SEMANTIC ANALYSIS BASED TEXT CLUSTERING BY THE FUSION OF BISECTING K-MEANS AND UPGMA ALGORITHM

G. Loshma[1] and Nagaratna P. Hedge[2]
[1]Jawaharlal Nehru Technological University, Hyderabad, India
[2]Vasavi College of Engineering, Hyderabad, India
E-Mail: loshmagunisetti15@gmail.com

ABSTRACT

Owing to the fastest data growth, this era can be claimed as the era of zettabytes. An effective mechanism is the need of this hour, to manage all the available data efficiently. Clustering is a technique to group relevant documents together. This work takes the semantics into account and clusters the document with the hybrid of bisecting k-means and UPGMA algorithm. The semantic analysis is made possible by the inclusion of wordnet, which is a lexical database. The outcome of this algorithm is more accurate, as the clusters are meaningful. The performance of the proposed algorithm is evaluated with respect to precision, recall, F-measure, accuracy and misclassification rate. The experimental results of the proposed work are satisfactory.

Keywords: clustering, bisecting k-means, UPGMA, wordnet.

INTRODUCTION

The Internet handles several EBs (exabytes) of data every single month. Thus, this era can be claimed as the era of ZBs (zettabytes). A single EB and ZB are equivalent to 1000000 TBs (Terrabytes) and 1000 exabytes, respectively. Thus, an effective mechanism is the need of this hour, to manage all the available data efficiently. Most of the present data are in textual format. Another problem to be addressed is the efficient data retrieval. The data retrieval speed is directly proportional to the growth of data. Delayed data retrieval is not tolerable in this fast paced world. Speedy data retrieval is possible only when the available data is organised properly.

Clustering is an efficient technique that aims at organising the similar data together in clusters. Thus, the required data can simply be retrieved from a vast collection of data in a shorter span of time. The main goal of a text clustering algorithm is to group relevant data together into clusters. The so formed clusters are easily controllable and results in faster data retrieval. An efficient clustering algorithm paves way for effective information retrieval with the least response time and thus improves the Quality of Service (QoS).

Several text clustering algorithms are presented in literature. However, most of the clustering algorithms do not deal with the semantic information of the text. For instance, Bag of Words (BoW) is the approach that does not consider the semantic relationship of the words. The cluster formation is more accurate when the clustering algorithm takes semantic relationship into account. This sort of clustering algorithm takes the textual meaning of the text into account, in order to cluster the text. The resultant clusters are meaningful because the clusters are formed on the basis of the literal meaning of the words [1-11]. Thus, the clustering efficiency can be improved.

This paper presents a semantic based clustering algorithm by incorporating wordnet and feature based semantic similarity measure. Wordnet is the most popular English thesaurus, which lists out the sematic relationship between terms. This work employs a feature based semantic similarity measure, which relies on the semantic relationship between the terms. Finally, the so formed clusters are labelled for easier recognition. Labelling is not a big deal in this work, as the semantic relationship between the terms is the central theme of this work.

The rest of the paper is systematized as follows. Section 2 presents the review of literature. The proposed algorithm is explained in the section 3. The performance of the proposed algorithm is analysed and the evaluation results are discussed in section 4. Finally, the concluding remarks are presented in section 5.

Review of literature

This section reviews the related clustering algorithms in the literature. Text clustering is the most researchable topic and several works are already present in the literature. This section aims at analysing the related works.

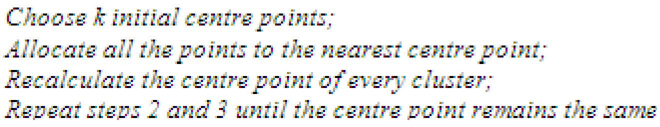
Wordnet

Wordnet is one of the largest thesauruses of English language. It connects all the terms to relevant terms, with respect to their meaning. It contains synonyms and the relationship of terms. Wordnet 2.1 consists of 1, 55,327 words in 1, 17,597 senses. Synset is a technical term of wordnet, which aggregates nouns, verbs, adjectives and adverbs to form synonym set. This lexical database is employed for text clustering applications to improve the accuracy based on semantics.

