International Journal of Engineering & Technology, 7 (3.24) (2018) 733-737

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

An approach towards Semantic Web Document Clustering using Background Knowledge

Sujata R. Kolhe[1]*, Dr. S. D. Sawarkar [2]

[1,2] *Department of Computer Engineering, Datta Meghe College of Engineering,Airoli Navi Mumbai, India*
**Corresponding author E-mail: crsujata@gmail.com*

Abstract

Extensive use of Information Technology in diverse applications has lead to massive online database. This database is often retrieved for various purposes. In order to smooth the progress of effective browsing and efficient searching for a user, one has to arrange the documents in a systematic manner. Clustering text documents is a vital step in organizing, management and indexing a huge text data on Web. Traditional approaches toil taking place in the keyword identical method, the routine destroys then the handler didn't change to the suitable upshot. Therefore, it grows into essential to signify the article semantically besides formerly remain managed. The background knowledge: Wordnet can be used for this purpose. In this paper Semantic weight is used to represent the document semantically and then a nature inspired meta heuristic algorithm : Cuckoo Search is implemented for automatic web search results clustering. Wordnet concepts are been identified from the text by considering the ontology of the text word. A novel method to find the semantic similarity is presented to represent the document as features.. These enhanced features are fed to the clustering algorithm. Cuckoo Search algorithm solves the problem of automatically defining number of clusters. Divide n conquer technique is used to avoid algorithm converging too quickly. The algorithm is tested on AMBIENT Dataset and shown good results.

Keywords— *Web Document clustering; k-mean; cuckoo search algorithm, Wordnet*

1. Introduction

The extensive growth of Internet has dramatically changed the way in which web documents are managed and accessed. So it becomes necessary to apply different techniques to organize documents to facilitate effective browsing and efficient searching. Data mining techniques have been applied to solve this Information Extraction. To achieve this, individual approaches that take stayed intensive is grouping, that clusters comparable information objects affording towards their contented, then edifices. The grouping can be functional to script data popular instruction to abstract the important information after the enormous data which is generated by various sources, such as E-books , news reports, Web pages, research articles, digital libraries, e-mail communications and diverse blogs. Documents grouping is programmed document association, text mining in demand to quick info recovery in an unsupervised method. Traditional script grouping procedure functioned improved to establish text article nevertheless organizational stuffing and additional sources of data can too be cast-off to improve grouping method

[1]. Researchers have established documents grouping, web grounded outcome grouping and data imagining constructed on numerous algorithms such by way of k-means algorithm, k-medoid algorithm, Bisecting k-means algorithm, Suffix Tree Clustering (STC) algorithm, Hierarchical clustering algorithm, Lingo besides optimization algorithm similar clustering by means of Ant colony optimization algorithm, Cuckoo Search algorithm, Fireflies algorithm then consequently. Web Grouping Engines remain the organizations that achieve clustering of web exploration outcomes. These schemes cluster the outcomes refunded through a pursuit engine addicted to a pyramid of categorized groups (likewise named groups). Roughly available clustering engines remain Clusty's, Grokker's, KartOO's, Lingo3G's, and CREDO's. The aforementioned systematizes hunt consequences through theme, therefore contribution a corresponding interpretation toward the flat ranked incline refunded by conservative exploration engine [2,3].

Document clustering involves 3 basic steps : Document pre-processing , Data representation Model using Feature Extraction and Clustering Model[4].

1. The steps involved in document preprocessing remain: Lexica's breakdown of the script file, Eradication of stop's word and Curtailing. Lexical analysis is used to separate the token words of the document. Some irrelevant words like a, is, an, the, do not contribute in clustering process. Such words are called stop words. Elimination of stop words aims to remove such words from the documents. Stemming is the process to bring down the token to its root word.
2. Data representation model deals with demonstrating the document into a term document matrix. Each file is characterized by a course of relationships seeming in that file. The relations in a document are represented by columns and documents as row. The weight of individual terms becomes each feature of the document. This

depiction is recognized by way of Vector's Space Models (VSM) [4]. Numerous approaches such by way of Latent's Semantic Analysis (LSA), Latent's Semantic Indexing (LSI), Singular's Values Decompositions (SVD) then Non-negative Matrix Factorization (NMF) remain cast-off aimed at improved illustration of the file and make it more easy for clustering[3].

