ISSN: 0976-3031

Available Online at http://www.recentscientific.com

CODEN: IJRSFP (USA)

International Journal of Recent Scientific Research
Vol. 8, Issue, 7, pp. 18717-18722, July, 2017

International Journal of Recent Scientific Research

DOI: 10.24327/IJRSR

Research Article

CLUSTERING BASED INFORMATION RETRIEVAL WITH THE ACO AND THE K-MEANS CLUSTERING ALGORITHM

Poonam Yadav*

D.A.V College of Engineering & Technology, Kanina, Haryana 123027, India

DOI: http://dx.doi.org/10.24327/ijrsr.2017.0807.0555

ARTICLE INFO

Article History:

Received 15th April, 2017
Received in revised form 25th May, 2017
Accepted 23rd June, 2017
Published online 28th July, 2017

Key Words:

Information Retrieval, K-Means Clustering, Aco, Query, Trec Database

ABSTRACT

Information retrieval has gained momentum due to increase in the various optimization algorithms. Information retrieval based on the user query needs challenging algorithms, since the features may vary between the user request and the documents in the database. This work introduces information retrieval system based on the clustering of the documents. The proposed approach performs the information retrieval through pre-processing, feature selection, and clustering. The proposed model uses the ACO algorithm for the feature selection and the k-means clustering algorithm for the clustering approach. When the query is provided by the user, the proposed model checks the similarity between the query and the each cluster and thus provides the documents in the cluster most similar to the query. The simulation uses two standard databases such as TREC, and the 20 newsgroup for the processing. From the simulation results, it is evident that the proposed ACO with the k-means algorithm has the improved accuracy 0.6105 and 0.4444 for the TREC and the 20 newsgroup database respectively. Similarly, the proposed model has reduced fallout value of 0.1505 and 0.165 for the TREC and the 20 newsgroup database respectively.

INTRODUCTION

Information retrieval is the emerging research field which allows the user to retrieve the required information from the database. When the database is larger in size, it is difficult for the user to get the required information from the database through the normal search process. This is where the information retrieval algorithms come in handy. Information retrieval has gained popularity due to the advent of the various data clustering and the feature selection algorithms. Information retrieval finds application in the various research fields, such as data mining, medical image retrieval, and the big data analyses. The retrieval of the information from the database depends on the query provided by the user. The information present in the database is searched based on two techniques. They are 1) navigational search, and 2) exploratory search. Information retrieval from the database has the following steps, 1) Pre-processing, 2) feature extraction and selection, and 3) Clustering the database. Many algorithms use the query based models to retrieve the information from the database. Clustering based models [5] successively divide the database and match the queries with the each cluster.

The query based schemes also use the implicit and the explicit based techniques for the information retrieval [6]. The query provided to the user is the combination of the words that specify the user interest. Matching of the queries with the documents in the clusters requires the exploratory search [10]. Because the documents from source articles for clustering may have been written by different groups, from different viewpoints, or have different writing style, clustering these textual materials is, therefore, a challenge due to the diversity of vocabulary used and the general lack of guidance regarding background knowledge that could provide domain information. Clustering [13, 16] with the high dimensionality of document representation usually causes poor performance of clustering results and affects the efficiency of clustering algorithms. Thus, optimizing the means of feature selection for low-dimensional document representation is very important in the document clustering task [2]. Some works [14] has utilized the ontology-based techniques for the document clustering. But, the ontology-based techniques focus on the single task rather than multitasking [14].

This work has introduced an information retrieval system based on the query based system. The proposed model achieves the information retrieval through, pre-processing, feature selection, and the clustering. The proposed model uses the traditional techniques such as stemming, stop word removal, tokenization

Corresponding author:* **Poonam Yadav
D.A.V College of Engineering & Technology, Kanina, Haryana 123027, India

and the VSM for the pre-processing of the documents. Then, the required features for the information retrieval are selected with the use of the ACO algorithm. Then, the features are subjected to the dynamic reduction scheme. Then, the documents present in the database are clustered using the k-means clustering algorithm. When the query is provided by the user, the query is matched with the each document in the cluster. Then, the documents in the cluster matching more similar to the query are retrieved. The metrics such as accuracy and fallout measure the proposed model.

