Hierarchy based firefly optimized k-means clustering for complex question answering

A. Chandra Obula Reddy, K. Madhavi
Department of Computer Science & Engineering, Jawaharlal Nehru Technological University, India

Article Info

Article history:

Received May 1, 2019
Revised Jul 2, 2019
Accepted Jul 16, 2019

Keywords:

Clustering
Complex question answering
Firefly optimization
Query segmentation

ABSTRACT

Complex Question Answering (CQA) is commonly used for answering community questions which requires human knowledge for answering them. It is essential to find complex question answering system for avoiding the complexities behind the question answering system. In the present work, we proposed Hierarchy based Firefly Optimized k-means Clustering (HFO-KC) method for complex question answering. Initially, the given input query is preprocessed. It eliminates the way of misclassification when comparing the strings. In order to enhance the answer selection process, the obtained keywords are mapped into the candidate solutions. After mapping, the obtained keywords are segmented. Each segmentation forms a new query for answer selection and various number of answers selected for each query. Okapi-25 similarity computation is utilized for the process of document retrieval. Then the answers selected are classified with K means clustering which forms the hierarchy for each answer. Finally the firefly optimization algorithm is used for selecting the best quality of answer from the hierarchy.

Corresponding Author:

A. Chandra Obula Reddy,
Department of Computer Science & Engineering,
Jawaharlal Nehru Technological University,
Ananthapur, Ananthapuramu - 515002, A.P., India.
Email: chandrajntuanantapur@gmail.com

1. INTRODUCTION

Semantic information published on the web is increased rapidly with linked data initiative. However it is typically complex for the user to search and query the vast amount of structured and heterogeneous semantic data [1]. It is essential to build a system which can able to answer from different domain. It is termed as open domain question answering system which should be access the knowledge in novel way [2]. When concern about the stored data, the volume is high and it increases the burden of filtering and browsing the result for retrieving precise information. Question answering system is a technology used to find, extract, and provide a proper answer to the user's query in the natural language format [3]. The repositories are specially made for accomplishing several tasks like question answering, knowledge mining and searching [4]. Data mining is a subfield of computer science that enables intelligent extraction of useful information [5].

Due to its large and growing structure of data, efficient and intuitive techniques are essential to deal with them. The complexity and ease of interference is taken into account while processing the data [6]. Instead of knowing the query language, the knowledge graph extracts the structure and relation between the question and answer [7]. In addition with collaborative information seeking and sharing, collaborative answers are also included. The community agreements among Question Answering (QA) pairs are obtained with micro collaboration and the enhancement of collective intelligence [8]. The keywords from the query are matched with the metadata in which sequence of answers are retrieved for the given query.

The semantic question answering system was developed in which uncertain words are the question. The fuzzy based ontology system is developed by the researchers in the text extraction level. The characteristics of data are analyzed to check the possibility of solving frequently posed questions [9]. The search facility is the main feature of CQA services which permits the members to search their archives. Normally, the information retrieval approaches are developed in which the member can construct and send arbitrary collection of questions until the old question for the current need is obtained [10]. The reuse of past QA pairs provides the benefit of enhancing user experience [11].

The efficiency of processing natural language questions are improved while heterogeneous data is utilized as an answer source. The usage of unique source is not straightforward because of pattern variation [12]. When mapping the question with the semantic content of knowledgebase, depth information is required [13]. Group based recommendations are developed with two techniques namely aggregation of interesting profile and aggregation of recommendation list [14]. The terminology used in NL question varies from the terminology used in knowledge base. The solution for conceptual disambiguation is essential for searching the matches from homogeneous or heterogeneous resources [15].

The machine learning paradigms are developed recently for classifying, organizing and extracting relevant information. Even though, the question classification is more accurate, it is required to make the QAS comprehension more understandable for easily obtaining the correct answer [16, 17]. It faces the difficulties such as linguistic gap between the documents and search queries and the unavailability of recently posed questions. Hence it is not possible for searching CQA achieves for obtaining web queries [18]. The similarity between question and matching words provide the extraction features for top ranked answer [19].

