Honey Bee Mating Optimization of K-Means Algorithm for Medical Document Clustering

P. Venkateshkumar and A. Subramani

KSR College of Engineering, Tiruchengode, India.

(Received: 18 February 2015; accepted: 03 May 2015)

Usage of digital medical documents and sharing by Web services has tremendously increased the size of document collections and increases the burden on the user for getting relevant document while searching. Many tools such as query-based retrieval and browsing are available to search a document of interest. Document clustering is widely used for efficient Information Retrieval (IR) and data mining applications. Traditional methods use 'bag of words' approach to find the relevant document for a query. But, high dimensionality of the features of a document and ambiguity in the natural language needs concept-based search instead of using bag of words. Ranking the features and expansion of concepts of ranked features will be helpful for efficient data retrieval and mining. This work proposes a Honey Bee Mating optimization with k-Means clustering (HBM-KM) algorithm for optimal clustering of documents. The proposed technique performs better than with Hierarchical Agglomerative clustering (HAC) and k-means algorithm.

Key words: Document Clustering, Concept Expansion, Hierarchical Agglomerative Clustering (HAC), k-Means clustering, Honey Bee Mating algorithm (HBM).

The usage of digital documents has increased in all the fields, and most of these documents are shared by World Wide Web (WWW). Many tools such as query-based retrieval and browsing are available to search a document of interest. Document clustering is widely used for efficient Information Retrieval (IR) and data mining applications by grouping related texts into clusters and labeling them will reduce the overhead of topic discovery so that health professionals can visualize the results of a search quickly[1]. The goal of IR is to get the documents from the entire collection of documents that matches best with the query given by the user.

Existence of huge information and ambiguity available in natural language increases the burden of finding best matching document for an average user. Therefore, new techniques are required to organize the information, store, and trace the documents. Clustering is one of the techniques used to solve this problem. Clustering is an unsupervised technique, which organizes a set of input data into clusters. Clusters are formed such that similar data are clustered together to form multiple clusters. But, before clustering the documents, the type of attributes or characteristics of the documents, which can be used for clustering must be chosen[2].

The two major types of clustering algorithms are hierarchical and partition algorithms[3, 4]. In hierarchical clustering, the given database D is divided into subsets in hierarchical manner. Agglomerative clustering and divisive clustering are two major types of hierarchical clustering. Agglomerative clustering starts with a single cluster and merges with one or more relevant clusters recursively. Divisive clustering takes entire data set into one cluster and splits the documents of initial cluster into multiple clusters recursively.

* To whom all correspondence should be addressed.
E-mail:

In both clustering methods, recursion is stopped when termination condition is reached.

Partition clustering partitions the data collection into predetermined number of clusters in a single step. Single pass methods and iterative methods are two popular types of partitioning clustering. Direct k-way clustering is a single pass method, which partitions documents into k number of clusters. Initially, k documents are taken as seeds for k clusters. Each seed is compared by a similarity measure with each document, and each document is assigned to a cluster, which has best matching between the document and associated seed of the cluster. In iterative method, direct k-way clustering form initial clusters, clustering is repeated until some optimization criteria are met. Therefore, reallocation is allowed in iterative method. Though, the hierarchical clustering technique is supposed to achieve better clustering, it is not optimal as reallocation of poorly classified objects cannot be carried out in the original hierarchical clustering algorithm.

Finding similarity among the documents is useful for the analysis of text, IR, clustering, and classification of the documents. Traditional methods use similarity calculation based on bag of words instead of semantic connections. In bag of words method, initially tokens separation is done by using white spaces and then stop words and other punctuation marks in the document are removed. Finally, the number of occurrences of each word is counted and represented along with the word. The detriment of this approach is that the ambiguity in the words of natural language misguides the retrieval and classification of documents. Therefore, semantic representation along with bag of word must be considered for clustering and classifying the documents.

Concept is an abstract unit of knowledge represented in a document or an article. Concept is unambiguous and represents a unique meaning. If there is an overlap of concepts between two documents, then these two documents are connected. When the semantic relationships among the concepts are used for clustering and classification, the search results will be more appropriate to the query[5]. In bag of concepts method, initially concepts are identified, and index vocabulary is extracted from concept systems such as Wikipedia or WordNet. Wikipedia developed concept system based on the phrases of the author, and WorldNet uses machine learning[6]. Advantages of using concepts are,

1) Redundancy is lessened

2) Dis-ambiguous words which have multiple meanings are identified

3) Semantic connections among the different concepts have been identified, quantified, organized, and traced when finding similarity connectedness between two documents.