The major contributions of this work is enlisted as follows

This work uses the ACO algorithm for the feature selection and the k-means clustering algorithm for the clustering of the documents.

The rest of this paper is organized as follows: Section 2 provides various literary works regarding the information retrieval. Section 3 provides the explanation about the proposed work. Section 4 provides the simulation results, and Section 5 concludes the paper.

Motivation

LITERATURE SURVEY

This section deals with the various literary works based on the information retrieval through the clustering algorithms.

Muhammad Kamran Abbasi and Ingo Frommholz [1] proposed the information retrieval system based on the cognitive framework. This work is based on the principle of poly representation along with document clustering. Various information such as citations is used for representing the documents during the clustering process.

Lin Yue *et al.* [2] presented the automatic document organization for the information retrieval from the set of documents. They have also utilized various features for clustering the documents. The optimal feature selection technique utilized here suffers from the disadvantage, since is does not consider the dimension of the features.

Malik Tahir Hassan *et al.* [3] presented the CDIM based information retrieval system. The proposed model allows the partitional clustering of documents to differentiate each document. The proposed methodology uses the semantic information in the documents to identify the discriminating features. Laith Mohammad Abualigaha *et al.* [4] proposed the feature weight scheme and dynamic dimension reduction for efficient utilization of the features in the documents. The method has used various optimization algorithms for the feature selection. The dynamic reduction scheme reduces the dimension of the selected featured to improve the retrieval process. The document clustering in the proposed scheme depends on the selected features using the k-means clustering scheme.

Challenges

The retrieval of the information from the database through the clustering algorithms faces various challenges. The challenges are enlisted as follows,

- Text documents contain high dimensional informative and uninformative 20 features, besides it has more noisy features [4].
- K-means algorithm has difficulties in dealing with high dimension data [17].

