Annals of Library and Information Studies
Vol. 66, September 2019, pp. 110-116

Document clustering for knowledge synthesis and project portfolio funding decision in R&D organizations

Abhishek Kumar[a], Naresh Kumar[b] and Richa Gupta[c]

[a] Scientist, Council of Scientific and Industrial Research, Anusandhan Bhawan,
Rafi Marg, New Delhi-110001, India, Email: abhishek_kumar@csir.res.in

[b] Senior Principal Scientist, CSIR-NISTADS, Pusa Gate, K. K. Krishnan Marg, New Delhi-110012, Email: nareshkr@nistads.res.in

[c] Stryker India Pvt. Ltd., Gurugram, Haryana, India

Received: 29 June 2019; accepted: 08 September 2019

The paper discusses a method of using document clustering for information/knowledge synthesis and decision facilitation in R&D organisations. The emerging methodologies of machine learning, artificial intelligence and data science in conjunction with fuzzy mathematics can be optimally exploited to catalyse development of information bank for research organisations. This knowledge ecosystem can be utilized by the proposed mechanism to accelerate and reinforce interdisciplinary research for R&D organisations and empower them to make efficacious information-driven decisions related to project portfolio selection and proposal funding.

Keywords: Decision Support System (DSS); Soft clustering; Information synthesis; Fuzzy clustering

Introduction

Research project portfolio selection, fund allocation and thereafter knowledge management involve complex processing in large R&D organizations. Numerous project proposals that are received need to be objectively evaluated. The general trend in R&D organization is to constitute expert committees for proposals' evaluation. It is assumed that the committee will shortlist proposal portfolio based on sound intuitive and rational cognitive grounds. However, it is known that human intuitive judgment and decision making might be erroneous, less optimal and which further deteriorates with complexity and stress.

Further to this, while making such decisions, R&D organizations have to deal with various strategic, tactical and operational issues which possess unique nature of challenges and demands critical examination and handling for a rational and informed choice selection. Other issues related with diverse cultural, political, social, economic and ecological environments also play prominent roles in shaping such decisions. One of the obvious reasons behind such challenges is that for any R&D organization, their spending not only represents a sizeable investment but also has a vital and significant impact on their stakeholders, current and future financial position as well as their ability to compete strategically and technically. Due to these reasons, appropriate project portfolio selection and their funding support become quite crucial and vital for any R&D organization.

The other crucial challenge faced by these organizations is the management, utilization and synthesis of knowledge originated from such supported projects. Knowledge is a general term that describes various products and services whose primary input is human creativity. These knowledge products are intangible, non-excludable, and in the current era the most valuable outcome of any research work. Knowledge retention and utilization provides a competitive advantage for any research organization.

Many progressive R&D organizations use customized Decision Support System (DSS) and data warehousing for knowledge management and decision making. Being customized in nature, these solutions are mostly designed to work in an in-silo approach and integrating these solutions to accrue a comprehensive benefit at futuristic decision and vision making level always remain a challenge.

The growing trend of interdisciplinary research requires well-planned knitting of different domains of scientific research fields. Many frontier research organizations in India have realized the fact that catalysing interdisciplinary research program drawn from two or more different disciplines often create a powerful research experience and result into integrative learning, critical thinking, and novel findings and solutions.

Review of literature

Document clustering is a verified method in the field of information retrieval. Using it in documents summarization helps in avoiding content overlap and ensure better information coverage[1-5]. The widely used document representation model is Vector Space Model (VSM) which represents documents as a collection of vital terms in a vector space. To effectively represent document in a vector space, the dimension reduction is achieved by means of removal of stop words, tokenization, stemming and lemmatization. The VSM is an established method for information retrieval and semantic level clustering as documents related semantically are empirically proved to contain many words in common and it is easy to find out by using popular VSM measures like cosine similarity[5].

The studies carried by various researchers proved the efficacy of document clustering in an efficient document classification[6-9]. However, all these studies majorly focus on hard clustering where it is assumed that a document can be a member of only one cluster at a time and varieties of *K* mean clustering are used.

The intention of this paper is to capture the fuzzy relation that exists between different clusters. By running a soft clustering (fuzzy clustering), we intuitively expect at least one of the clusters to be closely related to the concept and idea represented by a document, whereas, other clusters may contain information indirectly related to the document some way hitherto unknown to us and in such case provide a chance to synthesis new knowledge[5].

