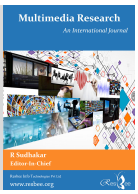
Vol.4 No.4 Oct 2021

Multimedia Research

Received 24 August, Revised 24 September, Accepted 14 October



Query Indexing and Cluster-based Indexing Model for the Document Retrieval

Sathish Vuyyala

Assistant Professor, Department of CSE,
MVSR Engineering College, Hyderabad, Telangana, India

Abstract: In the research community field, query optimization plays an important role to retrieve the important and the appropriate documents on the basis of query indexing. In the documents, using the query retrieval process the information is retrieved on the basis of the distance measured. Although several methods are present in the query processing scheme as well as indexing, extracting the matched as well as appropriate documents still outcomes in numerous confronts in the research community. Hence, to retrieve the appropriate documents competently an effective cluster-based inverted indexing model is adopted. By exploiting stop word removal and stemming approaches, unnecessary and redundant words are removed. By cluster-based inverted indexing approach, document indexing is carried out that is the integration of Possibilistic fuzzy c-means (PFCM) clustering approach to index the documents. For user queries, such as multigram queries or semantic queries, on basis of Bhattacharyya distance to generate an enhanced query outcome, query matching is processed. By exploiting the Pearson correlation coefficient, the query optimization is carried out and the appropriate documents are retrieved efficiently. The achievement of a developed cluster-based indexing approach is carried out in this paper. The developed cluster-based indexing approach performance is calculated by exploiting measures, namely precision, recall, as well as F-measure.

Keywords: Cluster-based, Documents, Indexing, Information Retrieval, Query Optimization.

Nomenclature

Abbreviations	Descriptions
PFCM	Possibilistic fuzzy c-means
FCA	Formal Concept Analysis
DSS	Decision Support System
CBDR	Content-based Document Retrieval
FA	Firefly Algorithm
SEMINDEX	Semantic indexing query framework
FGMG	Fuzzy Generalized Median Graph
TF-IDF	Term Frequency-Inverse Document Frequency
FCM	fuzzy c-means
MBO	Monarch Butterfly Optimization
BPC	Bits Per Character
FARG	Fuzzy Attributed Relational Graph
CBDR	Content-based Document Retrieval
EM	Expectation Maximization algorithm

1. Introduction

Because of the fast Internet advancement, there are enormous numbers of documents that are saved either in an electronic format or hard copy, to be exploited daily on the basis of the requirements [1]. For the last few years, an attempt was carried out, for the advancement of a combined document filling as well as retrieval system that aids in several fields namely categorization, classification, and retrieval, storage as well as document reproduction. Additionally, it aids in extracting, browsing, synthesizing as well as data retrieving from several documents of a known application domain. Users are required to present details of document content to retrieve documents as well as data from these systems. Because of

https://doi.org/10.46253/j.mr.v4i4.a2

this motivation, the efficacy and the effectiveness of retrievals as well as document of data are affected, as systems do not regard user knowledge regarding documents [2].

On the basis of the FCA, the Concept of lattice-based retrieval techniques is a kind of unsupervised classification that presents a deliberate explanation for clusters that gives enhanced consideration. Using the FCA concept lattice is generated as well as shown its utility in document indexing as well as navigation approach in IR domain. For example, to drive transformation amid query representation as well as each document representation concept lattice can be exploited and the navigation is provided in conceptual document space. In the meantime, few techniques were presented to attain semantic information amid formal concepts. These techniques only regard whether terms happen in documents as well as queries, other than concerning all terms evenly might considerably minimize the quality of the retrieved results since diverse terms might have diverse significance degrees for those documents as well as queries. By means of fuzzy information, this kind of issue can be undertaken [3].

There are various types of document retrieval issues [11]. The majority fundamental one, document listing, aspires to retrieve all documents wherein a pattern emerges. By exploiting the natural language text collections which act as a fundamental action in the web as well as trade search engines, document retrieval is done [4]. By exploiting inverted index variants, the issue of this circumstance is resolved which is a hugely doing well approach. These disadvantages are not included as an issue in several instances as "natural language text collections are indexed". Additionally, they construct utilization of easy index organization that is more capable and scalable, as well as referred as an explanation to the achievement in Web-scale information retrieval [12]. In a few cases, these restrictions affect the employ of the inverted index in several other kinds of string collections, in that text segregated into words as well as restriction of questions to word series are also hard or pointless [5].

