Improving the Effectiveness of Short Text Understanding by Using Web Information Mining

Ms.Aparna M. Katekar, Department of Computer Science and Engineering, University of Nagpur,
aparna.katekar24@gmail.com

Abstract—Short texts are always more difficult to understand. These short text are produced including social posts, conversations, keywords etc. and contains limited context. Short text consist more than one meaning and very complicated to understand. It consist limited amount of data because of that various methods like Text segmentation, part-of-speech tagging, and concept labeling are used. For better accuracy and result we used clustering. We use HMM(Hidden Markov Model)for text segmentation and POS tagging, Semantic matching algorithm for concept labeling as well as clustering algorithm (K-Means) for best results of Disambiguation. These all traditional methods are used for Short text understanding.

***Keywords*-- Text segmentation, HMM, Clustering, Concept labeling, Part-of-speech tagging.**

I. INTRODUCTION

Understanding short texts is very crucial to understand meaning of particular text. It is alsochallenging to number of applications. It isimportant that short text must be disambiguated, easy to implement and understandable.Short texts consist limited context and produces result with ambiguity. Also consist multiple meanings with same word.Manyapplications, like social media, conversations, micro blogging services, web search etc., are handle lots of short texts. Find hidden semantics from that particular text is most important task of text understanding. Sometimes same word consist of different meaning. So many efforts have been taken tothis field because if we don't know proper meaning of particular text then all information may be wrong or useless. Therefore, to understand different concepts of short text [1] we define short text understanding.We use various methods like text segmentation, POS tagging, concept labelingand as shown in Figure of structure overview. We also used clustering [6] for better accuracy and efficiency in less time and get better disambiguated result. With various algorithms like HMM, Semantic matching and K-Means [2] [7] [10] [15] algorithm which provides us better results and make short text more easy and understandable. The results show that knowledge is important for short text understanding and knowledge intensives are effective to collect meaning of short text.

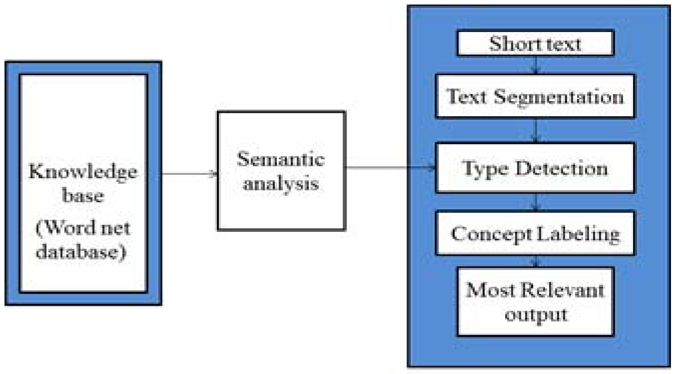


Fig: 1. Structure Overview of Short Text Understanding

II. PROBLEM STATEMENT

We have some common problems for short text understanding like Meaning, Vocabulary, Segmentation, Term, Type, Knowledgebase, Short text Understanding.

Meaning means definition of particular text or word. Vocabulary is a collection of phrase and words. A term is an entry in the vocabulary. Segmentation means divide a text in terms. Type denotes the lexical or semantic role a term plays in a text. Lexical types include verb and adjective. We consider lexical types in this work for two reasons. First, verbs and adjectives can help with instance disambiguation. A knowledgebase stores mappings between instances and concepts. Some existing knowledge bases also associate each concept with certain attributes. Short text is written in natural language so it is difficult to understanding. So terms like Text segmentation, Type detection, Concept labeling are used.

III. RELATEDWORK

A. Text Segmentation:- Text segmentation means divide a sentence into word. It converts long text into short text by using HMM (Hidden Markov Model)[4] [5] [9] [11] [12].The HMM works on multiple states and describes the probabilities of going from one state to another. The state is indirectly visible to the user, but the state dependent output is visible .It assign the state of word with highest probability. Probability distribution has taken in each state over the output tokens. In this we useRI WordNet as our database. Fig: 4, Fig: 5and Fig: 6 shows semantic analysis of input data[3] [8] [13] [14].

*B. POS Tagging:-*POS tagging consist POS tags of words in a text.Rule-based POS taggersassign POStagsto ambiguous terms and words [1].It also consist of HMM which defines the POS of the sentence and helps to understand the type of text. In means texts should fulfill sequential relations or tagging rules between consecutive tags. Below figure show POS tagging of input data after semantic analysis of input data.