The outline of this paper is described as follows. Section 2 briefly explains the proposed method of complex question answering system. Section 3 describes the Research Method, Hierarchy based Firefly Optimized k-means Clustering (HFO-KC). In Section 4, the experimental results are analyzed. Section 5 discusses the significant aspects of the work and concludes.

2. THE PROPOSED METHOD

In the proposed method of complex question answering system, initially the input query is preprocessed. After preprocessing, the keywords are obtained and they are segmented. For each segment, number of answers are extracted.

In order to select the correct answer for the given input query, the collected answers are classified with k means clustering and the best answer is selected using firefly optimization algorithm. K-Means is one of the promising and effective clustering algorithm [20]. Clustering plays a wide role in the recent development of computer science [21]. In Machine learning, supervised learning known as classification and unsupervised learning known as clustering [22]. The flow diagram of proposed CQA system is shown in Figure 1.

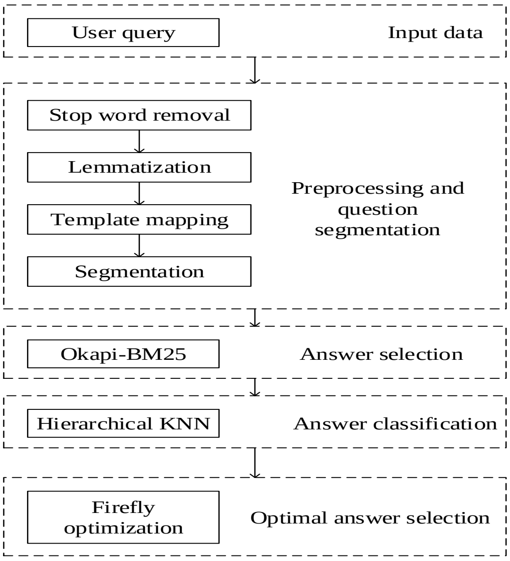

Figure 1. Optimal hierarchy based k means clustering for complex question answering

2.1. Preprocessing and Question Segmentation

The preprocessing can be applied to the input query and the collection of documents. Initially, the individual keywords are extracted and the stop words are removed. After stop word removal, word lemmatization is applied for the remaining keywords. Each keyword is mapped with its corresponding templates. After preprocessing, the input keywords contain n tuples $Q=\{a_1,a_2...a_n\}$. Each keyword is mapped with set of templates denoted as $a=\{t_1,t_2...t_m\}$. Then the templates are grouped to form segments and the number of answers are selected for each segment. The block diagram for preprocessing and question segmentation is shown in Figure 2.

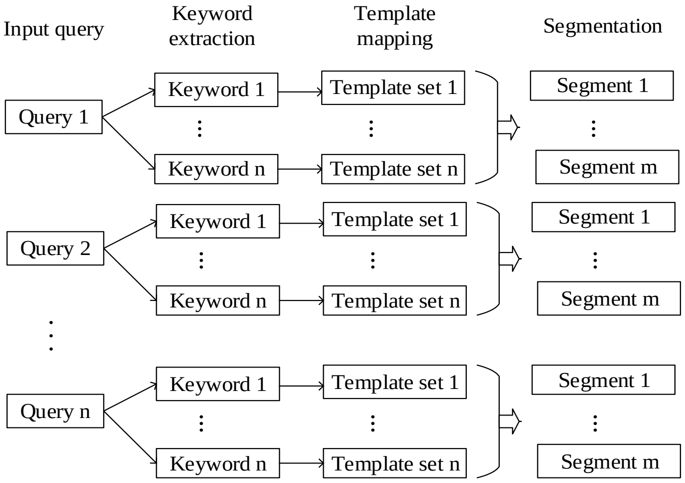

Figure 2. Preprocessing and question segmentation

2.2. Answer Selection for the Given Input Query

For the segmented questions, the answer is selected based on the Okapi-BM25 score. The score is computed for each answer. It selects initial set of relevant answers based on similarity and it can process efficiently than cosine similarity measurement. By using the following formula, the Okapi-BM25 score is computed:

$$Okapi(Q,A_i)=\sum_{t\in Q\cap A_i} w\frac{(b_1+1)atf}{B+atf}\times\frac{(b_3+1)qtf}{b_3+qtf} \qquad (1)$$