Though the idea of utilizing vector space model with fuzzy clustering means is examined in a controlled environment[10,11], the area of decision facilitation in the field of R&D funding using such intervention is still largely unexplored.

Objective of the study

- To design a methodology to help R&D organizations in rational and informed project proposal portfolio selection and funding decisions;
- To empower in efficacious knowledge management using artificial intelligence, machine learning and data science mechanism.

Methodology

Project proposals contain comprehensive information about the submitted projects. Clustering the proposals and calculating fuzzy relation among the pre-identified clusters can not only help in decision facilitation but also provide insight in knowledge synthesis and scope of interdisciplinary research.

Fuzzy clustering is considered as soft clustering in which each element has a probability of belonging to other clusters. The scope of such relationship is represented by set of membership coefficients which empirically reflect the degree of membership of element within a given cluster. This membership score varies between 0 and 1.

Fuzzy clustering algorithm works by assigning membership to each data point corresponding to each cluster centre based on distance between the cluster centre and the data point. More the data is near to the cluster centre more is its membership towards the particular cluster centre. Mathematically, for a given document and cluster sets, main objective of fuzzy cluster means is to minimize:

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^{m} \left\| x_i - c_j \right\|^2 \quad , \quad 1 \leq m < \infty$$

where m is any real number greater than 1, u_{ij} is the degree of membership of x_i in the cluster j, x_i is the $i^{th}$of d-dimensional measured data, c_j is the d-dimension center of the cluster, and ||*|| is any norm expressing the similarity between any measured data and the centre. However, as per *Hathaway and Bezdek 2001* m=2.0 is a good choice for fuzzy c clustering[12] .

The hard document clustering (*K means clustering and its variants)* presumes that each document can be a member of exactly one cluster and thus in principle implements set theory basics in clustering i.e:

If D = $\{d_1, d_2, d_3, \ldots\ldots, d_n\}$ and C= $\{c_1, c_2, c_3, \ldots\ldots, c_m\}$

Where D is the set of proposal documents and C is the set of distinct clusters. Hard document clustering presumes that,

$\forall\ d_i \in c_j$ and $d_{i+1} \in c_{j+1},$

a) $S_s (d_i \wedge d_{i+1}) = 0$, where S_s is the similarity score between documents; and

b) These statements cannot hold true:

$d_i \in c_j \wedge d_i \in c_{j+1}$ and $d_{i+1} \in c_j \wedge d_{i+1} \in c_{j+1}$ at the same time.

Whereas, the proposed solution assigns grade of membership for each proposal inside a fuzzy set. This grade corresponds to degree to which the proposal is similar with the concept of a particular research domain represented by its corresponding fuzzy set. Under this clustering principle based on fuzzy relations, it can be safely presumed that:

$\forall\ d_i \in c_j$ and $d_{i+1} \in c_{j+1,}$

a) $S_s (d_i \wedge d_{i+1)} = n$, where $n \in R+$; and

b) Statements $d_i \in c_j \wedge d_i \in c_{j+1}$ and $d_{i+1} \in c_j \wedge d_{i+1} \in c_{j+1}$ may hole true at the same time.

Statement (a) reflects that documents from different clusters may have a scope to share similar concepts with each other whereas statement (b) reflects that under the proposed clustering one document can be a member of two different clusters at the same time. Both the statements provide empirical scope for catalysing interdisciplinary research from different domains and schematically represented in Figure 1.

Data processing and execution

The received proposal documents are divided into two corpora namely training data set and testing data set. The training data set proposals are uniquely classified under various clusters (research domains). It is pertinent to mention that a hard clustering is carried out for training data set labelling (i.e., each document strictly belongs to one single cluster) to efficiently establish the centroid for each cluster which represent a unique research domain.The dimension reduction of these documents is done by data cleaning, stemming and tokenization. After this the Bag of word (BoW) model is used to represent these documents in VSM (Vector Space Model). The TF-IDF of these documents are calculated as follows:

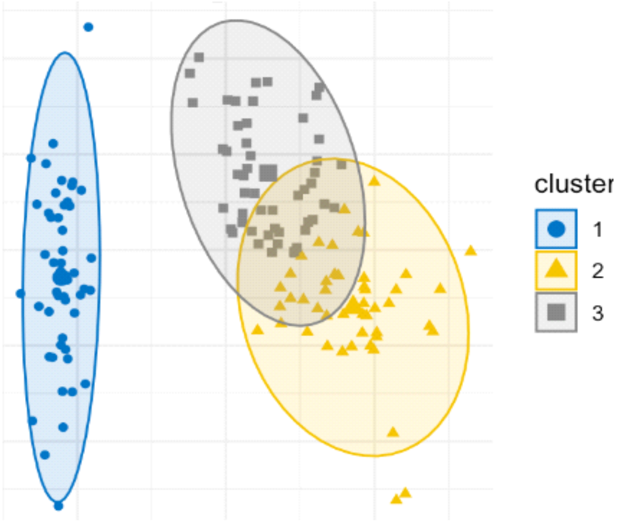

Fig. 1 — Schematic representation for proposed soft clustering proposal documents based on information embedded inside them

TF = Number of times a term appears in a proposal document / Total Number of terms in the proposal document

IDF = Log (Total No. of proposal document / No of document with term T in it)

The TF-IDF value for each proposal document is evaluated. Further to this, fuzzy c means clustering is applied on the testing data set to calculate the distance of each document from identified clusters.Following decision criteria is applied while finalizing the labelling if

$0.4 \leq fms(d_i) \leq 0.6$, *where fms is proposal fuzzy membership score calculated by fuzzy c means algorithm*

The document can be a considered as weak association member for that particular cluster and if

$0.6 < fms(d_i)$

The document exhibits a strong membership association with that particular cluster.

The weak fuzzy membership relation can be exploited to identify scope of interdisciplinary research whereas strong fuzzy membership can reflect a proposal specific research area focus. Further to this, the cosine similarity will be evaluated for all proposal documents from the same cluster group to find the scope of incremental / collaborative research among them. This all efforts will in parallel also develop a data bank which will be used for all suchlater on activities and knowledge synthesis.

Based on aforementioned approach and criteria creation, the steps for proposal portfolio selection are as follows:

1. Do a soft clustering of received proposals against a pre-trained data sets (training data set) of different research domains (fuzzy clusters/ sets);
2. If a proposal shows a strong membership/association with one of the research data set (cluster), the proposal should be marked as dedicated research work belong to that particular domain and move to step 4;
3. If a proposal shows considerable association scores with 2 or more research domain sets (clusters), there can be potential to cultivate an interdisciplinary research. The proposal should be evaluated in two different ways; for its primary evaluation it should be associated with such

research domain where it is exhibiting strongassociation score and for interdisciplinary research it should be evaluated with other associated sets as per step 5;

4. The primary evaluation should be carried with the expert team members where the novelty, objectives and deliverables should be considered and evaluated empirically. Further a cosine similarity of the proposal document should be evaluated with all the others pre-existing proposal documents of that particular research domain (fuzzy cluster/set) to examine similarity score. This evaluation will help selection team in two ways: the high cosine similarity rate can help team in detect incremental/ associative and/or duplicated research work. A rational decision can be taken accordingly by the team;
5. Proposals exhibiting scope of interdisciplinary research should be evaluated against associated research sets (with considerable membership score as per step 3) and a cosine similarity can be calculated to find those pre-existing research proposals with which new proposals shows similarity. Based on the empirical merits of new proposal and development of pre-existing proposal, the expert team can take rational decision on scope of interdisciplinary research diffusion;
6. The above-mentioned methodology will empower expert team members to take a data driven rational and informed decision. Further, the research domains' data sets formulation catalyses information bank creation which can be exploited either by AI and ML methodologies proposed above to facilitate decision or by area experts as a mean of knowledge references and synthesis. These information banks can play crucial role in empowering R&D organization towards a robust data ecosystem.

The pictorial diagram of the proposed soft clustering is depicted in Figure 2:

Observations

The POC (proof of concept) for the proposed methodology is executed with 8 research proposals from information, biological and physical science research streams. The selected proposals were as follows:

- Research proposals from Information Sciences : 2 (DOC 4 and DOC5);