Same word may discuss different concepts depending on the document where the word is used. Concept expansion is used to identify list of concepts from a document. While expansion, disambiguation is required for proper semantic representation. To prune the irrelevant concepts from a concept list of document, a measure salience is used. Let *U* be the list of concepts of a document, and C_i be a concept within the list *U*. Salience of concept C_i is calculated by the equation (1).

$$SAL(C_i) = \frac{\sum_{C_j/U, j \neq i} SIM(C_i, C_j)}{|U|} \qquad ...(1)$$

where C_j represents other concepts of the same document and available in the list *U* and $|U|$ represents total number of concepts identified in the document. $SIM(C_i, C_j)$ represents semantic connectedness between two concepts $C_i$and C_j. Whenever a query is given by the user to perform efficient IR, list of concepts is created based on the surrounding words.

The problem of clustering has been studied extensively using machine learning algorithms, statistics but fail to produce satisfactory results. The limitations are due to non-validation of the clusters produced, assumptions made on the data distributions or due to the criterion used during the formation of the clusters. Most of the existing clustering algorithms are inefficient when the number of feature vector is large.

The feature space for document is high and sparse which affects the performance of clustering algorithms. Some of the features are redundant and not relevant, and irrelevant features will misguide the classification algorithms[7, 8]. In the proposed approach, features are extracted and ranked to identify most important features and avoid irrelevant features. Then concepts of the document are expanded to include semantics of the document and used in clustering. The Use of concepts in clustering the documents will be helpful for accurate document classification, information retrieval, and data mining. The study proposes Honey Bee Mating optimization with k-Means clustering (HBM-KM) algorithm for optimal clustering of documents. The HBM optimization is used to find the centroids of the clusters and then extended to k-Means algorithm. The subsequent sections discuss the related work, proposed methodology and finally conclude with result and discussion.

Related works

K-means clustering is an iterative clustering algorithm and has been widely used for document clustering. Slowness, scalability, sensitive to the selection of initial seed and convergence to local minima were some of the drawbacks of k-means clustering algorithm. Yang[9] proposed greedy k-means clustering algorithm to solve these problems. This algorithm gave superior performance by giving most relevant documents when user searches the large database with a query. Cui and SeokChae[10] applied 18 different clustering techniques for component identification for legacy systems. Various similarity and weighing factors were used, and performance was compared. Result showed that hierarchical clustering produced better results than other clustering algorithms.

Sreemathy *et al*.,[11] proposed an efficient document clustering technique using multiple features. Dimensionality reduction was done at preprocessing, and weightage for each feature was assigned by the measures such as document frequency, mutual information, information gain, term and document frequency. kNN and Bayesian classifiers were used, and performance was compared by precision and recall values of classification. kNN clustering proved as superior to Bayesian classifier. In kNN clustering, if the training samples are uneven then classifier will decrease the precision of classification. Therefore, an adaptive adjustment in each iteration of identifying neighborhood was proposed by Zhou[12]. This method reduced uneven classification and mis adjustment at the boundaries of clusters obtained from samples.

To solve the problems associated with high dimension feature space, Sharma and Dhir[13] proposed Frequent Item set Hierarchical Clustering (FIHC) with Word Set based Clustering (WDC). When the number of documents in the cluster was large, hierarchical clustering struck with local minima. Therefore, WDC used with hierarchical clustering for the documents which had common word sets. Results revealed that combining hierarchical clustering with WDC clustering provided efficient retrieval of data and improved reliability and scalability when compared to k-means clustering. Traditional clustering algorithms used vector model for comparing the similarity of the documents, but, recent algorithms use suffix based similarity for the documents. Rafi *et al*., [14] used two methods such as efficient phrase extraction and frequent word or word meaning extraction to covert the document into a compact representation. Then, hierarchical based agglomerative clustering was used subsequently. The purity and entropy measures were used and results show that efficient phrase based method was superior to frequent word approach.

Removing noisy redundant features and selection of important features improves the accuracy of the classification, but it affected the integrity of the document. Fu [15] presented a novel approach to satisfy integrity of the documents and improve the accuracy of the classification. Maximum-Relevance and Minimum-Redundancy Analysis (MR2A) was used to find index terms and concepts of a document. Concept Similarity Mapping (CSM) was used to merge the predictive power of concepts and index terms. Experiments revealed that combining MR2A and CSM improved accuracy when comparing to information gain and chi-square. Constructing the domain knowledge about the particular collection of documents was useful to improve the precision of data retrieval.