The major objective of this research is to apply the removal of stop words as well as the stemming approaches in the documents to evade unnecessary as well as redundant words. By exploiting the interactive query optimization on the basis of the Pearson correlation coefficient appropriate documents are efficiently retrieved. To produce document indexing on basis of keywords, a cluster-based inverted indexing approach that combines the PFCM clustering approach, as well as inverted indexing, is exploited.

2. Literature Review

In 2019, Hector Ferrada and Gonzalo Navarro [1], worked on the retrieval document structure index, which was a string documents collection of string documents. To retrieve the documents, which were appropriate to query strings, a document listing retrieves all documents. Generally, traditional structures exploit extra space. The majority of recent research exploits compressed suffix arrays, however, speed indices still exploit 17–21 BPC, wherein small ones acquire milliseconds per returned answer. In this paper, initial document retrieval structures on the basis of the Lempel–Ziv compression, precisely LZ78 were presented. In 2021, Raymon van Dinter et al [2], worked on open-source DSS which aids the dataset preprocessing, document retrieval step, as well as citation classification. The DSS was domain-independent; as it was shown to cautiously choose an article's significance based exclusively on title as well as abstract. In addition, for reviewers, the DSS was modeled to run in the cloud without any necessary programming knowledge. The Multi-Channel Convolutional Neural Network model was simulated to aid the citation screening course. Reviewers can fill in the research scheme as well as manually label only a subset of citations with the provided DSS. The residual unlabeled citations were mechanically classified and sorted on the basis of the probability. In 2019, Mamta Kayest and Sanjay Kumar Jain [3], developed a document retrieval model by exploiting an optimization approach named MB-FF, which was a combination of the MBO and FA. By exploiting the stop word removal as well as stemming techniques, from the documents, the keywords were recognized from the pre-processed document. The TF-IDF was exploited in the keyword extraction as well as holoentropy concept was exploited in the important keywords selection. The chosen keywords make sure the appropriate documents retrieval that at first was processed via cluster-based indexing by exploiting the MB-FF, which was pursued with the two-level mod-Bhattacharya distance match. In 2017, Ramzi Chaieb et al [4], developed a novel approximate approach for the FGMG based on FARG computation which embeds in an appropriate vector space to detain the utmost information in graphs and to enhance accurateness as well as document image retrieval processing speed. Here, FGMGs to the CBDR application issue were focused on. In 2018, Shufeng Hao et al [5], developed an idea coupling relationship analysis technique to aggregate as well as learn the intra-as well as inter-concept coupling relationships. The widespread terms of proper concepts were exploited by intra-concept coupling relationships to identify explicit semantics of proper concepts. The formal concept's partial order relation was adopted by the inter-concept coupling relationship to capture the implicit dependency of formal concepts. A concept lattice-

based retrieval framework was proposed on basis of the concept coupling relationship analysis model. This model indicates documents, as well as user, queries in a concept space on the basis of the fuzzy formal concept analysis.

3. Adopted Model for Cluster-Based Inverted Indexing

The query plays an important role in indexing as well as retrieving documents on basis of the query in the document processing as well as retrieval field. Fig. 1 exhibits the architecture diagram of the adopted model. The adopted cluster-based indexing approach includes 4 phases namely pre-processing, complex query matching, document indexing, as well as query optimization. The database comprises numerous documents, as well as these input documents, are fed to pre-processing phase, by stop word removal as well as stemming approaches the needless, as well as the redundant words, are removed. Further, the pre-processed documents are subjected to a document indexing phase that uses a clustering-based indexing approach called cluster-based inverted indexing via integrating inverted indexing with the PFCM clustering approach to index documents. In the complex query phase, the clustered documents are subjected. For user queries namely multigram queries else semantic queries query matching is done on basis of "Bhattacharyya distance" to generate a superior query matching outcome. Based on a minimum distance measure, to identify appropriate documents the Bhattacharya distance is exploited. At last, by exploiting the Pearson correlation coefficient, the query optimization is carried out based on the interactive query optimization that ascertains an efficient manner to retrieve the documents