3. Clustering Algorithm are broadly categorized as Partitioning approach (representative methods like:- k_means, k_medoids, Clarans), Classified approachs (typical methods:- Diana's, Agnes's, BIRCH's, CAMELEON's), Densitys_built method (typical methods are:- DBSACN's, OPTICS's, DenClue's), Grid based approaches (typical_methods:- STING's, Wave_Cluster's, CLIQUE's) ,Model_based (typical_methods:- EM's, SOM's, COBWEB's), Users_guided's or constraints_based (typical_methods:- COD's obstacle's, constrainer's clusterings) [5].

Entirely these outmoded approaches effort scheduled on the keyword's identical method, the presentation damages then the handler didn't become the appropriate consequences. Consequently, it developed essential to signify the file semantically then it can be administered. A perception grading can remain formed toward deliver the contextual information aimed at grouping. The usage of Word_Net consumes exposed inspiring consequences in clustering.

2. Related Work

A state of art in data grouping reports huge quantity of research with diverse methods. Some of them concentrate in representation of document. Some are worried about Methodology used and limitations of existing algorithms.

Apart from these conventional clustering_algorithm's, Different Web Pursuit Outcome Grouping Algorithms are proposed by various researchers. Those are broadly classified into following categories [6] :

Data- Centric Algorithms: Such kind of grouping algorithm's remain cast-off aimed at data clusterings which follows Partitioning approach resulting in superior solution but are poor in signifying the cluster label. K_Mean's clusterings algorithm's remains one and only of the maximum current clustering approaches in technical and engineering applications aimed at grouping a usual of information vectors. In these algorithm's, to each group is signified through the so_called group centroid. An impartial purpose is distinct by the summary of the inconsistencies among the opinions then its centroid articulated over suitable detachment. Usually Euclidian's detachment is cast-off in K_Mean's as the detachment quantity [8]. Though, for assumed randomly designated initial opinions, the algorithm describes a defiend mapping after preliminary points to the resolution. The aforementioned designates that original worth could disturb the answer to be attained. Consequently, suitable range of preliminary themes in K_Mean's is significant in creation a healthier grouping [9]. Similarly, these algorithms are delicate towards outlier.

Description Aware Algorithms: They are much worried about the classification problematic then stab to guarantee that the structure of group accounts remains that feasible besides it harvests consequences interpretable to a hominoid. Suffix Tree Clustering (STC) is an example for such types of algorithms. The aforementioned usages a tree construction to signify communal suffixes amongst forms. Created taking place in these shared suffixes, they classify dishonorable groups of forms, which remain then joint hooked on final groups created on an associated_components chart algorithm. The STD's archetypal delivers a supple n_grams technique to recognize then abstract very overlapping idioms in the forms [10]. A phrase_based document similarity created taking place the Suffix's Tree Documents (STD) prototypical. Through charting separately knob in the suffix bush of STD classical hooked on an exceptional features period in the Vector's Space Documents (VSD) archetypal, the phrase_built file resemblance obviously receives the period tf_idf's allowance preparation in totaling the documents resemblance through expressions. The compensations of two text representations (STD's then VSD's) continue shared to munch an expression constructed resemblance in text gathering [11]. A Documents Indexing's Graph (DIG) which is cast-off to signify a text precisely the phrase extant in document. The phrase_based guide of the file is built increments to arrange a DIG. The cosine's resemblance events over TF-IDF's (Term Frequency_Inverse Document Frequency) is cast-off as a solitary period resemblance quantity. Formerly, Parallel Histograms_created Grouping technique (SHC), an incremental histogram created grouping algorithms is cast-off to intensification the grouping competence [12].