K-Means algorithm

K-means algorithm is the popular clustering algorithm, and this term was first introduced by James Macqueen in the year of 1967 [12]. The standard algorithm was presented by Stuart Lloyd in the year 1957. K-means algorithm is easy to implement and thus many clustering problems employed k-means algorithm.

This algorithm consumes minimal time for execution [13-16]. The pitfall of this algorithm is its dependency on cluster point [17-20]. The steps involved in k-means algorithm are presented below.

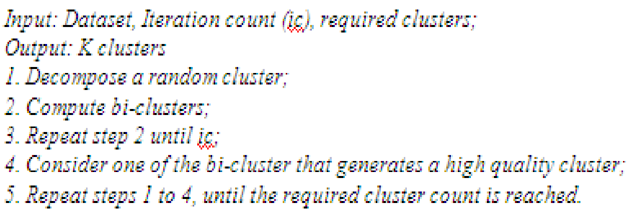


```
Choose k initial centre points;
Allocate all the points to the nearest centre point;
Recalculate the centre point of every cluster;
Repeat steps 2 and 3 until the centre point remains the same
```

The initial centre points are needed to be chosen and then all the points are allotted to the nearest centre point. The centre point of all the clusters is calculated again and this step is repeated until there is no change in the centre points.

Bisecting k-means algorithm

Bisecting k-means algorithm is the improved version of k-means algorithm. This algorithm iterates by selecting a cluster and follows a principle to divide the cluster. This process gets over as soon as the required count of clusters is attained or when the whole hierarchical tree is formed. The standard bisecting k-means algorithm is presented below.



```
Input: Dataset, Iteration count (ic), required clusters;
Output: K clusters
1. Decompose a random cluster;
2. Compute bi-clusters;
3. Repeat step 2 until ic;
4. Consider one of the bi-cluster that generates a high quality cluster;
5. Repeat steps 1 to 4, until the required cluster count is reached.
```

The clustering quality of this algorithm depends on the selected stopping criteria. This can be achieved by splitting the largest cluster and then to bisect the cluster by taking the cluster centroid into account.

UPGMA algorithm

UPGMA algorithm is the Unweighted Pair Group Method with Arithmetic Mean algorithm, which follows the bottom-up approach. This algorithm tends to construct a dendrogram by clubbing the two nearer clusters. The clustering process is achieved by the exploitation of distance or similarity matrix.

Proposed approach

This section presents the proposed text clustering algorithm. The proposed text clustering algorithm has its underlying roots on feature based semantic similarity measure, which yields the fruits of accuracy. The entire work is compartmentalized into phases such as pre-processing; feature based semantic similarity, clustering operation and cluster labelling. All these phases are described as follows.

Pre-processing

Pre-processing is the most important step in any sort of application, as it makes the data more suitable for further process. In this work, this step expedites the clustering process and enhances the quality of the algorithm. This phase eliminates stop words and stems the terms available in the document and represents the text document in a suitable format.

The initial step aims at eliminating or removing the words which has no meaning on its own. For instance, conjunctions, prepositions, articles and pronouns are meaningless by themselves. These words are called as stop words and are removed from the documents, in order to save memory and time. The sample stop words are listed in Table-1.

Table-1. List of stop words.

List of stop words			
A	an	The	About
Above	Across	Afore	After
against	Along	Aside	Beside
Among	Except	Include	In
Out	Despite	During	Below
Beyond	From	Into	Unto
To	Out	Until	Till
With	Than	Through	Upon
And	But	Because	For
So	Or	Yet	I
You	My	Me	He
She	Who	Myself	herself

The second step of term stemming is triggered after the completion of stop words removal. The task of stemming aims at clipping the words by removing ‘ing’, ‘ed’, ‘es’, ‘s’ and so on. This phase removes stop words and stems the available words, so as to save memory and execution time.