*C. Semantic Labeling:-*It discovers hidden meaning from a natural language text [1]. It consist of Semantic matching algorithm[3] [8] [13] [14] identifies those nodes in two structures which semantically equivalent to each otherand find out whether the two structures are completely match to each other or not. If match then the result display in the form of list which indicates related domains. That domain is our created database for checking. When the particular domain formed then we check the distance of particular word and input text .If it is greater than 0.7 then we link the input base word to the domain in database. If it match to that domain then link that domain and display in list. So, it helps for harvesting of best match domains from the Semantic data. It matches words and meaning of words. In this we create domains present in the database. Fig: 7 and Fig: 8 shows harvesting of best match domains from the semantic data.

*D. Score Evaluation Using Jaccard Distance Matrix:-*When we find recommended words after that we find domain and connect to the Google to search main text. When we connect to the Google and get text files we evaluate score of related sentences formed by Google by usingjaccard distance matrix. It calculatesjaccard distances between the columns of 0-1 matrix. The jaccard distance between two species is 1-(number of regions consist both species)/(number of regions consist at least one species).

*E. Clustering:-*Clustering is the task of grouping a particular set of objects in same group called as cluster. When information from Google is formed then we get text file and after thatwe compare content with domain and action word by using K-Means [2] [7] [10] [15]. We get two clusters one for matching text and other for non- matching text. All sentences from matching texts are compared with each other and we get score by using jaccard distance matrix. After that we find mean of all scores. If score is greater than mean then discard else show output of summarized input sentence and we get detail information of input text.

IV. METHODOLOGY

A.Hidden Markov Model

It is simple Markov model used to check probability and works on multiple states. It describes the probabilities of going from one state to another. It assign the state of word with higher probability. HMM is used in Text Segmentation and POS tagging.

*1) Text Segmentation:-*In Text Segmentation we used Hidden Markov Model for short text conversion by dividing the text.

Algorithm:-

It consist of components:-
M = m1m2 ……mN =N states set.
B = b11b12 ……...bn1…..bnn= a transition probabilitymatrix B, each bij representingstate I to state j moving probability, s.t.Summation of n to j=1 bi j = 1 yeniS = s1s2 ……sT= a sequence of T observations, each observation drawnfrom a vocabulary.V = v1;v2……v*V*.A first-order hidden Markov model consists two simplifying assumptions.
First, the probability of a particular state dependsonly on the previous state:
Markov Assumption: P(mi|m1….mi-1) = P(mi|mi-1)
Second, the probability of an output observation oinot depends on any other state but on state that produce observation mi:
Output Independence: P(si|m1,…..,mi,……,mT ;s1,……,si,……,sT) = P(si|mi).

Example:-

In below example we see that the input i.e. m1 = i , m2 = want, m3 = to, m4 = go, m5 = to, m6 = Kashmir, m7 = for, m8 = holidays. After defining states and probabilities we get processed text i.e. s1 = want, s2 = Kashmir, s3 = holidays. After feature reduction there is only 3 words are present. It is calculated in 1766 ms.

*2) POS Tagging:-*Hidden Markov model is also used in POS Tagging defines the POS of the sentence and helps to understand the type of text.

Algorithm:-

We assume that we have a set of examples, (a(i); b(i)) for i =1 ……n, where each a(i) is a sentence a(i)1 …..a(i)ni , and each b(i) is a tag sequenceb(i)1 ……b(i)ni (we assume that the i'th example is of length ni). Hence a(i)j is the j'thword in the i'th training example, and b(i)j is the tag for that word.

Example:-

We use a1 …an for input to the tagging model. In the above example we have the length n = 8, anda1 = i, a2 = want, a3 = to, a4 = go, a5 = to, a6 = Kashmir, a7 = for, a8 = holidays.
We use b1 ……bn to denotethe output of the tagging model. In the above example we have b1 = v, b2 =v , b3 = n, b4 = n and so on.
This type of problem task is to map a sentence a1 ….an to a tag sequencey1 …..yn, is often referred as a taggingproblem or sequence labeling problem.

B. Semantic Matching Algorithm

*1) Semantic Labeling:-*It matches the meaning of words and identifies semantically corresponding nodesand checks whether the two structures are fully match or not. If match then it displays the result in the form of list which specifies related domains.