Where, Q represents the query, A represents the answer for the given query, qtf is the question term frequency, atf is the answer term frequency, and b_1,b_3 represents the constant parameters. The value of B is computed as:

$$B=b_1(1-c)+\left(c.\frac{al}{aval}\right) \qquad (2)$$

Where, c represents the constant parameter, al represents the answer length and $aval$ represents the average answer length. The weight value used in equation (1) is defined as

$$w=\log\frac{(P-p+0.5)}{(p+0.5)} \qquad (3)$$

Where, P is the number of answers, p represents the number of answers having term t. The top relevant answers are selected from the document based on Okapi-BM25 method [23]. These top relevant documents are utilized for further processing in terms of document classification.

3. RESEARCH METHOD

This section describes Hierarchy based Firefly Optimized k-means Clustering (HFO-KC) method for complex question answering.

3.1. Answer Classification with hierarchical K-Nearest Neighbor

By using the score obtained, hierarchical K-Nearest Neighbor (KNN) is utilized in which the top relevant answers are separated. The answer with highest and lowest score is separated as different groups. The KNN classification can be accomplished based on the centroid score and each time the new group is formed hierarchically. KNN is one of the simplest and popular supervised learning algorithm for classification [24]. The input data taken by the KNN are the input value k and the collection of answers used for classification. The classification problem is solved by the number of nearest neighbors which are taken for the input parameter k. It is the straightforward approach for classification. For each group the k nearest neighbors are computed based on the centroid value. Initially the answers are randomly divided into two groups. From each group, the centroid value is chosen based on the score. Then the distance between the centroid value and the remaining tuples are computed. The tuples are added into the group which produces less distance when compared with the other group. The KNN algorithm for answer classification is described as follows.

```
Algorithm 1: KNN algorithm for answer classification
Input: Answer collection with score, k value
Output: classified set of answers
Step1: The score from each answer is taken into consideration for answer selection.
Step 2: The answers are divided into k groups randomly
Step 3: Select centroid from each group.
Step 4: For each answer, compute the distance between the answer and centroid.
Step 5: The answer is added with the group which produce minimum distance when compared with
the other groups.
Step 6: Similar to that all the answers are added to the relevant group.
Step 7: After dividing into k groups again the centroid value is selected and new group is formed.
Step 8: The process is repeated until the centroid is same for the proceeding iterations.
```

The collection of answers can be considered as a data point in n dimensional space. The number of attributes are denoted as n. In order to compute the distance between two data points the Euclidean distance is used. The Euclidean distance between data points x and y is calculated as

$$d = \sqrt{\sum_{1 \le i \le n} (x_i - y_i)^2} \tag{4}$$

Where, n represents the number of attributes in data set x_i and y_i are values of attribute i in data tuples x and y respectively. Instead of using Euclidean distance, Minkowski distance and Manhattan distance also be used. The simplest case of this algorithm is attained with setting the value of k to one. The specific property of this algorithm is predicting the continuous valued attributes instead of using categorical attributes.

3.2. Optimized Answer Selection with Firefly Algorithm

After grouping the answers, the accurate answer relevant to each query is selected based on firefly optimization. It is a meta-heuristic algorithm for finding optimal solution for the optimization problem. The concept behind this firefly optimization is the flashing behavior of each firefly. Set of assumptions were made for this firefly optimization. They are

a) It is assumed that all fireflies can be attracted by the other fireflies.
b) The attractiveness is represented by its brightness. The firefly which has lower brightness is attracted by the firefly which has higher brightness.
c) The fireflies having same brightness are moved randomly.

