IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 3, Ver. IX (May-Jun. 2014), PP 34-40
www.iosrjournals.org

Obtaining the Name aliases from the Web, using them to Cluster Text Documents with Cuckoo Algorithm and Comparing Results with K-Means Algorithm

Pornima Deshpande[1], Smita Chaudhari[2]
[1]*(Computer Engineering, DYPSOE (Pune), India)*
[2]*(Computer Engineering, DYPSOE (Pune), India)*

Abstract: *There is an increase in the searching where name aliases are concerned. Approximately 30 percent of searches are based on aliases; hence it becomes important to obtain correct aliases. Lexical pattern based method is used to obtain the aliases of any personal or place from the web The aliases obtained are ranked and filtered based on the co-occurrence frequency and web dice methods These final aliases are then used to cluster the text documents present in a huge database. To get the best cluster cuckoo method of clustering is used. This method is based on the reproduction system of the cuckoo bird. According to the studies this clustering method when used with levy flight concept gives the best results when huge data is concern and also outperforms particle swarm optimization algorithm and genetic algorithm. The result will be compared with the result of k-means clustering method.*

Keywords: *Co-occurrence frequency, Genetic algorithm, K-means clustering method.*

I. Introduction

A PERSON or a place is known by many different names on the web referred to as aliases. This is the reason to fine the aliases. The aliases of any person or place should be accurate which helps to obtain maximum information related to them easily. This will help to absorb maximum possible information about that person on web. When the data obtained from the web search is concern it's huge. It is crucial to identify accurate aliases of a name, which makes extracting relationships among different entities easy.

The technique used for extraction of aliases is automatic extraction of lexical patterns [4]. The method obtains patterns which give accurate aliases for a personal name or name of a place. The obtained patterns are used to get the accurate aliases which are filtered to get best aliases. These aliases are further ranked with the help of a threshold value obtained from the web search engine.

These ranked aliases then are used for clustering of text documents which are present in a database. The cuckoo method of clustering is used with levy flight technique to obtain the best clusters. This clustering algorithm is based on the reproduction system of the cuckoo bird. The cuckoo bird lays her egg in the nest of bird belonging to different species and the surrogate bird unknowingly or raises the brood [17]. This concept is used for clustering the text documents using the aliases.

.

II. Related Work

The work is done on the automatic alias extraction and this work will be further used for clustering the text documents.

2.1 Lexical Pattern Based Automatic Alias Extraction Method

With the base of [4], the method of automatic name alias extraction method is implemented to extract the aliases from web. There are three parts for obtaining the aliases. In first part the best patterns (e.g. is known as, also called as etc.) are obtained which are refined with the help of co-occurrence frequency method. Lexical pattern based approach is used for obtaining the patterns. The method works by feeding a training data set of name or place with their aliases to the web. With the help of Ngram algorithm using a query name*alias the patterns are obtained [4]. This algorithm searches for the patterns and returns the snippets. The second part deals with obtaining the candidate aliases. It also uses the Ngram algorithm with query Namep* where p stands for patterns obtained in the previous part. It obtains n number of words after the pattern p, these n words are the candidate aliases extracted. The third part filters and ranks the aliases for obtaining the best aliases. The ranking is done base on the threshold value obtained from the web search engine. For filtering the co-occurrence frequency method and web dice algorithm are used [4]. The aliases obtained by this method are the best of the lot.

2.2 Clustering of Text Documents using Cuckoo Method and Comparison with K-means

As the data present in any database is always huge cuckoo method is used to form the clusters. This method is selected based on the studies of [10] and [19] which does the comparison between PSO, GA and CS and the other does the comparison between CS, ABC and DE. Both studies shows CS i.e. Cuckoo Clustering method outperforms the other methods. When the cuckoo bird lays eggs in the nest of the other species bird. The other species bird may, raise the brood, or it may destroy the egg, or it may leave the nest. These reactions are observed by cuckoo bird and the nest where best care of eggs is taken and are raised, are passed to the next generation as the best nest. This concept is powered by levy flight method [16], which finds the best and optimal clusters.

The algorithm form groups of the data points having same features. The class of the input data is k=not known during the process. The solutions obtained will be modified by levy flight technique. The obtained clusters are logically similar and different from each other. This means that the text documents in one cluster do not match to the documents in the other cluster but documents in a cluster are of similar type.

III. Method

As a huge amount of data is present on the related to a person or place. This information may be available with the reference to the different aliases of that name.