Deepa and Revathy[16] proposed a fuzzy C-means clustering algorithm which was not based on hard clustering. Clustering was validated by both internal and external criteria. External criteria used prior knowledge about the domain of the text documents. Purity, Entropy, F-measure, and NMI measure were used for external evaluation. Internal evaluation was based on the intrinsic information available in the document. Silhouette index, Bic index were some of the internal validation measures.

Applying swarm based intelligence to clustering algorithms was proposed by Alam *et al*.,[17]. Hierarchical Particle Swarm Optimization

(HPSO) was proposed for data mining applications and efficiency was compared with bench mark clustering algorithms such as k-means, PSO clustering and HAC. HPSO proved better efficiency than these bench mark classification algorithms. Chellatamilan and Suresh|[18] used TDF- IDF, concept expansion, and language modeling for query likely hood to extract the additional terms when user gave a query. By varying number of additional terms extracted for a query, it was seen that maximum precision was achieved.

METHODOLOGY

The PubMed document sets and the Reuters dataset were used in this investigation. Stop words and Stemming were used clean the data. Figure 1 shows the flowchart of the proposed technique. The subsequent sections describe each module.

Inverse Document Frequency (IDF)

Figure 1 shows the flow of steps of the proposed framework for document clustering. During preprocessing, stop words are removed from the documents to collect the bag of words. Stop words are non-significant words available in a document. Stop words need not be considered before starting indexing process[19]. Stemming is a procedure used to prefixes and suffixes in order to solve the variants of the words with same root or stem. If there are N documents in a data set, a term is denoted as t_i and the number of occurrence of t_i is represented as n_i then Inverse Document Frequency (IDF) is calculated by using equation (2)[20],

$$idf(t_i) = log\frac{N}{n_i} \qquad ...(2)$$

All clustering techniques use the vector space model for representing a document. In the vector space model each document d is a vector in the term-space. In TF-IDF term weighting model each document is represented by equation (3):

$$tf_1 \log(n/df_1), tf_2 \log(n/df_2), \ldots\ldots, tf_m \log(n/df_m) \qquad ...(3)$$

Where tf_i the number of occurrence of ith term in the document and df_i is the number of documents that contain the ith term. For managing documents of different number of terms, each document vector's length can be normalized which holds unit length given by equation (4):

$$(\| d_{tfidf} \| = 1) \qquad ...(4)$$

which shows that each document can be represented as a vector in unit hyper sphere.

After extracting IDF, features are ranked by using information gain value. Information gain indicates the importance of a feature to resulting clusters. Information Gain (IG) of feature 'A' is computed by the difference between expected information for classification of data set D and actual information needed for classification by selecting a feature 'A'[21]. The mathematical formula for Information Gain is given in equation (5),

$$\text{GAIN}(\text{A}) = \text{Info}(\text{D}) - \text{Info}_\text{A}(\text{D}) \qquad ...(5)$$

Expected information is computed by equation (6),

$$Info(A) = -\sum_{i=1}^{m} P_i log_2(P_i) \qquad ...(6)$$

where p_i is the probability of a document in D belongs to Class C_i. The actual information needed can be computed using equation (7),

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} * Info(D_j) \qquad ...(7)$$

where D_j is the set of documents in D that have value a_j for the feature 'A'. Feature with high Once IG is computed for all the features, concepts are extracted for top ranked features and documents are clustered according to the concepts.

Table 1. Analogy between the natural honey bee colony and the artificial honey algorithm

Natural honey bee	Artificial honey bee
Queen	Best solution
Drones	Incumbent solutions
Broods	New trial solutions
Worker	Heuristic search
Mating Breeding	Crossover

Table 2. HBM Parameters

No. of Drones	30
No. of Mating Flights	1000
Size of Queen Spermtheca	15
No. of Broods	15
No. of selected genes in Crossover	10
Number of Iteration	1000

Concept Expansion

A term can have many meanings when it is not domain specific. For example the term Orange can have many meanings based on the usage domain. Using domain concepts can improve the document clustering. If the list of concepts in documents could be extracted and annotated the clustering accuracy improves. To ensure this, the meaning of a term needs to be disambiguated to identify the concept it refers to. Concepts can also include synonymous terms in the document[22] to improve document clustering. When a concept is important, it usually has other related concepts. For example, 'charge' has at least two distinct meanings: electric charge and financial charge. When a document speaks about electric charge, the document will have other terms like current, electricity, etc. while in the case of financial charge; document will have terms like payment, amount, etc. The idea is to score a concept by looking at it and references to related concepts.

How each term in document is mapped to its corresponding concept and how each concept's significance is estimated regarding the current document are important for efficient clustering. A term 'charge' can map to electric charge, financial charge or criminal charge. Out of mapped concepts, the most appropriate concept for a specific domain should be located. Related concepts occurrences are looked into for this. Captured inter-concept relationship in ontology is used.