The attractiveness of a firefly is calculated using following function:

$$\beta(r) = \beta_0 . e^{-\gamma, r^2} \tag{5}$$

Where, β_0 is the attractiveness of the firefly when $r = 0$ and γ is light absorption coefficient. The firefly's movement totally depends on its attractiveness. Firefly i would move towards firefly j if and only the attractiveness of the firefly j is greater than that of firefly i. In that case, the movement is shown by following formula:

$$x_{ik} = x_{ik} + \beta_0 , e^{-\gamma , r_{ij}^2} .(x_{ik} - y_{jk}) + \alpha .S_k (rand_{ik} - 0.5) \tag{6}$$

x_{ik} and y_{jk} are values of attribute k. k takes values from $1,2,...n$, where n is the dimension of the data set. $rand_{ik}$ is a random number between 0 and 1. α is called randomization parameter which will decide how much to move and takes value between $0 \& 1$, S_k is scaling parameter which is calculated for each attribute. S_k is calculated as

$$S_k = |u_k - l_k| \tag{7}$$

u_k and l_k are the upper bound and lower bound of the attribute k respectively. r_{ij} is the distance between the fireflies i and j which calculated from:

$$r_{ij} = \sqrt{\sum_{1 \le i \le n} (x_i - y_i)}^2 \tag{8}$$

The value of attractiveness in optimization problems is calculated using an objective function. The algorithm for standard firefly algorithm is given below:

```
Algorithm 2: Firefly optimization
Input: Objective function $f(x)$ and algorithm parameters $\alpha_0, \beta_0$, and γ
Output: Minimized function value position
Step 1: Initialize firefly population $p$ randomly.
Step 2: Initialize algorithm parameters $\alpha_0, \beta_0$, and γ.
Step 3: Calculate fitness value using the objective function $f(x)$ for each firefly.
Step 4: while $t < \max generation$
    for $i = 1 : p$
        for $j = 1 : i$
            if ( $f(x_j) < f(x_i)$
                move firefly $i$ towards $j$ using (3)
                calculate fitness value again of all fireflies
            end if
        end for
    end for
end while
Step 5: Rank the fireflies to find the current best firefly.
```

In present paper, the preprocessing can be accomplished initially and it makes easier for further processing. After preprocessing the relevant answers are collected and they are classified with KNN classifier. Finally, in order to improve the classification accuracy and for finding the correct answer, the optimization algorithm firefly is used. In this CQA system, the complexity of the processing is reduced with the help of simplest algorithm. When compared with the existing literatures, the trade-off between complexity and accuracy can be attained.

4. RESULTS AND DISCUSSION

The Factoid Q&A Corpus is used as a dataset in our work for complex question answering [25]. It consists of 1,714 factoid questions which are created manually. The answer for the question is collected from Carnegie Mellon University and University of Pittsburgh in between 2008 and 2010. For KNN algorithm the K value is defined as 2 and the constant parameters are $b_1 = 1.2$, $c = 0.75$ and $b_3 = 7.0$ .The proposed HFO-KC is compared with the existing approaches such as JAIST, ICRC and RCNN [26]. The performance metrics such as precision, recall, f-measure, accuracy and complexity are evaluated for the proposed approach and compared with the existing approaches. The improved performance of the proposed approach shows the efficiency of the technique.