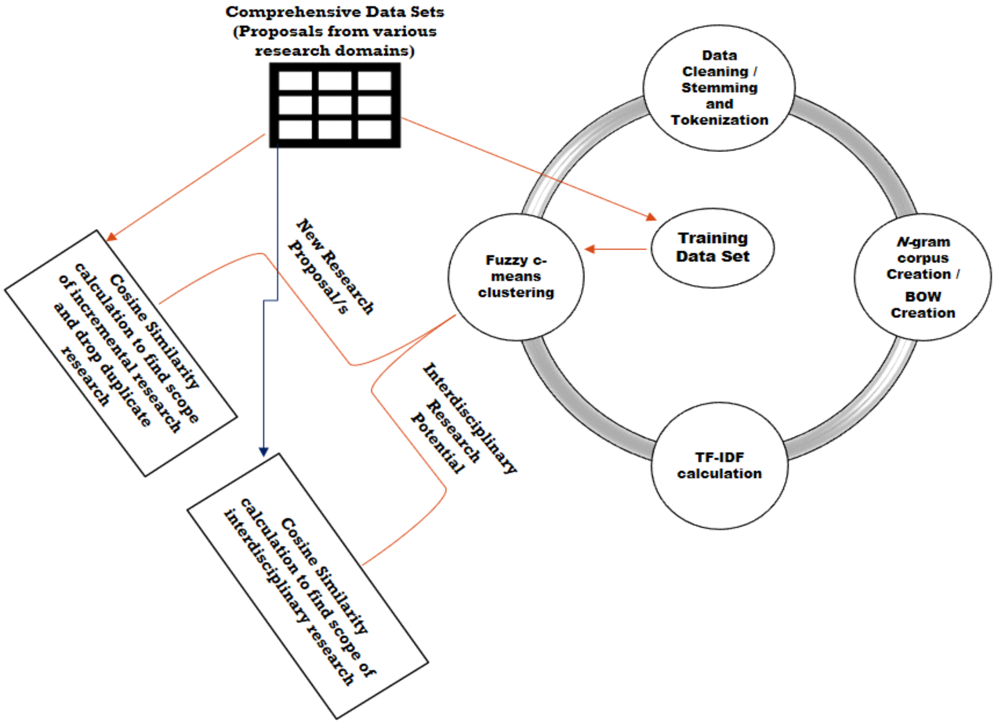
Comprehensive Data Sets
(Proposals from various
research domains)
Data Cleaning / Stemming and Tokenization
Training Data Set
Fuzzy c-means clustering
N-gram corpus Creation / BOW Creation
TF-IDF calculation
New Research Proposal/s
Interdisciplinary Research Potential
Cosine Similarity calculation to find scope of incremental research and drop duplicate research
Cosine Similarity calculation to find scope of interdisciplinary research

Fig. 2 — Pictorial representation of proposed clustering

- Research Proposals from Biological Sciences: 3 (DOC 1, DOC2 and DOC3); and
- Research Proposals from Physical Sciences: 2 (DOC 6, DOC 7 and DOC8)

Spyder IDE(Scientific Python Development Environment)) is used to run and execute the codes whereas MySQL relational database is used to store and examine data. In the first step the data cleaning is carried out, and one comprehensive corpus is created consisting all the major terms from the proposal documents. After this stemming and tokenization is carried on the developed corpus and tf-ifd vector is calculated to find the relation among the documents. In our case the tf-idf matrix generated. The corresponding python code along with arguments is as follows:

```
def calc_tfidf(corpus):
tfidf = TfidfVectorizer(stop_words=stop_words, lowercase=True , token_pattern=u'(?u)\b\w*[a-zA-Z]\w*\b',analyzer='word', tokenizer=stemming_tokenizer ,ngram_range=(1,1) )
tfs = tfidf.fit_transform(corpus)
 vocabulary = tfidf.get_feature_names()
 #print(vocabulary)
store_sparse_mat(tfs,"tfmat")
pickle.dump(tfidf,open("data/Vocabtfidf","wb"))
 return tfs
```

The code generates following tf-idf matrix:

```
[[1, 0.15194641, 0.07485966, 0.0779501, 0.03494827, 0.01284879, 0.01921408, 0.03383354]
[0.15194641, 1, 0.11803776, 0.05118157, 0.03765689, 0, 0.02668147, 0.03070638]
[0.07485966, 0.11803776, 1, 0.11294924, 0.13706709, 0, 0.0350161 , 0.01990437]
[0.0779501, 0.05118157, 0.11294924, 1, 0.20388325, 0.00974244, 0.02697318, 0.02747143]
[0.03494827, 0.03765689, 0.13706709, 0.20388325, 1, 0.02275601, 0.02951931, 0.03015326]
[0.01284879, 0, 0, 0.00974244, 0.02275601, 1, 0.03939233, 0.00741722]
[0.01921408, 0.02668147, 0.0350161, 0.02697318, 0.02951931, 0.0393923, 1, 0.1954298 ]
[0.03383354, 0.03070638, 0.01990437, 0.02747143, 0.03015326, 0.00741722, 0.1954298 , 1 ]]
```