The algorithm for identifying concepts for each term t_i in document D proposed by Gong *et al* is given in[23]. A corpus with concept expansion using Word Net is formed.WordNet is a lingual database for English, the link language and is termed as an abounding lexical database for English constituting groups of nouns, verbs, adjectives and adverbs called synsets, contrived on conceptual semantic and lingual relations. A corpus with proposed concept expansion using Word Net is formed.

Agglomerative Algorithm

An agglomerative algorithm builds the solution with initial assign to each document to its own cluster and uses a bottom up tree and hence also named as bottom up clustering method. That is, it starts with set of documents $d_1, d_2, ..., d_N$ and considers each document as a cluster. Based on the similarity measure each cluster is merged with one or more relevant clusters recursively until the termination condition is reached. After merging clusters i and m, the similarity of new cluster j is calculated by the function $\mathrm{SIM}(i, m, j)$ and only active clusters are stored by deactivating empty clusters. Merging is stopped when similarity between any two clusters is low[24]. Similarity is measured using equation (9):

$$sim_{complete-link}(S_r, S_t) = \min_{d_i \in S_r,\, d_j \in S_t} \{\cos(d_i, d_j)\} \qquad ...(9)$$

Pseudo code for a simple HAC[23] is given in the following,

```
$HAC(d_1, d_2 .... d_N)$
$for\ n \leftarrow 1\ to\ N$
    $do\ for\ I \leftarrow 1\ to\ N$
        $do\ C[n][i] \leftarrow SIM(d_n, d_i)$
    $I[n] \leftarrow 1$ (keeps track of active clusters)
$A \leftarrow []$ (assembles clustering as a sequence of merges)
$for\ k \leftarrow 1\ to\ N - 1$
    $do\ \langle i, m \rangle \leftarrow \arg\max_{\{\langle i,m \rangle : i \neq m \wedge I[i]=1 \wedge I[m]=1\}} C[i][m]$
    $A.APPEND\ (\langle i, m \rangle)$ (store merge)
    $for\ j \leftarrow 1\ to\ N$
        $do\ C[i][j] \leftarrow SIM(i, m, j)$
        $C[i][j] \leftarrow SIM(i, m, j)$
    $I[m] \leftarrow 0$ (deactive cluster)
$return\ A$
```

Proposed Honey Bee Mating optimization with k-means clustering (HBM-KM) algorithm

K-meansclustering is a popular unsupervised clustering suboptimal technique. The number of clusters is pre-defined in the k-means algorithm. When test data is given, it calculates similarity measures to form concept. Widely used similarity measure is Euclidean distance. Clustering steps is summarized as:

1) Initially k number of seeds is randomly assigned as centroids.

2) Based on some similarity measure the neighborhoods for each centroid is searched and assigned into a cluster, thus k clusters are formed.

$$\mathrm{DIST}(X, Y) = \sqrt{\sum_{i=1}^{D} (x_i - y_i)^2} \qquad ...(10)$$

where D denotes the number of parameters in each vector.

3) Now center point at each cluster is recalculated as

$$y_i = \frac{1}{N_{cj}} \sum_{\forall x_i \in C_j} x_i \qquad ...(11)$$

where N_{cj} is number of vectors in cluster C_j.

Step 2 and 3 are performed until convergence is reached

Many algorithms in various domains are optimized by simulating the natural behaviour and intelligence of fishes and animals. Honey Bee Mating (HBM)[25] has been successfully applied to solve nonlinear constrained and unconstrained optimization optimization problems. Behaviour of Honey bee mating is applied to clustering algorithm to improve the efficiency. In a hive, there is only one queen bee and many drone bees. Usually drone bee searches for food in parallel and return to hive. Queen bee flies randomly in air in high speed and if she meets a drone bee, then mating happens and queen bee stores the genotype. Using stored genotype of every mating, queen bee produces a drone bee after completing the flight. This mating behaviour of queen bee with drone bees can be simulated in the clustering algorithm. Honey bee includes a queen (best solution), drones (incumbent solutions), workers (heuristic) and broods (trial solutions). In the initial phase the queen gets initialized with some energy. The mating ritual starts and ends only when the Queen's energy level drops below set threshold[26]. The mating of queen with drone is shown by equation (12):