Description Centric Algorithms : Such type of clustering balances between clustering quality and description of cluster. In such kind of clustering algorithms, Cluster label comes first and then the documents are allocated to the cluster according to the description. LINGO is an example for such types of algorithms. These algorithms syndicate a mutual phrase detection besides concealed semantical indexing's approaches to discrete exploration outcomes into expressive collections. The procedure requirement to safeguard that tags are meaningfully dissimilar while casing greatest of the themes in the contribution scraps. Towards invention such applicants, we custom the vector's space model's (VSM) then hidden semantics' indexing's (LSI) methods [13]. The semantics method aimed at text grouping by means of WordNet besides lexical shackles is projected. This method signifies, modified WordNet_based semantics resemblance quantity for word's sense disambiguation's (WSD) besides verbal manacles remain busy to abstract essential semantics structures that direct the theme of forms [14].

Cobos Proposed a novel description's_centric algorithms aimed at the web article grouping which utilizes the basic K-means algorithms, Global_Best Harmony exploration algorithms, Frequent phrase then Baysian Information Criterion. The harmony search is used as a global search strategy in which the clusters are optimized using Bayesian Information Criterion. It results into promising experimental grades in standard datasets [15]. Cui Proposed Particle Swarm Optimization (PSO) which provides the global solution in the complete solution space in assistance with K-means algorithm which provides local solution. Apiece partiale's resolution is estimated using a fitness's functions ADDC (average distance of documents) is used. The tentative consequences prove that by means of the mixture PSO's algorithms goods developed dense grouping than by means of whichever the PSO's before the K_Mean's unaccompanied [16].

Boura's proposed an enrichment in typical K-Mean's algorithms by means of the appropriate information after WordNet's hypernym's in dual way. The new weighing scheme is suggested for calculating the weight of the feature by generating the ontology graph for each feature using WordNet. Formerly, recognized clustering methods then distance trials are estimated on an update dataset by means of an evaluative standards of Clustering's Index (CI) besides F_measures. The aforementioned, is considered that K_means nearly continuously outpaces another clustering method. Also, cosine's resemblance then Euclidian's distances shows healthier for K_mean's, subsequently the intra groups likeness is a smaller amount than thru the city_block distances. Additional assumption is that the number of groups consume a result happening the CI's metrices. [17]

Wei's planned a semantics' method aimed at manuscript gathering by means of Word_Net then lexica's chain. Lexica's chains remain cast-off to receipts out concepts_based geographies which municipal the topic of text. The novel resemblance amount supportive WordNet

aimed at word intelligence disambiguation's is familiarized. The clustering outcome is appraised founded on F1_measures, cleanliness then entropies. The outcomes prove hopeful grouping technique [18].

Large number of executions of universal text clustering algorithms, may be found in some toolkits such as BagOfWords toolkit which gives overall idea of Text clustering[19].

3. Background

I. *Cuckoo's Search Algorithm:*

Cuckoo Search algorithms remains a meta heuristic's algorithm driven through some clutch parasitism's of a insufficient cuckoo's classes which is obtainable in landscape. It mimics the egg pattern and color of additional swarm natures (of additional kind). Respectively, egg in a shell is represented by a resolution, besides a cuckoo egg's characterizes a novel resolution. The objective remains to utilize a newly generated cuckoo solution to substitute weak solutions in the nest[20].

The purpose is to optimize the clustering process then towards discovery the finest conceivable resolution for clustering. The Cuckoo Search algorithm's is blended with the traditional k-means algorithm to automatically identify sum of groups. Balanced Bayesian's Information's Criterions (BBIC) as an Objective Function allows in the course of repeatedly choose the quantity of clusters. [21] The split and merge operations reduces execution time with better results. The flow for the same is given in Fig 1.

