mutation, other operators are available in genetic algorithm such as inversion, genetic edge recombination and dominance. These are used to process the constrains of specified problems. Evaluation step in the cycle is related to actual systems, and the algorithm is optimizing the needs of individuals. The individuals are coded with strings, and based on any of this parameter optimization is performed. However, the resulting strings are big and the process is little bit slow to get the optimization characteristics.

Having this in issue in mind, the obtained results are further optimized using inverted ant colony optimization. This model is developed without changing the results of genetic algorithm and their individual behaviours. This second stage of optimization is classifying the results further based on the distance measures and optimization factors. A criterion used to calculate the cost function of each edge is obtained by adding the distance of the calculated edge though Dijkstra's algorithm, and it is given as

$$\text{edge}\ \ \cos t(i,j) = \text{distance}(i,j) + \text{edge pheromones}(i,j) \tag{1}$$

In the beginning of analysis, there are no pheromones in the model and inverted ant colony optimization is used in this stage to materialize the ants without any defined path. To avoid cold start problem, this algorithm is used which allows the ants to determine the optimal path to reach destination. Once the shortest path is selected using Dijkstra's algorithm, IACO performs its initial process. Depositing the pheromone and update process is initiated. IACO differs from conventional ACO model in its pheromone characteristics. The pheromone level in IACO does not attract ants rather it repels, and this repulsion simulates relevant information required for the proposed model.

Based on the query intensity, the pheromone level changes: if the intensity is high, the pheromone value is high. If the intensity is low, the pheromone is low and IACO will behave normally. The minimum travel time is obtained as

$$\min \text{travel time} = \frac{\text{Edge length}}{\text{Edge maximum allowed pace}}. \tag{2}$$

The pheromone evaporation is obtained based on this travel time and the deposited pheromone constant, and it is given as

$$P_{\text{edge}} = \begin{cases} P_{\text{edge}} + \text{Dep}_{\text{Pher}} \\ P_{\text{edge}} - \min \text{travel time} \times \text{Dep}_{\text{Pher}} \end{cases} \tag{3}$$

It is important the updated pheromone level and its routes must be updated accordingly. The route recalculation which allows to calculate the cost is given as

$$\text{edge}\ \ \cos t(i,j) = \text{distance}(i,j) + \text{edge pheromone variation}(i,j) \tag{4}$$

Based on the mathematical formulations, the proposed model retrieves the relevant information from the database. Though various optimization algorithms are available the proposed model is implemented using genetic algorithm since the characteristic features of genetic algorithm in classification performance are much better to other optimization algorithms. The unique genomic nature improves the accuracy based on the index terms. In the case of second stage optimization, inverted ant colony optimization is used. Compared to other nature inspired algorithms, data processing performance is better in ant colony optimization model. The proposed inverted ant colony optimization is suitable for improved analytic models which has been applied to all the data-oriented applications. Summary of the proposed double algorithm is given as follows:

Initialize GA parameters
Begin
Population = GA initial solutions
Repeat until maximum criteria
Select parents for offspring production
Employ cross over and mutation operator
Enhance offspring using IACO

Initialize IACO parameters

Possible movements initialization for searching the document in database

While target not found

do for each index

do

generate all the possible index and its location as list
acquire index information to all locations
generate probabilities of each index terms in database

select the next best location

update the current position and index value

end

Replace destination with new documents based on query logo

Summarize the information
end

---

4 Result and discussion

The proposed model is experimented and compared with conventional ACO and PSO model. MATLAB 14.1 Simulink toolbox is used to compute the operation on a 2.6 GHz Intel Core i7 processor.

Flickrlogos-32 database is used in the experimentation, and the dataset contains brand logos of 32 classes with 6000 negative images. It includes images, annotations, ground truth and precomputed visual features for 8240 images from Flickr. Figure 6 depicts the images in database. The parameters used in the proposed model are listed in Table 1.

Classification accuracy of the proposed model is compared in Fig. 7. This accuracy is tested based on relevant accurate document retrieval for the user query, and it is obtained as

$$\text{Classificaiton Accuracy} = \frac{\text{Number of relevant documents retrieved}}{\text{Total Number of documents in database}} \times 100 \quad (5)$$

From the database, three logos are selected and it is given as query to all the models and it is clear that the proposed model performs better and retrieves information relevant to query logo.

Figure 8 shows a comparison chart for precision values and total recall values for the proposed model and existing models. The precision is obtained though the below mathematical formulation:

$$\text{precision} = \frac{\text{No. of relevant documents}}{\text{No. of relevant documents} + \text{No. of Irrelevant documents}} \quad (6)$$

The graph indicates that the proposed model has high precision values, compared to other two models which have lower precision values for the same data set.

Computation time for the proposed model is compared with other two models in Fig. 9. It is observed that proposed GA-IACO model consumes less time than other models. The computation time is calculated for all the three logo based document retrieval process. The performance of proposed model is compared with conventional models and listed in Table 2.