Text document representation

The text documents are represented in such a way that the documents are represented as vectors. Two documents are claimed to be similar if those documents have high degree of correlation between them. All the documents are organised as vectors in the vector space as matrix. The term weights of all documents are given by

$$doc_i = wt_{1i}, wt_{2i}, .. wt_{hi} \tag{1}$$

Where doc_i is the specific document, wt_{1i} is the weight of first term in the i^{th} document, wt_{hi} is the weight of the h^{th} term in the i^{th} document.

$$Vdoc_i = \{wt_{1,i}, wt_{2,i}, .. wt_{h,i}\} \tag{2}$$

Equation 2 notifies the vector space model of the documents. $wt_{1,i}$, $wt_{2,i}$ are the term weights of the documents and are computed by

$$wt_{h,i} = tf_h * IDF \tag{3}$$

$$\text{IDF} = \log(\frac{Doc}{docf_h}) \tag{4}$$

tf_h is the occurrence frequency of h in the i_{th} document, $docf_h$ is the total count of documents that possesses the term h, Doc is the total number of documents in the dataset. The weight of the document is fixed on the basis of the importance of term. However, the above equations from 1 to 4 focus on the occurrence frequency of the terms alone.

This work formulates the vector space model by taking the semantic of the term into account and is presented below.

Semantic similarity

The semantic similarity between terms is computed by the incorporation of wordnet [21-25]. Wordnet is a lexical database which accumulates the terms called as synsets. The semantical relationship between terms is calculated by taking the semantic correlation between the terms. Every word is checked for the semantic relationship of another word in wordnet.

Let $\alpha_{h1,h2}$ is the semantic relationship between two terms wd_1 and wd_2. In case, if wd_2 is present in the synset of wd_1, then $\alpha_{h1,h2}$ is set to 1; otherwise $\alpha_{h1,h2}$ is set to 0 and is represented in (5).

$$\begin{cases} wd_2 \in wd_1 & \alpha_{h1,h2} = 1 \\ wd_2 \notin wd_1 & \alpha_{h1,h2} = 0 \end{cases} \tag{5}$$

The weight wd_{ij1} of term t_{i1} in document doc_x is given by (6).

$$wd_{ij1} = wd_{ij1} + \sum_{\substack{h2=1 \\ h2 \neq h1}}^{i} \alpha_{h1,h2}\ wd_{ij2} \tag{6}$$

By this way, the semantic relationship between every pair of terms is computed. This is followed by the computation of similarity measure. This work exploits the cosine similarity between the documents and is presented in (7, 8).

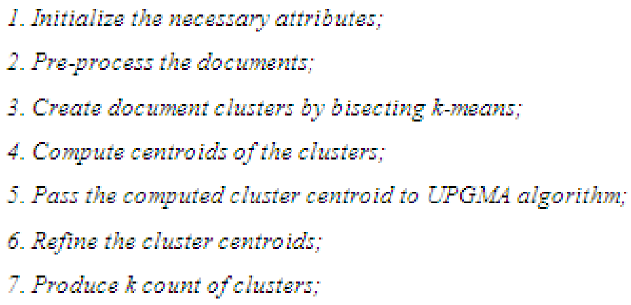
$$Sim(Doc_a, Doc_b) = cosine(Doc_a, Doc_b) \tag{7}$$

$$cosine(Doc_a, Doc_b) = \frac{\sum_{i=1}^{n} wd_{ij1} wd_{ij2}}{\sqrt{\sum_{i=1}^{n} wd_{ij1}^2 . wd_{ij2}^2}}, a \tag{8}$$

Thus, the semantic similarity between the terms and the documents are found out. This step is followed by the process of clustering.