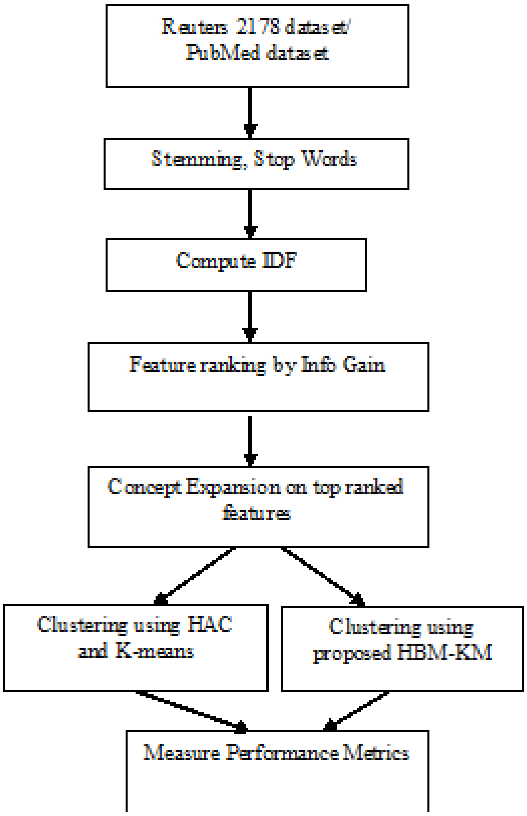

Reuters 2178 dataset/
PubMed dataset
Stemming, Stop Words
Compute IDF
Feature ranking by Info Gain
Concept Expansion on top ranked features
Clustering using HAC and K-means
Clustering using proposed HBM-KM
Measure Performance Metrics

Fig. 1. Flow chart of proposed document clustering method

$$p(Queen, Drone_i) = e^{[\frac{-\Delta(f_i)}{energy(t)}]} \qquad ...(12)$$

Where *p(queen, Drone$_i$)* computes the probability of taking the *i*th drone for mating. *Ä(f)* is the fitness value and given by

$$\Delta(f_i) = | f(Queen) - f(Drone_i) | .energy(t) \qquad ...(13)$$

In[27] it is explained that the queen's energy is high at the beginning of her flights. Mating probability increases when the energy is high for both drone and queen. The decrease in queen energy can be given by (14):

$$\begin{aligned} &energy(t+1) = \alpha \times energy(t) \quad \text{where } t \in [0,1,2,....t] \text{ and decay rate } \alpha \text{ within } [0,1] \\ &speed(t+1) = energy(t) - \beta \quad \text{where } t \in [0,1,2,....t] \text{ and decay rate } \beta \text{ within } [0,1] \end{aligned} \qquad ...(14)$$

Where á is the decay rate. In this work swarm represents the number of possible clustering solutions. The fitness is given by the average distance between the document and the cluster centroid[28] as shown in equation (15):

$$f = \frac{\sum_{i=1}^{N_c} d(o_i, m_{ij})}{N_c} \qquad ...(15)$$

Where i represents the cluster and j represents the document. Initial population is configured by the number of drone bees and placed randomly in the solution space. When searching solution, if drone bee mates with queen bee, queen bee stores the partial solution of drone bee and returns to hive to create a new drone bee. If totally C number of classes are available, then confusion matrix of size C x C is formed. The entry of confusion matrix C_{ij} denotes number of true classes C_i classified as C_j. Therefore, diagonal entries of confusion matrix shows correctly classified samples and the remaining entries shows misclassified samples. Fitness function is set to maximize the classification accuracy of the overall system[29]. Overall classification accuracy is calculated by using formula in equation (16),

$$OCA = 100 * \frac{\sum_{i=1}^{c} C_{ii}}{\sum_{i=1}^{c} \sum_{j=1}^{c} C_{ij}} \qquad ...(16)$$