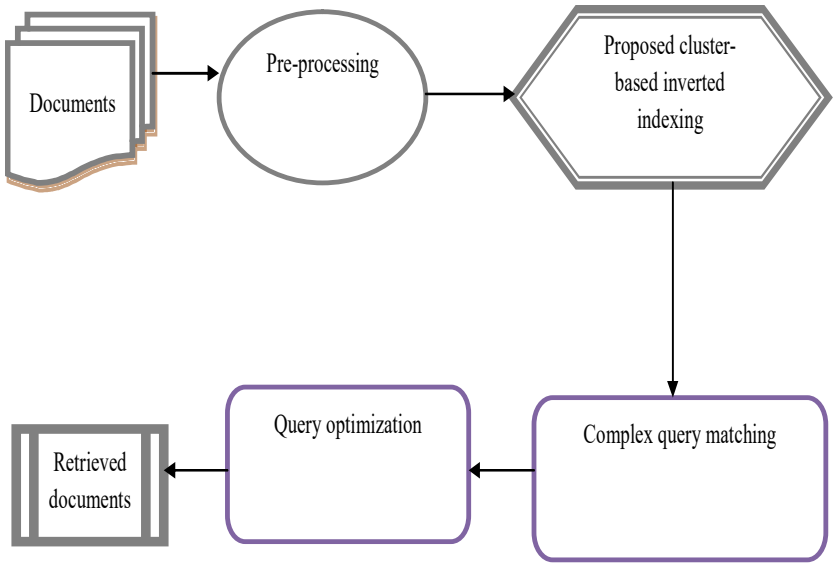


Fig.1. *Architectural diagram of the adopted model*

3.1 Pre-Processing

The database comprises several documents, as well as each of these input documents, is pre-processed by exploiting stemming approaches as well as stop words to evade removing unnecessary as well as redundant words. To minimize the user seeking time, the stemming process, as well as stop word removal, is used. At first, to get rid of the duplicate and the unnecessary words, the input documents are pre-processed, and subsequently, to retrieve the optimal information, the indexing and the clustering are used in the pre-processed documents. Each and every document transmitted to the indexing process has to be pre-processed previous to it being transmitted to the subsequent phase. Nevertheless, the query matching criterion is not used for the documents with the redundant words to produce the retrieval outcomes. In addition, documents with no related information might also subsist in the database. Hence, to carry out the effectual documents retrieval it is needed be pre-processed documents. The performance of the retrieval is affected by the retrieval model without appropriate pre-processing. The indexing operation is used for the pre-processed documents that exclusively recognized the matched documents on basis of query keywords. The stop word is used for the pre-processing phase, as well as the stemming approaches reduce efficiently the unnecessary words in documents.

3.2 Document Indexing

In the document indexing phase, the ensued pre-processed documents are subjected. The adopted cluster-based inverted indexing approach is carried out for the document indexing which is the integration of the clustering approach as well as the inverted indexing. The clustering is performed by the adopted cluster-based indexing approach as well as on the basis of the appropriate document information the indexing process is performed. For the simple retrieval of data, documents are indexed on basis of query keywords. By exploiting the piFCM [8] clustering approach, the documents are clustered on the basis of the information present in documents. The same documents are grouped collectively by the clustering; therefore the documents with the same features are clustered in a similar cluster as well as subsist in diverse count of cluster groups else, in other words, here each cluster group comprises several documents however the information in document should be same while it is clustered in the similar cluster group.

3.2.1 piFCM Clustering Approach

The documents with appropriate contents are clustered in the clusters. On basis of the document's content, in several cluster groups documents might subsist. Based on the matching documents, a diverse cluster group comprises diverse documents and it will be retrieved. For the user query, to retrieve the information in an easy manner "the documents are clustered in a group [10]". Every group cluster group comprises any count of data objects with appropriate information. The adopted cluster-based inverted indexing exploited the PFCM approach that is on the basis of the FCM algorithm to efficiently cluster data objects mutually. PFCM is a clustering approach that improves indexing performance by exploiting the objective function of membership data.

Let unlabeled data sets represents as $Y = \{y_1, y_2, \ldots y_n\} \in R^p$ ($p = n \times s$) is clustered into a fuzzy subset of cl ($1 < cl < n$) clusters [9].

$$J_{m,T}(U,T,V;Y) = \sum_{k=1}^{c}\sum_{k=1}^{c}\left(\alpha u_{i,k}^{m} + \beta t_{i,k}^{\tau}\right)d_{i,k}^{2} \sum_{i=1}^{c}\delta_i \sum_{i=1}^{c}(1-t_{ik})^{T} \quad (1)$$

The steps for the proposed PFCM approach are stated below:

Step 1: Fix $m > 1$, $\tau > 1$, and $1 < cl < n$. Choose $v^{(0)} \in R^s, v^{(0)} Rs$, as well as v(0), can be selected arbitrarily from $Y = \{y_1, y_2, \ldots y_n\} \in R^p$. Subsequently at step l, l = 1; 2; · · ·