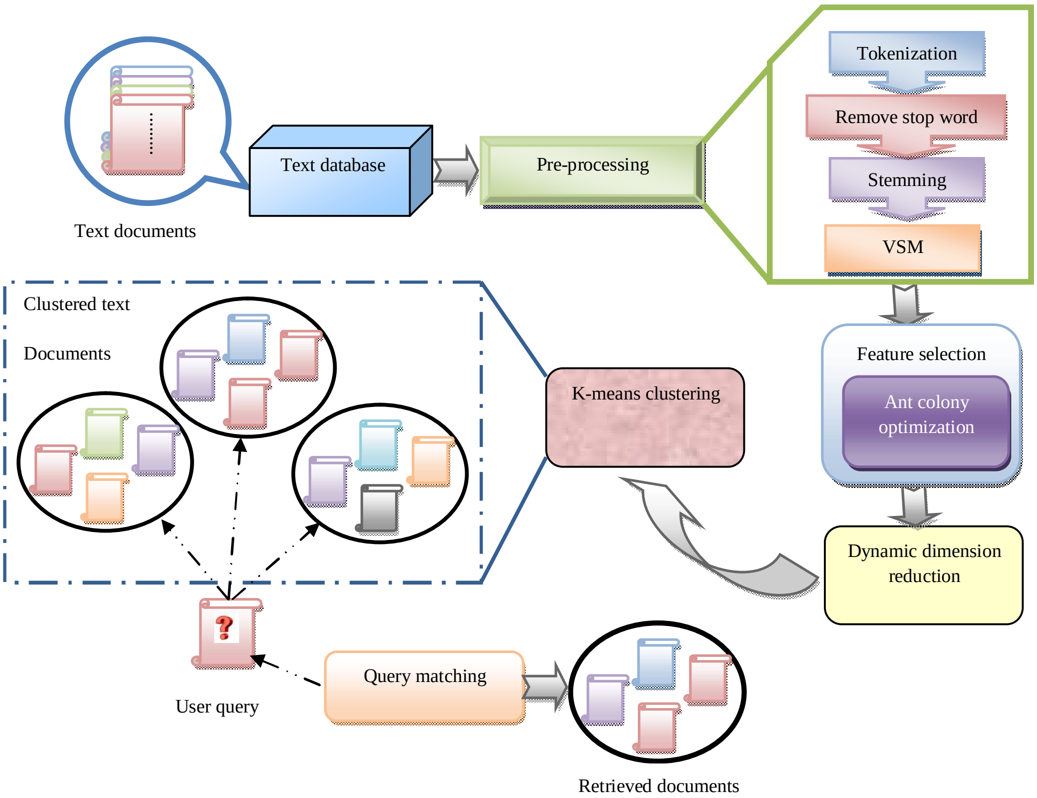


Figure 1 Architecture of the information retrieval system with the ACO

Architecture of the proposed information retrieval system

The information retrieval algorithms from the large dataset have seen tremendous increase nowadays due to the development of the meta-heuristic algorithms. This paper has introduced the information retrieval system through the clustering approach. Figure 1 shows the block diagram of the proposed information retrieval system. The steps involved in the proposed system for the retrieval of the information from the database based on the user query is provided below,

1. The data present in the text database are pre-processed to make it suitable for the feature extraction. Traditional pre-processing techniques on the text such as tokenization, stop words removal, and stemming are applied to the documents present in the database.
2. In the next step, the pre-processed data is provided to the ACO for selecting the appropriate features for clustering the documents in the database.
3. The selected features are provided to the dynamic dimension reduction scheme. This eliminates the unwanted features from the selected features.
4. The k-means clustering algorithm clusters the documents present in the database based on the features.
5. When the user provides the query, the query is matched with the each document of the cluster. And the relevant documents to the user query are retrieved.

Consider the text database T with N number of documents. The database is represented as follows,

$$T = T_t \quad 1 \le t \le N \tag{1}$$

where, the term T_i represents the i-th document in the text database. The text document in the database contains a collection of various text documents. The text documents present in the database is the various collections of the texts.

Pre-processing

The initial step in the information retrieval is the pre-processing. The pre-processing of the each document in the database improves the retrieval process. The pre-processing is done in the following steps,

1. Tokenization
2. Stop word removal
3. Stemming and
4. VSM

Applying each pre-processing steps to the text documents provides text documents with no connecting words. Next, the pre-processed words are sent for the feature selection.

Feature selection using ACO

In this step, the required features for the information retrieval are selected with the use of the existing ACO algorithm. The ACO algorithm is the nature-inspired algorithm which finds the optimal solution based on the movement of the ant for food search. The movement of the ant from their colony for the search of the food can be modeled as a bidirectional problem. The ACO algorithm is more suitable for a minimization problem. The ACO algorithm best suited for the global search. The features required for the information retrieval primarily depends on the frequency of the words occurring in the text documents.

Dynamic dimension reduction

This work utilizes the dynamic reduction scheme [4] to eliminate the unnecessary features selected by the ACO algorithm. The dynamic reduction scheme uses the document frequency to eliminate the features. This is done in three steps as explained as follows,

1. Find the frequency values of the document based on the average document frequency.
2. Find the dynamic document frequency features in the each document.
3. Initialize threshold value, and according to that threshold eliminate the features with the low dynamic frequency.

Grouping of similar documents using K-means clustering

This work uses the k-means clustering algorithm [11] to cluster the similar featured documents into a single cluster. The database is sent to the k-means clustering algorithm to identify the similar featured documents. The k-means clustering algorithm performs the clustering in two steps. They are

1. The k-means clustering algorithm defines the number of cluster centers for performing the clustering
2. The cluster point in the algorithms are placed near the cluster centers, and based on that new cluster centroids are found.