Where information retrieval is concern all aliases should be known for best results. This is the reason why it is necessary to work on the discovery of name or place aliases on the web. When all the aliases are known maximum information can be secured.

The system architecture in fig 3.1 shows the working of the implemented method. The method made up of two important parts. First part obtains the aliases and the second part does the clustering using both cuckoo method and k-means method and the comparison is done.

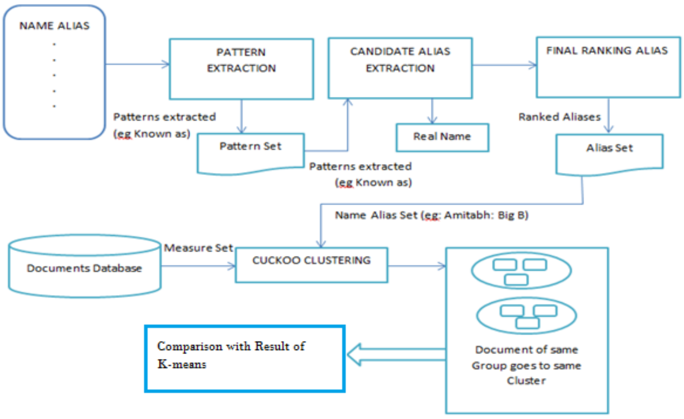

NAME ALIAS
PATTERN EXTRACTION
CANDIDATE ALIAS EXTRACTION
FINAL RANKING ALIAS
Patterns extracted (eg Known as)
Pattern Set
Patterns extracted (eg Known as)
Real Name
Ranked Aliases
Alias Set
Name Alias Set (eg: Amitabh: Big B)
Documents Database
Measure Set
CUCKOO CLUSTERING
Comparison with Result of K-means
Document of same Group goes to same Cluster

Fig. 3.1 System Architecture

The modules which are implemented with their sub modules are as follows:

1. Pattern Extraction Module
1.1 Search Engine phase
1.2 Snippet parser
1.3 N Gram algorithm
2 Candidate Name Aliases Module
2.1 Search Engine phase
2.2 Snippet parser
2.3 Extract candidate aliases
3 Final Name Aliases Ranking Module
3.1 Ranking
3.2 Final name aliases
4 Clustering and Comparison Module
4.1 Get dataset and parse
4.2 Clustering by cuckoo method
4.3 Clustering by K-means method
4.4 Comparison of results obtained by both methods

2.3 Obtaining the Aliases

2.3.1 Pattern Extraction Phase

The pattern extraction phase is sub divided into three steps as search engine part, snippet parser part and the N gram algorithm part. The search engine part deals with the searching for the name and alias which are present in the training data set provided to the search engine. The snippets are returned by the engine. The Ngram algorithm is used to find the patterns. The snippets obtained are replaced by [NAME] and [ALIAS] so as to obtain patterns. The wildcard operator "*" is used to relate the Name and Alias it helps to get the patterns. For example "Sachin is called the God of cricket" here "is called" is the pattern which will be obtained and a file of patterns is created.

Pseudo code for Ngram Algorithm:

```
a.    Algorithm: Obtain-patterns(T)
b.    T is a set of (Name, Alias) pairs
c.    P = Null
d.     for (Name, Alias)
a.    X = snippets(Name*Alias)
b.    for snippet yϵX
i.    P <- P+ Pattern(y)
e.    return(P)
```

2.3.2 Candidate Alias Extraction Phase

This part also has three steps. Which are same to the first part as again the query is given to the search engine i.e. "Name p*". Here Name any person or place name for which the aliases are to be obtained and p are the patterns. Here it will return "the God of Cricket" as the output. It again uses N gram algorithm but in different way the pseudo code is given below [4], here the n stands for the number of words to be obtained after the wildcard operator "*" .

```
a.    Algorithm:
Candidate-aliases(Name, P)
b.    P  is a set of patterns
c.    For(pattern p ϵ P)
d.     for (Name, Alias)
a.    X = snippets(Namep*)
b.    for snippet yϵX
 i.   C<- C+ GetNgrams(y, Name,p)
e.    return(C)
```

2.3.3 Final Alias Ranking Phase

This is the last part containing two steps as ranking and filtering. Filtering is necessary as the anchored texts may contain some noisy data due to which exact results are not obtained. Different ranking scores are defined for measuring the relation between name and alias. Co-occurrence frequency algorithm and web dice algorithm are used. Co-occurrence frequency is nothing but the number of urls in which an anchored text appears. Co-occurrence algorithm is used to find frequency of the name and alias pair appearing in different anchor texts of a url. The name and alias occurring in same urls are ignored. Co-occurrence frequency is inclined towards highly frequent words. After ranking and filtering the final file of ranked aliases is obtained.