The pseudo code of HBM algorithm is

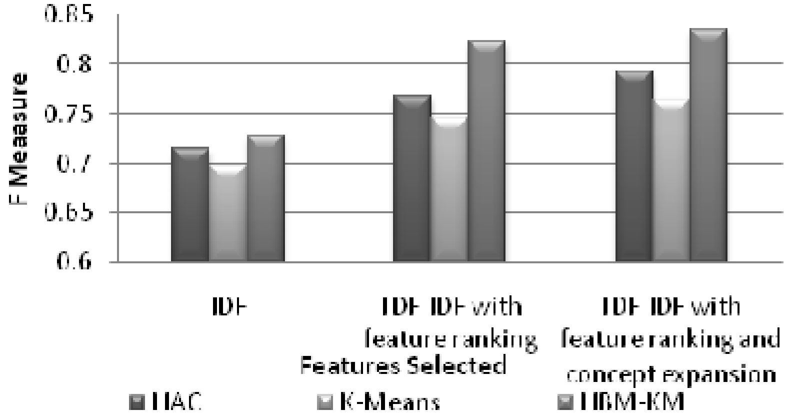


Fig. 2. F-measure for Reuters

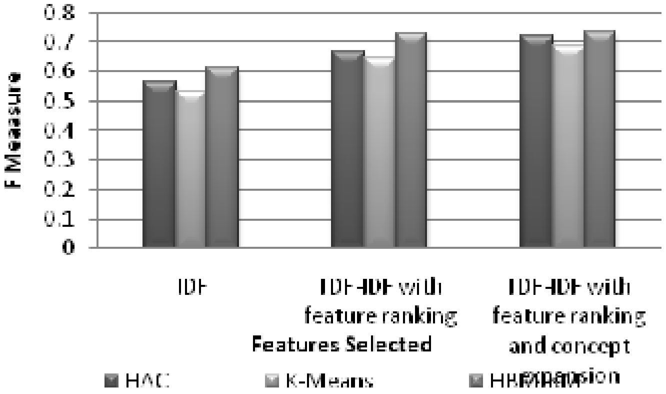


Fig. 3. F-measure for PubMed dataset

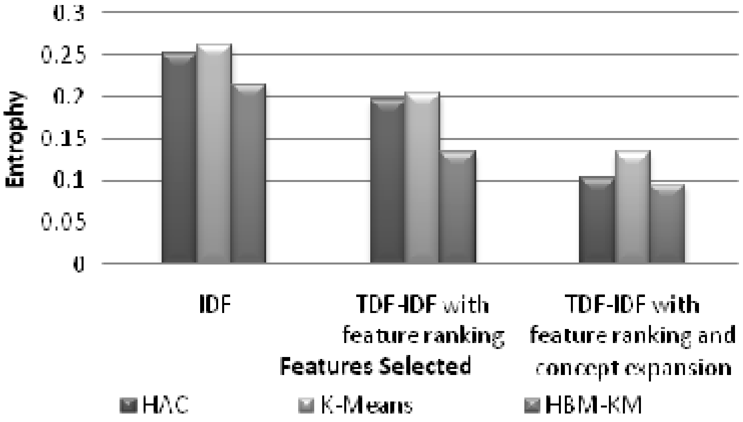
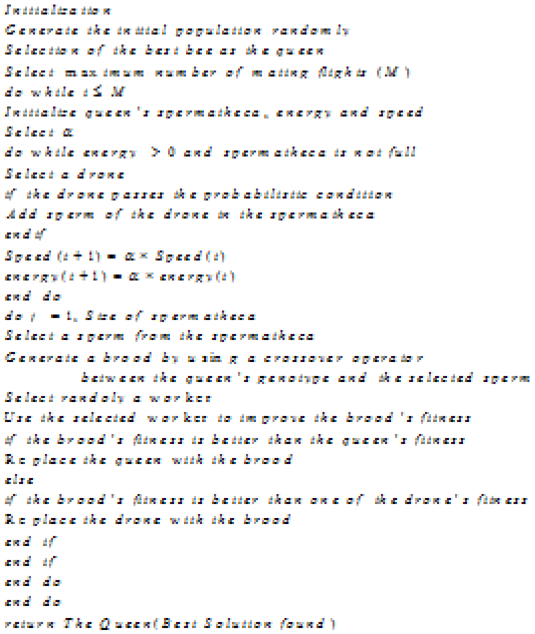


Fig. 4. Entropy for Reuters

shown in Figure 2. Table 1 tabulates the mapping parameters.

Table 1 Analogy between the natural honey bee colony and the artificial honey bee algorithm



```
Initialization
Generate the initial population randomly
Selection of the best bee as the queen
Select maximum number of mating flights (M)
do while i ≤ M
Initialize queen's spermatheca, energy and speed
Select α
do while energy > 0 and spermatheca is not full
Select a drone
if the drone passes the probabilistic condition
Add sperm of the drone in the spermatheca
end if
Speed (i + 1) = α × Speed (i)
energy(i + 1) = α × energy(i)
end do
do j = 1, Size of spermatheca
Select a sperm from the spermatheca
Generate a brood by using a crossover operator
            between the queen's genotype and the selected sperm
Select randoly a worker
Use the selected worker to improve the brood's fitness
if the brood's fitness is better than the queen's fitness
Replace the queen with the brood
else
if the brood's fitness is better than one of the drone's fitness
Replace the drone with the brood
end if
end if
end do
end do
return The Queen(Best Solution found)
```

RESULTS AND DISCUSSION

Documents from Reuter's dataset and PubMed dataset are used for evaluating the proposed HBM-KM. The Reuters dataset contains 21578 documents and tagged with keywords[30]. The dataset is tagged using 445 different keywords. The experiments were conducted using a subset of these dataset. The proposed technique is tested and compared with HAC and k means algorithms. Features are extracted from each document using IDF and are ranked by IG. Concepts are extracted for top ranked features. Parameters used for evaluating the efficacy of the algorithms were entropy and F-measure.