Efficiency of the proposed model is compared in Fig. 10. The efficiency of the proposed model is obtained through the equation

$$\text{Efficiency} = \frac{\text{Number of documents retrieved for user query}}{\text{Total Number of documents in database}} \times 100 \tag{7}$$

The overall efficiency is calculated based on the accurate document retrieval and the ratio of irreverent information while retrieval in process. The proposed model attains better efficiency compared to conventional models on average of 97.88% for the given three query logos.

5 Conclusion

Logo-based document retrieval system helps an organization to manage their documents in an easy way. Document retrieval using other methodologies like keywords search, string search lags in performance by displaying irrelevant documents in search process. This reduces the searching accuracy and retrieval efficiency. This research model identifies such issues, and it is overcome by using two-stage double optimization using genetic algorithm and

Table 1 Parameter used in proposed model

S. no	Parameter	Value
1	Population size	100
2	Number of generation	200
3	Chromosome length	20
4	Mutation	0.15
5	Crossover	0.85
6	Pheromone ratio	0.16
7	Positive pheromone ratio	0.2
8	Negative pheromone ratio	0.3
9	Evaporation ratio	0.45
10	Maximum of tour	15

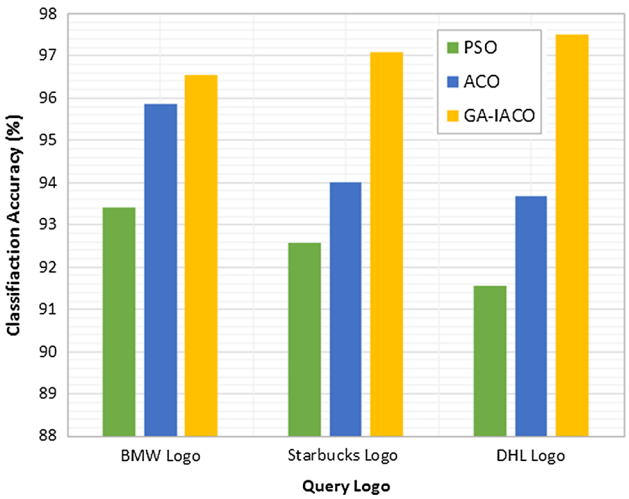


Fig. 7 Classification accuracy comparison

Fig. 6 Logos in database



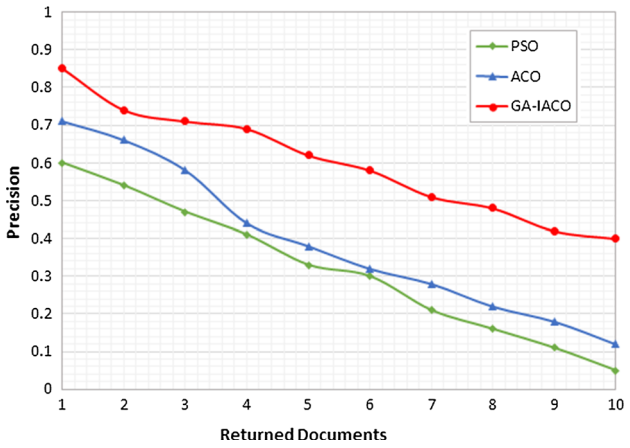


Fig. 8 Precision comparison for proposed model

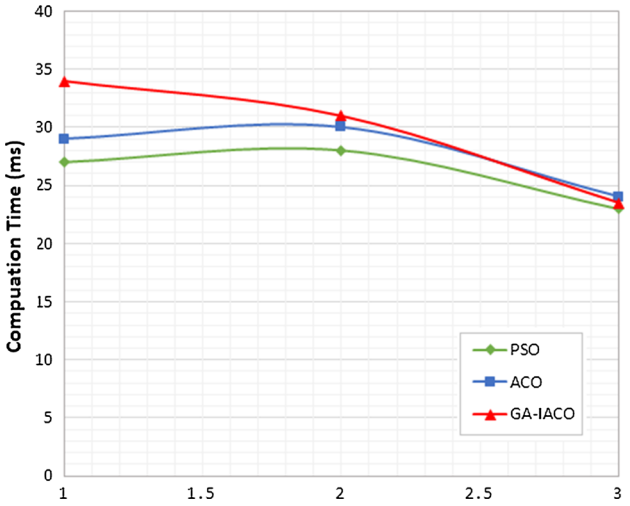


Fig. 9 Computation time comparison

inverted ant colony optimization. This improves the classification accuracy and retrieval efficiency in the document retrieval system. The further improvement of this research considers the mixed optimization model to improve the efficiency.