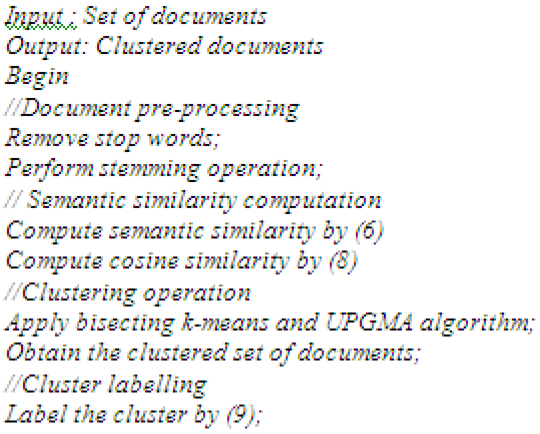
Clustering algorithm

This work exploits the fusion of bisecting k-means algorithm and UPGMA algorithm. Thus, the proposed work inherits the merits of both the algorithms. Bisecting k-means algorithm follows top-down approach, whereas UPGMA algorithm utilizes the bottom-up approach. This paves way for the generation of refined clustering results. The algorithm is presented below.



```
1. Initialize the necessary attributes;
2. Pre-process the documents;
3. Create document clusters by bisecting k-means;
4. Compute centroids of the clusters;
5. Pass the computed cluster centroid to UPGMA algorithm;
6. Refine the cluster centroids;
7. Produce k count of clusters;
```

The outcome of the above presented algorithm is the refined clusters of documents. The related documents are clustered together, thus the documents present in a cluster are highly correlated. On the other hand, the documents in the different clusters show lesser degree of similarity. By this way, the text documents are clustered effectively based on the semantics. The overall flow of the algorithm is presented below.



```
Input : Set of documents
Output: Clustered documents
Begin
//Document pre-processing
Remove stop words;
Perform stemming operation;
// Semantic similarity computation
Compute semantic similarity by (6)
Compute cosine similarity by (8)
//Clustering operation
Apply bisecting k-means and UPGMA algorithm;
Obtain the clustered set of documents;
//Cluster labelling
Label the cluster by (9);
```

The outcome of the above presented algorithm is the refined clusters of documents. The related documents are clustered together, thus the documents present in a cluster are highly correlated. On the other hand, the documents in the different clusters show lesser degree of similarity. By this way, the text documents are clustered effectively.

Cluster labelling

In this step, the degree of recurrence of all terms in all documents of every cluster is found out. This is followed by arranging the terms in descending order with respect to the degree of recurrence. The term with greatest occurrence is picked up and the cluster is labelled with that term. This is given by

$$lbl = Clt(Doc\left(Tm\left(DoR(Tm_1, Tm_2, Tm_3, \ldots Tm_n)\right)\right)) \tag{9}$$

Where cluster is denoted by Clt, Doc is the documents present in the cluster, Tm is the terms present in the cluster, DoR is the degree of recurrence of all terms present in the document.

This is followed by the arrangement of terms in descending order with respect to the degree of recurrence in the entire cluster. The term which is ranked first is declared as the label for the corresponding cluster. The main objective of cluster labelling is to enhance the readability. Any user can come to a conclusion about the substance of the cluster, at a streak. Thus, the intention of this work to cluster the similar documents and to label the cluster with meaningful term is achieved, successfully.

Experimental analysis

This section evaluates the performance of the proposed algorithm in terms of precision rate, recall rate, F-measure, accuracy and misclassification rate. The proposed work is compared with the outcome of k-means, bisecting k-means and UPGMA algorithms. The proposed work which is based on semantic analysis proves accurate results.

The dataset being exploited for evaluating the performance of the proposed work is Reuters-21578 R8, which has got 8 classes [26]. On the whole, the dataset contains 7674 documents, which consists of 5485 training documents and 2189 testing documents.

Precision rate

Precision rate is the total number of documents whose actual label is x, but misclassified with label y.

$$P_{rate} = \frac{doc_{xy}}{doc_y} \times 100 \tag{10}$$

Where doc_{xy} is the total number of documents with actual label x, but wrongly classified as y. doc_y is the documents which as correctly labelled as y. Thus, a clustering algorithm works well with greater precision rates.