Algorithm:-

The algorithm takes two frameworks as a input and calculates output as a set of mapping.

It consist of four elements:

Step 1: Compute concepts of labels C_L, for all labels *L* in two structures.
Step 2: Compute concepts at nodes C_N, for all nodes *N* in two structures.
Step 3: Compute relations among C_L*'s*, for all pairs of labels in two structures.
Step 4: Compute relations among C_N*'s*, for all pairs of nodes in two structures.
Preprocessing phase represent by first two steps, while the element level represent by third and fourth steps and it also represent structure level matching respectively[3].*Step 1* and *Step 2* of the specific matching problem done independently once. Once the two match trees have been chosen*Step 3* and *Step 4* can be done at run time.

Example:-

First of all we calculate concept of labels and concept of node .In below example we find POS tagging and action words i.e. L1 = like, L2 = picture, L3 = camera, L4 = quality, L5 = phone. And element level matching takes place. After that node matching takes place i.e. N1 = camera, N2 = laptop, N3 = mobile phones, N4 = tablets, N5 = TVs, N6 = video surveillance, N7 = car. Single level belongs to multiple nodes.

*C. Jaccard Distance Matrix:-*It computes jaccarddistances between the 0-1 matrix column. The jaccard distance between two species is 1- (number of regions consist both species)/(number of regions consist at least one species).

*Arguments:-*Regret matrix 0-1-matrix. Columns= species, rows= regions.

Example:-

S1:- Hello there, S2:- Hi there, S3:-How are you, S4:- Hello madam. S1S2=1, S1S3=0, S1S4=1, S2S3=0, S2S4=0, S3S4=0.Therefore S=2. So Mean= 1+1/6=2/6=1/3=0.33. If S>Mean then discard the sentence otherwise show. Therefore we take sentence S2, S3, S4.

*D. K-Means:-*For clustering we used K-Means algorithm to split given dataset in to fixed number of clusters. In cluster we have to define centroid i.e. k .This algorithm minimizes an objective function and uses squared error function.The objective function,

number of clusters
number of cases
case i
centroid for cluster j

$$\text{objective function} \leftarrow J = \sum_{j=1}^{k}\sum_{i=1}^{n} \underbrace{\left\| x_i^{(j)} - c_j \right\|^2}_{\text{Distance function}}$$

Fig: 2. Objective function

Algorithm:-

1. Clusters the data into k number of groups where k is predefined.
2. At randomly select k points as a cluster center.
3. According to the Euclidean distance function assign objects to their closest cluster center.
4. Calculate the mean of all objects in each center of object.
5. Until the same points are assigned to each cluster repeat steps 2, 3 and 4 in consecutive rounds.

*Example:-*Suppose we want to group visitors by using their age (a one-dimensional space) as follows: 15,15,16,19,19,20,20,21,22,28,35,40,41,42,43,44,60,61,65Initial clusters: Centroid (C1) =16 [16], Centroid (C2) =22 [22].Iteration 1: C1= 15.33 [15,15,16], C2=36.25 [19,19,20,20,21,22,28,35,40,41,42,43,44,60,61,65]. Iteration 2: C1=18.56 [15,15,16,19,19,20,20,21,22] , C2=45.90 [28,35,40,41,42,43,44,60,61,65] . Iteration 3: C1=19.50 [15,15,16,19,19,20,20,21,22,28], C2=47.89[35,40,41,42,43,44,60,61,65]. Iteration 4: C1=19.50 [15,15,16,19,19,20,20,21,22,28],C2=47.89 [35,40,41,42,43,44,60,61,65].

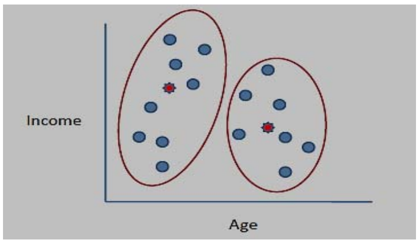


Fig:3. Clusters

No difference in between iteration 3 and 4 has been noted. Two groups have been identified 15-28 and 35-65 by using clustering. Run this algorithm multiple times and gate proper output with different starting conditions for better result.