BBIC criterion to select finest nest is given by the equation 1.

$$\text{BBIC} = n * \text{Ln}\left(\frac{\text{SSE}}{n*\text{ADBC}}\right) + k * \text{Ln}(n) \quad (1)$$

n is total number of documents present in collection

ADBC- Average distance between all centroids

k - no. of clusters

$$\text{SSE}=\sum_{j=1}^{k}\sum_{i=1}^{n} P_{i,j} \; ||X_{i-Cj} \; ||^2 \quad (2)$$

$P_{i,j}$ - Its equals 'one' when docu. X_i belongs to cluster C_j. otherwise 0.

$$ADBC = \frac{2}{k*(k-1)}\sum_{j=1}^{k-1}\sum_{l=j+1}^{k} ||1 - Sim_{cos}(C_l, C_j)|| \quad (3)$$

k- no. of clusters

C_l,C_j are the Centroids of clusters l and j

II. *Wordnet*

Wordnet is an electronic dictionary for English language which includes infromation about 155000 nouns, adjectives, verbs and adverbs[22]. It describes hierarchical semantic association between the words which can be used to improve the clustering performance. Wordnet also provides an Ontology: a vocabulary of terms and relations in a graphical way. It is represented by a directed Tree which has a set of nodes and edges[22]. An ontology for terms 'Indigo' and 'blue' term is represented in Fig 2. The hypernyms and hyponyms can be added in the data representation as those can uncover the unknown similarity between the two terms. Based on the above ontology, we could compute the semantics resemblance among relations t1 and t2 by Wu Parmer as [23],

$$s(t1,t2) = \left[2 * \frac{d(lcs)}{d(t1)+d(t2)}\right] \quad (4)$$

where, d(lcs)= depth of longest common sub summer from root node d(t1)= depth of term t1 from root node

The semantic similarity plays vital role in Word Sense Disambiguation. According to the Wu Parmer the similarity between the terms 'blue' and 'indigo' is s(t1,t2)= [2*5/7+7]=0.14