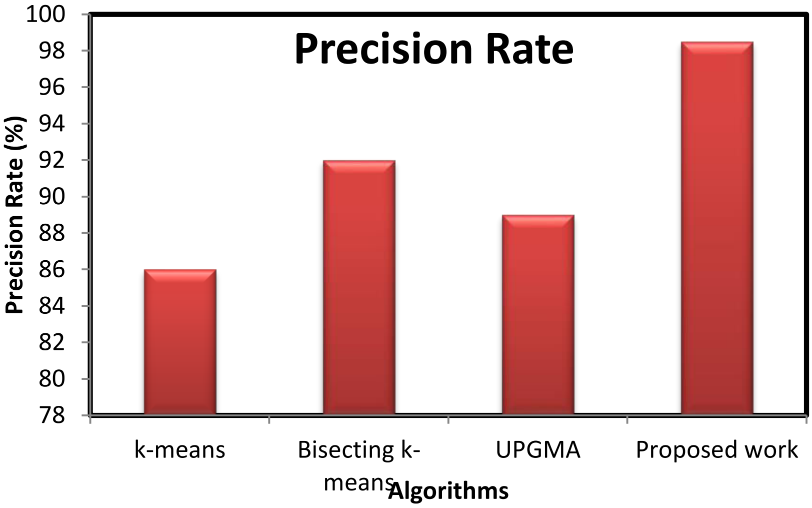


Figure-1. Precision rate.

From the experimental results, it is obvious that the proposed work shows greater precision rate with 98.5%. Thus, the documents are clustered in a better way.

Recall rate

Recall rate is the total number of documents whose actual label is x, but misclassified with label y.

$$R_{rate} = \frac{doc_{xy}}{doc_x} \times 100 \qquad (11)$$

Where doc_{xy} is the total number of documents with actual label x, but wrongly classified as y. doc_x is the documents which as correctly labelled as x. Thus, a clustering algorithm works well with greater recall rates.

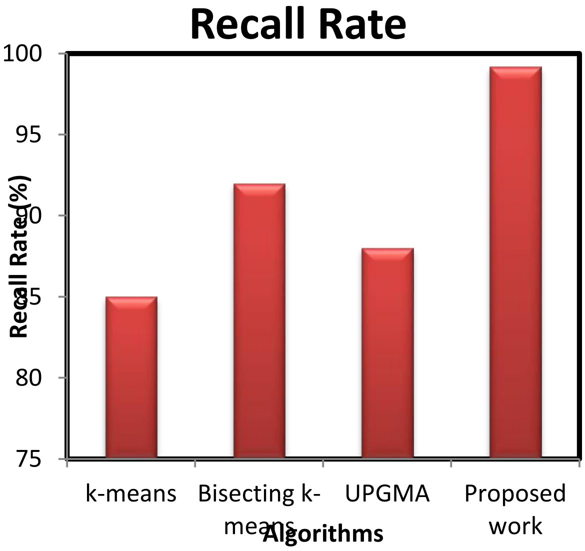


Figure-2. Recall rate.

The experimental results show that the recall rate of the proposed work is 99.2%, which is comparatively greater than other algorithms.

F-measure

F-measure is computed by taking precision and recall rate into account. F-measure of a cluster and a class is given by

$$F(cls, cltr) = \frac{2*P_{rate}*R_{rate}}{P_{rate}+R_{rate}} \times 100 \qquad (12)$$

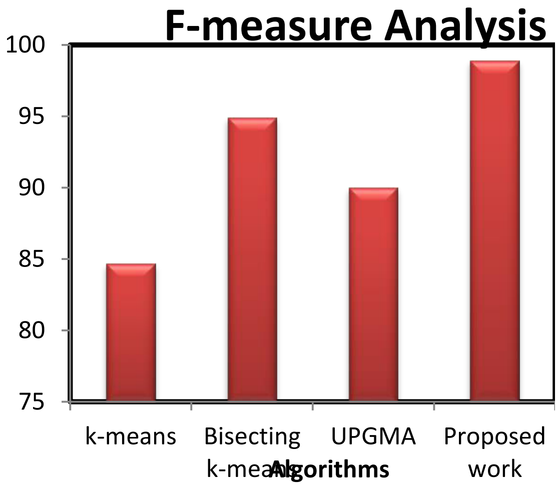


Figure-3. F-measure analysis.

The greater the value of F-measure, the higher is the quality of the cluster. On observing the experimental results, the proposed work shows the maximum quality of cluster with 98.9%.