Here each row consists document similarity scores with other documents in the corpus. This is a diagonal matrix as a document is always similar to itself in similarity score. We can also print the term vocabulary for all the documents separately by uncommenting the “#print(vocabulary)” statement. With the calculation of tf-idf matrix, now we can calculate fuzzy c means clustering score for all 8 documents for available 3 clusters. Scikit-Fuzzy library is used to calculate fuzzy c means score. To perform the clustering, Scikit-Fuzzy implements the *cmeans* method which requires following mandatory parameters: data, which must be an array $D \in \square^{N \times \quad M}$ (*N* is the number of features; therefore, the array used with Scikit-Learn must be transposed); *c*, the number of clusters; the coefficient *m*, *error*, which is the maximum tolerance; and *maxiter*, which is the maximum number of iterations. Another useful parameter (not mandatory) is the seed parameter which allows specifying the random seed to be able to easily reproduce the experiments. The *cmeans* function returns many values, but for our purposes, the most important are: the first one, which is the array containing the cluster centroids and the second one, which is the final membership degree matrix[15]. As per above mentioned guidelines, our function will be

```
fc, W, _, _, _, _, _ = cmeans(X_train.T, c=10, m=1.25, error=0.005, maxiter=10000, init = None)
```

When we run this code on the above calculated tf-idf matrix and pre-developed comprehensive corpus in Spyder IDE

```
filename="corpus_train.json"
 with open("data/"+filename, "rb") as input_file:
  commands=pickle.load(input_file)
tfs = tf.load_sparse_mat("tfmat")
X= (tfs * tfs.T).toarray()
fc, W, _, _, _, _, _ = cmeans(X, c=3, m=1.25, error=0.005, maxiter=10000, init=None)
print(W)
```

The code will yield the final fuzzy relation matrix of documents with identifies clusters (research domain) as:

```
[[0.96592899, 0.97935594, 0.90348289, 0.00521029, 0.00436757, 0.07330123, 0.01089759, 0.01524899]
[0.0191015, 0.01195292, 0.03249906, 0.00268782, 0.00278236, 0.8791362, 0.98267572, 0.97595212]
 [0.01496951, 0.00869114, 0.06401804, 0.99210189, 0.99285007, 0.04756257, 0.00642669, 0.00879889]]
```

These results yield following observations:

i. Documents 1, 2 and 3 exhibit strong membership towards first cluster, documents 4 and 5 incline towards third cluster whereas documents 6,7 and 8 show affinity towards second clusters. This shows that the provided proposal documents tend to make 3 distinct clusters.

ii. As per set assumptions, proposals 1, 2 and 3 belong to same research domains whereas proposals 4 and 5 and proposals 6,7 and 8 belong to a different research domain. The manual verification valorises this membership association as described previously.

iii. The selected documents mostly oriented towards core research to a dedicated research area (as indicated by membership score) and in the given data set there does not seems to be a scope of interdisciplinary research as the document shows weak membership with other clusters.

iv. As a sample examination for the scope of incremental/collaborative research, cosine score between proposal documents of third cluster (proposal document number 4 and 5) is evaluated tofindlevel of similarity between them.

```
import fetch_db
import tfidf_cosine as tf
import pickle
import collections
def calc_tfidf_corpus():
result_dict={}
 commands={}
 filename= fetch_db.create_corpus()
 with open("data/"+filename, "rb") as input_file:
  commands=pickle.load(input_file)
corpus_index = [n for n in commands];
 corpus = list(commands.values())
tfs= tf.calc_tfidf(corpus)
 d=collections.OrderedDict(commands)
 index=list(d.keys()).index("DOC4")
 for index,score in tf.find_similar(tfs,index,10):
result_dict[corpus_index[index]]=str(score)
 print(result_dict)
calc_tfidf_corpus()
```

The output for above yields the cosine similarity score fordocument 4 with all other proposal documents.

```
{'DOC5':          '0.15194641152221908',          'DOC7':
'0.07795009704407282',                            'DOC6':
'0.07485966294584795',                            'DOC8':
'0.03494826612203501',                            'DOC3':
'0.03383353866344254',                            'DOC2':
'0.019214075871820804',                           'DOC1':
'0.012848787680101706'}
```

The considerable affinity and empirical inclination of document 4 towards document 5 justifies their fuzzy relation association for the same cluster but the low cosine similarity between them indicates that both the research proposals are quite distinct in nature and potential scope of incremental/collaborative research does not lie between them. This observation strengthens the practicableness of the methodology which effectively calculate the fuzzy association for provided proposals in identical research clusters (which help in effective fuzzy association mapping of proposals for research domains) but this cannot guarantee a scope of collaborative research as the proposals can be of distinct nature and targeting different areas of same research domains.