V. SYSTEM SOFTWARE

In this project, Flow of software is as follows:

In first module semantic analysis of input data takes place as shown in Fig.4, Fig.5 and Fig.6 and in second module action words (noun, pronoun, adverb,etc.) are formed by using text segmentation and POS tagging as shown in Fig.7.

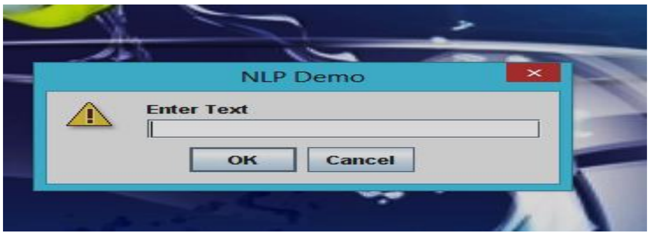


Fig: 4. NLP demo for enter text

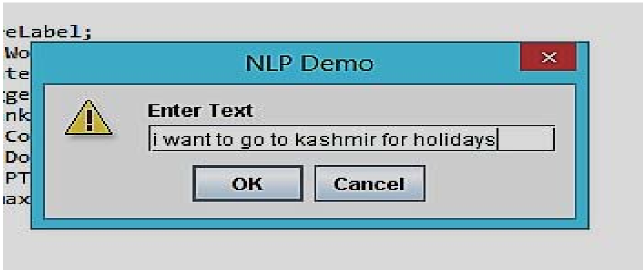


Fig: 5. NLP demo after entered text

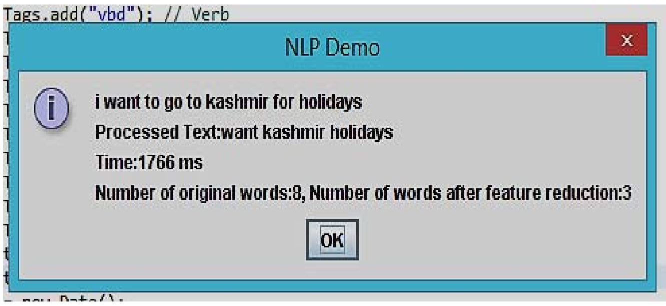


Fig: 6. NLP demo of action words

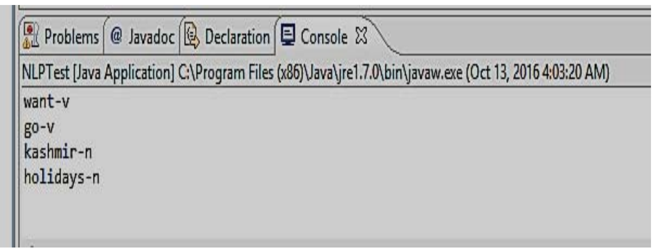


Fig: 7. POS tagging

In third module harvesting of best match domains from the Semantic data takes place by using concept labeling as shown in Fig.8 and Fig.9.

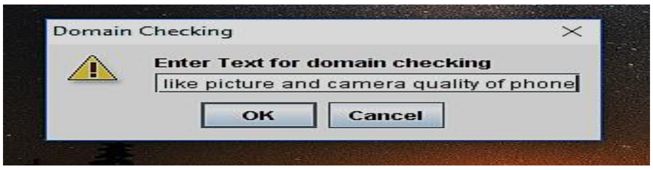


Fig: 8. Text entered for domain checking

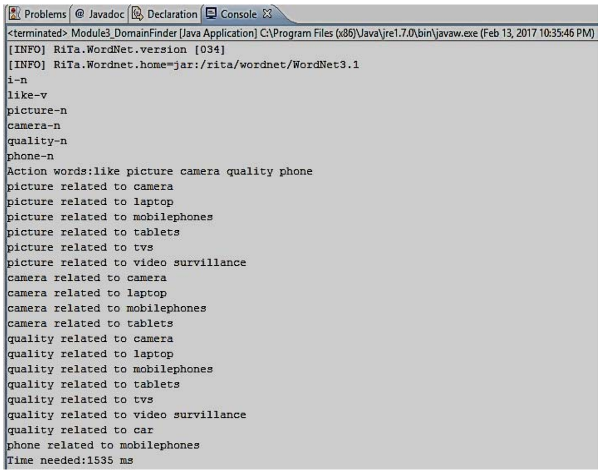


Fig: 9. Related domains

In fourth module score evaluation of harvested data takes place by using jaccard distance matrix as shown in Fig.10.