4.1. Precision

Precision computes the correct prediction of positive observations from the total number of predictions with positive observations. The performance comparison of the proposed CQA is shown in Figure 3 and Figure 4. In Figure 3, the precision value is compared by varying the number of documents to 300, 500, 700 and 1000. The precision value is reached near 1. That is near optimal performance is obtained with our proposed method. When the numbers of documents are 300, the precision values obtained for the existing methods are 0.58, 0.57, 0.57, and 0.59. For 500 documents, the precision values are 0.55, 0.56, 0.55 and 0.58. The number of documents are increased to 700 and 1000 then the existing precision values are 0.54, 0.53, 0.54, 0.56 and 0.53, 0.52, 0.53, 0.55. But in case of proposed algorithm the precision value is improved as 0.99, 0.98, 0.96 and 0.94 for the number of documents 300, 500,700 and 1000. The average precision values computed by RCNN, ICRC, JAIST, A-ARC I and HFO-KC are 0.545, 0.545, 0.5475, 0.57 and 0.9675 as shown in Figure 4. The improved precision values shows the efficiency of the proposed approach.

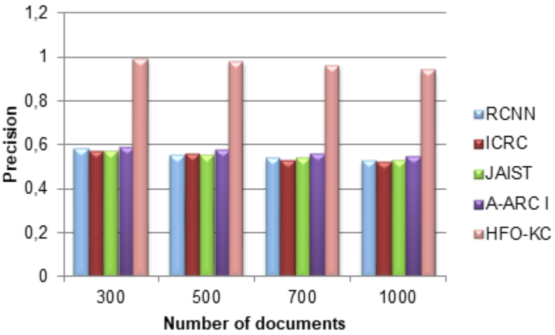

Figure 3. Precision comparison for varying the number of documents

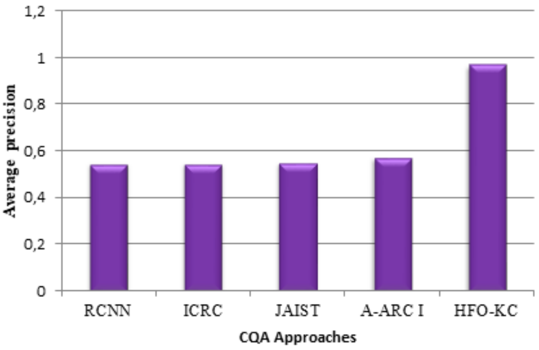

Figure 4. Average precision values for various CQA approaches

4.2. Recall

Recall computes the correctly predicted positive observations from the total number of observations. The recall values obtained by RCNN are 0.56, 0.55, 0.54 and 0.53 for the number of documents 300, 500,700 and 1000. For ICRC these values are 0.56, 0.5, 0.53 and 0.52, JAIST produces the precision values as 0.57, 0.565, 0.56 and 0.555. The existing A-ARC I have the precision values 0.58, 0.57, 0.56 and 0.55. For our proposed CQA system, the recall values produced are 0.93, 0.9, 0.89 and 0.87 as shown in Figure 5. The average recall values computed by RCNN, ICRC, JAIST, A-ARC I and HFO-KC are 0.545, 0.53, 0.55, 0.57 and 0.9 as shown in Figure 6. When the numbers of answers are increased, then the recall value is reduced. For less number of answers, the recall value obtained is high.