Accuracy rate

The accuracy rate of the algorithm is determined by the sum of correctly clustered documents and the correctly rejected documents (as they are not relevant) to the total number of clustered documents.

$$acc = \frac{ccd+crd}{total\ clustered\ documents} \quad (13)$$

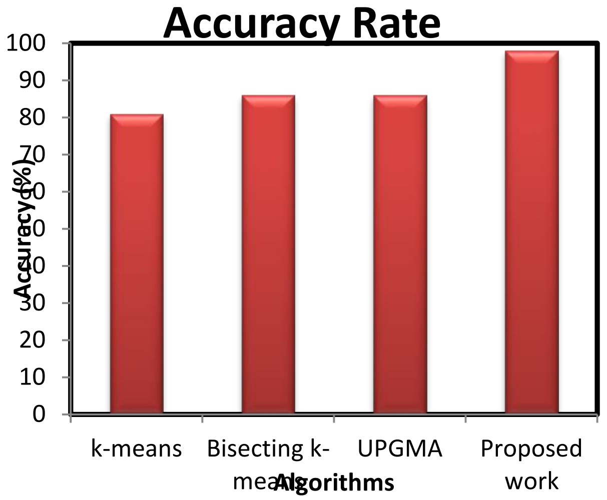


Figure-4. Accuracy rate analysis.

The accuracy rate of the proposed work is comparatively better than other algorithms, whereby the objective of the work is fulfilled.

Misclassification rate

Misclassification rate is the rate of wrong clustering of documents. The misclassification rate must relatively be low and is calculated by

$$mis_{rate} = 1 - acc \quad (14)$$

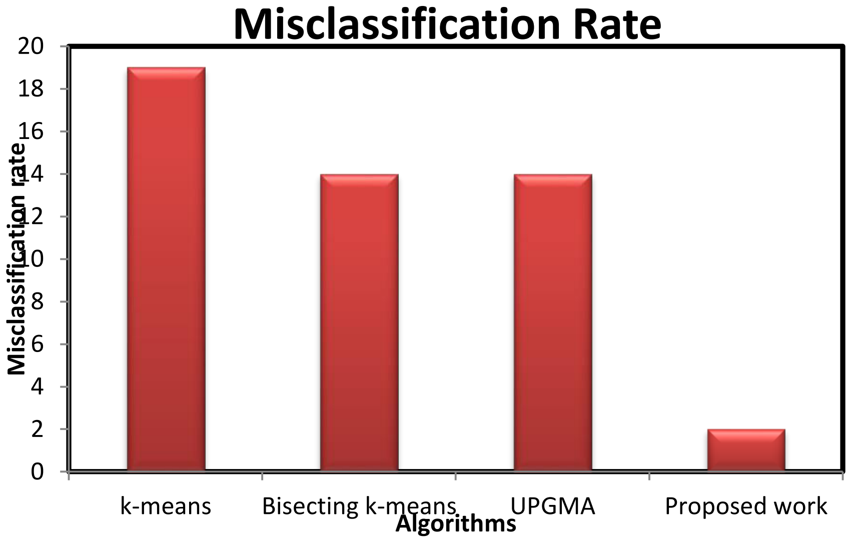


Figure-5. Misclassification rate.

Thus, the misclassification rate of the proposed work is the least, which when compared with all the other algorithms. Thus, the power of fusion of bisecting k-means and UPGMA algorithm along with semantic analysis is proven by the experimental results.

CONCLUSIONS

This paper presents a semantic approach based hybrid clustering algorithm. The main goal of clustering is to group related documents together in a cluster. This work employs wordnet to perform semantic analysis. The cluster formation is accurate, as the clusters are framed semantically. Bisecting k-means and UPGMA algorithms are employed for the process of clustering. Finally, the clusters are labelled on the basis of term weight. The performance of the proposed algorithm is evaluated and compared with several existing algorithms. The experimental results prove the efficiency of the proposed algorithm.

REFERENCES

[1] Amine A., Elberrichi Z. and Simonet M. 2010. Evaluation of text clustering methods using WordNet. The International Arab Journal of Information Technology. 7(4): 349-357.

[2] Bouras C. and Tsogkas V. 2012. A clustering technique for news articles using WordNet. Knowledge-Based Systems. 36, 115-128.