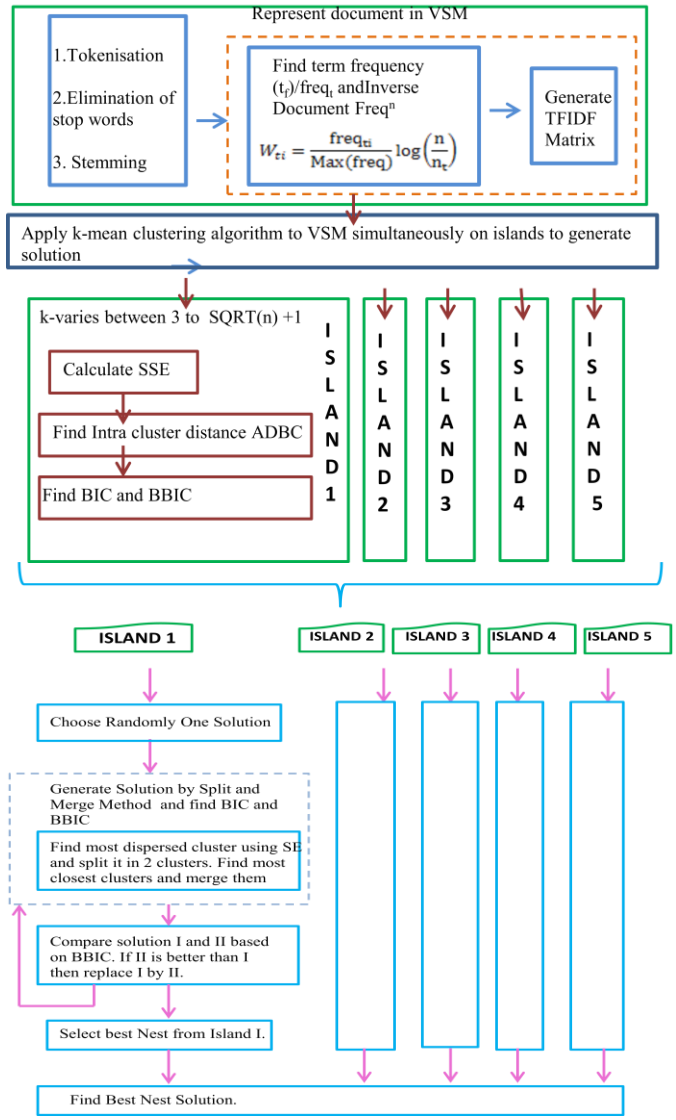

Fig 1. Flow of Cuckoo Search Algorithm

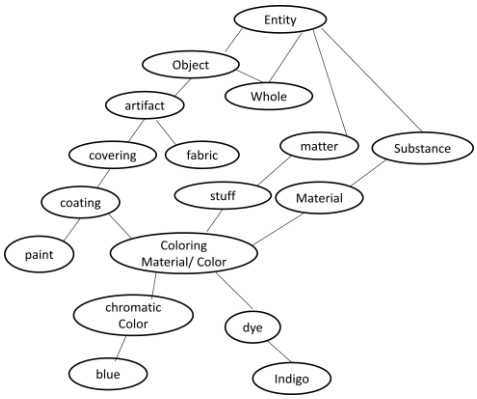

Fig 2. Ontology for the word blue and indigo

4. The Novel Algorithm: Semantic Web Document Clustering (Semantic Cuckoo Search)

In the proposed semantic Cuckoo Search algorithm, features are enhanced by using background knowledge in the preprocessing phase. Consider the Ontology for Indigo and Blue words given in figure 4 and the Term frequency for the terms appearing in documents as given in Table 1.

Table 1: Term Frequency for Documents D1 to D5

Document /Term	Indigo	Dye	color	blue	fabric	Paint
D1	0.2	2.19				
D2	0.2		1.24	0.987		
D3				1.44		
D4	0.2				1.09	
D5	0.2					1.09

Though all documents do not have common terms, they all belong to same category. If we consider document D3 , according to the ontology of the word blue the hypernyms of the words are chromatic color, color, material , substance etc. The hypernyms of the word paint in document D5 are coating, covering etc. Even the word Indigo has common hypernyms as color, material , substance etc. So the common words i.e. hypernyms which are present in other documents can add more weight to the terms. So we can update the original weights of a term by including the weights of all hypernyms present in all the documents. The updated weight plays important part in calculating the similarity.
The updated weights can be calculated as per below formula.

$$Wt_i = Tfidf_{(ti)} + \sum_{j=1}^{k} Tfidf_{(thj)} * S(t_i, th_j) \quad (5)$$

Wt_i - New Weight of term ti
$Tfidf_{(ti)}$= Old weight of term ti
$Tfidf_{(thj)}$ = Tfidf of hypernym hj of term ti
$S(t_i, th_j)$= Semantic similarity between t_i and its jth hypernym th_j present in all documents
k= total number of hypernyms
The semantic similarity is calculated according to the vital formula as below

$$S(t1,t2) = \frac{1}{d_{lcs(t1,t2)}} * \frac{1}{2}\left(\frac{d(t1) - d_{lcs(t1,t2)}}{path(t1)} + \frac{d(t2) - d_{lcs(t1,t2)}}{path(t2)}\right) \quad (6)$$

d_lcs(t1, t2)= depth of longest common subsummer from root node
d(t1)= depth of term t1 from root node
path(t1)= number of occurrences of the node going through the path towards root.
The semantic similarity between the terms 'blue' and 'indigo' is s(t1,t2)=1/5*1/2[(7-5)/3+(7-5)/3]=0.132.
Here maximum depth for longest common subsummer is considered.

The new weight for the term 'blue' is 1.44+(1.24*0.4)=3.08. The flow for updating the weights is given in Fig 3.