2.3.4 Clustering and Comparison Phase

The web is a huge and widely distributed global information service center. Finding the relevant document from a data base is a huge challenge due to increase in information. This is why it creates the necessity to develop new technique which will help users to effectively navigate, curtail and establish the scared information. This can be done by using the technique of document clustering. This paper aims to develop such a technique using cuckoo clustering algorithm and apply it for text document.

2.4 Cuckoo clustering Method

As stated earlier the cuckoo method gives the best and approximate results in less time as compared to nature inspired algorithms. As K-means being the oldest and the sturdiest method the results obtained by cuckoo method are compared with that of the k-means algorithm. This phase three part, first part gets the data set and the documents which are to be clustered the documents are then parsed. All the stop words or noisy words are removed from the document. Remaining words are collected and represented in a vector space model. Cosine similarity formula [16] is used to calculate the distance between the document which is centroid and other

documents to find the minimum distance between the documents. The failed centroids will be removed and they will be replaced by new centroids using levy flight method. Clusters are then formed which are the best clusters.
There are few guidelines which are to be followed when using cuckoo method they are:

a. Only one egg is laid at a time, which is dumped randomly into nest of other species.
b. Best nest where the egg is taken care of will be forwarded to the next generation.

The numbers of available host nests are fixed, if the egg is discovered by the other bird say with pa probability [16]. Then pa is the probability that the surrogate bird may throw the egg or it may abandon the nest, building a completely new nest. This helps obtain the best nest or best centroid for clustering. For future clustering purpose the best nest results are stored and then used further. The pseudo code for cuckoo algorithm is as follows [16]:

```
1. Set the initial condition to n number of nests.
2. Repeat till the end of loop criteria is met.
a. Use fitness function (Fc) by levy flight to select cuckoo at random..
b.  Nest is randomly select.
c. Calculate fitness of nest using fitness function (Fn) .
d.  If we have (Fc < Fn) then replace the cuckoo is replaced by the nest.
e. A pa fraction of nest is replaced by new nests.
f. Again calculate fitness and keep best nests.
g. Optimal fitness value is that of the best nest which is stored for further use.
3. The center of cluster will be the best nest position.
```

This is how the clusters are formed using cuckoo method.

2.5 K-means Clustering Method

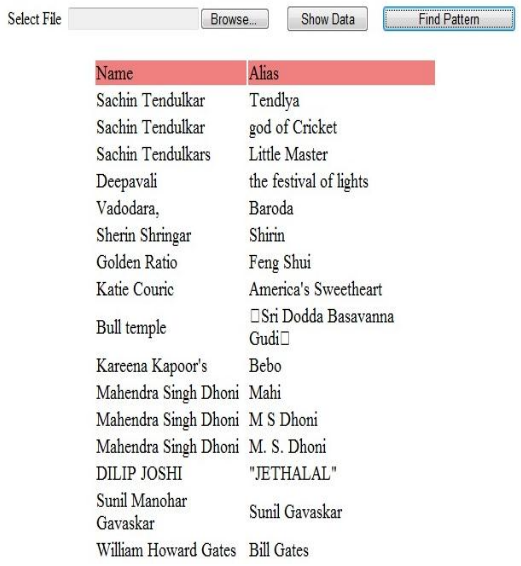
As k-means is the oldest clustering method which is used on a wide scale, this method is also implemented and the results obtained by cuckoo method and k-means method are compared on the basis of iterations required by both the methods to form clusters.

K-means clustering is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. Clustering is defined as grouping similar objects either physical or abstract. Each created group is called a cluster. The objects inside one cluster have most similarity with each other and maximum diversity with other groups. The algorithm is as follows:

```
Initialize clustering:
Loop until done
a. Compute mean of each cluster
b. Update clustering based on new means
c. End loop.
```