The parameters used for HBM are tabulated in Table 2.

The experimental results are shown graphically by the following figures 3 to 6.

Figure 3 shows the F-measure of classification for the Reuters dataset. From the figure it is observed that F-measure for proposed HBM-KM clustering increases up to 4.76 % when comparing HAC and also increases up to 7.88 % when comparing k-Means clustering.

Figure 4 shows the F-measure of classification for the dataset collected from PubMed dataset. Figure 4 reveals that F-measure for proposed HBM-KM clustering increases up to 6.72% when comparing HAC and also increases up to 11.79% when comparing k-Means clustering.

Figure 5 shows the entropy measure of classification for the Reuters dataset. From the figure it is observed that entropy measure for proposed HBM-KM clustering decreases up to 17.32 % when comparing HAC and also decreases up to 26.99 % when comparing k-Means clustering.

Figure 6 shows the entropy measure of classification for the dataset collected from PubMed dataset, and it is observed that entropy

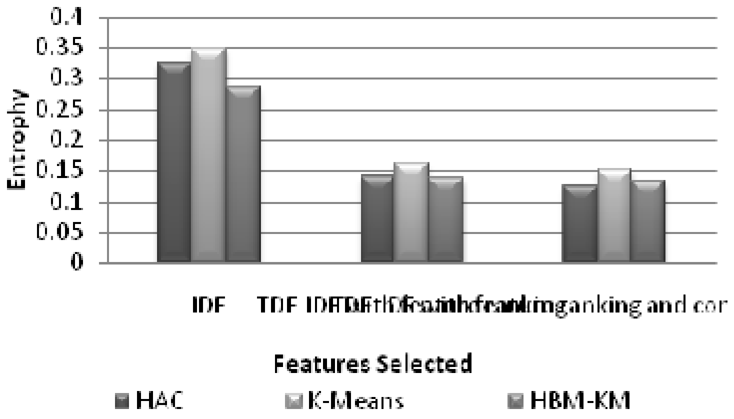


Fig. 5. Entropy for PubMed dataset

measure for proposed HBM-KM clustering decreases up to 2.76 % when comparing HAC and also decreases up to 15.16 % when comparing k-Means clustering.

CONCLUSION

Information retrieval and data mining applications need to get most relevant document from large data set. In this paper, k-Means clustering with Honey bee optimization is proposed for efficient clustering of the documents in a large database. Feature extraction from the documents and reduction is used to solve the problem of high dimensionality of text documents. Expanding the concepts for selected features will be useful for getting relevant document during information search and retrieval. Proposed HBM-KM clustering of the documents is compared with HAC and k-means clustering by entropy and F-measure. F-measure shows the effectiveness of clustering and entropy shows the homogeneity of clusters. Results show that proposed HBM-KM increased the F-measure and decreases entropy significantly.