The new cluster centroids found by k-means clustering algorithm reduces the squared error function. The equation 10 expresses the squared mean function.

$$\arg\min_{C} \sum_{t=1}^{T} \sum_{y_m \in B_t} \left\| T_a^n - \mu_t \right\|^2 \tag{2}$$

where, the term μ_t in the expression indicates the mean of the cluster, and the term $\left\| T_a^n - \mu_t \right\|$ expresses the distance. The distance is calculated as the Euclidean distance. It is the measure the distance between the data point, and the cluster mean in the document. The k-0means clustering algorithm aims in reducing the error for the various values of the centroid point. The clusters provided are expressed by the equation 3.

$$B = \{B_Q, Q = 1,2,......., q\} \tag{3}$$

where, the term B expresses total clusters, q expresses the number of clusters and B_Q is the Q^{th} cluster.

Algorithmic steps for k-means clustering

The algorithmic steps of the k-means is expressed as follows, Let $T = \{T^1, T^2, .., T^R, .., T^n\}$ be set of all the data points and $B_b = \{b_1, b_2, ..., b_q\}$ be the set of all the cluster centres. In the initialization the cluster centres randomly selected.

Step 1: Random selection of the cluster centre C_c*:* The equation 4 expresses the random selection of the clusters for the processing.

$$B_b = \{b_1^R, b_2^R, \ldots, b_q^R\} \tag{4}$$

where, the cluster points are expressed by the following equation.

$$b_1^R = \left[w_n^T \| S_n^T \|\right]_1^R \tag{5}$$

$$b_2^R = \left[w_n^T \| S_n^T \|\right]_2^R \tag{6}$$

$$b_q^R = \left[w_n^T \| S_n^T \|\right]_q^R \tag{7}$$

where, the term $b_1^R, b_2^R, \ldots, b_q^R$ represents the randomly selected cluster centres. They carry the features such as keywords, semantic words.

Step 2: Calculation of the distance between each data point and the cluster centre: The distance between the data points is calculated as follows,

$$dist(T_a^n, B_q) = \left(\left[1 - \left(w_n^T \cap w_n^{T(R)}\right)\right] + \left[1 - \left(S_n^T \cap S_n^{T(R)}\right)\right]\right) \tag{8}$$

where, $dist(D_b^n, C_q)$represent the distance between the data point and the cluster, $\left[1 - \left(w_n^T \cap w_n^{T(R)}\right)\right]$ and $\left[1 - \left(S_n^T \cap S_n^{T(R)}\right)\right]$represents that the distance between the data point and the cluster is minimum,

Step 3: Assign a data point to the cluster center: Now calculate the distance between the various data poi9nts and the cluster center obtained through the random process.

Step 4: Compute the new cluster centre: Now select the new cluster point based on the frequency of the various document features. The features with the maximum frequency are selected as the new cluster centroid point.

Step 5: Compute the new distance based on the new cluster centre: The distance between the data point and the new cluster centre is calculated using equation 8.

Step 6: Repeat steps 3 to 5, or stop if no reassignment: Repeat the steps 3-5 until no new groups are formed.

Finally, the grouped features are formed into clusters and are given by the equation 9.

$$B_b = \{b_1^f, b_2^f, \ldots., b_q^f\} \tag{9}$$

where, $b_1^f, b_2^f, \ldots., b_q^f$ are the clusters after grouping and q represents the number of clusters with key features.

Matching of Query document with the clustered centroid

In this step, the query document provided by the user is verified with the each cluster centroid. The features of the query are matched with each cluster. Then the clusters with the more similar documents related to the query are provided to the user. The query matching is done based on the similarity measure between the documents. For the each matching, the similarity measure is calculated. Then the documents having the higher similarity measures are retrieved.

RESULTS AND DISCUSSION

This section presents the results obtained through the proposed information retrieval process. The metrics such as accuracy and fallout measure the performance of the proposed model. Various standard databases are used for the performance analysis.

Experimental setup

The simulation tool MATLAB 2015a finds the performance analysis of the proposed information retrieval system against various existing works such as GA, HS, and PSO. The evaluation metrics analyzing the performance of the each model is explained as follows,

Accuracy: The accuracy of the information retrieval process defines the ratio o f the number of accurately classified instances to the total number of instances in the model. The accuracy metric is defined in term of the true positive, true negative, false positive, and false negative.