Similarly, the cosine similarities for other proposals associated with same cluster can be evaluated to find the scope of collaborative research. The observation emanates from our sample set that though the proposals are associated with three different research areas (represented by distinct clusters) they are of very unique in nature and focus on different issues of research areas. They are of focused research proposals and should be evaluated accordingly. Further, the scope of interdisciplinary research or collaborative research could not be identified due to the specific nature of research proposals.

Conclusion

The proposed study and thereafter empirical observations prove that data science has a huge and efficacious potential to facilitate R&D organizations in many aspects. AI and ML based methodologies can play a vital and crucial role in empowering R&D organizations in making rational and eminent decisions regarding funding decisions and valorise the same.

The work carried in this paper also indicates that apart from usual VSM (vector space model) representation of research document, the scope of representing them in the form of distinct information nuggets for better information retrieval (IR) can be explored. This approach can play imminent role in knowledge synthesis which is a much-sought requirement for R&D organisation. The effectiveness of fuzzy relation association rather than hard clustering in research document classification is realised effectively by the proposed methodology. However, the optimal exploitation of this approach still takes some time. Development of well-defined data ecosystems is another aspect which organisations needs to look after for fruitful realisation of aristocratic and remunerative outcomes associated with theseinitiatives.

References

1 Hatzivassiloglou V, Klavans J L, Holcombe M L, Barzilay R, Kan M and McKeowm K R, “SIMFINDER: A flexible clustering tools for summarization,” *Proc. NAACL Workshop Automatic Summarization*, 2001, 41-49.

2 Zha H, Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering, *Proc. 25th Ann. Int’l ACM SIGIR Conf. Research and Development in Information Retrieval*, 2002, 113-120.

3 Aliguyev R M, A new sentence similarity measure and sentence based extractive technique for automatic text summarization, *Expert Systems with Applications*, 36 (2009) 7764-7772.

4 Radev D R, Jing H, Stys M and Tam D, Centroid bases summarization of multiple documents, *Information Processing and Management*, 40 (2004) 919-938.

5 Skabar A, Abdalgader K, Clustering sentence-level text using a novel fuzzy relational clustering algorithm, *IEEE Transaction on Knowledge and Data Engineering*, 25 (1) (2013) 62-75.

6 Kathiravan A V and Kalaiyarasi P, Sentence-similarity based document clustering using birch algorithm, *International Journal of Innovative Research in Computer and Communication Engineering*, 3(5) (2015) 4385-4390.

7 Renukadevi D and Sumathi S, Term based similarity measure for text classification and clustering using fuzzy c mean algorithm, *International Journal of Science, Engineering and Technology Research*, 3(4) (2014) 1093-1097.

8 Thakare R D and Patil M R, Extraction of template using clustering from heterogeneous web documents, *International Journal of Computer Applications*, 119 (11) (2015) 23-31.

9 Chiranjeevi P, Supraja T and Rao P S, A survey on extension techniques for text document clustering, *International Journal of Advanced Research in Computer Science and Software Engineering*, 5 (7) (2015) 1251-1256.

10 Hadi R M, Hashem S H and Maolood A T, Proposed method to enhance text document clustering using improved Fuzzy c mean algorithm with named entity tag, *AI-Mansour Journal*, 28 (2017) 43-61.

11 Abinaya V, Vennila M and Padmanabhan, Sentence level text clustering using a hierarchical fuzzy relational clustering algorithm, *International Journal of Communication and Computer Technologies*, 2(10) (2014) 50-55.

12 Fuzzy c-means clustering algorithm, Available at: https://sites.google.com/site/dataclusteringalgorithms/fuzzy-c-means-clustering-algorithm

13 A tutorial on clustering algorithms, Available at: https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/cmeans.html", accessed on June 21,2019

14 Document clustering, Available at: https://en.wikipedia.org/wiki/ Document_clustering”

15 Example of fuzzy C-means with Scikit-Fuzzy, Available at: https://learning.oreilly.com/library/view/mastering-machine-learning/9781788621113/6967d36f-e04e-46d3-8c99-e30e2193d464.xhtml