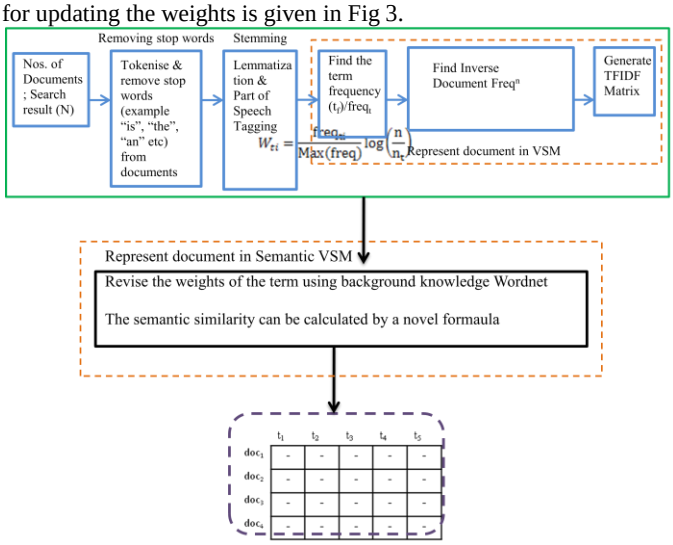

Fig.3 Feature enhancement using WordNet

5. Experimentation and Result

In this section we illustrate the results Semantic Cuckoo Search Algorithm and analyze how well it performs compared with k means algorithm and Cuckoo Search Algorithm. AMBIENT's [AMBIgous ENTries] is a dataset intended aimed at assessing subtopic info retrieval. The aforementioned includes 44 topics each topic with group of subtopic and 100 snippets (ranked search results) of each topic. ODP-239 is an Open Directory Project dataset having 239 topics, each of having 10 subtopics and 100 search results of single subtopic. All the topics, subtopics and snippets are chosen from Open Directory Project (www.dmoz.org). We have used 3 datasets by considering Open Directory Project (ODP) and Ambient datasets.
The clustering quality can be defined by different evaluation measures like BBIC, ADBC, Bayesian Information Criterion (BIC) Davies-Bouldin Index (DB). They are calculated and associated aimed at dissimilar amount of groups as given in Table.2

Table 2 Quality of cluster evaluated based on BIC, BBIC, ADBC and DB for different number of clusters

No. of clusters	BIC	BBIC	ADBC	DB
k=3	779.782	1.355	110.527	6.598
k=4	744.171	2.048	97.567	6.836
k=5	740.183	2.537	93.238	7.544
k=6	717.018	3.195	87.379	10.529
k=7	661.435	3.916	80.738	10.438
k=8	621.263	4.624	76.08	8.915

It is observed that as we increase the value of k (no. of clusters) the value of objective functions increasing. The superiority of clustering is improved as we increase the value of k.

The effectiveness of Semantic Cuckoo Search Algorithm is evaluated using three quality procedures Precisions, Recalls then F_scores. Precisions are intended by way of the proportion of the sum of relevant forms repossessed to the total amount of documents. Recalls is considered by way of the quantity of sum of suitable documents recovered to the entire sum of applicable documents in the class. F_measures is the harmonic means of precisions then recalls. Table.3 describe the accuracy of three different algorithms.

Table 3 Accuracy of clustering Algorithm

Algo	Dataset 1			Dataset 2			Dataset 3		
	P	R	F	P	R	F	P	R	F
K-means	70	40	50	65	45	53	62	56	59
CS	77	45	56	68	60	64	70	64	**67**
Semantic CS	75	55	**63**	70	60	**65**	70	56	62

It is observed that semantic cuckoo search outperforms k means then Cuckoo Search Algorithm (CS). Dataset 3 provides unlike result. Cuckoo Search outperforms other 2 algorithms.

6. Conclusion

The main result of this work is to demonstrate that use of background knowledge (Wordnet) can improve the results over traditional clustering algorithms. Also addition of hypernyms weight signifies the weight of the term and consequently the enhancement of features to represent the document is accomplished. Also it is observed that too many addition of hypernyms causes to represent the term in very general concept. The vital similarity measure improved the results. Three different algorithms are executed and the experimental results shows an improvement of Semantic Cuckoo Search algorithm concluded the Cuckoo's Search Algorithm's in addition K-mean's algorithms.