[3] Chen C.-L., Tseng F. S. and Liang T. 2010. An integration of WordNet and fuzzy association rule mining for multi-label document clustering. Data and Knowledge Engineering. 69(11): 1208-1226.

[4] Dang Q., Zhang J., Lu Y. and Zhang K. 2013. WordNet-based suffix tree clustering algorithm. In Paper presented at the 2013 international conference on information science and computer applications (ISCA 2013).

[5] Fodeh S. J., Punch W. F. and Tan P. -N. 2009. Combining statistics and semantics via ensemble model for document clustering. In Paper presented at the proceedings of the 2009 ACM symposium on applied computing.

[6] Hotho A., Staab S. and Stumme G. 2003. WordNet improves text document clustering. In Paper presented at the in SIGIR international conference on semantic Web Workshop.

[7] Jing L., Zhou L., Ng M. K. and Huang J. Z. 2006. Ontology-based distance measure for text clustering.

[8] Kang B.-Y., Kim D.-W. and Lee S.-J. 2005. Exploiting concept clusters for contentbased information retrieval. Information Sciences. 170(2): 443-462.

[9] Recupero D. R. 2007. A new unsupervised method for document clustering by using WordNet lexical and conceptual relations. Information Retrieval. 10(6): 563-579.

[10] Sedding J. and Kazakov D. 2004. WordNet-based text document clustering. In Paper presented at the proceedings of the 3rd workshop on robust methods in analysis of natural language data.

[11] Song W., Li C. H. and Park S. C. 2009. Genetic algorithm for text clustering using ontology and evaluating the validity of various semantic similarity measures. Expert Systems with Applications. 36(5): 9095-9104.

[12] J.B. MacQueen. 1967. Some methods for classification and analysis of multivariate observations, in: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistic and Probability, University of California Press, Berkley, CA. pp. 281-297.

[13] S. Lee, W. Lee. 2012. Evaluation of time complexity based on max average distance for K-means clustering, Int. J. Security Appl. 6: 449-454.

[14] C. Manning, P. Raghavan, H. Schütze. 2008. Introduction to Information Retrieval, Cambridge University Press, Cambridge, England.

[15] D. Reddy, P.K. Jana. 2012. Initialization for K-means clustering using voronoi diagram, Procedia Technol. 4, 395-400.

[16] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. McLachlan, A. Ng, B. Liu, P. Yu, Z.-H. Zhou, M. Steinbach, D. Hand, D. Steinberg. 2008. Top 10 algorithms in data mining, Knowl. Inf. Syst. 14: 1-37.

[17] P. Berkhin. 2002. Survey of Clustering Data Mining Techniques, Accrue Software Inc.

[18] J. Han, M. Kamber, A.K.H. Tung. 2001. Spatial clustering methods in data mining: a survey, in: Geographic Data Mining and Knowledge Discovery, Taylor and Francis. pp. 1-29.

[19] A.K. Jain, M.N. Murty, P.J. Flynn. 1999. Data clustering: a review, ACM Comput. Surv. 31: 264-323.

[20] G.H.O. Mahamed, P.E. Andries, S. Ayed. 2007. An overview of clustering methods, Intell. Data Anal. 11: 583-605.

[21] D. Hindle. 1990. Noun classification from predicate-argument structures, Proc. of the Annual meeting of

the association for computational linguistics. pp. 268-275.

[22] S. Caraballo. 1999. Automatic construction of a hypernymbased noun hierarch from text, Proc. of the Annual meeting of the association for computational linguistics. pp. 120-126.

[23] P. Velardi, R. Fabriani, and M. Missikoff. 2001. Using text processing techniques to automatically enrich domain ontology, Proc. of the international conference on Formal ontology in information systems. pp. 270-284.

[24] P. Cimiano, A. Hotho, and S. Staab. 2005. Learning concept hierarchies from text corpora using formal concept analysis, Journal of Artificial Intelligence Research. 24: 305-339.

[25] C. Fellbaum. 1998. WordNet: an electronic lexical database, MIT Press.

[26] http://www.csmining.org/index.php/r52-and-r8-of-reuters-21578.html.