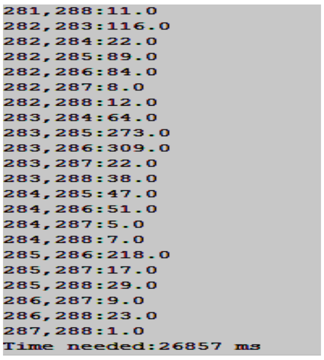
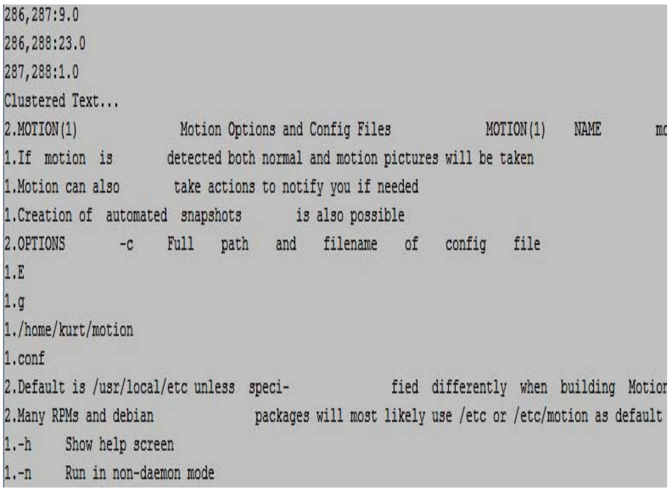


Fig: 10. Score evaluation

VI. EXPERIMENTAL RESULTS

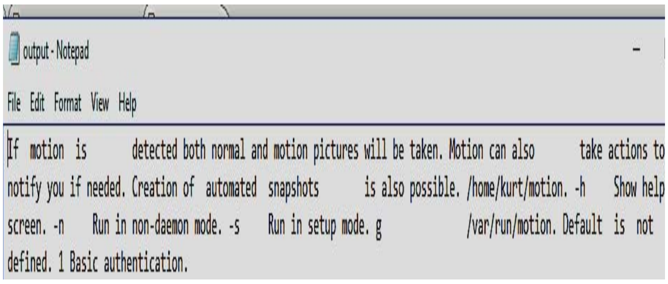
The correct output efficiency and accuracy of short text understanding is extremely critical. Therefore we use our framework to get correct output in less time period and we describe HMM, Semantic matching algorithm and K-Means for better result.After using all algorithms we get disambiguated results and output which show summarized input sentences.

- Development of clustering algorithm takes place for best results of disambiguation and show summarized output of input sentences.



```
286,287:9.0
286,288:23.0
287,288:1.0
Clustered Text...
2.MOTION(1)            Motion Options and Config Files            MOTION(1)    NAME       mc
1.If  motion  is       detected both normal and motion pictures will be taken
1.Motion can also       take actions to notify you if needed
1.Creation of  automated  snapshots       is also possible
2.OPTIONS       -c     Full    path    and    filename    of    config    file
1.E
1.g
1./home/kurt/motion
1.conf
2.Default is /usr/local/etc unless  speci-            fied  differently  when  building  Motior
2.Many RPMs and debian             packages will most likely use /etc or /etc/motion as default
1.-h     Show help screen
1.-n     Run in non-daemon mode
```

Fig: 11. Output of clustered text



```
output - Notepad
File  Edit  Format  View  Help
If  motion  is       detected both normal and motion pictures will be taken. Motion can also       take actions to
notify you if needed. Creation of  automated  snapshots       is also possible. /home/kurt/motion. -h     Show help
screen. -n     Run in non-daemon mode. -s     Run in setup mode. g             /var/run/motion. Default  is  not
defined. 1 Basic authentication.
```

Fig: 12. Output display on notepad

```
285,286:218.0
285,287:17.0
285,288:29.0
286,287:9.0
286,288:23.0
287,288:1.0
Summarized Text...
If  motion  is       detected both normal and motion pictures will be taken
Motion can also       take actions to notify you if needed
Creation of  automated  snapshots       is also possible
/home/kurt/motion
-h     Show help screen
-n     Run in non-daemon mode
-s     Run in setup mode
g              /var/run/motion
Default  is  not defined
1 Basic authentication
Time needed:25443 ms
```

Fig: 13. Output of summarized text

VII. Future Work

As a future work, the concept of image understanding can be added in future. That is the entire scenario of the image can be explained based on one image.We attempt to analyze and incorporate the impact of spatial-temporal features in to our framework for short text understanding.