Step 2: Compute a fuzzy membership degree $u_{i,k}^{(l)}$ that reduces the objective model $J_{m,T}$ is stated as follows:

$$u_{i,k}^{(l)} = \left(\sum_{j=1}^{c}\left(\frac{d_{ik}^2}{d_{jk}^2}\right)^{\frac{1}{m-1}}\right)^{-1}, 1 \le i \le cl; 1 \le k \le n \quad (2)$$

Step 3: Compute possibilistic representative $\delta_{i,}^{(l)}$ that reduces objective function $J_{m,T}$ is stated as follows:

$$\delta_{i,}^{(l)} = \frac{\sum_{k=1}^{n}\left(u_{i,k}^{l}\right)^{m} d_{i,k}^{2}}{\sum_{k=1}^{n}\left(u_{i,k}^{l}\right)^{m}} 1 \le k \le n \quad (3)$$

Step 4: Compute a possibilistic membership degree ($t_{i,k}^{(l)}$) that reduces objective function $J_{m,T}$ as stated as follows:

$$t_{i,k}^{(l)} = \left(1 + \left(\frac{\beta}{\delta_i} d_{ik}^2\right)^{\frac{1}{\tau-1}}\right)^{-1}, 1 \le i \le cl; 1 \le k \le n \quad (4)$$

Step 5: Update cluster center $v_{i,}^{(l)}$ that minimizes objective function $J_{m,T}$ which is stated as follows :

$$v_{i,}^{(l)} = \frac{\sum_{k=1}^{n}\left(\left(\alpha u_{i,k}^{l}\right)^{m} + \left(\beta t_{i,k}^{l}\right)^{T}\right) x_K}{\left(\left(\alpha u_{i,k}^{l}\right)^{m} + \left(\beta t_{i,k}^{l}\right)^{T}\right)} 1 \le k \le n \quad (5)$$

Step 6: Compare $v_{i,}^{(l)}$ to $v_{i,}^{(l-1)}$ by exploiting $|v_{i,}^{(l)} - v_{i,}^{(l-1)}|$. If true, subsequently halt. Else, set l = l + 1 as well as return to n Step 2.

3.2.2 Inverted Indexing

To carry out the indexing model, the inverted indexing procedure is exploited by the adopted cluster-based inverted indexing approach. Moreover, to the clustered groups input query keyword is transmitted. From diverse cluster groups, documents are retrieved on basis of a keyword query. While the matched keyword is available in one or more documents in diverse cluster groups, subsequently every document is retrieved. Therefore, from several cluster groups, diverse documents are retrieved via inverted indexing. In an efficient manner to carry out the document indexing, inverted indexing is exploited besides with clustering approach. On the basis of the associated matched keyword, the documents are indexed. For each cluster group, the keyword of the query is forwarded to search to match documents wherein every cluster group comprises several documents. Hence, from several cluster groups, the whole documents are searched for each query keyword, and by exploiting the indexing process only matched documents are retrieved.

3.3 Bhattacharyya Distance for Complex Query Matching

Furthermore, the inverted indexing outcome phase is fed to the query matching phase. For user queries such as semantic or multiple queries complex query matching is carried out to produce the outcomes. Generally, the semantic queries are referred to as contextual and associative in nature. To retrieve the documents explicitly as well as implicitly the semantic query is used which depends upon the semantic, structural, and syntactic information that is available in the information. Using the query matching criterion, it is modeled to deliver the outcomes. On the basis of the semantics of unstructured data, the relationship amid the document is processed by the semantic query. The Bhattacharyya distance is used by the query matching process to generate superior query matching outcomes. On basis of the least distance measure, the Bhattacharyya distance measure identifies the same documents. In the retrieval process, Bhattacharyya distance measures document differences on basis of the Bhattacharyya coefficient. The Bhattacharyya distance is calculated as follows:

$$\mathrm{X}(\mathrm{k},\mathrm{l}) = \frac{1}{4}\ln\left(\frac{1}{4}\left(\frac{\lambda_{\mathrm{k}}^2}{\lambda_{\mathrm{l}}^2}+\frac{\lambda_{\mathrm{l}}^2}{\lambda_{\mathrm{k}}^2}+2\right)\right)+\frac{1}{4}\left(\frac{(\sigma_{\mathrm{k}}-\sigma_{\mathrm{l}})^2}{\lambda_{\mathrm{k}}^2+\lambda_{\mathrm{l}}^2}\right) \tag{6}$$

wherein, σ_{k} indicates the mean of the k^{th} documents, $\mathrm{k,l}$ indicates two diverse documents, λ_{k}^2 indicates the variance of k^{th} documents, $\mathrm{X(k,l)}$ indicates the Bhattacharyya distance between the two documents, correspondingly. On the Bhattacharyya distance, a complex query matching process is carried out that exploits the query of the keyword to search appropriate documents in cluster groups.