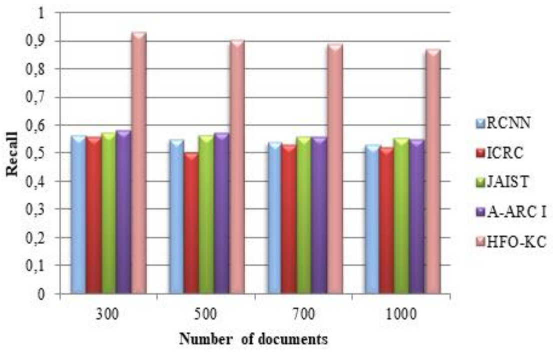

Figure 5. Recall comparison for varying the number of documents

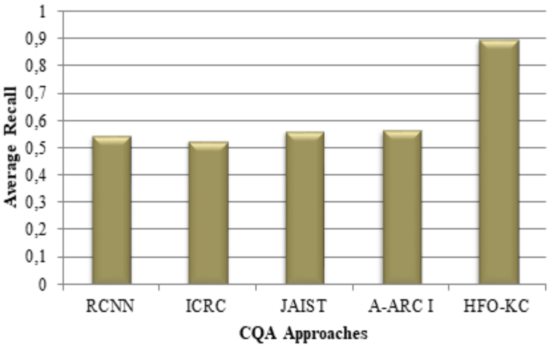

Figure 6. Average recall values for various CQA approaches

4.3. F-measure

The weighted average between precision and recall is termed as f-measure. For 300 documents, the RCNN, ICRC, JAIST, A-ARC I and HFO-KC have the f-measure values 0.55,0.56, 0.57, 0.58 and 0.957. For 500 documents, the RCNN, ICRC, JAIST, A-ARC I and HFO-KC have the f-measure values 0.55, 0.555, 0.56, 0.57 and 0.941. For 700 documents, the RCNN, ICRC, JAIST, A-ARC I and HFO-KC have the f-measure values 0.545, 0.5, 0.55, 0.56 and 0.957. For 1000 documents, the RCNN, ICRC, JAIST, A-ARC I and HFO-KC have the f-measure values 0.54, 0.45, 0.54, 0.55 and 0.9 as shown in Figure 7. The average f-measure values obtained by RCNN, ICRC, JAIST, A-ARC I and HFO-KC are 0.548, 0.516, 0.555, 0.565 and 0.928 as shown in Figure 8. The f-measure values obtained by the proposed method is high when compared with the other existing approaches.

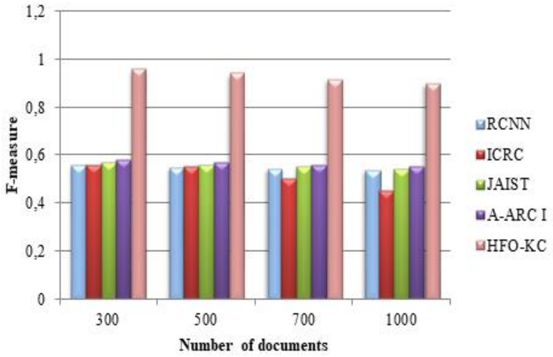

Figure 7. F-measure comparison for varying the number of documents

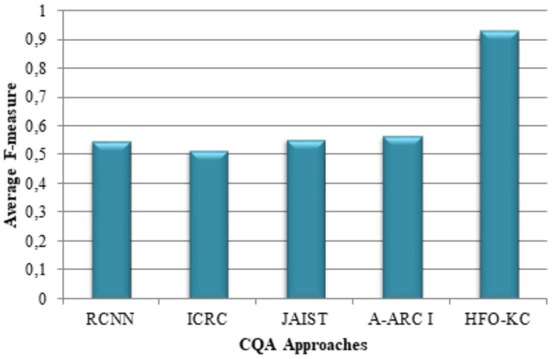

Figure 8. Average F-measure for various CQA approaches

4.4. Accuracy

Accuracy computes the correct observations from the total number of observations. The accuracy of the proposed approach is evaluated and compared with the existing approaches. When compared with the existing approaches, the accuracy of the proposed technique is high. The accuracy value obtained for RCNN, ICRC, JAIST, A-ARC I and HFO-KC is 0.72, 0.68, 0.72, 0.76, and 0.991 for 300 documents. When the documents are 500, the accuracy value obtained for RCNN, ICRC, JAIST, A-ARC I and HFO-KC is 0.71, 0.67, 0.715, 0.75 and 0.982. For 700 documents, RCNN, ICRC, JAIST, A-ARC I and HFO-KC produces 0.705, 0.665, 0.71, 0.74 and 0.972. By increasing the number of documents to 1000, the accuracy is 0.7, 0.65, 0.7, 0.73 and 0.962. The improved performance is obtained with our proposed approach as shown in Figure 9. The Average accuracy for various CQA approaches as shown in Figure 10.