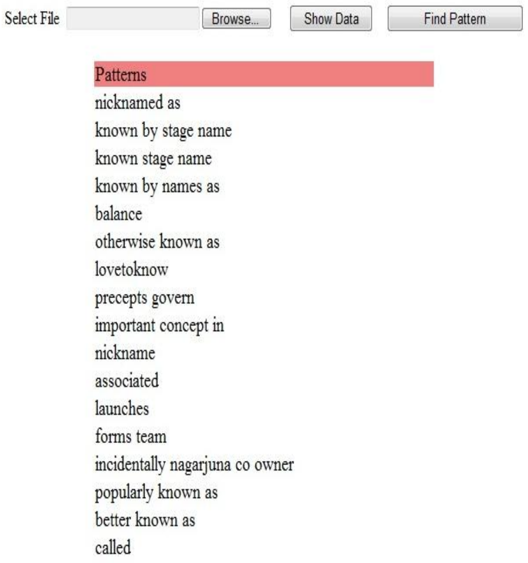
IV. Result

All the phases are implemented successfully. The snapshots of all the phases are given below. Figure 4.1 is the input to the first phase of patterns extraction where the training data set is displayed. Figure 4.2 is the output of the first phase where the patterns obtained are displayed. Figure 4.3 is the output of the second phase of the first part where the candidate aliases are obtained. Figure 4.4 is the output of the third phase where the final ranked aliases are obtained. Figure 4.5 is the output of the cuckoo clustering method and figure 4.6 is the output of cluster obtained by k-means algorithm.



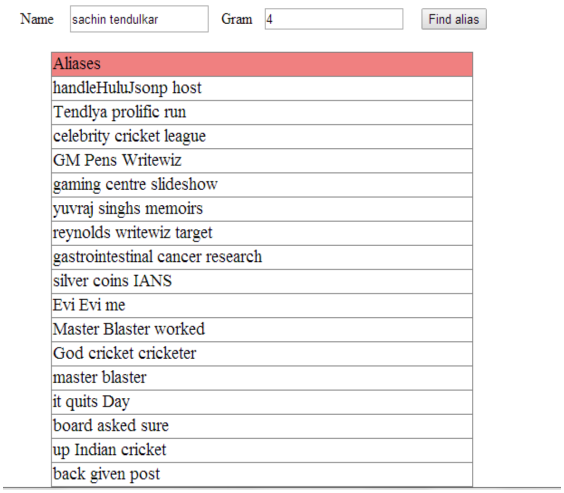
Name	Alias
Sachin Tendulkar	Tendlya
Sachin Tendulkar	god of Cricket
Sachin Tendulkars	Little Master
Deepavali	the festival of lights
Vadodara,	Baroda
Sherin Shringar	Shirin
Golden Ratio	Feng Shui
Katie Couric	America's Sweetheart
Bull temple	□Sri Dodda Basavanna Gudi□
Kareena Kapoor's	Bebo
Mahendra Singh Dhoni	Mahi
Mahendra Singh Dhoni	M S Dhoni
Mahendra Singh Dhoni	M. S. Dhoni
DILIP JOSHI	"JETHALAL"
Sunil Manohar Gavaskar	Sunil Gavaskar
William Howard Gates	Bill Gates

Fig. 4.1 Training Data Set.



Patterns
nicknamed as
known by stage name
known stage name
known by names as
balance
otherwise known as
lovetoknow
precepts govern
important concept in
nickname
associated
launches
forms team
incidentally nagarjuna co owner
popularly known as
better known as
called