The query matching once more searches the documents, if the appropriate documents are retrieved that is retrieved using a cluster-based inverted indexing approach. On basis of user preference, query matching retrieves precise documents amid all the chosen documents. Hence, by exploiting the Bhattacharyya distance the query matching generates superior matching outcomes.

The input keyword is transmitted to cluster groups in the complex query matching phase. Here, 2 cluster groups are taken into account which are indicated as cluster group-1 as well as 2. The cluster group-1 possesses documents as, "s_1, s_2, s_3, s_4, s_5, and s_6" and the cluster group-2 comprises documents as, "z_1, z_2, z_3, z_4, z_5, and z_6" correspondingly. To retrieve appropriate documents, the keyword is equivalent to each document in both cluster groups. The document content that is equivalent to input keywords is chosen from both clusters. In addition, if the searching keyword is equivalent with groups of the cluster, documents s_2 and s_6 in cluster group-1, as well as document z_4 in cluster group-2 is retrieved. Therefore, on basis of optimal matching outcomes, the keyword searches the same documents between both cluster groups as well as documents s_2, s_6 and z_4 are retrieved.

3.4 Document Retrieval using Query Optimization

Using the query optimization phase, the matching outcome attained from complex query matching is furthermore processed. At last, using interactive query optimization, query optimization is performed to determine a competent manner to perform a query with diverse probable query plans to retrieve appropriate documents. To retrieve documents, on basis of the Pearson correlation coefficient the query optimization exploits the similarity measure. Pearson correlation coefficient is stated as correlation measure amid documents, it is stated as,

$$\nu(\mathrm{p},\mathrm{q}) = \frac{\mathrm{cov}(\mathrm{p},\mathrm{q})}{\sigma_{\mathrm{p}}\sigma_{\mathrm{q}}} \tag{7}$$

wherein, cov signifies covariance measure, p and q signifies random variables, σ_q signifies a standard deviation of q and σ_p signifies a standard deviation of p, correspondingly. The similarity measure to documents is used by query optimization that is retrieved using the query matching process to efficiently recognize appropriate documents. Hence, associated documents are retrieved on basis of the Pearson correlation coefficient similarity measure.

4. Result and Discussion

In this section, the experimentation of the adopted model was described by exploiting clustering and inverted indexing. The performance of the adopted model was analyzed with the conventional models, like SemIndex and EM and these were evaluated on the basis of the measures, namely precision, recall, as well as F-measure.

Fig 2 exhibits performance analysis of the adopted model with the conventional models via taking into account the utmost values for metrics, namely precision, recall, as well as F-measure, correspondingly. Here, the proposed model is 10% better than the SemIndex and 22% better than the EM for precision. For EM, the proposed model is 15% better than the SemiIndex and 18% better than the EM. Likewise, overall the proposed model namely EM obtained the utmost values for metrics, such as precision, recall, as well as F-measure.