References

[1] Aggarwal C.C., Zhai C. (2012) A Survey of Text Clustering Algorithms. In: Aggarwal C., Zhai C. (eds) Mining Text Data. Springer, Boston, MA

[2] Carrot2. http://project.carrot2.org/.

[3] Carpineto, C., Osi´nski, S., Romano, G., Weiss, D.: A survey of web clustering engines. ACM Comput. Surv. 41(3), 17:1–17:38 (2009)

[4] R. Baeza-Yates, A.B. Ribeiro-Neto, Modern Information Retrieval, Addison-Wesley Longman Publishing Co., Inc., 1999.

[5] C. Carpineto, S. Osin´ ski, G. Romano, D. Weiss, A survey of Web clustering engines, ACM Comput. Surv. 41 (2009) 1–38.

[6] Data Clustering: Algorithms and Applications. CRC Press; 2014.

[7] McCallum, Andrew Kachites. "Bow: A toolkit for statistical languag modeling, text retrieval, classification and clustering." http://www.cs.cmu.edu/~mccallum/bow. 1996.

[8] Jain, Anil K., M. Narasimha Murty, and Patrick J. Flynn. "Data clustering: a review." ACM computing surveys (CSUR) 31.3 (1999): 264-323.

[9] P.S.Bradley,U.M.Fayyad. Refining initial points for K- Means clustering. In Proc. 15th International Conf. on Machine Learning, pages 91–99. Morgan Kaufmann, San Fran- cisco, CA, 1998

[10] O. Zamir and O. Etzioni, "Web Document Clustering: A Feasibility Demonstration," Proc. 21st Ann. Int'l ACM SIGIR Conf., pp. 46-54, 1998.

[11] Chim, Hung, and Xiaotie Deng. "Efficient phrase-based document similarity for clustering." IEEE Transactions on Knowledge and Data Engineering 20.9 (2008): 1217-1229.

[12] Hammouda, Khaled M., and Mohamed S. Kamel. "Efficient phrase-based document indexing for web document clustering." IEEE Transactions on knowledge and data engineering 16.10 (2004): 1279-1296.

[13] Osinski, Stanislaw, and Dawid Weiss. "A concept-driven algorithm for clustering search results." IEEE Intelligent Systems 20.3 (2005): 48-54.

[14] Ahmed, M.S., Amar, M.K.: Semantic Web Search Results Clustering Using Lingo and Wordnet. In: IJRRCS: Kohat University of Science and Technology (KUST), Vol. 1, No 2, pp. 71–76. , Pakistan (2010)

[15] Cobos, Carlos, et al. "Web document clustering based on global-best harmony search, K-means, frequent term sets and Bayesian information criterion." Evolutionary Computation (CEC), 2010 IEEE Congress on. IEEE, 2010.

[16] Cui, Xiaohui, Thomas E. Potok, and Paul Palathingal. "Document clustering using particle swarm optimization." Swarm Intelligence Symposium, 2005. SIS 2005. Proceedings 2005 IEEE. IEEE, 2005.

[17] Bouras, Christos, and Vassilis Tsogkas. "A clustering technique for news articles using WordNet." Knowledge-Based Systems 36 (2012): 115-128.

[18] Wei, Tingting, et al. "A semantic approach for text clustering using WordNet and lexical chains." Expert Systems with Applications 42.4 (2015): 2264-2275.

[19] McCallum, Andrew Kachites. "Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering." http://www. cs. cmu. edu/~ mccallum/bow (1996).

[20] Yang, Xin-She. Nature-inspired metaheuristic algorithms. Luniver press, 2010.

[21] Cobos, Carlos, et al. "Clustering of web search results based on the cuckoo search algorithm and Balanced Bayesian Information Criterion." Information Sciences 281 (2014): 248-264.

[22] Fellbaum, Christiane. WordNet. John Wiley & Sons, Inc., 1998.

[23] Wu, Zhibiao, and Martha Palmer. "Verbs semantics and lexical selection." Proceedings of the 32nd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 1994.