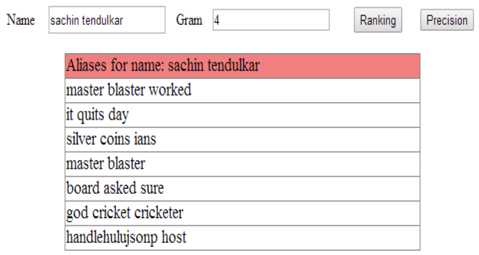
Fig. 4.2 Output of Patterns Extraction



Name sachin tendulkar Gram 4

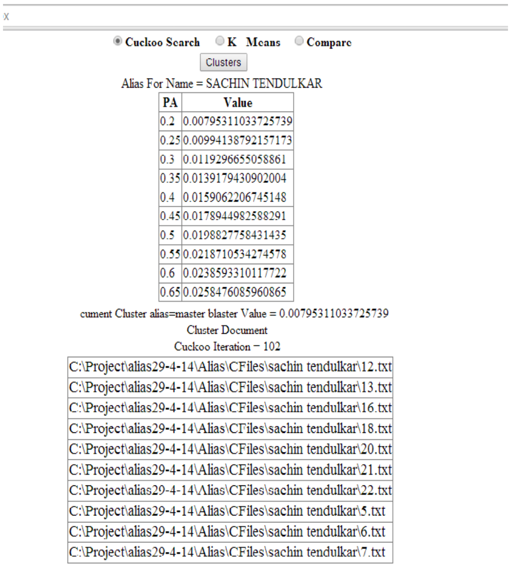
Aliases
handleHuluJsonp host
Tendlya prolific run
celebrity cricket league
GM Pens Writewiz
gaming centre slideshow
yuvraj singhs memoirs
reynolds writewiz target
gastrointestinal cancer research
silver coins IANS
Evi Evi me
Master Blaster worked
God cricket cricketer
master blaster
it quits Day
board asked sure
up Indian cricket
back given post

Fig. 4.3 Output of Ranked Aliases



Aliases for name: sachin tendulkar
master blaster worked
it quits day
silver coins ians
master blaster
board asked sure
god cricket cricketer
handlehulujsonp host

Fig. 4.4 Output of Ranked Aliases



Alias For Name = SACHIN TENDULKAR

PA	Value
0.2	0.00795311033725739
0.25	0.00994138792157173
0.3	0.0119296655058861
0.35	0.0139179430902004
0.4	0.0159062206745148
0.45	0.0178944982588291
0.5	0.0198827758431435
0.55	0.0218710534274578
0.6	0.0238593310117722
0.65	0.0258476085960865

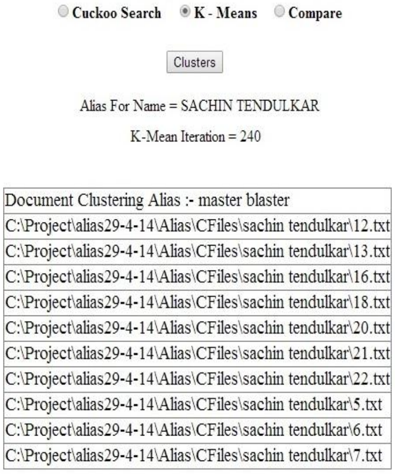
cument Cluster alias=master blaster Value = 0.00795311033725739

Cluster Document

Cuckoo Iteration − 102

C:\Project\alias29-4-14\Alias\CFiles\sachin tendulkar\12.txt
C:\Project\alias29-4-14\Alias\CFiles\sachin tendulkar\13.txt
C:\Project\alias29-4-14\Alias\CFiles\sachin tendulkar\16.txt
C:\Project\alias29-4-14\Alias\CFiles\sachin tendulkar\18.txt
C:\Project\alias29-4-14\Alias\CFiles\sachin tendulkar\20.txt
C:\Project\alias29-4-14\Alias\CFiles\sachin tendulkar\21.txt
C:\Project\alias29-4-14\Alias\CFiles\sachin tendulkar\22.txt
C:\Project\alias29-4-14\Alias\CFiles\sachin tendulkar\5.txt
C:\Project\alias29-4-14\Alias\CFiles\sachin tendulkar\6.txt
C:\Project\alias29-4-14\Alias\CFiles\sachin tendulkar\7.txt

Fig. 4.5 Output of Cuckoo Clusters



Alias For Name = SACHIN TENDULKAR

K-Mean Iteration = 240

Document Clustering Alias :- master blaster
C:\Project\alias29-4-14\Alias\CFiles\sachin tendulkar\12.txt
C:\Project\alias29-4-14\Alias\CFiles\sachin tendulkar\13.txt
C:\Project\alias29-4-14\Alias\CFiles\sachin tendulkar\16.txt
C:\Project\alias29-4-14\Alias\CFiles\sachin tendulkar\18.txt
C:\Project\alias29-4-14\Alias\CFiles\sachin tendulkar\20.txt
C:\Project\alias29-4-14\Alias\CFiles\sachin tendulkar\21.txt
C:\Project\alias29-4-14\Alias\CFiles\sachin tendulkar\22.txt
C:\Project\alias29-4-14\Alias\CFiles\sachin tendulkar\5.txt
C:\Project\alias29-4-14\Alias\CFiles\sachin tendulkar\6.txt
C:\Project\alias29-4-14\Alias\CFiles\sachin tendulkar\7.txt

Fig. 4.6 Output of Cuckoo Clusters

V. Conclusion

The automatic method to extract aliases of a personal or name of place is implemented successfully. Also the clusters are obtained using both cuckoo clustering method and k-means clustering method. The comparison between the both the methods is done based on the iterations required by both the methods to form the clusters. The comparison is done graphically as follows:

Here in the figure 5.1 it is observed that the when the number of files are increasing the number of iterations required are not increasing with a big difference. Here the number of iterations required for clustering 8 files does not differ with more number as compared to the iterations required to cluster 13 files.