REFERENCES

1. A. Murua, W. Stuetzle, J. Tantrum, and S. Sieberts, "Model Based Document Classification and Clustering", *International Journal of Tomography & Statistics, Winter* 2008; **8**.
2. K. Sridevi, R. Umarani, and V.Selvi, "An Analysis of Web Document Clustering Algorithms", *International Journal of Science and Technology,* 2011; **6**.
3. K.Sathiyakumari, G.Manimekalai, and V.Preamsudha, " A Survey on Various Approaches in Document Clustering ", International Journal of Computer and Technical Applications, 2011.
4. K. Raja, and C. Prakash Narayanan, "Clustering Technique with Feature Selection for Text Documents", Proceedings of the International Conference on Information Science and Applications ICISA 2010 6 February 2010.
5. Anna Huang, David Milne,Eibe Frank, and Ian H. Witten, "Clustering Documents using aWikipedia-based Concept Representation".
6. Lan Huang, David Milne, Eibe Frank and Ian H. Witten, "Learning a Concept-based Document Similarity Measure", September 2011.
7. FabrizioSebastiani "Machine Learning in Automated Text Categorization" *ACM Computing Surveys,* 2002; **34**(1).
8. Luiz G. P. Almeida, Ana T. R. Vasconcelos and Marco A. G. Maia," A Simple and Fast Term Selection Procedure for Text Clustering "Seventh International Conference on Intelligent Systems Design and Applications, 2007 IEEE,.
9. Yang, H., "A Document Clustering Algorithm for Web Search Engine Retrieval System", IEEE Conference on e-commerce, e-business, Jan 2010.
10. JianFeng cui and SeokChae, "Applying agglomerative hierarchical clustering algorithms to component identification for legacy systems", Journal of Information and Software Technology, volume 53, Issue 6, June 2011.
11. J.Sreemathy *et al.*, "An Efficient Textclassification Usingknn And Naive Bayesian "*, International Journal on Computer Science and Engineering (IJCSE),* 2012; **4**(3).
12. Lijuan Zhou, Linhusan Wang, XubeinGe, and Qisnsh ," An clustering based KNN improved algorithm CLKNN for text classification", Second International Conference on Informatics in Control, Automation and Robotics, 2010, Volume 3.
13. Sharma and Dihir R, "A Wordsets based document clustering algorithm for large datasets" International Conference on Models and Methods in computer Science, 2009.
14. Muhammad Rafi, Mehdi Maujood ,Murtaza Munawar Fazal, and Syed Muhammad Ali, "A comparison of two suffix tree-based document clustering algorithms", *IEEE transactions* , 2010.
15. XiangHua Fu, LianDong Liu, TianXue Gong, and Lan Tao ," Improving Text Classification with Concept Index Terms and Expansion Terms", *Advances in Neural Networks,* 2011; 6677.
16. Deepa M and Revathy P, "Validation of document Clustering based on Purity and Entropy Measures", *International Journal of Advanced Research in computer and Communications engineering,* 2012; **1**(3).
17. Alam S, Dobbie G, Riddle P and Naeem M.A, "Particle Swarm Optimization Based Hierarchial Agglomerative Clustering ", IEEE Conference on Web Intelligence and Intelligent Agent technology, 2010.
18. ChellatamilanT , Dr. Suresh R. M , "Concept Based Query Expansion and Cluster Based Feature Selection for Information Retrieval "*,Life Science Journal* 2013;**10**(7s).
19. Jacques Savoy, " A Stemming Procedure and Stop word List for General French Corpora ", *Journal of the American Society for Information Science,* 1999; **50**(10): pp 944-952.

20. Stephen Robertson, “Understanding Inverse Document Frequency: On theoretical arguments for IDF “, *Journal of Documentation,* 2000; **5**: pp 503–520.

21. Jiawei Han & Micheline Kamber Data Mining – Concepts and Techniques -, Morgan Kaufmann Publishers, Elsevier,2nd Edition, 2006

22. Roy. R.S.andToshniwal. D, “Fuzzy Clustering of Text Documents Using Nave Bayesian Concept”, International Conference on Recent Trends in Information, Tele Communication and Computing (ITC), Pp. 55-59, 2010.

23. Gong, Z., Cheang, C. W., & Hou, U. L. (2005, January). Web query expansion by wordnet. In *Database and Expert Systems Applications* (pp. 166-175). Springer Berlin Heidelberg.

24. Hierarchical Clustering, Chapter 17, Cambridge University Press, April 2009.

25. Abbass, H.A., A monogenous MBO approach to satisfiability. In: Proceeding of the International Conference on Computational Intelligence for Modeling, Control and Automation, CIMCA’2001, 2001a. Las Vegas, NV, USA

26. Afshar, A., Haddad, Bozog, O. Marino, M. A., Adams, B. J., Honey-bee mating optimization (HBMO) algorithm for optimal reservoir operation. *Journal of the Franklin Institute,* 2007; **344**: 452-462.

27. Sabar, N. R., Ayob, M., Kendall, G., & Qu, R. (2012). A honey-bee mating optimization algorithm for educational timetabling problems. *European Journal of Operational Research,* **216**(3): 533-543.

28. Cui, X., Potok, T. E., & Palathingal, P. (2005, June). Document clustering using particle swarm optimization. In Swarm Intelligence Symposium, 2005. SIS 2005. *Proceedings* 2005 IEEE (pp. 185-191). IEEE.

29. Magadalene Marinaki, Yannis Marinaki and Constantin Zopoundidis, “Honey Bee Mating Optimization algorithm for financial Classification Problem”, Journal of Applied Soft Computing, September 2009.

30. Steinbach, M., Karypis, G., & Kumar, V. (2000, August). A comparison of document clustering techniques. In KDD workshop on text mining (Vol. 400, N Table 1 Analogy between the natural honey bee colony and the artificial honey algorithm