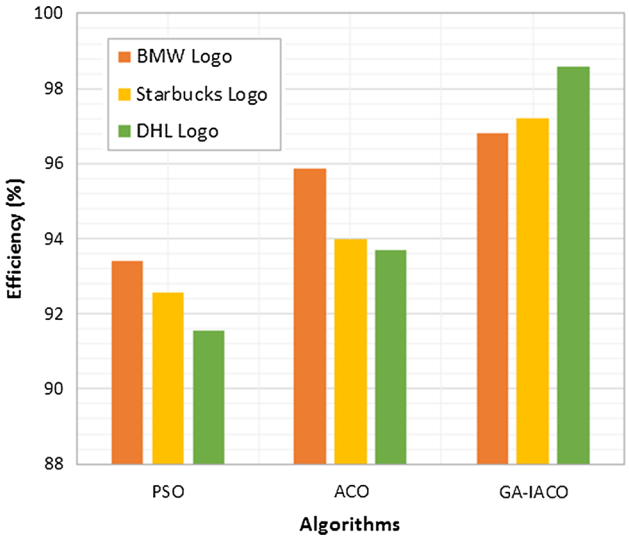


Fig. 10 Efficiency comparison

Compliance with ethical standards

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

References

Alaei A, Roy PP, Pal U (2016) Logo and seal based administrative document image retrieval: a survey. Comput Sci Rev 22:47–63

Alaei F, Alaei A, Pal U, Blumenstein M (2019) A comparative study of different texture features for document image retrieval. Expert Syst Appl 121:97–114

Bai C, Huang L, Pan X, Zheng J, Chen S (2018) Optimization of deep convolutional neural network for large scale image retrieval. Neuro Comput 303:60–67

Bianco S, Buzzelli M, Mazzini D, Schettini R (2017) Deep learning for logo recognition. Neuro Comput 245:23–30

Chaieb R, Kalti K, Luqman MM, Coustaty M, Amara NEB (2017) Fuzzy generalized median graphs computation: application to content-based document retrieval. Pattern Recognit 72:266–284

Costa R, dos Santos M, Machado AS (2018) Fuzzy information retrieval for document recommendation. Procedia Comput Sci 139:56–63

Table 2 Performance comparison of proposed model

Sample logo from Flickr dataset	Classification accuracy (%)			Computation time (ms)			Efficiency (%)		
	PSO	ACO	GA-IACO	PSO	ACO	GA-IACO	PSO	ACO	GA-IACO
BMW logo	93.42	95.86	96.55	27	29	34	93.42	95.86	96.8
Starbucks logo	92.56	94	97.1	28	30	31	92.56	94	97.2
DHL logo	91.56	93.68	97.5	23	24	23.5	91.56	93.68	98.6

Cristani M, Bertolaso A, Scannapieco S, Tomazzoli C (2018) Future paradigms of automated processing of business documents. Int J Inf Manage 40:67–75

Dang QB, Coustaty M, Luqman MM, Ogier JM, De Tran C (2018) New spatial-organization-based scale and rotation invariant features for heterogeneous-content camera-based document image retrieval. Pattern Recognit Lett 112:153–160

Deng D, Wang R, Hefeng W, He H, Luo X (2018) Learning deep similarity models with focus ranking for fabric image retrieval. Image Vis Comput 70:11–20

Dixit UD, Shirdhonkar MS (2018) Face-based document image retrieval system. Procedia Comput Sci 132:659–668

Djenouri Y, Belhadi A, Belkebir R (2018) Bees swarm optimization guided by data mining techniques for document information retrieval. Expert Syst Appl 94:126–136

Ge L-W, Zhang J, Xia Y, Chen P, Zheng C-H (2019) Deep spatial attention hashing network for image retrieval. J Vis Commun Image Represent 63:1–9

Haijiao X, Huang C, Wang D (2019) Enhancing semantic image retrieval with limited labeled examples via deep learning. Knowl-Based Syst 163:252–266

Hao W, Bie R, Guo J, Meng X, Wang S (2018a) Optimized CNN based image recognition through target region selection. Optik 156:772–777

Hao W, Li Y, Bi X, Zhang L, Wang Y (2018b) Joint entropy based learning model for image retrieval. J Vis Commun Image Represent 55:415–423

Joby PP (2020) Expedient information retrieval system for web pages using the natural language modeling. J Artif Intell 2(02):100–110

Nagy G (2016) Disruptive developments in document recognition. Pattern Recognit Lett 79:106–112

Pratheek VK, Vijaya Kantha V, Govindaraju KN, Guru DS (2016) Features fusion for classification of logos. Procedia Comput Sci 85:370–379

Qureshi R, Uzair M, Khurshid K, Yan H (2019) Hyperspectral document image processing: applications challenges and future prospects. Pattern Recognit 90:12–22

Tang P, Peng Y (2017) Exploiting distinctive topological constraint of local feature matching for logo image recognition. Neuro Comput 236:113–122

Yang S, Zhang J, Bo C, Wang M, Chen L (2019) Fast vehicle logo detection in complex scenes. Opt Laser Technol 110:196–201

Zhou M, Zeng X, Chen A (2019) Deep forest hashing for image retrieval. Pattern Recognit 95:114–127

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.