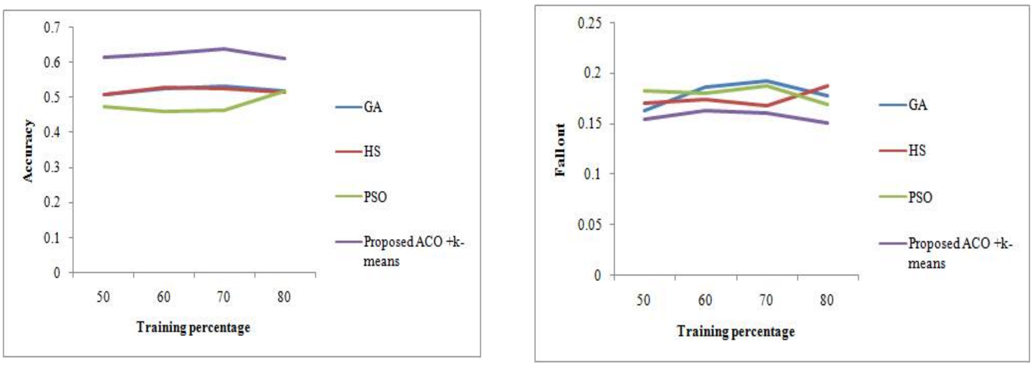
Fallout: The fall out metric defines the inverse of the specificity. It defines the total number of non-relevant documents retrieved by the algorithm. Hence, the value of the fallout must be very low for the increased performance.

Performance analysis of the proposed model with the TREC database

This section shows the performance analysis of the proposed work with the TREC database. The analysis is done for the various training values of the TREC database vs. i) accuracy and ii) fallout. Figure 2.a shows the performance analysis of the proposed information retrieval system for the TREC based on the accuracy metric. When 50 % data of the TREC database is trained, the existing GA, HS, and the PSO algorithms have the accuracy value of 0.5073, 0.5077, and 0.4718. The proposed ACO + k-means approach has increased accuracy value of 0.6145. While increasing the training percentage to 80, the existing algorithms GA, HS, PSO has the accuracy value of 0.5187, 0.5134, and 0.5782. The proposed model has the overall increased accuracy value of 0.6105. Hence, the increase in the training percentage has negligible changes in the accuracy. Figure 2.b shows the performance analysis of the proposed information retrieval system for the TREC based on the fallout metric. The method with the reduced fallout can be considered as the better model. When 50 % data of the TREC database is trained, the existing GA, HS, and the PSO algorithms have the fallout value of 0.1634, 0.1706, and 0.183. The proposed ACO + k-means approach has reduced fallout value of 0.1545. For the training percentage = 80 %, the existing algorithms GA, HS, PSO has the fallout value of 0.1777, 0.1875, and 0.1694. The proposed model has the fallout value 0.1505.

Performance analysis of the proposed work with the 20 newsgroup database

This section shows the performance analysis of the proposed work with the 20 newsgroup database. The analysis is done for the various training values of the 20 newsgroup database vs. i) accuracy and ii) fallout. Figure 3.a shows the performance analysis of the proposed information retrieval system for the 20 newsgroup based on the accuracy metric.