VIII. Conclusion

The short text conversion is necessary for understanding .Text segmentation and type detection with Hidden Markov Model algorithmprovides semantic analysis of input data and formed action words while Concept labeling with Semantic matching algorithm is provides Harvesting of best match domains from the Semantic data. Clustering gives better accuracy, efficiency and provides more disambiguated results. It improves efficiency in less time and provides disambiguated result.

Acknowledgment

This work is supported by my guide and I would like to thank my guide for her valuable guidance. You definitely provide me with valuable guidance that is necessary to work on proper way and Isuccessfully completed my work.

References

[1] Wen Hua, Haixun Wang, Kai Zheng, Zhong yuan Wang, and Xiao fang Zhou, "Understand Short Texts by Harvesting and Analyzing Semantic Knowledge", IEEE Transactions on Knowledge and Data Engineering, 2016.

[2] Sumya Akter, Aysa Siddika Asa, Md. Palash Uddin, Md. Delowar Hossain, Shikhor Kumer Roy, Masud Ibn Afjal"An extractive text summarization technique for Bengali document (s) using K-means clustering algorithm", IEEE International Conference on Imaging, Vision & Pattern Recognition, 2017.

[3] Gerard Deepak, J Sheeba Priyadarshini, M S Hareesh Babu"A differential semantic algorithm for query relevant web pagereco mmendation", IEEE International Conference on Advances in Computer Applications, 2016.

[4] Kostadin Damevski, Hui Chen, David Shepherd, Lori Pollock "Interactive Exploration of Developer Interaction Traces using a Hidden Markov Model", IEEE/ACM 13th Working Conference on Mining Software Repositories, 2016.

[5] Qiang Zhang, Baoxin Li, "Relative Hidden Markov Models for video-based evaluation of motion skills in surgical training", IEEE transaction on pattern analysis and machine intelligence,2015.

[6] Annalisa Appice and Donato Malerba,"A co-training strategy for multiple view clustering in process mining", IEEE transactions on services computing, 2015.

[7] Kazuki Ichikawa and Shinichi Morishita ,"A simple but powerful heuristic method for accelerating k-means clustering of large-scale data in life science", IEEE/ACM transactions on computational biology and bioinformatics, 2014.

[8] Bowen Li ,Lin Zhang, Anrui Hu, Yifan Mai, "A simulation management system with the application of semantic matching algorithm based on ontology", IEEE 3rd International conference on cloud computing and intelligence systems, 2014.

[9] Tongwei Lu , Ling Peng, Shaojun Miao, "Human Action Recognition of Hidden MarkovModel Based on Depth Information", 15th International Symposium on Parallel and Distributed Computing, 2016.

[10] Jai Puneet Singh, Nizar Bouguila, "Proportional data clustering using K means algorithm: A comparison of different distances", IEEE International Conference on Industrial Technology, 2017.

[11] Jun Du , Zi-Rui Wang, Jian-Fang Zhai, Jin-Shui Hu, “Deep neural network based hidden Markov model for offline handwritten Chinese text recognition”, 23rd International Conference on Pattern Recognition, 2016.

[12] Petya Dinkova, Petia Georgieva, Agata Manolova, Mariofanna Milanova, “Face recognition based on subject dependent Hidden Markov Models”, IEEE International Black Sea Conference on Communications and Networking, 2016.

[13] Shuoyan Liu, Kai Fang, Li Jiang, “A novel template matching algorithm based on the contextual semantic information”, IEEE International Conference on Information and Automation, 2015.

[14] Bowen Li, Lin Zhang , Anrui Hu , Yifan Mai, “A simulation knowledge management system with the application of semantic matchingalgorithm based on ontology”, IEEE 3rd International Conference on Cloud Computing and Intelligence Systems, 2014.

[15] Nur Ulfatur Roiha, Yoyon K. Suprapto, Adhi Dharma Wibawa, “The optimization of the weblog central clusterusing the genetic K - means algorithm”, International Seminar on Application for Technology of Information and Communication, 2016.

978-1-5090-4890-8/17/$31.00 ©2017 IEEE