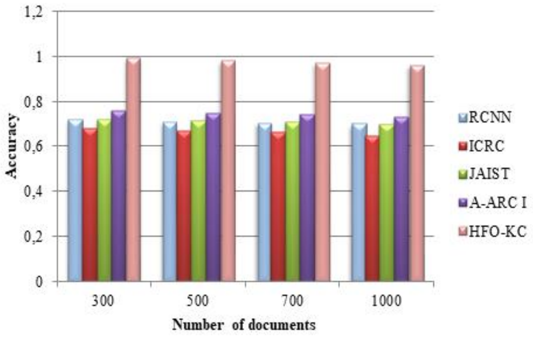

Figure 9. Accuracy comparison for varying the number of documents

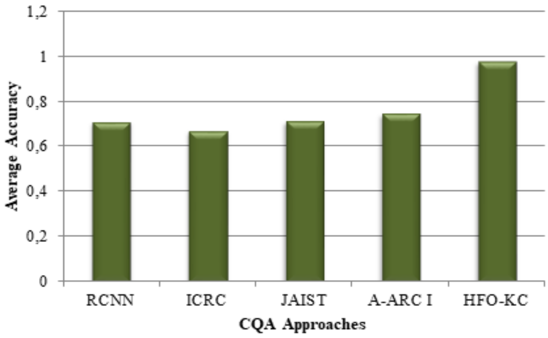

Figure 10. Average accuracy for various CQA approaches

The complexity of the proposed HFO-KC complex question answering system has the complexity of $O(ndk)+O(m^2t)$. Where, d represents the dimension of each answer, n represents the cardinality of the document, m represents the population size and t is the number of iterations and k represents the number of groups used on KNN algorithm. The computation time for the proposed work is 15ms.The proposed HFO-KC approach for complex question answering can be evaluated with the performance metrics like precision, recall, accuracy, f-measure and complexity. When compared with the existing approaches, the performance of the proposed approach is high. The proposed approach provides the trade-off between complexity and accuracy.

5. CONCLUSION

In this paper, initially the input query is preprocessed. It includes stop word removal and word lemmatization. Then individual keywords are extracted from the query and the extracted keywords are segmented. The process of segmentation is accomplished with the collection of keywords. The candidate solutions are mapped from the obtained keywords. The correct answer is retrieved from the database using the segmented query. It can be obtained with Okapi-25 similarity computation. Based on the similarity score, the large number of answers are selected for the given question. Then the selected answers are clustered with K means clustering in which it eliminates the incorrect answer selection. The hierarchy is formed with the algorithm which simplifies the process of answer selection. From the hierarchy, the optimized result is obtained with firefly optimization.

REFERENCES

[1] Bordes A, Weston J, Usunier N. Open question answering with weakly supervised embedding models. *In Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. 2014; 8724: 165-180.

[2] Lopez V, Unger C, Cimiano P, Motta E. Evaluating question answering over linked data. *Journal of Web Semantics.* 2013; 21: 3-13.

[3] Utomo FS, Suryana N, Azmi MS. New instances classification framework on Quran ontology applied to question answering system. *Telecommunication Computing Electronics and Control (TELKOMNIKA).* 2019; 17 (1): 139-146.

[4] Zhang K, Wei W, Haocheng W, Zhoujun L, Zhou M. Question retrieval with high quality answers in community question answering. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. 2014; 371-380.

[5] Adekitan AI, Adewale A, Olaitan A. Determining the operational status of a three phase induction motor using a predictive data mining model. *International Journal of Power Electronics and Drive System (IJPEDS).* 2019: 10 (1): 93-103.

[6] Unger C, Freitas A, Cimiano P. An introduction to question answering over linked data. *Reasoning Web*, 2014; 8714:100-140.