a) Accuracy **b) Fallout**

Figure 2 Performance analysis of the proposed work with the TREC database

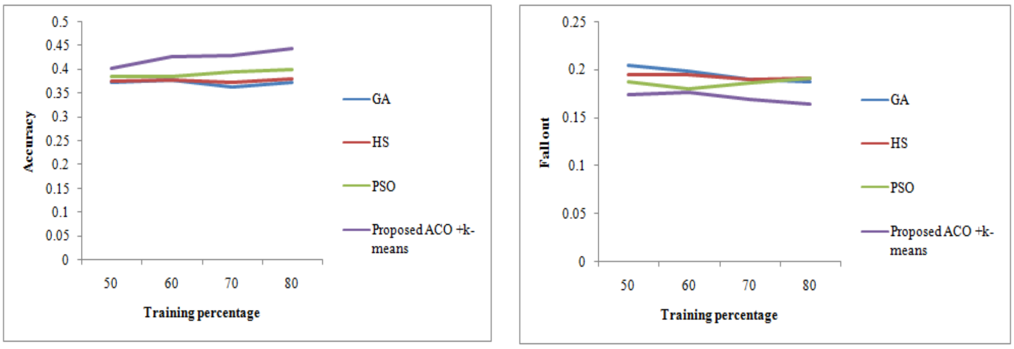
When 50 % of the 20 newsgroup database is trained, the existing GA, HS, and the PSO algorithms have the accuracy value of 0.3745, 0.3748, and 0.3854. The proposed ACO + k-means approach has increased accuracy value of 0.4035. While increasing the training percentage to 80, the existing algorithms GA, HS, PSO has the accuracy value of 0.3741, 0.3798, and 0.401. The proposed model has the overall increased accuracy value of 0.4444. Figure 3.b shows the performance analysis of the proposed information retrieval system for the 20 newsgroup based on the fallout metric. When 50 % data of the TREC database is trained, the existing GA, HS, and the PSO algorithms have the fallout value of 0.2051, 0.1945, and 0.1875. The proposed ACO + k-means approach has reduced fallout value of 0.1747. For the training percentage = 80 %, the existing algorithms GA, HS, PSO has the fallout value of 0.1879, 0.1914, and 0.1911. The proposed model has the fallout value 0.165.

Table 1 Comparative analysis of proposed model based on the evaluation metrics

Database	Evaluation metrics	Comparative models			
		GA	**HS**	**PSO**	**Proposed ACO + k-means**
TREC	Accuracy	0.5187	0.5134	0.5182	0.6105
	Fallout	0.1777	0.1875	0.1694	0.1505
20 news group	Accuracy	0.3741	0.3798	0.401	0.4444
	Fallout	0.1879	0.1914	0.1911	0.165

CONCLUSION

This work introduces a query-based information retrieval system. The proposed model achieves the information retrieval through, pre-processing, feature selection, and the clustering. The proposed model uses the traditional techniques such as stemming, stop word removal, tokenization and the VSM for the pre-processing of the documents. Then, the required features for the information retrieval are selected with the use of the ACO algorithm. Then, the features are subjected to the dynamic reduction scheme. Then, the documents present in the database are clustered using the k-means clustering algorithm. When the query is provided by the user, the query is matched with the each document in the cluster. Then, the documents in the cluster matching more similar to the query are retrieved. The proposed retrieval process is analyzed with the TREC and the 20 newsgroup database. From the simulation results, it is evident that the proposed ACO with the k-means algorithm has the improved accuracy 0.6105 and 0.4444 for the TREC and the 20 newsgroup database respectively. Similarly, the proposed model has reduced fallout value of 0.1505 and 0.165 for the TREC and the 20 newsgroup database respectively.



a) Accuracy **b) Fallout**

Figure 3 Performance analysis of the proposed work with the 20 newsgroup database

DISCUSSION

This section analyzes the various simulation results obtained by the proposed models and the existing works. Table 1 provides the comparative results of the proposed models and existing works such as GA, HS, and PSO. For the TREC database, the existing works have achieved an accuracy value of 0.5187, 0.5134, and 0.5782. The existing algorithms GA, HS, and PSO have the fallout value of 0.1777, 0.1875, and 0.1694. The proposed information retrieval system with the ACO and the k-means has the overall increased accuracy value of 0.6105 and 0.4444 for the TREC and the 20 newsgroup database respectively. Similarly, the proposed model has reduced fallout value of 0.1505 and 0.165 for the TREC and the 20 newsgroup database respectively.