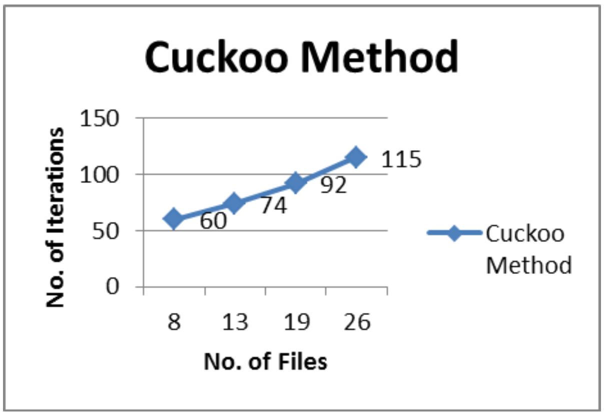


Fig. 5.1 Cuckoo Cluster Graph

The figure 5.2 shows the graph of iterations for k-means algorithm here it is observed that the number of iteration required for 8 files is more as compared to that of cuckoo algorithm. Similarly, the iterations are differing with huge number when the number of files is increasing.

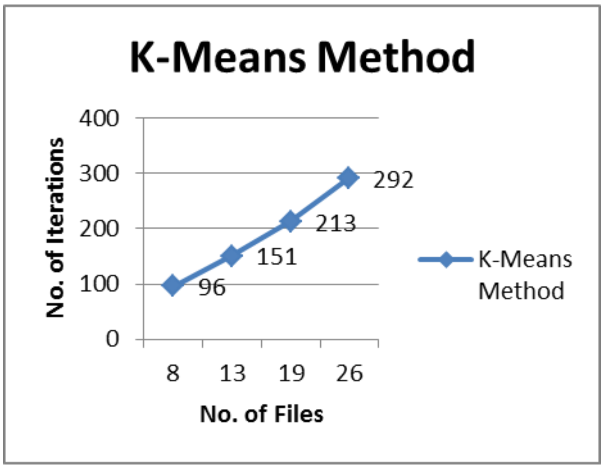


Fig. 5.2 K-means Cluster Graph

From both the graphs it is observed that when number of files are increasing k-means takes more number of iteration to obtain the cluster, more number of iterations actually means more time consumption. Also it happens that when huge number of data is concerned k-means fails to obtained the optimum solution in some cases. So it is observed that cuckoo is faster and accurate as compared to the k-means algorithm when huge data is concerned.

Future work: Document data time line generation which shows the clusters according to name aliases and sort those documents according to timeline.

References

Journal Papers:

[1] Danushka Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka, Automatic Discovery of Personal Name Aliases from the Web, IEEE Transactions on knowledge and data engineering, Vol. 23, No. 6, June 2011.

[2] Moe Moe Zaw and Ei Ei Mon, Web Document Clustering Using Cuckoo Search Clustering Algorithm based on Levy Flight, Faculty of Information and Communication Technology, University of Technology (Yatanarpon Cyber City), Myanmar, International Journal of Innovation and Applied Studies ISSN 2028- 9324 Vol. 4 No. 1 Sep. 2013, pp. 182-188 2013 Innovative Space of Scientific Research Journals http://www.issr-journals.org/ijias/.

[3] J. Senthilnath, Vipul Das, S.N. Omkar, V. Mani, “Clustering using Levy Flight Cuckoo Search”, J. C. Bansal et al. (eds.), Proceedings of Seventh International Conference on Bio-Inspired, Computing: Theories and Applications (BIC-TA 2012), Advances in Intelligent Systems and Computing 202, DOI: 10.1007/978-81-322-1041-2_6 Ó Springer India 2013.

[4] Pinar Civicioglu, Erkan Besdok “A conceptual comparison of the Cuckoo-search, particle swarm optimization, differential evolution and artificial bee colony algorithms”, Artif Intell Rev DOI 10.1007/s10462-011-9276-0, Springer Science+ Business Media B.V. 2011

[5] M. Mitra, A. Singhal, and C. Buckley, “Improving Automatic Query Expansion,” Proc. SIGIR ’98, pp. 206-214, 1998.