[7] Lukovnikov D, Fischer A. Neural network-based question answering over knowledge graphs on word and character level. In *Proceedings of the 26th international conference on World Wide Web*. 2017; 1211-1220.

[8] John BM, Kurian JC, Chua AY, Goh DHL, Lien NH. Social question answering: Analyzing knowledge, cognitive processes and social dimensions of micro-collaborations. *Computers & Education.* 2013; 69: 109-120.

[9] Mans RS, Aalst WMP, Vanwersch RJB, Moleman AJ. Process mining in healthcare: Data challenges when answering frequently posed questions. *Springer-Verlag Berlin Heidelber.* 2013; 140-153.

[10] Niemelä J. Ecology of urban green spaces: The way forward in answering major research questions. *Landscape and Urban Planning.* 2014; 125: 298-303.

[11] Figueroa A, Neumann G. Context-aware semantic classification of search queries for browsing community question–answering archives. *Knowledge-Based Systems.* 2016; 96: 1-13.

[12] Liu K, Zhao J, Shizhu H, Zhang Y. Question answering over knowledge bases. *IEEE Intelligent Systems*. 2015; 5: 26-35.

[13] Sharef NM, Noah SAM, Murad MAA. Issues and Challenges in Semantic Question Answering through Natural Language Interface. *Journal of Next Generation Information Technology.* 2013; 4(7): 50-60.

[14] Liu DR, Chen YH, Huang CK. QA document recommendations for communities of question–answering websites. *Knowledge-Based Systems*. 2014; 57: 146-160.

[15] Hazrina S, Sharef NM, H. Ibrahim, Murad MAA, Noah SAM. Review on the advancements of disambiguation in semantic question answering system. *Information Processing and Management*. 2017; 53: 52-69.

[16] Islam MS, Liu C, Li J. Efficient answering of why-not questions in similar graph matching. *IEEE Transactions on Knowledge and Data Engineering*. 2015; 27: 2672-2686.

[17] Gharehchopogh FS, Lotfi Y. Machine learning based question classification methods in the question answering systems. *International Journal of Innovation and Applied Studies*. 2013; 4 (2): 264-273.

[18] Figueroa A. Automatically generating effective search queries directly from community question-answering questions for finding related questions. *Expert Systems With Applications.* 2017; 77: 11-19.

[19] Sarrouti M, Ouatik SEA. A passage retrieval method based on probabilistic information retrieval model and UMLS concepts in biomedical question answering. *Journal of Biomedical Informatics*. 2017; 68: 96-103.

[20] Mahboub A, Arioua M, En-Naimi EM. Energy-efficient hybrid k-means algorithm for clustered wireless sensor networks. *International Journal of Electrical and Computer Engineering (IJECE)*. 2017; 7 (4): 2054-2060.

[21] Girsang AS, Cenggoro TW, Huang KW. Fast Ant Colony Optimization for Clustering. *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*. 2018; 12(1): 78-86.

[22] Azim MA, Bhuiyan MH. Text to emotion extraction using supervised machine learning techniques. *Telecommunication Computing Electronics and Control (TELKOMNIKA)*. 2018; 16 (3): 1394-1401.

[23] Amoli PV, Sh OS. Scientific Documents clustering based on Text Summarization. *International Journal of Electrical and Computer Engineering (IJECE)*. 2015; 5 (4), 782-787.

[24] Mustakim, NK Sari, Jasril, Kusumanto I, Reza NGI. Eigenvalue of Analytic Hierarchy process as the Determinant for Class Target on Classification Algorithm. *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*. 2018; 12 (3): 1257-1264.

[25] Yang Z, Jones I, Hu X, Liu H. Finding the right social media site for questions. In *Proceedings of the 2015 IEEE/ACM International Conference*. 2015; 639-644.

[26] Xiang Y, Chen Q, Wang X, Qin Y. Answer Selection in Community Question Answering via Attentive Neural Networks, *IEEE*, 2017; 24 (4): 505-509.