References

1. Muhammad Kamran Abbasi, and Ingo Frommholz, “Cluster-based poly representation as science modeling approach for information retrieval,” *Scientometrics*, Volume 102, Issue 3, pp. 2301-2322, March 2015.
2. Lin Yue, Wanli Zuo, Tao Peng, YingWang, and Xuming, “A fuzzy document clustering approach based on domain-specified ontology,” *Hand Data & Knowledge Engineering*, Volume 100, pp. 148-166, November 2015.
3. Malik Tahir Hassan, Asim Karim, Jeong-Bae Kim, and Moongu Jeon, “CDIM: Document Clustering by Discrimination Information Maximization,” *Information Sciences, volume* 316, pp. 87-106, 2015.
4. Laith Mohammad Abualigaha, Ahamad Tajudin Khader, Mohammed Azmi Al-Betar, and Osama Ahmad Alomaria, “Text feature selection with a robust weight scheme and dynamic dimension reduction to text document clustering,” *Expert Systems with Applications*, Volume 84, pp. 24-36, October 2017.
5. Rana Forsatia, Andisheh Keikhab, and Mehrnoush Shmasfarda, “An Improved Bee Colony Optimization Algorithm with an Application to Document Clustering,” Neurocomputing, Volume 159, pp. 9-26, July 2015.
6. Irad Ben-Gal, Yuval Shavitt, Ela Weinsberg, and Udi Weinsberg, “Peer-to-peer information retrieval using shared-content clustering,” *Knowledge and Information Systems*, Volume 39, Issue 2, pp. 383-408, May 2014.
7. Deepanwita Datta, Shubham Varma, Ravindranath Chowdary C., and Sanjay K. Singh, “Multimodal Retrieval using Mutual Information based Textual Query Reformulation,” *Expert Systems With Applications*, volume 68, pp. 81-92, 2017.
8. Yong-Bin Kang, Shonali Krishnaswamy, and Arkady Zaslavsky “A Retrieval Strategy for Case-Based Reasoning Using Similarity and Association Knowledge,” IEEE transactions on cybernetics, vol. 44, no. 4, pp. 473-487, April 2014.
9. Larbi Guezouli, and Hassane Essafi, “CAS-based information retrieval in semi-structured documents: CASISS model,” *Journal of Innovation in Digital Ecosystems*, Volume 3, Issue 2, pp. 155-162, December 2016.
10. Roberto Pérez-Rodríguez, Luis Anido-Rifón, Miguel Gómez-Carballa, and Marcos Mouriño-García, “Architecture of a Concept-Based Information Retrieval System for Educational Resources,” *Science of Computer Programming*, Volume 129, pp. 72-91, November 2016.
11. Shi-xia Ma, Jian-hua Liu, and Dan Liu, “Research of Case Retrieval Strategy Based on Partitional Clustering,” The 2nd International Conference on Computer and Automation volume 2, pp. 307-310, 2010.
12. Shi-xia Ma, Qing-yun Ru, Dan Liu, and Zu-hua Guo, “A Case Retrieval Algorithm Based on Ant Colony Clustering,” IEEE International Conference on Computer Science and Information Technology, August 2009.
13. Adrian-Gabriel Chifu, Florentina Hriste, Josiane Mothe, and Marius Popescu, “Word sense discrimination in information retrieval: A spectral clustering-based approach,” Information Processing and Management, Volume 51, pp.16-31, 2015.
14. Balasubramaniam K, “Hybrid Fuzzy-Ontology Design using FCA based Clustering for Information Retrieval in Semantic Web,” *Procedia Computer Science*, Volume 50, pp.135 - 142, 2015.
15. Raya Horesh, Kush R. Varshney, and Jinfeng Yi, “Information Retrieval, Fusion, Completion, and Clustering for Employee Expertise Estimation,” IEEE International Conference on Big Data (BigData), 2016.
16. Surendra Shetty, and Sarika Hegde, “Clustering of Instruments in Carnatic Music for Content based Information Retrieval,” IEEE 6th International Conference on Advanced Computing, 2016.
17. Du Hui, “Case retrieval method based on clustering Algorithm,” Second International Workshop on Computer Science and Engineering, 2009.

How to cite this article:

Poonam Yadav.2017, Clustering Based Information Retrieval With The Aco And The K-Means Clustering Algorithm. *Int J Recent Sci Res.* 8(7), pp. 18717-18722. DOI: http://dx.doi.org/10.24327/ijrsr.2017.0807.0555

****