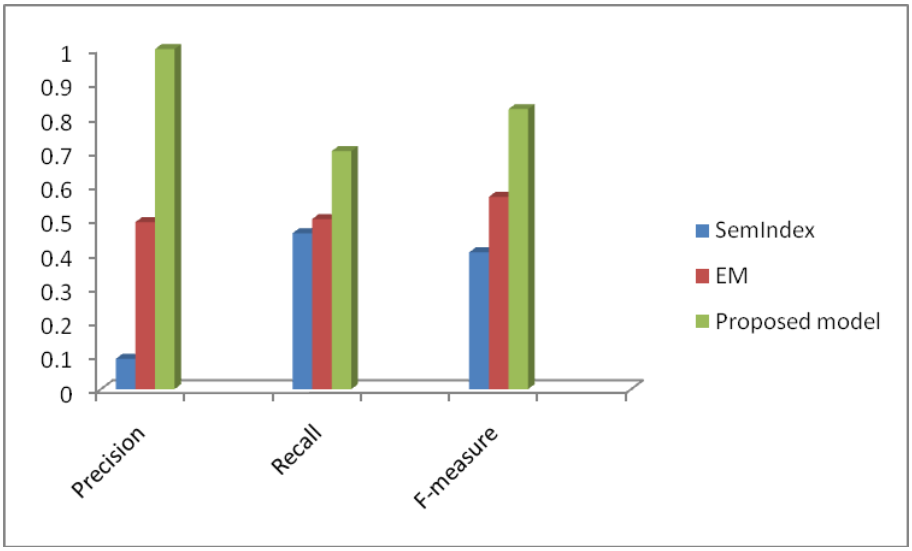


Fig.2. *Performance analysis of the proposed model over conventional models*

5. Conclusion

The cluster-based inverted indexing, which was a clustering approach was adopted to retrieve the appropriate documents in this paper. By integrating the inverted indexing with piecewise Fuzzy C-Means clustering approach adopted a cluster-based inverted indexing model, which effectually retrieves appropriate documents. By exploiting the removal of stop words as well as stemming approaches documents were pre-processed to evade the unnecessary as well as the redundant words. Using a cluster-based inverted indexing approach, document indexing was performed that employs the pre-processed documents, as well as indexing, was generated on basis of clustered documents keyword. Furthermore, the ensued documents were processed using complex query matching processing, wherein user queries namely multigram queries or semantic queries were matched by exploiting the Bhattacharyya distance. On the basis of the least distance measure or Bhattacharya distance, the enhanced query matching outcomes were obtained. The Pearson correlation coefficient was used by the query optimization on the basis of the interactive query optimization and retrieves appropriate documents competently. The developed cluster-based inverted indexing approach obtains enhanced performance with the measures, such as recall, precision, as well as F-measure values.

Compliance with Ethical Standards

Conflicts of interest: Authors declared that they have no conflict of interest.

Human participants: The conducted research follows the ethical standards and the authors ensured that they have not conducted any studies with human participants or animals.

Reference

[1] Hector FerradaGonzalo Navarro,"Lempel–Ziv compressed structures for document retrieval", Information and Computation31 January 2019.

[2] Raymon van DinterCagatay CatalBedir Tekinerdogan,"A decision support system for automating document retrieval and citation screening", Expert Systems with Applications25 May 2021.

[3] Mamta KayestSanjay Kumar Jain,"Optimization driven cluster based indexing and matching for the document retrieval", Journal of King Saud University - Computer and Information SciencesAvailable online 1 March 2019.

[4] Ramzi ChaiebKarim KaltiNajoua Essoukri Ben Amara,"Fuzzy generalized median graphs computation: Application to content-based document retrieval", Pattern Recognition29 July 2017.

[5] Shufeng HaoChongyang ShiLongbing Cao,"Concept coupling learning for improving concept lattice-based document retrieval", Engineering Applications of Artificial Intelligence3 January 2018.

[6] Tekli J, Chbeir R., Traina A.J. and Traina Jr, C., "SemIndex+: A semantic indexing scheme for structured, unstructured, and partly structured data," Knowledge-Based Systems, vol.164, pp.378-403, 2019.

[7] Shufeng Hao, Chongyang Shi, Zhendong Niu, Longbing Cao,"Modeling positive and negative feedback for improving document retrieval, "Expert Systems With Applications, vol.120, pp.253–261, 2019.

[8] Wu J., Wu Z., Cao J., Liu H., Chen, G. and Zhang, Y., "Fuzzy consensus clustering with applications on big data", IEEE Transactions on Fuzzy Systems, vol. 25, no. 6, pp.1430-1445, 2017.

[9] Rustam, Koredianto Usman, Mudyawati Kamaruddin, Dina Chamidah, Nopendri, Khaerudin Saleh, Yulinda Eliskar, Ismail Marzuki,"Modified Possibilistic Fuzzy C-Means Algorithm for Clustering Incomplete Data Sets",mputer Science > Artificial Intelligence on 9 Jul 2020 (v1), last revised 15 Jul 2020.

[10] Gunjan Chandwani, Anil Ahlawat, Gaurav Dubey, "An approach for document retrieval using cluster-based inverted indexing", Journal of Information Science, 2021.

[11] Neenavath Veeraiah and Dr.B.T.Krishna,"Intrusion Detection Based on Piecewise Fuzzy C-Means Clustering and Fuzzy Naive Bayes Rule", Multimedia Research, vol. 1, no. 1, October 2018.

[12] Amolkumar Narayan Jadhav,Gomathi N,"DIGWO: Hybridization of Dragonfly Algorithm with Improved Grey Wolf Optimization Algorithm for Data Clustering", Multimedia Research, vol. 2, no. 3, July 2019.